

The Great Transformation of AI and Healthcare



Attorney Kim Kyung-jin

Table of Contents

Part 1. Medical Records and Clinical Workflow Automation

Chapter 1. AI Scribes and Documentation Work

1.1 Clinical Voice Recognition and Reduction of Documentation Time

1.2 Case Studies of Adoption at Yale and Cleveland Clinic

1.3 Clinician Burnout and Off-Hours Documentation Work

Chapter 2. Clinical Decision Support Systems (CDSS) and Diagnosis

2.1 Large Language Model (LLM)-Based Antibiotic Prescription Support

2.2 Medical World Models and U2-Med's Diagnostic and Testing Requests

2.3 Diagnostic Reasoning of AMIE and MIRA

Part 2. AI Drug Discovery and Digital Twins

Chapter 3. Bio Foundation Models and Drug Targets

3.1 Medical Multimodal Foundation Models

3.2 Octopus's Exploration of Colorectal Cancer Biomarker (IGF2)

3.3 Cellular World Model (Chreode) and Molecular Design

Chapter 4. Health Digital Twins and Clinical Trials

4.1 Digital Twins Reflecting Patient Characteristics

4.2 Synthetic Control Arm Clinical Trials

4.3 Model-Informed Drug Development and Animal Testing Alternatives (FDA Modernization Act)

Part 3. Global Case Studies of Medical AI Adoption

Chapter 5. AI Adoption in U.S. Healthcare Institutions

7

9

10

17

23

30

31

40

49

58

60

61

74

82

94

95

105

114

125

127

kimkj.com

3

5.1 Mayo Clinic's VoiceCare AI and Administrative Automation	128
5.2 Tempus's Support for Precision Oncology	137
Chapter 6. Adoption of Healthcare AI in China	146
6.1 AI Operational Outcomes at the Affiliated Hospital of Hebei Medical University ¹⁴⁷	
6.2 WiNGPT 3.0 and the Ant Afu Service	154
6.3 Mindray Qiyuan (启元) and the Digitalization of Primary Healthcare	162
Part 4. Regulation, Safety, and Economics of Medical AI	170
Chapter 7. AI Medical Device Regulation and PCCP (Predetermined Change Control Plan)	172
7.1 PCCP Guidance and AI Model Change Management	173
7.2 FDA Competency Assessment and Risk Assessment	182
7.3 UK AI Airlock and EU AI Act	191
Chapter 8. Patient Safety and the Economics of Medical AI	199
8.1 Interpreter Error Propagation and Speech Recognition Bias	200
8.2 Automation Bias and Human-in-the-Loop Review	208
8.3 The Healthcare Agentic AI Market and CPT Reimbursement	216
Part 5. Medical AI Agents and Collaborative Care	225
Chapter 9. Healthcare Agentic AI	227
9.1 Limitations of Single Prediction Models	228
9.2 Sense-Think-Act and Memory	235
9.3 Taxonomy and Technological Maturity of Medical Agentic AI	244
Chapter 10. Multi-Agent Systems and AI Hospitals	251
10.1 Multi-Agent Collaborative Design	252
10.2 MAI-DxO's Division of Roles and Complex Diagnosis	260
10.3 ClinVerse and the Clinical Care Process of AI Hospitals	266

Part 6. The Future of Medicine and AI	275
Chapter 11. Collaboration Between Physicians and AI Agents	277
11.1 AI Screening, Human Review, and Comprehensive Diagnosis	278
11.2 The Role of Physicians and Comprehensive Clinical Judgment	288
Chapter 12. Technical and Ethical Challenges of Future Medical AI . . .	296
References	308
Book Information	360
Sponsorship Information	361

Part 1. Medical Records and Clinical Workflow Automation

Chapter 1. AI Scribes and Documentation Work



1.1 Clinical Voice Recognition and Reduction of Documentation Time

TXP Medical's product announcement details what type of records SpeechER generates from voice input and identifies who conducted the survey behind the time-saving figures. This document enables readers to distinguish between functional descriptions and efficacy claims.

<https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZFOBI/>

Medical records are essential materials for communicating a patient's symptoms, tests, and treatment plans, and for continuing subsequent care. However, data entry during consultations and documentation after consultations can increase the workload of medical staff. The term "pajama time," used in the U.S. medical community, refers to electronic health record (EHR) documentation work that continues even after returning home from work.

An ambient AI scribe is a technology that recognizes clinical conversations and creates draft documentation. "Scribe" refers to a documentarian, and "ambient" indicates the method in this technology of processing ambient conversation without separate manual input. Medical staff converse with the patient, and the system organizes the necessary information into document format.

Speech recognition and documentation are distinct tasks. Speech recognition transcribes spoken words into text. Documentation selects from that text the details necessary for clinical care and arranges them into appropriate sections. For example, when recording a patient's statement that "it hurt yesterday, but it is fine now," the chronological progression of the symptom must be preserved. Reducing it simply to "pain present" can misrepresent the current status. This example is intended to illustrate the mechanics of documentation errors.

The Process of Converting Clinical Voice into Documentation

The first step is the collection of dialogue. Audio is recorded using a microphone or a smartphone app. Patient consent and the processing method for recorded materials must be determined in accordance with local laws and healthcare institution policies.

Automatic Speech Recognition (ASR) converts the collected voice into text. In this process, medical terminology, drug names, speaker differentiation, and ambient noise can affect accuracy.

Large Language Models (LLMs) organize symptoms, medications taken, clinical findings, and treatment plans from the text. They can also structure documentation according to

the SOAP format, divided into Subjective, Objective, Assessment, and Plan. Depending on the product, processing takes place during the consultation or generates a draft after the consultation.

This is also why SOAP separates the symptoms perceived by the patient from the values measured by the medical staff. A complaint of "feeling like having a fever" and the result measured with a thermometer have different sources. If AI mixes the two, it becomes difficult for someone reading the record later to know what was actually confirmed. A good draft must preserve the source and timing of information in addition to making sentences smooth.

Finally, medical staff review, edit, and finalize the draft. A document generated by AI is not a verified record in itself. In a human-in-the-loop structure where a human participates in the processing workflow, the content and accuracy of the record must be verified.

Changes in Data Entry Time and Review Time

An AI scribe reduces direct data entry time but creates the task of verifying drafts. Therefore, overall time savings must take into account both the reduction in data entry and the time spent on review and editing. The effect may also vary depending on the documentation format, medical specialty, and the characteristics of the patient dialogue.

For example, even if manual writing time decreases from 10 minutes to 4 minutes, if 3 new minutes are required to verify the draft, the total documentation task takes 7 minutes. These numbers are a hypothetical example illustrating the calculation method. Comparing pre-implementation by measuring only manual entry time and post-implementation by measuring only draft generation time inflates the savings effect. Time must be measured across the identical scope from the start of documenting one patient to finalization.

Japan's SpeechER and nursing documentation support tool Caretomo are examples of organizing voice input into documentation.

The developer of SpeechER is TXP Medical of Japan. In the award announcement released by the company on October 15, 2025, it states that it has already been launched as a commercial app. This date is the date of the announcement in

question and should not be read as the initial launch date. Caretomo is a nursing documentation support product provided by SIND of Japan. In its official product description, the sequence of voice recognition, SOAP classification, and registration after verification and editing can be confirmed, but the initial launch date is not specified in that product overview. Both names are not the names of general-purpose AI models, but product names configured for medical workflows. The vendors announced reductions in documentation time in their self-surveys and short-term trials. However, it is difficult to generalize these results into a rate of reduction that can be expected across all institutions.

Comparative studies have also revealed differences depending on the product and metric. In a UCLA study, documentation time decreased for one product, but another product showed no significant difference compared to the control group. A study analyzing operational data from multiple healthcare institutions reported a reduction in documentation time, but off-hours work did not decrease alongside it.

Therefore, presenting a specific 70% reduction figure as a universal effect of the technology as a whole is not appropriate. Record entry, document drafting, EHR usage, and clinical decision-making are distinct tasks. For example, the

reduction in assessment time of a remote electrocardiogram platform must be distinguished from documentation time.

Conditions Affecting Time Savings

The actual effect is influenced by usage frequency and workflow patterns. Results may differ between clinicians who use drafts frequently and those who rarely use them. If templates and terminology do not fit the medical specialty, the burden of editing increases.

In some cases, clinical procedures and data management change alongside AI. Even if hospital efficiency improves, it is difficult to conclude that the entire transformation occurred solely due to AI. When examining the effects of adoption, one must identify which workflows changed and what accompanying support was provided.

Documentation Accuracy and Safety

Omissions, erroneous additions, or reversals of meaning can occur during speech recognition and content summarization. In a simulated clinical study involving interpretation, issues arose where interpretation errors were propagated into the AI documentation. There were also instances where a symptom stated by the patient was documented as absent in the record.

Automation bias is the tendency of users to place excessive trust in AI outputs. If sentences are written fluently, errors such as medication dosages, allergies, or the presence or absence of symptoms can be missed. Draft review must not be operated as a perfunctory sign-off procedure.

The goal of adopting an AI scribe is to maintain the accuracy of necessary information while reducing the burden of documentation. Healthcare institutions must collectively evaluate documentation time, editing burden, patient communication, and record quality. Workflow procedures and review conditions must also be established so that the effect of reducing repetitive data entry translates into improvements in clinical care.

1.2 Case Studies of Adoption at Yale and Cleveland Clinic

Cleveland Clinic's 2025 operational overview explains the developer of the selected product and the actual documentation generation process. The findings of the Yale research team can be confirmed in the original paper referenced at the beginning of the next section.

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

To deploy an AI scribe hospital-wide, user training, support, and department-specific templates must be prepared alongside product performance. Features that were usable in small-scale trials give rise to new demands when applied to diverse workflows by numerous clinicians.

Yale New Haven Health and Cleveland Clinic are case studies illustrating this adoption process. Yale participated in research evaluating clinician workload, and Cleveland Clinic made public its product comparison and large-scale rollout process.

Abridge, used in the Yale study, is a service developed by Abridge in the United States. The company's official launch announcement was made on October 6, 2020, focusing at the time on a service allowing patients to record clinical conversations and recheck medication instructions. The

functional scope of the clinician documentation product adopted later by hospitals is not identical to that of the early patient-facing service. Company founding, service unveiling, and adoption by a specific hospital must be distinguished as separate events.

Yale's Adoption and Workload Evaluation

The Yale research team surveyed the work experiences of clinicians across multiple health systems before and after using an AI scribe (Abridge). While trends of reduced burnout and documentation burden were reported, because this was a pre- and post-implementation comparison, the entirety of the changes cannot be concluded as an effect of AI.

Another multicenter study in which Yale participated also demonstrated a decrease in documentation time. However, off-hours work time did not decrease alongside it. This result indicates that the effects of product adoption must be evaluated both in terms of documentation speed and the actual working environment.

As Yale expanded ambient AI into clinical practice, it also examined the issue of residency and training education. Even if they rely on documentation drafts, medical students and residents must maintain the ability to independently interpret

content and write records. Education to prevent de-skilling remains a necessary priority even after implementation.

Cleveland Clinic's Product Selection

After piloting several AI scribes, Cleveland Clinic selected the product from Ambience Healthcare. It involved clinicians from diverse specialties to examine whether it suited real-world workflows.

The developer selected by this hospital is Ambience Healthcare of the United States. According to the hospital's overview on August 14, 2025, after obtaining patient consent and capturing conversations via a smartphone app, it generates clinical documentation and after-visit summaries in the Epic electronic health record. The year 2025 marks the point of confirming this hospital's adoption progress and does not mean that the product itself was first developed in that year. The hospital assumes responsibility for care and documentation, while the developer provides software supporting those tasks. It included users heavily burdened by documentation tasks as well as those highly interested in new technology.

The pilot was a process for evaluating feature comparisons and user experience. Subsequently, rollout expanded across outpatient care, and it was reported that thousands of clinicians began using it within a relatively short period.

However, operational reports from institutions and vendor companies must be distinguished from independent clinical efficacy evaluations.

Operational Governance for Large-Scale Adoption

Governance is a system that establishes decision-making and safety standards. Cleveland Clinic structured it so that relevant departments—including executive management, clinical leadership, health informatics, IT, and medical billing—participated together. This is because even a documentation tool affects multiple operations.

For example, if the phrasing of a diagnosis in a draft changes, it can influence not only the subsequent physician's judgment but also the billing specialist's coding classification. Therefore, matters to be established by the hospital include which fields to populate automatically, who finalizes them, and where to report errors when detected. Because a feature convenient for one department may increase verification work for another, the overall workflow is examined comprehensively.

Training was operated as a prerequisite procedure prior to use. Rollout was phased across medical specialties, and online training provided guidance on basic usage and review responsibilities. Complex clinical departments required time to adjust models and templates.

User support was utilized to resolve issues such as configuration, draft editing, and hospital policies. In cases where multiple clinicians conduct consultations together or specialized templates are required, the product's default settings alone may not suffice.

An improvement process reflecting operational data and user feedback was also established. Training and usage status by clinical department were reviewed, and documentation formats were refined through user groups. Even after deploying the product, actual conditions of use were continually adjusted.

Interpreting Usage Scope and Satisfaction

Institutional reports demonstrate that AI drafts were generated across many consultations and that users experienced convenience. However, initial adoption differs from sustained use, and the mere fact of high satisfaction does not prove documentation accuracy or reductions in turnover.

Differences across medical specialties must also be taken into account. Depending on the length and complexity of dialogue and the required documentation format, the burden of draft editing varies. Furthermore, effects confirmed in outpatient settings cannot be assumed to apply directly to emergency departments, inpatient wards, or nursing workflows.

Sutter Health is likewise an example of expanding AI scribes into outpatient, emergency, and inpatient areas. While this trend shows that the scope of usage is broadening, cumulative record volume and improvements in care quality must be evaluated separately.

Common Implications of Adoption Cases

The Yale case illustrates the necessity of measuring clinician work experience, while the Cleveland Clinic case demonstrates the importance of an organizational framework supporting large-scale adoption. Beyond technical capabilities alone, product selection, training, support, and template refinement influence real-world usage.

To operate an AI scribe reliably, clinicians must be able to verify drafts, and when issues arise, they must be able to receive support and make corrections. Even as the scale of adoption expands, it is essential to monitor documentation quality and review burden concurrently.

1.3 Clinician Burnout and Off-Hours Documentation Work

The original JAMA Network Open paper published by Yale researchers and others presents pre- and post-implementation surveys along with burnout estimates. Reading it while checking the number of participants, the number of individuals included in the analysis, and the presence or absence of a control group makes it easier to understand the meaning of the figures.

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

The purpose of introducing AI scribes is not limited to reducing documentation time. Reducing the burden of documentation tasks so that healthcare professionals can focus on patient care is also an important goal. Paperwork that continues even after clinical care ends increases fatigue among medical staff and makes work-life balance difficult.

Burnout refers to a state of exhaustion resulting from work. For healthcare professionals, it can manifest as persistent fatigue, diminished work motivation, and emotional detachment from patients. The causes vary, including working hours, staffing shortages, and organizational environment, and clinical documentation is also a factor contributing to work burden.

Changes in Documentation Burden and Burnout

In a study of six US health systems led by Yale researchers, burnout and documentation workload showed a tendency to decrease after the introduction of an AI scribe. Healthcare professionals used an AI scribe from the same company (Abridge) and evaluated their work experience before introduction and after 30 days of use. The estimated burnout rate decreased from 51.9% to 38.8%.

This value is not a simple proportion of all survey participants, but an estimate from a model adjusted for various characteristics. While 263 participants completed both the pre- and post-surveys, 184 individuals were included in that adjusted analysis. This is because having responded to the survey is different from having all items required for a specific analysis. However, since these are survey results comparing before and after implementation without a control group, the entirety of the change cannot be conclusively attributed to the effect of AI.

The significance of this study lies in the fact that the criteria for evaluating medical AI are expanding from documentation speed to the work environment of healthcare professionals. It has begun to examine not only how quickly records are written, but also whether it has become easier to focus during

consultations and whether paperwork continues even after completing shifts.

Studies from University of Wisconsin Hospitals (UW Health) and UCLA Health also reported a tendency for documentation burden or burnout to improve after using AI scribes. However, because tools, usage periods, and work environments differ across studies, it is difficult to expect effects of the same magnitude.

The Process of Workload Reduction

When AI records conversations and creates drafts, physicians can be relieved of the burden of typing while listening to patient explanations. As the frequency of looking back and forth between the screen and the patient decreases, it also becomes easier to focus on the conversation.

The process of repeatedly switching between listening to and assessing clinical details and documenting them also imposes a burden. Such task switching entails context switching costs. When AI takes charge of drafting records, physicians gain room to focus on reviewing patient symptoms and treatment directions. However, the specific pathways through which these changes lead to burnout reduction require additional research.

Whether documentation can be completed before leaving work is also important. When after-hours paperwork decreases, securing rest time becomes easier. However, just because time spent writing clinical notes has decreased does not mean that after-hours work decreases to the same degree, as other administrative duties may remain or time may be needed to review AI drafts.

Perceived Burden and Actual Working Hours

It is necessary to distinguish between the burden perceived by healthcare professionals and the working hours logged in the system. In studies analyzing electronic health record (EHR) usage data across multiple medical institutions, documentation time decreased, but after-hours work time did not significantly decline. In a Boston hospital system as well, the time spent writing clinical notes decreased, but there was no difference in completing documentation outside of work hours.

This discrepancy shows that the effect of AI cannot be explained solely by time reduction. Clinicians may feel that work has become easier because the burden of repetitive data entry was reduced, while conversely, some time spent modifying AI drafts might not have been captured in measurements. Therefore, user experience and actual working hours must be examined together.

Surveys can capture the burden perceived by medical staff, but they are influenced by respondents' expectations or work conditions at the time. Access logs in electronic health records show traces of actual usage, but may not accurately distinguish time spent talking with patients while keeping the screen open. Rather than treating either single metric as the definitive answer, it is preferable to check whether both data sources point in the same direction.

The effects over prolonged use are also not yet clear. In a study surveying at multiple time points after introduction, burnout did not improve concurrently even though paperwork burden decreased. Because burnout is influenced by organizational and work environments in addition to documentation tasks, it is difficult to resolve through AI adoption alone.

The New Task of Draft Review

When AI scribes are introduced, clinicians' documentation tasks partly shift from writing directly to reviewing and correcting drafts. Even if drafts are written in natural sentences, clinicians must verify whether medication dosages, presence or absence of symptoms, and test results are accurate. Errors in the interpretation process may also be directly reflected in the records.

Hospitals must consider the workload required for review alongside the reduction in documentation time. Operational approaches are needed that provide templates tailored to specialties, facilitate easy error correction, and clearly define review responsibilities.

Review time must be built into clinic schedules. If that time is instead filled with additional patient appointments simply because drafts are generated quickly, reviewing records can be pushed back until after work once again. Hospitals must monitor when documentation is finalized and the volume of revisions even after adoption. Where the time saved by healthcare providers ends up being spent is not determined by software speed alone.

AI scribes are demonstrating the potential to reduce the documentation burden on healthcare professionals. To translate this into sustained improvement of the work environment, work processes, training, and staffing management must be examined alongside technology adoption. The goal of implementation is to secure time for medical staff to dedicate to patient care and rest while maintaining record accuracy.

Chapter 2. Clinical Decision Support Systems (CDSS) and Diagnosis



2.1 Large Language Model (LLM)-Based Antibiotic Prescription Support

The original paper published by researchers from Capital Medical University, Tsinghua University affiliated hospitals, and other institutions describes the process of linking patient records, clinical guidelines, and prescription rules. It can be read while keeping in mind that it was an evaluation study using historical data before deployment in actual prescribing.

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

Antibiotic prescription is a clinical judgment that must consider both the cause of infection and the patient's condition. Physicians select drugs based on suspected bacteria, infection site, severity, and regional antibiotic resistance patterns. There are also cases where dosage or dosing intervals must be adjusted according to liver and kidney function, age, and concomitant medications.

Clinical Decision Support Systems (CDSS) provide information necessary for these judgments. They check patient test results and clinical guidelines, presenting considerations regarding drug selection or dosage. The role of AI lies in helping physicians review multiple conditions without omission.

At this point, the timing of the information entered into the system is crucial. Even for the same patient, test results at the time of admission and results a few days later can have different clinical significance. The screen recommending medications must clearly show which tests and records were used as evidence. Confirming only whether the patient's name is correct is not sufficient. If recent results are missing or medications used in other clinical encounters remain listed, even a plausible response may conflict with current prescribing decisions.

From Rule-Based Systems to Language Models

MYCIN, developed at Stanford University in the 1970s, was an early expert system for infectious disease diagnosis and antibiotic recommendation. Because it presented conclusions according to rules formulated by experts, explaining the reasoning process was straightforward. However, the rules required continuous updating, and there were limitations in processing clinical notes written in free-form text.

MYCIN was researched primarily by Edward Shortliffe and others at Stanford and was systematically organized in a 1976 book. Unlike today's generative AI, it linked expert-authored conditions to conclusions. For example, if predefined conditions were met, it prompted the consideration of a specific clinical judgment. The historical research system

MYCIN and recent language model products share the same CDSS category, yet their technological starting points differ.

Subsequently, machine learning models were used to find patterns in patient data and predict risks. However, because it was often difficult to explain the reasoning behind predictions, additional verification was required to apply them to prescribing decisions.

Large language models (LLMs) can read clinical records and explain recommendation rationales in sentences. They differ from conventional systems in that physicians can ask questions in natural language and review the basis of responses. However, a natural explanation does not guarantee that the prescription is accurate. Hallucinations that generate factually incorrect information, omissions of dose adjustment conditions, and variations in responses depending on question phrasing can occur.

Issues of Accuracy and Consistency

An Italian infectious disease research team presented infection cases to multiple language models and compared their antibiotic recommendations. Performance varied widely across models, and subsequent evaluations also revealed incorrect or unnecessary recommendations. Inconsistencies where answers varied upon repeated queries of the same case were also identified.

In an experiment comparing a medical tool that provides literature evidence (OpenEvidence) with a general-purpose conversational model (ChatGPT-5.2), the tool that provided more accurate answers was not necessarily more consistent. Accuracy refers to whether the answer is appropriate, while consistency refers to whether similar answers are maintained under identical conditions. In prescription support, both criteria must be evaluated together.

OpenEvidence is a medical information service provided by the US company of the same name, and ChatGPT is a conversational service from OpenAI. The model version used in the paper must be distinguished from the overall service, because changes to the service can also alter retrieved materials and answering behavior. While corporate announcements for OpenEvidence were documented in 2023, the initial release date of the version used in the comparative

experiment is a matter to be verified separately from the testing timeframe stated in the paper.

Integration of Patient Data, Clinical Guidelines, and Prescribing Rules

Chinese researchers evaluated a language model-based antibiotic prescription support system using past electronic medical records of hospitalized pneumonia patients.

This system is a joint study by Capital Medical University, Beijing Tsinghua Changgung Hospital, Beijing Friendship Hospital, and others, and is not a case released under a specific commercial product name. The preprint was made public on April 14, 2026, and the peer-reviewed paper on August 4 of the same year. This does not mean the date when hospitals began applying this system to all patients. The core approach was that, rather than presenting only patient conditions to the model, it simultaneously provided prescriptions of similar patients, clinical guidelines, and rules defined by physicians.

The first category of information is clinical records of similar patients. After converting patient records into vectors—arrays of numbers—cases similar to the current patient are retrieved. An embedding model is used in this process. While referring to

prescriptions from past cases, whether they can be applied as-is to the current patient must be judged separately.

A vector is a list representing multiple features as numbers. A computer compares the relationships between these numbers to find similar records. The fact that sentences are similar differs from the fact that patients are clinically identical. For example, even if symptom descriptions resemble each other, following the same prescription is difficult if allergies or organ functions differ. This is why patient-specific constraints must be reapplied to search results.

The second category of information is a knowledge graph organizing clinical practice guidelines. It structures relationships among diseases, causative pathogens, drugs, and contraindications, enabling the model to find necessary evidence. For example, it can link medications to avoid under certain conditions or requirements for dose reductions.

The third category of information is prescription rules compiled by clinicians. These specify drug selection according to disease severity, contraindicated drug combinations, and dosage adjustments based on hepatic and renal function. Referencing this information, the language model proposes the drug name, dosage, administration frequency, and route,

while explaining its rationale. The physician evaluates the proposal and rationale together.

In this architecture, a design that bridges the hospital's clinical experience and guidelines is just as important as the general knowledge of the language model. Even when the model generates answers, the data and constraints necessary for clinical judgment must be systematically provided by the healthcare institution.

Outcomes and Limitations of Prescription Support

In an evaluation using records of pneumonia patients, providing similar cases, guidelines, and prescribing rules together yielded higher prescription concordance than using a base language model. Excluding parts of this information degraded performance, showing that similar patient retrieval and clinician rules played critical roles.

The study evaluated the degree of agreement between recommended prescriptions and baseline prescriptions using F1 score and Jaccard accuracy. An ablation study is a method of examining the contribution of each component by removing them one by one. While such evaluations help compare system architectures, they do not directly demonstrate therapeutic efficacy in actual patients.

The F1 score reflects both how comprehensively necessary items were identified and how appropriate the recommended items were. The Jaccard index examines the proportion of overlap between two sets of items. Even if the drug name matches the baseline, the dosage may be incorrect, and an alternative that differs from the baseline prescription may still be medically acceptable. Therefore, when interpreting the numbers, one must examine what was defined as the correct answer and who reviewed the errors.

Recommendation performance was relatively low in elderly patients and those with impaired hepatic or renal function. Review of error cases also revealed remaining dangerous recommendations that failed to sufficiently account for organ function. This is why clinician verification cannot be bypassed even when dose adjustment rules are provided.

This study represents a stage where the system was evaluated using historical records. How prescriptions change in actual clinical practice, and whether adverse effects or treatment outcomes improve, must be further confirmed. Because patient demographics, patterns of resistant bacteria, and prescribing practices differ across hospitals, verification is also required when introducing it to other institutions.

Antimicrobial Stewardship and the Role of Medical Staff

Antimicrobial stewardship programs are activities that manage the appropriate selection and administration of antibiotics. In this process, language model-based CDSS can contribute to organizing relevant information and presenting prescribing rationales and precautions.

In actual implementation, it must be structured so that recommendation results can be easily cross-checked against clinical practice guidelines and patient status. Healthcare providers check patient-specific conditions such as liver and kidney function, allergies, and concomitant medications, and determine the final prescription. The developmental direction of antibiotic prescription support lies in improving language model performance while connecting reliable data, managing rules, and refining review procedures for medical staff.

2.2 Medical World Models and U2-Med's Diagnostic and Testing Requests

The technical introduction by the developer Unisound explains the roles of U2-Med and its training environment, U2-Med-World, separately. Because this material is a corporate research presentation, it should be read distinctly from clinical trial results.

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

Diagnosis is not a task concluded with a single question. Physicians listen to the patient's symptoms, verify medical history and current medications, and perform tests as necessary. As new information arrives, they re-examine possible diseases and subsequent actions. In medical AI research as well, attempts to implement such sequential decision-making processes are increasing.

Conventional medical AI evaluations often provided organized patient information all at once and required choosing the correct answer. In contrast, in actual clinical practice, one must decide which information to check first. Medical world models connect patient states, clinical actions, and resulting observation outcomes to bridge this gap.

Concept of Medical World Models

A world model is a model that predicts what outcomes will occur when a specific action is taken in a current state. In the medical field, it can be applied to simulate changes in patient state or information obtained according to clinical actions such as medical interviews, examinations, and treatments.

Chinese AI company Unisound (云知声) announced the release of U2-Med on July 17, 2026, and explained its connection architecture with U2-Med-World in a presentation on October 21 of the same year. U2-Med-World is a virtual environment that represents the relationship between patient states and clinical actions, while U2-Med is an agent that selects questions and tool usage within this environment. The company explained that through this architecture, it learns the ability to revise judgments over multiple rounds of dialogue.

U2-Med-World organizes medical history, dialogue, laboratory tests, imaging, vital signs, and prescription information in chronological order. When an agent selects a question or a test, the environment provides the relevant information, and the agent determines subsequent actions based on this. At this point, it is crucial to prevent data leakage, where information verified later is mixed into earlier judgments.

For example, if a discharge summary containing the correct diagnosis is presented from the initial interview, the AI can arrive at the correct answer without exercising test selection capabilities. In sequential evaluation, only information that medical staff could have actually known at that specific point in time should be provided. Rules must also be established in advance regarding whether results are revealed when a test is requested, and how requests for tests without results are handled. If such environmental design is flawed, even achieving a high score makes it difficult to consider that clinical competence has been evaluated.

Distinction Between Questions, Tests, and Treatments

Medical interviews and diagnostic tests are processes that alter the information known about a patient. Treatment and disease progression alter the patient's actual state. A medical world model must distinguish between the two. This is because the mere fact that a test was performed does not signify improvement in the patient's condition, nor does initiating treatment confirm a diagnosis.

In the virtual patient case presented in the company's announcement, based on initial symptoms such as fatigue and loss of appetite, the agent selects what to verify first among symptom duration, accompanying symptoms, past medical history, and recently taken medications. It then adjusts the

order of questions and its judgments based on the provided information. By comparing multiple interview sequences for the same patient, one can evaluate how effectively necessary information was acquired and warning signs were identified.

This environment generates not only the clinical processes preserved in actual records, but also synthetic trajectories.

The term 'trajectory' refers to the sequence in which certain questions and actions were followed. In the record of a single patient, only the path actually chosen is observed. When a computer creates and compares alternative paths, it must indicate up to what point confirmation can be drawn from observed facts. This is because even if generated sentences are detailed, actual observations have not increased. A synthetic trajectory is a hypothetical clinical process when different questions or tests are selected from the identical initial state. While it can be used as training material to compare multiple choices, it is not evidence that proves the efficacy of treatments that were not actually administered.

Methods for Learning the Entire Dialogue Process

The training of U2-Med is described as a method of generating dialogues in a virtual environment, evaluating the entire process, and selecting appropriate cases for subsequent training. In addition to the final diagnosis, information gathering, risk identification, and communication processes are also subjects of evaluation.

The Rubric Reinforcement Learning (Rubric RL) introduced by the company is a method of reflecting detailed criteria, such as medical accuracy or completeness of information, into training. Online Preference Distillation (OPD) incorporates interview strategies selected through evaluation into model training. This approach is significant in that it expands the scope of evaluation from the accuracy of a single response to multi-stage clinical processes.

Discrepancies Between Performance Evaluation and Actual Clinical Practice

Unisound announced that U2-Med achieved high scores on HealthBench, a medical dialogue benchmark. HealthBench is an examination that evaluates whether necessary information was asked, whether danger signs were recognized, and whether explanations were provided accurately and safely.

However, the test results announced by the company must be distinguished from actual clinical efficacy. Even if good answers are presented in an evaluation environment, there is no guarantee that the same results will appear when actually used by patients and medical staff. The clinical efficacy of U2-Med, testing costs, and impact on patient safety must be verified separately in real clinical environments.

Test Selection and the Value of Information

Diagnostic tests provide information necessary for diagnosis, but entail costs, risks, and patient discomfort. Therefore, one must examine to what extent a subsequent test can change clinical judgment or treatment. The concept explaining this is the value of information.

The sequence of tests varies according to urgency, suspected diseases, test accuracy, and the impact of results on treatment. For AI to assist in test selection, it must consider both the benefits of securing more information and the necessity of reducing unnecessary tests. However, it is difficult to determine from public descriptions alone how these general principles are implemented in the internal computations of U2-Med.

Similar Research Directions

Microsoft's MAI-DxO is a framework in which multiple agents diagnose while requesting additional information. In an evaluation environment called SDBench, it selects interview and test information while examining complex clinical cases step by step. Published research demonstrates the feasibility of diagnosis through sequential information acquisition, but results from controlled evaluations cannot be regarded as actual clinical outcomes.

MAI-DxO is a diagnostic coordination framework created by Microsoft researchers. The initial arXiv disclosure of the related research was on June 27, 2025. SDBench is the name of the evaluation environment used to test this framework and various models. Distinguishing between the system and the benchmark prevents confusion between what AI was created and what problems were used to test it.

ClinVerse is a multi-agent diagnostic environment constructed based on real electronic health records. An agent in the role of a physician starts from initial symptoms, gathers information through medical interviews and tests, and makes a diagnosis. Information gathering, test selection, and diagnostic accuracy are evaluated together. This, too, is research intended for training and evaluation in a virtual environment and must be distinguished from implementation in actual clinical practice.

The paper on ClinVerse was first released on arXiv on October 26, 2026, by a research team including Sirry Chen et al. This does not mean that a completed product for treating patients in hospitals has been launched. By separating the patient role from the test result provider role, researchers can control the sequence in which information is disclosed while comparing the choices made by physician-role agents.

U2-Med-World, SDBench, and ClinVerse all investigate the process by which AI gathers additional information and revises its judgments. They demonstrate that medical AI is evolving from a tool that presents correct answers into a system that supports parts of the diagnostic process.

Considerations Required for Implementation

In AI that supports test selection, both over-testing and the omission of necessary tests must be managed simultaneously. Recommending excessive tests increases costs and patient burdens, whereas omitting necessary tests can lead to delayed diagnoses.

Current autonomous test requesting primarily refers to the capability of AI to propose test candidates within virtual environments. It does not mean that authority to independently prescribe tests in actual hospitals has been established. In clinical application, procedures are required in

which medical staff review and approve the recommendation rationale and patient status. The scope of liability must also be assessed according to regulatory clearance status, modalities of use, review procedures, causes of adverse events, and respective national legal systems.

The significance of medical world models lies in the ability to study not only diagnostic outcomes, but also the process of acquiring information and revising judgments. The task ahead is to verify whether capabilities learned in virtual environments operate accurately and safely in actual clinical practice.

2.3 Diagnostic Reasoning of AMIE and MIRA

Reading Google's early introduction of AMIE alongside the original Nature paper on MIRA reveals how medical dialogue research differs from behavioral research within electronic health records. Both sources explain their research evaluation environments and limitations.

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

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

Clinical practice guidelines are recommendations summarizing research evidence regarding diagnostic tests and treatments for each disease. Guidelines vary across specialties and are revised based on emerging evidence. Medical staff must identify recommendations suited to patient conditions while also taking into account hospital practice environments and conditions for drug usage.

Google's AMIE and MIRA from German researchers are medical AI studies supporting such tasks. AMIE focuses on reviewing diagnoses and management plans through dialogue with patients, while MIRA investigates methods of performing clinical actions, such as tests and prescriptions, within electronic health records. Both systems demonstrate that the

capabilities of medical AI are expanding from providing information to supporting clinical workflows.

Medical Dialogue and Management Planning in AMIE

AMIE (Articulate Medical Intelligence Explorer) is Google's research-oriented medical dialogue AI.

Google Research officially introduced AMIE on January 12, 2024. Research on diagnostic dialogue was released first, and the introduction of research extending it to disease management was published on June 17, 2026. AMIE is the name of the research system, and Gemini, used within it, is the name of a foundation model family developed by Google. Distinguishing between the foundation model and the system configured for medical tasks makes it easier to follow developments in research. Initial research evaluated the capability to make diagnoses through text-based dialogue with participants acting as patients. Subsequent research broadened the scope to disease management, which involves checking symptoms and treatment responses and adjusting plans across multiple clinical encounters.

Two agents participate in this process. The dialogue agent asks questions of the patient and explains necessary details. The management reasoning (Mx) agent formulates a management plan by reviewing clinical practice guidelines and drug

information based on the dialogue content. It is an architecture where roles are divided so that relevant evidence can be reviewed separately while maintaining continuous dialogue with the patient.

For example, if a patient reports discomfort experienced after taking medication, the dialogue agent asks further about the timing and manifestation. The management reasoning agent creates a plan to review by cross-referencing the acquired information with relevant guidelines. This example is intended to illustrate the distinction between the two roles. Dividing roles does not mean errors will automatically correct each other, so areas with insufficient evidence must be documented and made verifiable by humans.

Google explained that it cross-references voluminous clinical practice guidelines and drug information using the long-context processing capabilities of Gemini. British National Formulary (BNF) data was also used in the evaluation. In virtual consultation studies, AMIE's management plans demonstrated potential in areas such as adherence to guidelines and specificity of planning. However, results from simulated dialogues do not imply therapeutic efficacy in actual patients.

Pre-Consultation Medical Interviews with Actual Patients

In 2026, a study was published in The Lancet in which AMIE participated in pre-consultation interviews with actual patients. Conducted jointly by researchers from Google and Beth Israel Deaconess Medical Center (BIDMC), patients conversed in text with AMIE at home prior to their appointments. AMIE inquired about symptoms and medical history and compiled possible diagnoses. Medical staff were able to review dialogue transcripts and summaries before the clinical encounter.

In this study, the AI did not independently administer treatment. Physicians monitored dialogues in real time and were able to intervene according to defined safety criteria. Although some healthcare providers evaluated that the summaries assisted in consultation preparation, there was also a need to rectify inaccurate information and explanations.

The significance of this case lies in testing what role medical conversational AI can play in the preparation process for actual medical care. By organizing the medical interview details, physicians can focus on matters requiring additional confirmation during in-person consultations. However, because these results were obtained under supervision, they do not confirm safety or therapeutic efficacy when used without medical staff.

MIRA's Electronic Health Record Integration

MIRA (Medical Intelligence for Reasoning and Action) is an agent developed by researchers from institutions including the National Center for Tumor Diseases at Heidelberg University Hospital and Dresden University of Technology in Germany. Moving beyond explaining patient information, it investigates ways to access electronic health records (EHR) within defined permissions and organize clinical actions.

The research was conducted in a sandbox environment separated from actual hospital systems. A sandbox is a controlled environment designed to prevent experiments from affecting actual operations. The simulated EHR was designed in accordance with FHIR, a standard for health information exchange.

The release date of MIRA's Nature paper is June 17, 2026. FHIR is a standard for exchanging information such as patients, test results, and medications in a consistent structure, not an AI model itself. An agent generating a sentence proposing a test is different from inputting a test order with specified items into the system. The latter must satisfy additional conditions such as patient identification, item selection, and permission verification.

MIRA converses with an agent playing the role of a patient to verify medical history, orders blood tests and imaging exams, and reviews differential diagnoses based on the provided results. It also formulates medication prescriptions or admission plans into action formats that can be processed within the EHR. These actions are performed inside a simulated system and are not executed as tests or prescriptions on real patients.

The patient agent was configured to respond based on actual records from MIMIC-IV, an open critical care and emergency database. In this environment, the researchers evaluated various care processes in the emergency department. Even though real records were used, this must be distinguished from clinical trials involving direct interaction with patients.

Criteria for Evaluating Clinical Actions

The MIRA study examined not only diagnostic accuracy but also treatment planning and medication safety. Review items included dose adjustment based on renal function, drug interactions, allergies, and electrocardiogram QT risk. This shows that when evaluating clinical agents, one must examine the process leading to that diagnosis and the follow-up actions in addition to the final diagnosis.

In the simulated environment, potential was demonstrated across several diagnostic and treatment planning items, but there were also performance differences across diseases. In actual hospitals, more diverse conditions arise, such as missing information, documentation errors, and work interruptions. Therefore, safety and operational methods must be evaluated separately before implementation in real clinical practice. The researchers also emphasized that medical AI agents must not be designed to replace healthcare professionals.

Roles of the Two Systems and Challenges for Adoption

AMIE evolved around patient dialogue and management planning, while MIRA developed centering on clinical actions linked to electronic health records. The ability to understand patient explanations and connect them to precise task processing are both necessary for medical AI. Although there is potential for these functions to be linked in the future, this does not mean the two systems have been validated as an integrated clinical care system.

In adoption, it must be clear which guidelines the AI references. This is because international recommendations, hospital formularies, and insurance criteria may differ. When reference materials are updated, they must be appropriately reflected in the system as well.

Review authority and responsibility for healthcare professionals must also be established. Even if a proposal aligns with guidelines, patient-specific exceptions may be missed; therefore, medical staff must be able to verify the evidence and modify or reject proposals. It is also necessary to distinguish between research-stage systems and approved clinical products. In these studies, AMIE and MIRA were evaluated as research systems.

If AI scribes assist with documentation, AMIE and MIRA expand the scope of support to the interpretation of clinical information, management planning, and clinical actions. For these changes to translate into improvements in actual healthcare, medical supervision, information updates, and integration with hospital workflows must be validated alongside technical performance.

Part 2. AI Drug Discovery and Digital Twins

Chapter 3. Bio Foundation Models and Drug Targets



3.1 Medical Multimodal Foundation Models

The release announcement of Evo 2 from the Arc Institute and Google DeepMind's introduction of AlphaGenome describe the input data and prediction targets of genomic AI. The developing institutions and initial release dates for both models can also be confirmed in the materials below.

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

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

In medicine and life sciences, the same patient or tissue is analyzed using multiple types of data. The genome reveals genetic characteristics, while histopathology images display cell and tissue morphology. Spatial transcriptomics measures gene expression by location within tissue, and medical records and diagnostic reports provide clinical context. Analyzing these data together yields information that would be difficult to discern using only a single type of data.

A multimodal foundation model is a model trained on large-scale data in various formats and applied to multiple analytical tasks. In healthcare, research is underway connecting pathology images with molecular information, and images with clinical records. The cases below primarily feature models that combine two or three types of data.

The foundation aspect of a foundation model means that it first learns representations that can be used commonly across multiple tasks. This distinguishes it from models that repeatedly learn the correct answers for only a single disease. After pre-training, additional training can be conducted or separate predictors can be attached to suit required tasks. However, training on large amounts of medical data does not automatically mean understanding every test and every patient. One must first verify what kinds of data were input and what outcomes the model was trained to produce.

Types of Medical Data

The genome is genetic information including DNA base sequences. It is used to analyze how genetic variations relate to diseases. However, genetic information alone cannot fully determine which genes are actually expressed in a specific tissue.

A histopathology slide is an image observed by staining thinly sliced tissue. In representative H&E-stained images, pathologists examine cell nuclear morphology, tissue architecture, and immune cell distribution. Because digital slides capture large tissue areas at high resolution, the size of the images to be analyzed is also large.

Spatial transcriptomics measures gene expression together with spatial location information within tissue. It allows analysis not only of which genes are expressed, but also where the relevant cells are located and what relationships they have with surrounding cells. Text data includes pathology reports, medical records, disease descriptions, and research papers.

Each data type differs in measurement target and format. Multimodal models handle these differences while learning relationships between data types. Through this, molecular characteristics can be estimated from pathology images, or imaging findings can be interpreted alongside clinical records.

Methods for Combining Different Data

An encoder transforms input data into a vector, which is an array of numbers. Sequence models can be used for DNA sequences, while image models such as vision transformers can be used for pathology images. Graph neural networks are applied when processing spatial relationships, and language models are used for text. Specific configurations vary by model.

A representative method for learning relationships between data types is contrastive learning. It trains the representations of related data—such as images and molecular information obtained from the same tissue—to be close, while separating

representations of unrelated data. The common representation space formed this way is called a shared latent space. Loss functions such as InfoNCE can be used.

A loss function is a calculation method that represents numerically how far a model's current output deviates from the learning target. The model adjusts its internal values in a direction that reduces this value. When linking imaging and molecular data from the same patient, patient identification must be accurate. If test results from different individuals are paired together, the model will learn incorrect relationships. Gathering large volumes of data and pairing them correctly must go hand in hand.

Cross-attention is a method of referencing relevant parts of another data type when interpreting one type of data. For example, a specific region of an image can be linked to a clinical description. However, combining data does not automatically guarantee improved performance, and one must verify which information actually contributed.

Advances in Genomic Models

Evo 2 is a genomic language model trained on vast nucleotide sequences across diverse organisms.

Led by the Arc Institute and NVIDIA in the United States, it was developed with researchers from institutions such as Stanford, UC Berkeley, and UC San Francisco, and was released on February 19, 2025. Here, the basic unit read by the language model is DNA bases instead of words in everyday sentences. After learning what patterns appear in sequences, it analyzes new sequences or variants. Its capabilities in processing long sequences and estimating the impact of genetic variations were studied. Google DeepMind's AlphaGenome focuses on predicting the impact of variants in non-coding regions on gene regulation.

The initial official introduction of AlphaGenome was on June 25, 2025. It takes long DNA sequences as input to predict molecular properties related to gene regulation. By comparing prediction values between sequences with and without variants, it is possible to examine which functions may potentially change. DNA outside regions that directly code for proteins can also be involved in when and to what extent genes operate, making it a subject of research.

These models are evolving into research tools for interpreting the functions of genetic information. However, computational results do not immediately constitute clinical evidence that can be used for determining pathogenic variants or for genetic counseling in patients. For specific tasks such as chromatin

accessibility, dedicated models trained on relevant data may be more suitable.

Connecting Pathology Images and Molecular Information

HistoMethyl is a research example that links pathology slides with DNA methylation information. DNA methylation is a chemical modification related to the regulation of gene expression. In this approach, images and methylation data are provided together during training, and H&E images are used to estimate molecular characteristics during deployment.

This study demonstrates the potential to prioritize and screen patients who require molecular testing using pathology images. Several cancer cohorts (patient groups established under common conditions) were used for evaluation. However, results estimated from images do not replace validated molecular tests. The researchers also focused on initial screening rather than definitive diagnosis.

HistoMethyl is a research model published by researchers including Dannong Wang and Yu Tian of the University of Central Florida in the United States. The corresponding author's affiliation registered on bioRxiv is the University of Central Florida, and the first preprint release date is June 14, 2026. When estimating molecular status from pathology images, 'measured' and 'predicted' must be distinguished in

documentation. Even if output in the same numerical format, results from analyzing actual specimens and values calculated by a model have different generation processes. The design and evaluation conditions of the study can be reviewed in the following original paper: <https://www.biorxiv.org/content/10.64898/2026.06.11.731518v1>

Fuji proposed a direction for jointly learning pathology images, RNA sequencing, and genomic data. Bulk RNA sequencing is a method that co-analyzes RNA from multiple cells contained in tissue, and whole genome data addresses information from the entire genome. Because the cited studies of HistoMethyl and Fuji are in the preprint stage, they must be distinguished from clinical products.

Fuji is a research model in which researchers from the Houston Methodist Research Institute, Weill Cornell Medicine of Cornell University, and the National Cancer Institute in the United States participated. Developed by Qing Li, Shengyu Li, and colleagues, it was released as a preprint on August 6, 2026. It should not be understood as a product of a medical device company simply because the name is identical or similar. The Fuji referred to in this study is a computational model that learns relationships between pathology images and molecular data. The design and evaluation conditions of

the study can be reviewed in the following original paper:
<https://pmc.ncbi.nlm.nih.gov/articles/PMC13484847/>

Combining Spatial Transcriptomics and Text

Although spatial transcriptomics allows detailed analysis of gene expression within tissues, it faces constraints of cost and measurement conditions. Accordingly, research is underway to link relatively widely available H&E images with spatial transcriptomics to estimate location-specific expression patterns from images. However, differences in resolution and measurement methods across platforms and a scarcity of training data remain challenges.

Tx-LLM and TxGemma process molecular structures, protein sequences, cell line information, and disease descriptions in text format to learn various tasks in drug development. This approach connects data formats through a shared language interface.

Tx-LLM is a model developed by researchers at Google and others, with its paper preprinted on June 10, 2024. Based on PaLM 2, it was trained for drug discovery research tasks. TxGemma is a model family created by Google researchers based on Gemma 2, and the related paper was published on April 8, 2025. Even though data expressing molecular structures as text and disease names can be entered together,

such text does not capture all biological states of actual patients. Furthermore, it is not a model that integrates both pathology images and spatial transcriptomics at the patient level.

Combining Imaging and Clinical Records

CGA-Net combined breast ultrasound images with standardized clinical descriptions such as shape and margin. It is a method of learning the correlation between the two types of data through cross-attention. In such studies, the combined model must be compared not only with models using imaging alone, but also with models using clinical information alone. This is because significant information can be obtained from only a subset of the input data.

CGA-Net was developed by a research team including Qin Sun in the Department of Ultrasound Diagnosis at the First People's Hospital of Taicang, Jiangsu Province, China. The research paper was published in *Frontiers in Medicine* on June 9, 2026. Because it is a neural network tailored to a specific breast ultrasound task, it is distinguished from models aimed at serving as a common foundation across multiple domains. Multimodal, which refers to inputting multiple types of data, and foundation models, which are reused across various tasks, are distinct classification criteria. The study's configuration and evaluation conditions can be examined in the following

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

Whether the region of interest (ROI) in an image was predetermined also influences real-world application. The conditions in a study where only cropped lesion areas are input differ from the conditions of processing raw images in clinical practice. Data leakage, where evaluation information becomes mixed into training, must also be prevented.

In cervical cancer research, a general-purpose multimodal model was used to estimate the risk of recurrence following chemoradiotherapy. However, this was a retrospective study analyzing completed medical records, and it also included information ascertainable only after treatment. Therefore, it should not be interpreted as a predictive tool prior to the initiation of treatment. Errors in the image interpretation process also persist, necessitating review by medical professionals.

Challenges in Performance Evaluation and Generalization

Large-scale models are not superior to existing methods across all tasks. In a study comparing single-cell foundation models, baseline models applying normalization and linear methods also demonstrated high performance. Similar results have been reported in tasks predicting responses to genetic perturbation—that is, altering gene function or the cellular environment.

Therefore, new models must be compared against appropriate baseline methods. Out-of-distribution evaluations involving cell types or biological species not present in the training data are also crucial. Confirming suitability for the actual purpose of analysis takes precedence over the scale of the model.

When combining multiple data modalities, a modality gap can also emerge. This is a phenomenon where representations of different data types are not sufficiently connected even when trained in a shared space. Architectural designs that preserve the characteristics of distinct data while learning their relationships are necessary.

Batch effects are measurement discrepancies arising from differences in institutions or experimental conditions. Staining reagents and scanners differ across hospitals, as do transcriptome measurement protocols. To apply a model

trained at one institution to another, these variations must be examined and evaluated using external data.

For instance, suppose one hospital imaged specimens of a particular disease predominantly with one scanner and other specimens with a different scanner. The test scores could be high even if the model merely distinguishes the image characteristics left by the machine rather than the disease itself. To verify whether it has learned true disease features, performance must also be examined on data where devices and institutions vary. This example illustrates the impact of data collection methods on evaluation.

Role in Precision Medicine

Multimodal foundation models can contribute to understanding diseases and prioritizing analyses by connecting disparate tests and records. An example is the approach of screening targets that require additional molecular testing from pathology images.

However, computational performance must be distinguished from patient clinical outcomes. Whether research predictions translate into reduced misdiagnoses or improved survival must be confirmed in clinical settings. It is appropriate to understand current cases as serving to support the judgments of medical staff and researchers. A human-in-the-loop

structure, where humans review the results, and reliable follow-up testing are both required.

3.2 Octopus's Exploration of Colorectal Cancer Biomarker (IGF2)

The preprint paper on Octopus by Christopher Baker and colleagues describes a framework connecting hypothesis generation with the analysis of public biological data. This study is an in silico candidate search and is distinct from clinical trials of new treatments.

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

In drug development, deciding which candidates to investigate first among numerous genes and molecules is critical. AI can support the candidate selection process by analyzing existing biological data and proposing hypotheses. Developed by researchers at Queen's University Belfast, Octopus is an example that combines multiple AI agents and predictive models to explore candidates associated with drug response in colorectal cancer.

The formal name of Octopus is Multi-Scale Autonomous Discovery Engine. Introduced in a 2026 preprint, it is a study co-authored by Christopher Baker, Tianyu Ren, Karen Rafferty, Hui Wang, and Simon McDade. Rather than denoting a single language model, the name refers more closely to a research framework that bundles multiple analytical tools and agents.

The candidate presented in this study is insulin-like growth factor 2, or IGF2. The researchers evaluated IGF2 as a candidate biomarker to predict response to 5-fluorouracil (5-FU). A biomarker is an indicator used to assess disease status or treatment response, and is distinct from a therapeutic target intended for direct treatment.

Variations in Colorectal Cancer Treatment Response

5-FU has long been used in chemotherapy for colorectal cancer. However, even when administering the same drug, responses vary from patient to patient. Identifying biological characteristics that account for these variations helps in selecting appropriate treatments or determining directions for follow-up research.

In conventional research, researchers have formulated hypotheses, analyzed gene expression data, and verified them through cellular and animal experiments. Because numerous genes interact, there are many candidates to analyze, and interpreting the association between a single gene and drug response is challenging.

While large language models (LLMs) can generate hypotheses based on literature and data, a fluent explanation does not guarantee biological validity. To reduce hallucinations that

generate counterfactual claims, a process of cross-checking hypotheses against real data is necessary.

Combining Multiple Agents and Predictive Models

In Octopus, multiple agents analyze colorectal cancer transcriptome data and propose candidate hypotheses. It was configured to run open, small-scale models within the laboratory's local environment. By adopting an architecture that avoids transmitting patient data to external services, it also relates to institutional data governance.

The generated hypotheses are evaluated using tools such as XGBoost and SHAP.

The landmark paper for XGBoost was preprinted by Tianqi Chen and Carlos Guestrin on March 9, 2016. SHAP is a prediction explanation method preprinted by Scott Lundberg and Su-In Lee on May 22, 2017. Rather than being newly invented by Octopus, these two tools are established, publicly available computational methods linked to suit research objectives. XGBoost is a machine learning method that predicts by combining multiple decision trees. SHAP is a technique that explains how much each input feature contributed to a specific prediction. In this structure, the language model proposes candidates, while a separate,

interpretable predictive model examines their relationship with the data.

A decision tree is a predictive structure that partitions data according to conditions. XGBoost combines the outputs of multiple trees to compensate for deficiencies in earlier predictions. SHAP calculates which inputs contributed to increasing or decreasing the predicted outcome. The fact that a feature contributed significantly to a prediction does not equate to biological causality. This is because one of several correlated genes might simply be acting as a proxy carrying information about another gene.

Public data at different levels were used for the evaluation. In cell line data, the impact of suppressing gene function on cancer cell survival was examined. In human colorectal cancer cohorts, the association between gene expression and survival was evaluated, while in patient-derived xenograft (PDX) mouse data, the association with tumor changes was assessed.

Cross-referencing candidate hypotheses proposed by researchers across multiple datasets can reduce conclusions that rely on only a single dataset. However, even if associations across datasets are confirmed, this does not establish causality proving that the gene directly causes the therapeutic response.

Cell lines, patient data, and mouse data answer different questions. It is uncertain whether responses observed in cell lines manifest to the same extent in humans, and whether features associated with patient survival are the cause of drug responses must be evaluated separately. When multiple datasets point in the same direction, the rationale for follow-up research strengthens, but the inherent limitations of each dataset do not disappear.

Significance of the IGF2 Candidate

Octopus suggested IGF2 as a priority candidate for review through unsupervised analysis. Associations between IGF2 expression and prognosis or drug response had previously been reported in existing patient and animal data.

IGF2 itself is not a gene discovered for the first time in this study. Prior studies had examined the link between loss of imprinting of IGF2 and disease risk in colorectal cancer. Imprinting is a phenomenon that regulates the expression of a specific copy among the two gene copies inherited from parents. The accomplishment of Octopus lies not in discovering a new gene, but in computationally selecting candidates for further research from among numerous possibilities.

Subsequent validation is required to utilize a biomarker candidate therapeutically. It must be confirmed whether the therapeutic efficacy of 5-FU varies according to IGF2 levels, and to develop it as a therapeutic target, it must also be investigated whether modulating the IGF2 pathway alters the disease trajectory.

This distinction is crucial when translating research findings into patient explanations. Prognostic indicators are features associated with future disease course, whereas predictive indicators of treatment response distinguish which patients will derive greater benefit from a particular therapy. While the two can overlap, they are not always identical. Therefore, an announcement that an associated gene has been identified must not be immediately translated into the claim that testing this gene enables drug selection.

Difference Between Analyzing Existing Data and New Experiments

The Octopus study is a retrospective computational study reanalyzing publicly available data. The researchers did not newly administer IGF2-inhibiting drugs to mice. Using existing PDX data, they evaluated whether predictions and actual observations aligned in the same direction.

Computational analysis (in silico) helps narrow down candidates and formulate experimental plans. However, it cannot replace new cell experiments, animal experiments, and clinical trials. As this study is also at the preprint stage prior to peer review, it should not be presented as an established therapeutic achievement.

Preceding Research and Advances in Study Design

The same research team had previously proposed the contextual invertible world model (CIWM). That study explored how prediction results change when hypothesizing genetic alterations using colorectal cancer cell lines and drug response data.

In such simulations, unexpected responses may emerge when a specific mutation or alteration is modified. This is because cellular responses result from the coordinated action of multiple pathways. However, changes calculated by a model must be distinguished from changes confirmed in actual cells. Taking these limitations into account, Octopus employed an approach of cross-checking multiple datasets and prioritizing candidates.

Role in the Drug Discovery Process

Octopus demonstrates an AI research framework that bridges hypothesis generation and data analysis. Researchers can review candidates proposed by AI and judge the necessity and priority of follow-up experiments. Running open models locally is also meaningful in terms of data management and choice of research environment.

Subsequent tasks include replicating the results on independent datasets and confirming candidate functions through cellular and animal experiments. Clinically, it is also necessary to distinguish whether it is an indicator predicting treatment response or a target that can be modulated through therapy. The role of AI does not lie in bypassing such validation, but in helping systematically select candidates to be validated.

3.3 Cellular World Model (Chreode) and Molecular Design

The paper published by the Chreode research team outlines input cellular states, intervention conditions, and the computational methods for state transitions. It details both the condition that the training data centered on mouse embryos and the scope across which it was tested when transferred to other tasks.

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

In drug development, alongside approaches that modulate the function of specific proteins, approaches that analyze changes in cellular states associated with disease are utilized. This is because in diseases involving multiple biological pathways, it is difficult to account for therapeutic responses through the action of a single target alone.

The transcriptome represents RNA information expressed in cells. Gene expression patterns associated with diseases are called disease signatures, and one can analyze how these patterns change after drug treatment. Connectivity Map and L1000 are representative datasets and measurement systems used in such research.

Interventions that administer drugs to cells or alter gene function are called perturbations. A cell world model or virtual cell is a model that attempts to predict changes in cellular state resulting from such interventions. It can be used to calculate what reactions candidate substances will induce, thereby guiding the direction of follow-up experiments.

Predicting Changes in Cellular State

There are measurement constraints in analyzing cellular states over time. Because single-cell RNA sequencing destroys cells during the measurement process, it is difficult to continuously observe the same cell. Researchers estimate the process of state change using data from different cells measured at multiple time points.

RNA velocity analysis and Neural Ordinary Differential Equations (Neural ODE) have been used to study such transition processes. However, computational costs and the burden of retraining on new data can arise. Recent models attempt to apply pretrained representations across multiple cells and analytical tasks.

Chreode's Approach

Chreode is a model that predicts the next state based on cellular state and intervention conditions.

It is a collaborative study by researchers from the University of North Carolina at Chapel Hill, the University of Texas at Austin, and the University of Pennsylvania, and was released as a preprint on May 27, 2026. The input includes not only the current state of the cell but also the hypothesized intervention and the time interval. The output is the predicted next cellular state under those conditions. Even though it is named 'world,' it does not mean that every molecule within a cell and the entire human body have been replicated as they are. The name originates from 'chreode,' which developmental biologist Conrad Waddington used to explain cellular differentiation pathways.

This model is designed to incorporate the computations required for state transitions into the training process, and when deployed, to generate the next state through a single-step transition operation. The transition operator represents directional changes, rotational components, and stochastic fluctuations together.

The transition operator is a computational rule that converts the current numerical representation into the numerical representation of the next state. Because responses can vary from cell to cell even under the same initial conditions, it is difficult to explain the changes with a single fixed answer. The model learns these patterns of change from data. The

explanation that it produces results in a single calculation also does not mean that the actual cellular changes conclude instantaneously, but refers to the computer's computational procedure.

Large-scale single-cell data from mouse embryos were used for training. An scVI encoder was applied to the part that represents cells as vectors, and a DiT (diffusion transformer) architecture was applied to the part that computes state transitions. The corresponding paper on Chreode is a preprint—that is, research disclosed prior to formal peer review.

The preceding research for scVI traces back to the single-cell gene expression model preprinted in 2017 by researchers including Romain Lopez and Nir Yosef. DiT is an architecture originating from diffusion transformer research published by William Peebles and Saining Xie on December 19, 2022. Chreode combined these existing components to suit the objective of cellular state transitions.

Application to Other Tasks

The Chreode study evaluated whether representations learned during developmental processes are also helpful for other cellular analyses. It predicted changes in cell distribution in hematopoiesis and pancreatic islet differentiation data, and provided learned embeddings to GEARS, a gene perturbation prediction model. In some evaluations, prediction error decreased.

GEARS is a gene perturbation prediction method developed by Yusuf Roohani, Kexin Huang, and Jure Leskovec's research team at Stanford University. The Nature Biotechnology paper was published online on August 17, 2023, and appeared in an issue in 2024. It predicts post-intervention expression changes by combining relational information between genes with deep learning.

Zero-shot evaluations predicting cellular differentiation outcomes without retraining were also conducted. Zero-shot is a method of applying previously learned representations without separate training for the task at hand. These results demonstrate the possibility that information obtained from cellular development can contribute to analyzing other state transitions. However, improvements in specific evaluations cannot be generalized to all cells or drug responses.

Diverse Designs of Cellular Perturbation Models

SCALE addresses the problem that it is difficult to directly pair and measure the same cell before and after treatment. It treats cells as a set to predict the state of the population after intervention. Various conditions such as genetic interventions, drugs, developmental, and immune responses were targeted for analysis. Tahoe-100M, a large-scale perturbation dataset, was also released as a research resource for training and evaluating related models.

SCALE was developed by researchers from institutions including Shanghai Artificial Intelligence Laboratory, Shanghai University, and Fudan University, and the paper was first disclosed on March 18, 2026. The revised version on August 31 also describes follow-up experiments. Tahoe-100M is a distinct data resource. Distinguishing between the name of the model and the name of the data used for training clarifies which study proposed a new computational method and which study provided new observational data.

PT-RAG applies retrieval-augmented generation (RAG) to cellular response prediction. When predicting the effects of a new intervention, it searches for and references relevant past experiments. Because the effect of the same gene can vary depending on the cell type, it retrieves data by considering

both the similarity of the intervention and the control cell state.

The first paper on PT-RAG was preprinted on March 7, 2026, by Andrea Giuseppe Di Francesco, Andrea Rubbi, and Pietro Liò, affiliated with Sapienza University of Rome and ISTI-CNR in Italy, and the University of Cambridge and the Wellcome Sanger Institute in the United Kingdom. It is a component attached to existing prediction models to search for relevant experiments. If the retrieved data do not match the current cellular conditions, they may be of little help, so retrieval and prediction must be evaluated together. The structure of the study and evaluation conditions can be examined in the following original paper: <https://arxiv.org/html/2603.07233v1>

These models commonly attempt to apply information from existing experiments to new conditions. In actual research, how well they respond to cells or interventions not included in the training data is crucial.

Evaluation Methods and Generalization

The Arc Institute's Virtual Cell Challenge is an evaluation that publicly compares cellular perturbation predictions. It predicts responses following knockdown, which weakens gene function, and examines whether models operate even under new cellular conditions.

Related comparative studies have reported that complex models do not surpass linear baselines across all metrics. On the other hand, follow-up research indicates that adjusting evaluation methods reveals the advantages of deep learning models. Therefore, it is difficult to judge performance based solely on model scale or a single score. Multiple aspects must be evaluated together, including not only expression levels but also response direction and the distribution of cell populations.

Molecular Design Tailored to Desired Cellular Changes

While cellular response prediction calculates the outcome of a specific intervention, molecular design is the process of finding candidates that will produce a desired outcome. An approach that generates new molecules conditioned on desired transcriptomic changes is called de novo molecular design.

CURE extracts functional features using transcriptomic information before and after drug treatment and connects them to molecular generation. A transcriptomic perturbation feature extractor (TFE) creates a representation of biological changes, and a diffusion model references this to generate molecular structures. It is significant in that it uses cellular responses as design conditions in addition to the three-dimensional structure of target proteins.

CURE involved researchers affiliated with the Institute of Automation of the Chinese Academy of Sciences, the University of Chinese Academy of Sciences, and the National University of Singapore, and the paper was preprinted on May 14, 2026. This name stands for Cellular Response Engine in the paper. Although it outputs candidate molecular structures using desired cellular changes as input conditions, follow-up verification is needed to determine whether the output structures are actually synthesizable and safe.

JoPMol jointly reflects cellular state, design strategies presented in text, and numerically represented chemical properties. This is because, even if a biologically desired change is induced, the chemical properties of the molecule might not be appropriate. It is an approach that seeks to generate molecules by simultaneously controlling multiple conditions.

JoPMol was released on July 13, 2026, as a collaborative study between researchers from South China Normal University and Osaka University. For example, if researchers simultaneously propose conditions for desired cellular responses and chemical properties, it searches for candidates that fit that combination. Because other properties may deteriorate in the process of optimizing only one condition, it is important to verify the balance among multiple conditions.

The results of CURE and JoPMol were primarily reported in computational evaluations. The computational evaluations in those papers alone cannot be regarded as confirming the synthesis, animal testing, or clinical efficacy of the generated molecules.

Challenges in Connecting with Experiments

Not all gene expression changes associated with a disease are the causes of the disease. Some may be defensive responses of cells. Therefore, it must be verified whether shifting expression patterns closer to a normal state is genuinely helpful for treatment.

Responses of cultured cell lines and responses inside the human body can also differ. The tissue environment, immune system, and drug distribution and metabolism influence actual efficacy. In data where a limited set of genes is measured and the rest are inferred, there is also the possibility that specific signals may not be sufficiently captured.

Connecting cell world models and molecular generation models can support candidate screening and experimental design. However, computational results are merely starting points for follow-up validation. A process of synthesizing candidates, verifying efficacy and toxicity in cells and animals, and clinically evaluating them is necessary. The advancement

of this technology can contribute to drug discovery when improvements in computational model performance and validation through experimental data take place together.

Chapter 4. Health Digital Twins and Clinical Trials



4.1 Digital Twins Reflecting Patient Characteristics

The original paper on PRECISE CURATE.AI by Singapore researchers contains the process of generating dosage recommendations from individual patient data and the results of examining whether it can be applied to actual clinical practice. It can be read by focusing on what decisions were supported, rather than on the notion of replicating the entire patient.

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

A digital twin is a technology that represents the state of a real-world object as a computer model and predicts changes or the outcomes of interventions by reflecting new data. In engineering, it has been used to analyze the condition of equipment and formulate maintenance plans. In healthcare, research is directed toward simulating disease progression, treatment response, and clinical trial outcomes by reflecting the characteristics of patients or organs.

Health digital twins can link genetic information, imaging, medical records, sensor data, and more. However, their intended purposes and implementation levels vary. It does not mean a system that completely replicates all biological characteristics of a patient.

When building a model, one first narrows down the question to be answered. A model intended to review drug dosage and a model intended to observe heart valve motion require different data. Even if there is a function that displays the patient in three dimensions on screen, it does not mean it can predict treatment responses. Conversely, a model that presents results solely through numbers and graphs without 3D illustrations can still be helpful for specific predictions.

Connecting the Model with the Actual Patient

In conventional classification, a digital model is a model built from past data and is not automatically linked to the real-world object. A digital shadow is a form where new information from the real-world object is transmitted to the model. A digital twin aims for a structure that links the states of the real-world object and the model, reflecting prediction results into decision-making.

In healthcare, components are described as the patient, data connectivity, patient-in-silico, user interface, and synchronization. It is critical that when the actual patient's condition changes, this must also be reflected in the model. However, the definition and implementation scope of digital twins vary across fields.

For reliable predictions, verification, validation, and uncertainty quantification (VVUQ) are required. This is the process of confirming whether the model is properly implemented, whether it suits the intended purpose of use, and what degree of uncertainty exists in the predictions.

The three questions must be verified separately: verifying whether mathematical formulations and programs compute as intended, comparing whether computational results match actual observations, and examining how much errors in data and assumptions perturb the results. Even if there are no errors in the program, predictions can be inaccurate if biological assumptions are incorrect. Passing just one test does not mean the remaining conditions have been met.

Scope of Application

Tumor models analyze genetic mutations, cellular characteristics, drug response, and resistance. Organ models simulate the structure and function of organs such as the heart or liver, as well as the effects of drugs or radiation. At the patient level, state changes can be estimated by linking medical records and wearable device data.

In clinical trials, they can be used to review patient recruitment, dose selection, and trial design. There are also approaches that assist in analyzing the actual control group by

predicting the trajectory in the absence of treatment. Linkage with adaptive designs is also being studied, but evaluations and regulatory reviews tailored to each use are required.

Data-Driven Models and Mechanism-Based Models

Data-driven models learn patterns from large volumes of patient data. While they are well-suited for analyzing complex imaging and genomics, they may have limitations in explaining the rationale behind judgments or in being applied to patients who differ from the training data.

Mechanism-based models mathematically express known biological relationships. These include pharmacokinetics, which deals with how drugs are distributed and eliminated in the body, and pharmacodynamics, which addresses the actions of drugs. Although their interpretation is relatively clear, it is difficult to express all the processes of complex diseases in mathematical formulas.

Hybrid models combine the two approaches. This method seeks to compensate for each other's limitations by using both known mechanisms and patient data. The choice of model depends on what clinical decision it is intended to support.

Individualized Dose Adjustment: CURATE.AI

CURATE.AI from a Singaporean research team is an approach that adjusts individualized drug doses using data obtained from a single patient.

The development was led by Dean Ho's research team at the National University of Singapore. The university had already introduced the clinical research of CURATE.AI on August 31, 2018, and the paper presenting the results of the PRECISE study was published on February 21, 2025. The appearance of a new paper does not mean the platform emerged for the first time in that year. The inputs are the patient's dosage and response data, and the output is an individualized dose recommendation for medical staff to review. The PRECISE study, conducted on advanced solid tumors, examined whether AI dose recommendations could be applied in real-world clinical care during capecitabine monotherapy or combination therapy.

This case represents an attempt to incorporate individual patient responses into treatment adjustments, rather than relying solely on the average response of a large population. However, because it was a feasibility study, it should not be interpreted as proving improvements in survival or treatment response.

Sufficient observation is also required for individualized models. The dose-response relationship can only be adjusted after measuring the response following a change in dosage. If the patient's condition or concurrent treatments change, it is also necessary to verify whether previously established relationships hold. For this reason, the process by which medical staff determine input data and acceptable ranges becomes part of operating the model.

Lifestyle Management: Twin Health

The Cleveland Clinic conducted a study applying the Twin Health platform to the management of patients with type 2 diabetes. This approach collects data on blood glucose, weight, blood pressure, activity, and sleep, providing personalized dietary and activity guidance.

The developer is the US-based Twin Health, and the hospital is the institution conducting the study. The Cleveland Clinic's announcement on August 20, 2025, refers to this program as Twin Precision Treatment and explains that it provides data obtained from devices alongside human remote coaching. Therefore, it is difficult to isolate and interpret the changes observed in the study solely as the effect of the AI algorithm. The announcement date should be viewed as the point in time when the study results were made public.

This approach integrates information repeatedly gathered from daily life into management plans. It is a lifestyle intervention study distinct from the clinical phases of new drug development, and the platform's functions must be evaluated separately from the effects observed in patients.

Organ Models: The Living Heart Project

Dassault Systèmes' Living Heart Project uses a three-dimensional model of the heart to study cardiac function and the impact of medical devices. Collaboration with the FDA was also conducted as part of research to examine virtual trial methods.

French software company Dassault Systèmes unveiled the heart simulation model of the Living Heart Project on May 20, 2014. This is an example of a computational model that had been developing long before the emergence of generative language models. Because the category of digital twins also includes models that compute physical and physiological laws, calling all digital twins language model-based AI is inaccurate.

Organ models can be helpful for reviewing designs and conditions prior to actual trials. However, research collaboration is distinct from adoption as an official review tool by regulatory agencies. To use a computational model as an

evidentiary basis for review, credibility assessment tailored to the intended purpose is necessary.

Clinical Trial Simulation: OPTIVIS

OPTIVIS, from the South Korean company OPRIMED, is a platform that simulates clinical trials using virtual patients reflecting patient characteristics and disease progression. A case study optimizing sample size using Alzheimer's disease clinical data was presented.

The company emphasizes an approach that incorporates Asian patient data, as data skewed toward a specific population may not adequately explain trajectories in other populations. However, the introduction of the platform must be distinguished from actual cost and time reduction effects in clinical trials.

In its official product introduction, OPRIMED presents virtual patient generation, external control arm design, and predictions of clinical trial duration and patient numbers. OPTIVIS is the company's platform name, and predictive models for specific diseases can be understood as individual application cases within it. Since the product introduction does not clearly state an initial release date, it is difficult to identify a specific year as the initial launch year.

Clinical Trial Analysis Support: PROCOVA

Unlearn's PROCOVA is a method that predicts the control-condition trajectory of patients participating in actual randomized clinical trials and uses it for statistical adjustment. The European Medicines Agency (EMA) issued a qualification opinion for this method.

This case shows that the application of digital twins is expanding beyond treatment recommendations to trial design and statistical analysis. It does not mean that all actual control groups are eliminated or that new drugs are approved based solely on synthetic data. Specific application conditions will be examined in the next section.

Challenges in Adoption and Validation

From a regulatory perspective, intended use, data quality, risk level, model performance, and management methods are critical. Different levels of evidence are required for models that directly influence treatment choices versus models that assist study design.

Domain shift, where performance varies when applied to different hospitals or populations, must also be evaluated. Data standards, integration with hospital systems, privacy protection, and continuous updates are also necessary conditions for operation. Even when real-world data is

abundant, differences in patient selection and measurement methods can alter comparative results.

Current cases focus on specific purposes such as individualized dose adjustment, lifestyle management, organ simulation, and clinical trial analysis. The clinical value of digital twins must be evaluated by clearly defining their scope of application and validating how predictions assist actual decision-making and patient outcomes.

4.2 Synthetic Control Arm Clinical Trials

Unlearn's official technical description distinguishes PROCOVA, which aids in the analysis of actual randomized trials, from external control arms. The scope of regulatory acceptance can be confirmed in the EMA's qualification opinion.

<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

The control group in a clinical trial serves as the baseline against which the effectiveness of a new treatment is compared. Depending on the disease and trial objectives, they receive standard care or a placebo, and researchers compare results between the treatment group and the control group to evaluate efficacy and safety. The randomized controlled trial (RCT) is an important method for assessing treatment effects by minimizing differences between groups.

However, recruiting control groups requires time and expense. In rare diseases with few patients, securing a sufficient number of participants is difficult, and the selection of comparative treatments also requires ethical scrutiny. To address these challenges, approaches using existing patient data and predictive models in trial design and analysis have been advancing.

Adjustment Using Predicted Values and External Control Arms

The term synthetic control arm encompasses different methods. One is an adjustment method that utilizes AI predictions in analysis while maintaining an actual randomized control group. The other is a method that uses external data, such as historical clinical trials or medical records, as the comparative baseline. The two methods differ in design and bias management requirements.

In the adjustment method, a prognostic score is created by predicting each patient's trajectory had they received the control treatment. Incorporating this into statistical analysis can account for some of the differences among patients. Unlearn's PROCOVA is a prime example.

PROCOVA, proposed by the US company Unlearn, is the name of a statistical method known as prognostic covariate adjustment. A distinction must be made between the machine learning model that generates the prognostic score and the method of incorporating that score into trial analysis. The AI calculates the anticipated trajectory for each patient, while the statistical analysis evaluates the observed differences between the actual treatment and control groups. It is not structured such that the AI's prediction is used as the definitive answer for treatment efficacy.

An external control or synthetic control constructs a comparator from data outside the current trial. For example, the results of a single-arm trial, in which all participants receive the new treatment, can be compared with the trajectories of historical patients. However, because the cohorts may differ from the outset, selection bias and measurement discrepancies must be addressed separately.

How PROCOVA Works

Clinical trial outcomes are influenced not only by treatment effects but also by patients' baseline status and the rate of disease progression. Predicting these differences in advance and incorporating them into analysis can reduce the variance of the results. Under appropriate conditions, it is possible to achieve the targeted analytical precision with fewer patients.

PROCOVA uses predicted values learned from historical control data for covariate adjustment. It is structured to analyze treatment effects while maintaining actual randomization. The actual number of patients required and the magnitude of efficiency gains vary depending on the disease, endpoints, predictive accuracy, and trial design.

Covariates are characteristics taken into account during analysis because they can affect the outcome. Information that can be known before trial initiation, such as age or baseline

pre-treatment status, can be included here. A prognostic score summarizes the expected trajectory from multiple characteristics into a single piece of information. If the extent to which the prediction helps is small, the benefit of reducing patient numbers may also diminish. This is why one cannot claim that the sample size will be reduced by a fixed percentage in any trial.

The European Medicines Agency (EMA) issued a qualification opinion for PROCOVA as a statistical method for analyzing Phase 2 and 3 randomized clinical trials with continuous outcomes. This signifies the endorsement of an analytical method under specific conditions and does not imply that new drugs are approved based solely on synthetic data.

The EMA document was published on September 20, 2022, and the date of adoption by the committee was the 15th of the same month. Continuous outcomes in the document refer to results with continuous numerical values, such as scores or measurements. It should not be broadly interpreted as a document endorsing all types of endpoints and all statistical methods at once.

The US FDA has also codified principles of covariate adjustment in randomized clinical trials into guidance. However, general analytical principles must be distinguished

from separate approvals of specific commercial methods. In practical application, consultation with regulatory authorities is required in accordance with the trial objectives and statistical analysis plan.

External Control Arms and Digital Twins

The approach of using external data as a comparator existed prior to AI. The international guideline ICH E10 also addresses the advantages and disadvantages of external controls. In situations where randomized comparison is challenging, such as in rare diseases, natural history data or historical trial results have been examined.

AI and digital twins attempt to use external data to predict control trajectories tailored to the characteristics of current trial participants. This approach incorporates individualized information rather than relying solely on group averages. However, even if predictions become sophisticated, biases arising from the absence of randomization do not automatically disappear.

For example, if patients in historical data had less advanced disease than patients in the current trial, part of the difference in trajectories might be unrelated to treatment effects. Even if age and laboratory test values appear similar, unrecorded health conditions or reasons for treatment selection may

differ. Unmeasured differences are also difficult to adjust for in statistical software. The mere fact that external data is abundant does not guarantee a valid comparison.

Regulatory authorities review the suitability of external data, disease trajectories, differences between comparative cohorts, and outcome measurement methods. Although relevant guidelines and discussion documents are being established, they do not serve as an interchangeable substitute uniformly applicable to all diseases and trials.

Limitations Revealed by Real-World Data

RCT DUPLICATE is a study that examined how well the results of existing randomized controlled trials could be replicated using real-world data. It constructed comparison cohorts using insurance claims data and other sources, adjusting for measurable differences using propensity scores.

While similar results were obtained in some trials, not all conclusions matched. Concordance was high when the eligibility criteria, treatments, and outcome measurement methods of the original trials could be closely emulated. Discrepancies could widen when essential information was missing from medical records or when it was difficult to replicate a placebo control.

These results demonstrate that data fitness is crucial even when constructing synthetic control arms. Beyond the performance of algorithms, one must verify whether the underlying source data can answer the trial's questions.

It is also important to establish outcome measures, comparison time points, and patient selection criteria prior to analysis. Selecting only favorable patients or altering measurement time points after observing results can make treatment effects appear exaggerated. Data used to train the model must be separated from trial outcome data, and an analysis plan must be documented so that other researchers can reproduce the same comparison.

Areas Under Consideration for Application

In rare diseases, there is a strong need for external control arms due to difficulties in patient recruitment. However, a small patient population does not eliminate the limitations of comparative data, meaning appropriateness must be assessed for each individual disease.

In neurological disorders such as Alzheimer's disease, approaches aimed at predicting individual differences in disease progression to enhance analytical efficiency are being considered. The clinical trial simulation introduced by OPRIMED of South Korea is a relevant example. However,

simulated designs utilizing historical data differ from outcomes actually achieved in new trials.

In oncology drug development as well, research applying synthetic control arms and digital twins to trial design is underway. Cases of inter-company collaboration or educational design examples must not be exaggerated as actual regulatory approval achievements.

Points to Verify in Trial Design

Historical data reflects standard care at that time. If treatment regimens or diagnostic criteria have evolved in the interim, direct comparison with current trials becomes difficult. While clinical trials measure outcomes at predefined time points and via specified methods, medical records may not. Missing information and differences in patient selection must also be taken into account.

Therefore, the purpose and generation method of the synthetic control arm, data sources, and analytical uncertainties must be clearly presented. The roles and review responsibilities of developers and trial-conducting organizations must also be established in advance.

The significance of synthetic control arms lies in reducing the burden of clinical trials and utilizing existing data more

systematically. The primary current approach is to supplement actual randomized trials with predictive information. When attempting to replace a control arm with external data, comparability and bias must be thoroughly examined. Securing both recruitment efficiency and the reliability of evidence remains the core challenge.

4.3 Model-Informed Drug Development and Animal Testing Alternatives (FDA Modernization Act)

The FDA's NAMs overview and MIDD Paired Meeting guidance outline the scope of new testing methods and the conditions for utilizing models in drug development decisions. Reading both documents together allows one to distinguish the distinct roles that human cell-based experiments and computational predictions play in drug development.

<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>

In drug development, substantial time and resources are required to verify efficacy and safety. The fact that results obtained in animals do not translate directly to humans also poses a development challenge. Interspecies differences exist in drug absorption and metabolism, immune responses, and the mechanisms of therapeutic targets.

As an approach to address these challenges, the use of human cell-based assays and computational models is expanding. New Approach Methodologies (NAMs) diversify nonclinical evaluation methods. Model-Informed Drug Development

(MIDD) combines data with mathematical models to support decision-making throughout the development process.

NAMs and MIDD are not AI product names released by a company. NAMs is a category of testing methods, while MIDD is an approach that integrates information obtained from models into development decisions. An organ-on-a-chip replicates partial functions of human tissue within a small device, and computational models compute data and assumptions. Because they provide different types of evidence, the advantages of one method cannot simply be substituted for another without validation.

Expansion of Alternatives to Animal Testing

The FDA Modernization Act 2.0 revised statutory language to allow nonclinical tests to include methods such as cell-based assays, organ-on-a-chip, and computational models. It did not ban animal testing, but rather clarified the statutory basis for employing suitable alternative methods. In actual drug development, animal studies continue to be used whenever necessary.

Subsequent regulatory adjustments represent a process of translating the legislative intent into detailed regulations and review practices. Discussions surrounding the FDA Modernization Act 3.0 and the FDA's roadmap for reducing

animal testing also belong to this trend. However, legislation, enacted regulations, and draft guidances differ in their legal status and must be distinguished from one another.

FDA materials clarifying the September 2026 terminology adjustments also state explicitly that this change does not mandate specific test methods or lower evidentiary standards. Policy directions aimed at reducing animal testing must be interpreted alongside determinations of the safety data required for individual new drugs. One should not assume that all animal studies have disappeared simply because terms like modernization or replacement appear in the title.

NAMs include computational models predicting toxicity and pharmacokinetics, organs-on-chips simulating human tissue functions, and organoids derived from stem cells. Their application is also being explored in fields where target differences between animals and humans are critical, such as monoclonal antibodies.

The purpose of these methods is not to eliminate specific tests unconditionally. Securing data appropriate to the risks being assessed remains the priority. New methods must also be validated for reproducibility, predictive capability, and applicability, and the required evidentiary standard is not lowered.

Concept of Model-Informed Drug Development

MIDD supports development decisions through modeling and simulation integrating nonclinical data, clinical data, and prior knowledge. It addresses diverse questions, including initial human starting doses, appropriate dosing intervals, dosing in special patient populations, and drug-drug interactions.

Pharmacokinetics addresses the processes by which the body absorbs, distributes, and eliminates a drug, whereas pharmacodynamics addresses what changes a drug induces in the body. Even when identical doses are administered, the resulting drug concentrations remaining in the body can vary, and identical concentrations can elicit different responses. Connecting these two relationships is necessary to interpret dosage numbers in conjunction with patient response.

Population pharmacokinetics and pharmacodynamics (PopPK/PD) estimate differences in drug concentration and response based on age, weight, organ function, and other factors. Physiologically based pharmacokinetics (PBPK) calculates internal drug distribution using information such as organ size, blood flow, and metabolic enzymes. Quantitative systems pharmacology (QSP) models interactions between drugs and biological signaling pathways. AI and machine learning (AI/ML) can also be utilized for related data analysis and prediction.

These methods help connect existing data to identify areas requiring additional testing and refine trial design. However, the extent to which model predictions can be used for specific decisions depends on data quality and the level of validation.

For example, in a model evaluating dosing intervals, how drug concentration changes over time following administration serves as core data. If a model was constructed using data from a specific patient population, one must examine whether the same relationship holds in other populations. Predictions that fall further outside the actually measured range may require additional supporting evidence. When reporting calculated values, the underlying assumptions and data scope must be documented alongside them.

Consultations with Regulatory Agencies

The FDA's MIDD Paired Meeting Program is a mechanism through which drug developers discuss model-informed approaches with regulatory experts. Consulting early in development regarding model objectives, assumptions, and evaluation methodologies aids in formulating subsequent evidence generation plans.

ICH M15 is a guideline establishing common standards for the evaluation and documentation of MIDD. It emphasizes an approach of determining the required level of validation based

on the context of use of the model and its impact on decision-making. AI guiding principles in drug development presented by the FDA and the EMA also highlight human-centric design, data governance, risk-based assessment, and ongoing lifecycle management.

The application of specific laws and regulations varies depending on whether documents are finalized and on jurisdiction. A company's consultation experience or introductory description of a methodology must not be broadly construed as general regulatory approval.

The Link Between NAMs and MIDD

NAMs provide data reflecting human biology, and MIDD can interpret such data to examine dosing and trial conditions. For instance, metabolism and toxicity data measured in human cells can be incorporated into pharmacokinetic models, and the models can be updated once clinical data becomes available.

In clinical stages, models can be applied to dose selection, patient population analysis, and trial design. Prognostic score adjustment and external control arms examined in the preceding section are also methods that can be considered depending on the purpose and conditions. However, NAMs and MIDD are mutually connectable approaches, not a

standardized single procedure mandated across all drug development.

Clinical Progress in AI Drug Discovery

Insilico Medicine's rentosertib (formerly ISM001-055) is a therapeutic candidate for idiopathic pulmonary fibrosis (IPF). The company explained that generative AI was used for target discovery and molecular design.

Rentosertib is the name of a therapeutic candidate, not an AI. Drug developer Insilico Medicine explained, upon announcing the initiation of Phase 3 on July 7, 2026, that PandaOmics was used for target selection and Chemistry42 for molecular generation and optimization. The two tools perform different functions even within the same company's research platform. According to Insilico Medicine's official technical descriptions, PandaOmics and Chemistry42 were both launched in 2020. This tool launch timing predates the start of clinical trials for rentosertib. Initiating Phase 3 must be distinguished from drug approval. In related research, a candidate targeting TNIK was developed and advanced to clinical evaluation. The release years of both tools and their target discovery and molecular design capabilities can be verified in the company's technical overview. <https://insilico.com/mediakit>

Early clinical trials evaluated safety, tolerability, and changes in lung function indicators such as forced vital capacity (FVC). This subsequently led to Phase 3 development evaluating efficacy over longer durations and in larger patient populations. Long-term safety and therapeutic efficacy cannot be definitively established based solely on early results.

This case illustrates that candidates discovered via AI must also undergo stepwise clinical evaluation. AI drug discovery cases should also be distinguished from specific use cases of MIDD. The fact that AI proposed a target or molecule does not mean that model-informed decision-making was applied across the entire development process.

Development Speed and Likelihood of Success

DSP-1181, an obsessive-compulsive disorder candidate from Exscientia and Sumitomo Dainippon Pharma, gained attention for AI-driven design and rapid clinical entry, but its development was subsequently discontinued. This serves as an example showing that rapid candidate generation does not guarantee clinical success.

In studies analyzing candidates developed by AI-focused companies, performance in early clinical trials appeared positive, but a consistent advantage through late-stage development has not been confirmed. Because the number of

candidates and the duration of observation are limited, it is difficult to conclude that the observed differences stem from the effect of AI itself.

As clinical trials progress, not only drug properties but also disease complexity, interpatient variability, and comorbidities become increasingly critical. While cellular world models and digital twins can contribute to research aimed at better understanding these variables, they do not obviate the need for clinical validation.

Research and Data Collaboration in South Korea

In South Korea as well, guidelines for applying AI to drug development and joint research projects are being pursued. K-MELLODDY is a federated learning-based approach that trains a shared model without directly aggregating companies' data externally.

According to the project consortium's overview, K-MELLODDY is an initiative jointly promoted by the Ministry of Science and ICT and the Ministry of Health and Welfare since April 2024. It is closer to a project name for building a shared learning infrastructure than the name of a single finished AI model. By pooling training parameters computed by each institution to refine the joint model, it reduces the burden of centralizing raw data. However, the condition of not transferring data does

not solve all security concerns, and shared training parameters and access permissions must also be managed.

This approach seeks to incorporate information from multiple institutions into model training while protecting sensitive research data. However, ensuring the performance of the joint model requires simultaneous management of data definitions, data quality, inter-institutional variations, and evaluation procedures.

Future Challenges

Even as the scope of utilizing human cells and computational models expands, not all animal testing or clinical trials will be immediately replaced. For novel therapeutic modalities or conditions not captured in training data, predictive uncertainty can increase. A process of biologically explaining model outputs and comparing them against real-world observations is also essential.

Shifts toward NAMs and MIDD signify that methods for generating developmental evidence are diversifying. They can assist in screening novel candidates and selecting appropriate dosages and trial designs. Ultimately, their value must be judged through validation tailored to the intended context of use and safety and efficacy data obtained from patients.

Part 3. Global Case Studies of Medical AI Adoption

Chapter 5. AI Adoption in U.S. Healthcare Institutions



5.1 Mayo Clinic's VoiceCare AI and Administrative Automation

Official commentary from the Mayo Clinic Platform explains agents used for hospital administration, patient education, and safety management. The scope of VoiceCare's pilot project can be confirmed in the developer's February 2025 announcement.

<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>

Hospital operations encompass not only patient care but also insurance benefit verification, prior authorization requests, and billing follow-up. Administrative affairs or back-office tasks involve verifying information across multiple institutions and recording the results in hospital systems. Because repetitive phone calls and document processing are involved, this area has attracted significant attention as an application domain for medical AI.

Administrative automation can reduce repetitive tasks for healthcare providers and administrative staff. However, because mishandling insurance information or authorization results can impact patients' financial burdens and treatment schedules, accuracy and verification procedures are crucial.

Voice Agents Supporting Calls with Insurance Companies

VoiceCare AI is the name of the development company, and Joy is the voice agent created by that company. The company's public launch and the announcement of its collaboration with Mayo took place on February 14, 2025. This does not signify the date development began, nor does it mean the date use began across all Mayo departments. It must also not be understood as a product that directly provides diagnoses to patients.

In February 2025, VoiceCare AI announced that it was pursuing a pilot project with Mayo Clinic for insurance-related administrative tasks. The targets were neurology, pediatrics, and medical and administrative support departments, with the verification of patient insurance benefits and prior authorization serving as the primary tasks.

The voice AI agent Joy was designed to speak with insurance companies, hold on the line, verify necessary information, and summarize and record the call details. The purpose of this case lies in automating repetitive communication with insurance companies. It does not mean autonomous handling of patient care and scheduling in their entirety.

Public materials explain the direction of the pilot project, but do not prove full enterprise-wide adoption across Mayo or

specific time and cost savings. Mayo is noted as having a financial interest in the technology. Therefore, adoption announcements and independent outcome evaluations must be read distinctly.

From Answer Generation to Task Execution

For example, suppose a hospital staff member is verifying insurance coverage prior to an examination. They must prepare the patient and examination information to be verified and listen to the insurance company's response. If necessary items are missing, they must ask again and record the verification results. Voice agents also take this sequence as their processing target. Functions that convert spoken audio into text, select the next question, and deliver answers via voice are connected. The fact that a call has ended is different from the fact that insurance verification has been accurately completed.

An agent is a system that connects the output of a language model to subsequent actions. Whereas a large language model (LLM) generates answers to questions, an agent can perform a series of tasks based on those results, such as making phone calls, looking up information, and updating records.

John Halamka and Paul Cerrato of Mayo Clinic Platform emphasized this connection of actions when explaining healthcare agents. In administrative operations, insurance company verification, questions and answers, recording results, and handing over exceptional circumstances to staff are configured into a single process.

In practical implementation, an API (Application Programming Interface), which serves as a communication pathway between systems, can be used, or voice and screen inputs can be processed. What is important is not only the method of connection, but also clearly defining the scope of what the agent can process and the conditions under which tasks should be handed over to a human.

Expansion into Patient Education

MedEduChat is a research system involving researchers affiliated with Mayo, Cornell University, and MIT. The relevant paper was published on December 18, 2025. It must be distinguished from an announcement that Mayo has fully launched a commercial service for patients. The starting point of the research lies in helping patients understand their own records even after a consultation has concluded.

The operating sequence can be understood as questioning, checking relevant records, and generating explanations. For

instance, suppose a patient asks for the meaning of a treatment term written in their record. The system consults the records related to the question to generate an easy-to-read explanation. It must not guess and fill in test results that do not exist in the record. If the record itself is outdated or missing, that limitation remains in the explanation as well.

Mayo's MedEduChat is a research case designed to aid the understanding of prostate cancer patients. It was developed to easily explain medical terminology and clinical details based on patient information from electronic health records (EHR).

In the initial study, patients and experts evaluated usability and responses. While the feature explaining details based on medical records showed promise, responses tailored to an individual's specific situation required further improvement. User perception of convenience, information accuracy, and improvement in treatment outcomes must each be evaluated separately.

Because this study is a small-scale usability evaluation, follow-up validation across diverse patient populations and regarding actual clinical effectiveness is necessary. Patient education agents can be seen as playing a role in supporting

clinicians' explanations and helping patients understand their care.

Patient Safety Operations and AI

PSVT is the name of a program that manages patient safety. It is difficult to view it as the name of a single language model or commercial product. Mayo's August 2026 commentary explained that the risk classification agent was in an operational phase, while the pattern exploration agent was under development. This means that readiness levels can differ across features even within the same program.

Mayo also introduced the Patient Safety Vital Signs (PSVT) program, which addresses falls, hospital-acquired infections, pressure ulcers, medication-related events, and clinical deterioration. This is an approach of developing agents that classify patient risks and search for patterns in data.

Such systems can help select targets for clinicians to review by organizing information accumulated during hospitalization. However, news of a program's operation or development is distinct from its actual accident reduction efficacy. The accuracy of alerts, unnecessary notifications, and the response burden on medical staff must also be evaluated together.

Corporate Collaboration and Application Areas

Mayo is also researching digital pathology and generative AI infrastructure in its collaboration with NVIDIA. However, this collaboration and VoiceCare's pilot project for calls with insurance companies represent different application areas.

Even within a single medical institution, the purposes and requirements of administration, patient education, safety management, and pathology analysis differ. To understand an institution's overall AI strategy, one must distinguish which tasks individual initiatives target.

Operational Conditions for Administrative Automation

The review screen must display the confirmation date and supporting evidence alongside the conclusion. For example, an authorization number may have been received, but it might only be valid for a specific period. If only the number is copied and the validity period is omitted, the next person in charge may misunderstand the scope of approval. Leaving an incomplete answer blank to request verification maintains workflow accuracy far better than filling it with a plausible value.

In prior authorization and insurance verification, insurer information, authorization numbers, and applicable conditions must be recorded accurately. If a call is disconnected or an

answer is unclear, a human staff member must be able to take over and process it. Appropriate verification procedures are required for outcomes that affect patient rights or costs.

Automation bias must also be considered. If system results are accepted without sufficient review, errors in input or interpretation may be carried over into subsequent tasks. Human-in-the-loop review points must be determined in accordance with operational risks.

Calls with insurance companies involve sensitive information such as names, diagnoses, and insurance policy numbers. VoiceCare explains HIPAA compliance and SOC 2 Type II, but information protection frameworks must be distinguished from clinical safety. HIPAA pertains to U.S. legal requirements regarding the protection of health information, while SOC 2 Type II is an audit framework that evaluates the operation of security controls over a specified period.

The Mayo case demonstrates that the adoption of medical AI is not limited to diagnostic algorithms. Its scope of application is broadening toward automating repetitive administrative tasks, explaining information to patients, and organizing safety management data. Verifying accuracy and actual efficacy in each task, and configuring the workflow so that humans can

handle exceptional situations, are fundamental prerequisites for adoption.

5.2 Tempus's Support for Precision Oncology

On Tempus's official product page for Next, one can examine features that identify next steps according to clinical guidelines and track missed care and follow-up targets. In Lens's 2026 announcement, one can read about features that link research questions to analysis plans and code execution.

<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/>

Precision oncology is an approach that selects treatments reflecting the molecular characteristics of a tumor and the patient's condition. Even within the same cancer type, appropriate treatments may differ depending on genetic mutations or protein expression. For this to happen, necessary tests must be performed, and the results must be linked to treatment decisions.

In actual clinical practice, issues can arise where candidates for testing are missed or test results are not adequately reflected. Tempus provides platforms that analyze clinical records and molecular information to reduce these care gaps and support research.

The Roles of Next and Lens

The developer of both products is Tempus in the United States. The official announcement introducing the platform named Next was released on April 16, 2024. The history of related algorithm development and collaboration prior to that date must be distinguished from the product release date. Lens was launched as a data analytics platform on May 20, 2021. The announcement on May 31, 2026, concerns a new generation that expands agentic capabilities on that platform.

Unstructured data refers to information left in sentences, unlike numbers placed in fixed fields. For example, a medical record might state that a test was referred to an external hospital. AI reads such sentences to organize the status of tests and results. One must not conclude that a test was not performed solely because the result is absent from the record. This is why clinicians who receive alerts must check external records and the patient's situation.

Tempus Next is a clinical decision support tool for healthcare institutions. It analyzes unstructured data from electronic health records (EHR) to identify patients who have not received guideline-recommended biomarker testing and provides notifications to medical staff. Rather than independently deciding on new therapies, it supports ensuring that necessary testing and review take place.

Tempus Lens is an analytics platform for pharmaceutical companies and researchers. It analyzes patient cohorts and data according to research questions, with the Co-scientist agent introduced to support analysis planning and code execution. This approach connects de-identified clinical data and molecular information to drug development research.

The two platforms are not former and new names of the same product. Next focuses on identifying omissions in the clinical care process, while Lens focuses on research analytics. They can be distinguished as cases where medical AI supports different tasks across clinical care and research.

Foundations for Connecting Diverse Data

In Lens, the researcher's question serves as input. After defining analysis targets and comparison conditions, it selects relevant data, computes calculations, and presents the results. For example, it can compare the records of patients with and without a specific mutation. If the age or prior treatments of the two cohorts differ, the interpretation of results also varies. Even if AI generates and executes code, the researcher's role in judging whether the comparison conditions are valid remains essential.

Genomics deals with DNA information, while transcriptomics deals with the state in which genes are transcribed into RNA.

Imaging can reveal tissue morphology. Connecting disparate data allows one to examine relationships that cannot be seen through a single type of test alone. However, one must first verify whether a patient's data are accurately linked to one another. Even as data volume grows, linkage errors and omissions do not vanish automatically.

Tempus announced that it possesses large-scale data encompassing clinical records, genomics, transcriptomics, and imaging. Such data can be utilized to co-analyze specific mutations and disease progression, or to construct patient cohorts tailored to research objectives.

Research on oncology multimodal foundation models is also conducted on this foundation. However, the volume of data possessed and the accuracy of the model are separate issues. Data representativeness, omissions, and institutional measurement variations must be verified, and the initial results presented by the company must be distinguished from independently validated performance.

Connecting Biomarker Testing with Treatment

Biomarkers are biological indicators referenced when assessing disease or treatment response. EGFR and ALK appear in gene-related testing, while PD-L1 is a marker evaluating protein expression. The mere appearance of the same name in a record cannot be seen as evidence that testing has concluded. This is because test ordering, specimen collection, result reporting, and clinician review are distinct stages. Systems that identify care gaps must be capable of distinguishing between these states.

A study presented at the American Society of Clinical Oncology (ASCO) compared patients with early-stage non-small cell lung cancer (eNSCLC) across several community healthcare institutions before and after the introduction of Next. Following adoption, testing rates for EGFR, ALK, and PD-L1 were reported to have increased.

The significance of this case lies in supporting the application of established clinical guidelines to actual practice. Continuously analyzing patient records makes it possible to discover cases where necessary tests were omitted and connect them to clinical review.

However, an increased rate of test execution must be distinguished from improvements in survival or recurrence.

Pre- and post-implementation comparisons can also be influenced by shifts in clinical practice during the same timeframe. It is appropriate to understand current findings as demonstrating the potential for improving the clinical care process.

Linking with Post-Treatment Follow-Up

Tempus has collaborated with Personalis, a molecular residual disease (MRD) testing company, and also announced an acquisition agreement. MRD testing is an approach that detects traces of disease that may remain after treatment. Linking tissue-based molecular analysis with post-treatment follow-up opens the potential for more continuous analysis of patient progress.

The announcement of an acquisition agreement must be distinguished from the completion of the transaction. A business plan alone cannot be taken as proof that an integrated care delivery system has already been finalized. Actual functionality and clinical efficacy must be verified separately.

Computing Infrastructure and Collaborative Partnerships

Analyzing large-scale clinical and molecular data requires high-performance computing. Tempus also describes related infrastructure, but the scope of cooperation with specific technology vendors should not be interpreted beyond what is publicly disclosed. This section focuses on the verified functions of the Tempus platform.

When understanding healthcare AI, it is necessary to distinguish the distinct roles played by data holdings, analytical models, system connectivity, and corporate partnerships. A partnership announcement does not substantiate the performance of a specific clinical function.

Clinical Judgment and Scope of Application

Tools like Next support clinical decision-making by medical staff. Even if test candidates are identified and notified from records, whether a patient will actually undergo testing and treatment must be decided in consideration of the clinical context. Regulatory clearance for an individual AI product must not be construed as authorization for autonomous care across the entire platform.

If data is skewed toward a specific country or demographic group, performance may vary when applied to other groups. It must be evaluated using data from external institutions, and

how notifications are handled in actual workflow must be verified. The accuracy of alerts and the response burden on medical staff are also included in operational assessments.

The case of Tempus demonstrates that, beyond generating new predictions, healthcare AI can support delivering necessary information to clinicians at the appropriate moment. The advancement of precision medicine requires not only molecular analysis technology, but also a workflow architecture connecting testing, results review, treatment decisions, and follow-up care.

Chapter 6. Adoption of Healthcare AI in China



6.1 AI Operational Outcomes at the Affiliated Hospital of Hebei Medical University

The operational paper by researchers at the First Hospital of Hebei Medical University describes which tasks AI was first introduced to. Distinguishing between trials assessing clinical accuracy and the hospital's operational experience clarifies the scope of the findings.

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

Establishing healthcare AI in actual workflow requires examining not only technological performance, but also clinical care pathways and organizational operations together. This is because there are cases where use is discontinued after completing a pilot project or remains confined to only a few departments. This phenomenon is referred to as the pilot trap.

The First Hospital of Hebei Medical University in China phased in AI across pre-consultation history-taking, oncology multidisciplinary team (MDT) support, and therapeutic drug monitoring. Drawing on approximately 18 months of operational experience, the related study describes how the hospital scaled and governed AI. It serves as a case study examining workflow and governance rather than single-model accuracy.

Phased Rollout According to Risk Levels

The paper was published on June 4, 2026. This marks the date the operational findings were made public and does not correspond to the initial development or commercial launch dates of the three systems. Pre-consultation history-taking, oncology multidisciplinary consultation support, and drug concentration management are names of specific clinical tasks. They should not be understood as three different names for the same model developed by a single company. The hospital's name denotes the operating institution, and that alone does not reveal the developer or the model version of the systems.

The hospital did not automate all clinical tasks from the beginning. It first introduced functions to support pre-visit information collection, and subsequently implemented features to assist professional judgment.

In intelligent pre-consultation history-taking, when patients enter their symptoms and medical history prior to a visit, AI drafts a structured medical history. Physicians can then review the draft and ask any necessary follow-up questions. This feature supports the routine information-gathering and documentation tasks shared across multiple clinical departments.

Multidisciplinary oncology decision support organizes imaging, pathology, and treatment history to present options for review in collaborative consultations. This method connects data required when multiple specialties discuss the same patient. Final treatment decisions remain the responsibility of the clinical team.

Duloxetine therapeutic drug monitoring assists in dose adjustments by referencing blood drug concentrations. Because it is a function requiring professional interpretation and patient-specific judgment, it was utilized within limited relevant clinical domains.

Differences in Scope of Implementation

Functions that organize patient-entered responses and functions that recommend drug dosages lead to very different consequences when errors occur. For example, if the onset of symptoms is omitted during pre-consultation history-taking, the physician can simply ask the patient again. Conversely, if units in medication-related data or blood draw timings are incorrect, the clinical implication of the recommendation can change completely. Both require verification, but the items to check and the specialists responsible for them are not the same. Phased implementation is the process of reflecting these distinctions in the actual division of responsibilities.

While pre-consultation history-taking was expanded across many clinical departments, oncology treatment and drug dosage support were applied primarily within relevant specialized fields. Although organizing information and drafting documentation can be used across multiple specialties, features that directly support treatment choices require more meticulous management of risk and responsibility.

This approach represents a form of staged rollout. It begins with common tasks and expands features carrying higher clinical risks while establishing clear scopes of use and review procedures. Within the hospital, the required levels of training and supervision also differ by function.

Reported Operational Changes

In performance scorecards, one must also examine what was calculated as the denominator. The time to produce a single document differs from a patient's total length of stay. If a satisfaction survey is answered only by staff familiar with AI use, it may diverge from the overall staff experience. If staffing was increased or registration traffic flows were altered during the same period, those effects are mingled into changes in wait times. Only by verifying these conditions can one estimate the anticipated effects when implementing the system in other hospitals.

The operational study reported improvements in wait times, documentation, clinical consultation efficiency, and patient satisfaction. However, documentation completion efficiency and overall clinical efficiency are distinct metrics. Improvement figures from a specific task should not be extrapolated to the performance of the entire hospital.

During the same period, outpatient traffic flow, data management, and quality control procedures were improved in tandem. Therefore, the reported changes should be understood as the outcome of an operational improvement process that included AI. The study does not separately demonstrate the isolated effect of AI alone.

The significance of this case lies less in the exact percentage increase in operational efficiency and more in the experience of integrating multiple functions into routine clinical care and operating them over a set period. However, hospital-wide adoption must be distinguished from sustained use by individual clinicians, and whether identical outcomes would emerge in other institutions requires further verification.

Technology Adoption and Organizational Capacity

NASSS is not the name of a specific AI product. It is an analytical framework that investigates reasons for non-adoption or abandonment of technology. It is valuable for observing whether actual usage continues after a hospital announces adoption. For example, if clinicians bypass a feature because data entry is cumbersome, the workflow impact remains minimal even if the model achieves high test scores. Whether there are designated personnel to receive and resolve on-site issues is also part of operational capacity.

The expansion of healthcare AI requires defined user roles, support systems, and exception-handling criteria. Operational protocols must determine who reviews pre-consultation history-taking drafts, how errors in consultation data are corrected, and who approves medication-related recommendations.

Technology adoption frameworks such as NASSS also examine the influence of patients, users, organizations, and institutional contexts in addition to the technology itself. For a hospital to sustainably operate new capabilities, these conditions must be configured to fit real-world workflow.

The First Hospital of Hebei Medical University can be viewed as a case where AI was applied according to task nature and risk,

rather than implemented as a single, blanket product rollout. It expanded its scope from information organization to supporting professional judgment, while preserving the decision-making authority of clinicians and formal review procedures.

Operational Outcomes and Clinical Outcomes

Improvements in wait times or document processing are important to the experiences of patients and healthcare providers. However, they must be evaluated separately from improvements in clinical outcomes such as survival, readmission, and adverse events. This study likewise primarily addresses implementation and operations.

Looking ahead, there is a need to examine the sustainability of actual usage, interdepartmental differences, the burden of error correction, and patient outcomes together. Whether healthcare AI becomes truly entrenched cannot be judged solely by adoption announcements or initial satisfaction rates. A process of verifying whether it operates stably within hospital workflows and governance structures is essential.

6.2 WiNGPT 3.0 and the Ant Afu Service

The technical report by the WiNGPT 3.0 research team discloses the model's training methodology and testing scope. Ant Group's launch materials for AQ help compare how hospital-oriented models differ from consumer health apps.

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

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

The application of healthcare AI in China is expanding into operational support for medical institutions and consumer healthcare services. Tools utilized by clinicians within hospital information systems and services through which residents check health information on mobile phones are evolving in tandem.

Behind these shifts lie the concentration of patients in tertiary hospitals and trust issues regarding local primary healthcare institutions. Attempts are underway to enhance local access to medical information and assist clinicians' workloads through AI. However, the dissemination of services and the resolution of structural imbalances across the healthcare system must be evaluated separately.

Hospital Information Systems and WiNGPT

The initial release date of the WiNGPT 3.0 technical report was May 23, 2025. A revised version was posted on June 5. This represents the publication history of the 3.0 research report and differs from the initial launch date of the broader WiNGPT product line. The model covered in the report possesses approximately 32 billion parameters. Parameters are internal values adjusted during training, and a larger count alone does not make medical judgment accurate.

The research team utilized supervised fine-tuning to learn from answer demonstrations and reinforcement learning to refine outputs based on rewards. Upon receiving a medical question, it generates answers based on learned knowledge and the question's conditions. Automatically reading hospital records requires separate system integration. Simply installing a language model does not grant access to every patient's medical records.

WiNGPT is a healthcare-specialized large language model from healthcare IT firm Winning Health (卫宁健康). The company connected it with its hospital information system WiNEX and configured it to support clinicians' workflows in formats such as WiNEX Copilot.

AI applied to hospitals needs to connect to existing workflows such as documentation, retrieval, and information synthesis, rather than remaining an isolated chat window. The objective of integration is to enable clinicians to receive necessary assistance within the systems they already use.

Winning Health announced that it had deployed AI products, including WiNEX Copilot, to approximately 150 medical institutions. This figure is not identical to the number of institutions adopting the standalone WiNGPT 3.0 model, nor does it denote actual frequency of use or clinical outcomes. A distinction must be drawn between the deployment footprint and the degree of on-site adoption.

The WiNGPT 3.0 technical report presented evaluation results for medical calculation and medical knowledge. While such tests help compare baseline capabilities, whether patient diagnosis and treatment improved must be verified in real-world deployment settings.

Distinction from Other Medical Models

Unisound's official release materials for U2-Med were issued on July 17, 2026. It is a specialized model targeting healthcare, medical insurance, and pharmaceutical operations. U2-Med-World, which the company subsequently described, is world-model research that simulates clinical care processes. One must distinguish between the underlying medical model and the virtual environment used to train and evaluate that model. HealthBench is also not a commercial medical AI, but an evaluation framework for scoring responses.

U2-Med is a model developed by Unisound (云知声), distinct from Winning Health's WiNGPT. U2-Med investigates methods of revising judgments through multi-turn questioning and incorporating additional information using a medical world model. Its detailed architecture was examined in Section 2.2.

Unisound announced that it achieved high ratings on HealthBench. HealthBench assesses medical conversations for accuracy, risk identification, and communication based on scoring rubrics. However, one must distinguish between corporate-announced benchmark scores and actual real-world effects in medical institutions.

Although WiNGPT and U2-Med are rooted in different companies and designs, both represent approaches aiming to

expand medical information processing and decision support. Rather than product names, it is important to examine what data they access and what workflows they support.

Consumer Healthcare Service Ant Afu

The entity developing and operating AQ is Ant Group. The app launch date announced by the company is June 26, 2025. Subsequent service name changes and functional expansions should be viewed separately from the initial launch. The app accepts user queries or uploaded materials, provides explanations, and connects to needed services. The interface of the app, the model generating answers within it, and actual physician consultations are each distinct roles.

For example, when an image of a lab test report is uploaded, a step to read text and numbers is first required. Subsequently, the meaning of the items can be explained. If the photo is blurry and numbers are misread, or if the test date is omitted, the premises of the explanation may be faulty. The quality of user-input data and whether actual medical records have been cross-checked must be examined together. A courteous and lengthy response does not solve these issues in their place.

After launching the health app AQ, Ant Group expanded the service into Ant Afu (蚂蚁阿福). Users can ask health-related questions and review information regarding lab results or

lifestyle habits. Integrations with daily health management—such as data from smartwatches and fitness bands, meal photos, and weight management features—were also introduced.

Such services can improve accessibility for checking information prior to hospital visits or managing health status in daily life. However, the company explains that the service does not make diagnoses or replace physicians. A distinction must be drawn between general health information and personalized medical judgment.

The company has announced growth in user numbers and inquiry volumes. However, cumulative users and monthly active users (MAU) are different metrics and cannot be directly compared. The mere fact that usage scale has expanded does not prove the accuracy of information or health improvement outcomes.

Connection with Local Healthcare Institutions

Ant AFu was also connected to digital family doctor services of some local governments and community medical institutions. In this model, residents can check examination results, view related explanations, or connect with family doctor services.

In this approach, AI-driven information provision and consultations with actual physicians are combined. It aims for a structure where the AI organizes residents' questions and data, and connects them with medical staff when professional judgment is required. In actual operations, the functions of individual institutions and the roles of staff members may vary.

To evaluate the effectiveness of community healthcare, one must examine whether patients are appropriately connected to necessary medical care, in addition to user numbers. Changes in visits to large tertiary hospitals, follow-up actions after consultations, and outcomes of chronic disease management can serve as relevant indicators. User numbers alone cannot be seen as confirmation of such effects.

Challenges Following Implementation

Systems for medical institutions and services for individuals have different operational environments, but they share the common challenges of accuracy and accountability. In hospitals, AI suggestions must be linked to reviews by medical personnel. In personal services, users must be guided to understand the limitations of the information and seek necessary medical care.

Business sustainability is also critical. Even if product deployment expands, the issues of who will bear the costs and

how operations will be supported must be resolved. Data management, information updates, and error responses are also tasks that require ongoing resources.

The case of China demonstrates that medical AI is expanding into both hospital information systems and personal health management platforms. Moving forward, along with the scale of dissemination, evaluations must address the quality of actual use, connectivity with medical staff, and outcomes observed in patients.

6.3 Mindray Qiyuan (启元) and the Digitalization of Primary Healthcare

Public materials from the Shenzhen Health Commission introduce the process of connecting the Qiyuan critical care medicine model to hospitals. They can be read with a focus on the work of aggregating device data and medical records, and features that assist the judgment of medical personnel.

https://wjw.sz.gov.cn/wzx/content/post_12075823.html

China considers supporting the clinical practice of community medical institutions a major task of its medical AI policy. In China, grassroots (基層) medical institutions refer to facilities responsible for primary care, such as township health centers, community health service centers, and village clinics. Their role is to provide examinations, chronic disease management, and basic clinical care in local communities.

Policies are being pursued in the direction of enhancing the information processing and clinical decision-support capabilities of these institutions and expanding collaboration with higher-level hospitals. Medical device companies are also developing products that connect AI to equipment and information systems. Mindray (迈瑞医疗)'s Qiyuan (启元) is a relevant example.

Directions of Primary Care AI Policy

China's implementation opinions on "Artificial Intelligence + Healthcare" presented phased goals to expand medical data foundations, specialized models, and agent applications. The direction is to establish data and application foundations by 2027, and to broadly disseminate intelligent assistive applications for primary care by 2030.

Application areas include assistive diagnosis and treatment, prescription review, follow-up management, chronic disease management, and family doctor services. A system that shares imaging, electrocardiogram (ECG), laboratory test, and pathology information at the regional level and connects with higher-level institutions is also treated as important. However, these goals should not be understood as currently achieved dissemination rates or clinical outcomes.

Local governments implement the central government's direction through specific projects. Beijing's action plan covers common inference services, computing infrastructure, and multi-agent platforms, and includes regional pilot projects for chronic disease management and medical record quality control. At the same time, it puts forward the principle that AI must not completely replace the professional judgment of healthcare professionals.

Connecting Equipment and Information Systems

The Qiyuan critical care medicine model is a case that Mindray built based on Tencent's Hunyuan. Shenzhen's March 18, 2025 document explains that it was deployed to select areas starting in late 2024. What this document confirms is the deployment history at that time. It does not mean that all Qiyuan products were released on the same day or that they were used across all community medical institutions.

In intensive care units, information on the same patient is divided across multiple devices. Values sent from patient monitors, laboratory test results, and records written by physicians must be aligned along the same timeline. Reading past values as present values can lead to misunderstanding a patient's condition. After the AI gathers and summarizes information, medical personnel must be able to verify the source and timestamp of the data.

Mindray is a company that supplies patient monitors, ultrasound, and in vitro diagnostic equipment. Qiyuan (启元) is introduced as an AI product family that connects such equipment with clinical information to support the judgment and workflow of medical personnel. The critical care medicine model is a collaborative case combining Tencent's Hunyuan (混元) model with Mindray's clinical support system.

In the ultrasound field, it supports the overall examination process by connecting equipment with information management systems. In the electrocardiogram (ECG) field, it links acquisition, analysis, review, referral, and quality control. Connectivity with hospital systems such as HIS, EMR, LIS, and PACS is a prerequisite for making examination results usable in actual clinical workflows.

In regional healthcare communities, primary institutions can perform examinations, while higher-level hospitals support interpretation and consultation. Features such as BI-RADS classification support for breast ultrasound and standard plane verification for abdominal ultrasound were also introduced. These functions can be understood as complementing the work of local medical staff.

Feature Introduction and Clinical Evaluation

The company has announced efficiency improvements in record writing, examination reporting, and early warning. However, the processing speeds presented by the corporation and the actual diagnostic and therapeutic outcomes for patients must be evaluated separately.

A comparative study using emergency intensive care unit records evaluated the diagnostic performance of Qiyuan (启元) and general-purpose models. Although the specialized model

showed higher results in some comparisons, it was difficult to consider it consistently superior to general-purpose models. The results also varied depending on the scope of provided information and the specific diseases.

This study is an evaluation using retrospective records and does not directly show the actual clinical performance of primary care facilities. Adding features to equipment, sustained usage by medical staff, and improvements in patient outcomes must be verified separately.

From Individual Features to Work Processes

HIS refers to a hospital information system that covers registration and administration. EMR stands for electronic medical records, LIS for laboratory information systems, and PACS for picture archiving and communication systems. Their names differ because the data and tasks they handle are different. To connect them, patient identification numbers, test names, units, and timestamps must align.

A standard plane in ultrasound is an image taken in a designated orientation necessary for comparison and interpretation. BI-RADS is a system that organizes breast imaging findings into consistent categories. Even if AI assists with image adequacy or classification, medical personnel examine whether sufficient images were acquired and

whether they align with other tests. Distinguishing between assistive features within devices and final clinical judgment also clarifies the criteria for evaluating each function.

The application of medical AI is expanding from providing a single function as a standalone program to integrating it into existing workflows. This is an approach aimed at handling test result interpretation, clinical conversation documentation, organization of test data from other institutions, and alerts for abnormal values directly within the systems used by medical personnel.

If medical staff have to launch a separate tool each time, adoption may be limited. Linking support features directly to existing screens and recording processes enhances accessibility. On the other hand, the semantics and formats of data across systems must align, and precautions must be taken to ensure that erroneous information is not propagated into subsequent workflows.

Mindray's integration of equipment and information systems and other companies' linkages to hospital systems involve different specific products, but they illustrate this common direction. However, one company's claims must not be generalized to the actual features of other companies.

Conditions for Sustainable Operation

In community healthcare, the capability to manage patients with multimorbidity and drug interactions is critical. If longitudinal health data is insufficient or dispersed across different institutions, constraints arise in AI's interpretation of a patient's condition.

Accountability and cost-sharing must also be determined. It must be clear who reviews AI suggestions and how errors are handled, as well as who bears the costs of system maintenance and training. A structure capable of sustaining operations even after pilot project funding ends is necessary.

China's digitalization of primary care is a broader trend that simultaneously advances policy, regional infrastructure, medical devices, and personal health management services. Future evaluations must focus not only on dissemination targets, but also on whether local medical staff can actually use these systems and whether necessary clinical care and consultations occur appropriately. Connecting AI capabilities into sustainable medical services is the core challenge.

Part 4. Regulation, Safety, and Economics of Medical AI

Chapter 7. AI Medical Device Regulation and PCCP (Predetermined Change Control Plan)



7.1 PCCP Guidance and AI Model Change Management

The FDA's final guidance on PCCP explains how to submit planned modifications and verification plans. Reading the initial scope of application together with the three components of a modification plan helps in understanding the conditions of the framework.

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

AI medical devices require improvements even after deployment. Models may be adjusted to incorporate new patient data or to process information acquired from different equipment. However, because modifications can affect diagnosis and treatment, performance improvements and safety validation must be managed together.

A Predetermined Change Control Plan (PCCP) is a regulatory mechanism through which the scope of future modifications and verification methods are reviewed in advance. Modifications that conform to an FDA-cleared or approved plan can be implemented under specified conditions without submitting separate premarket filings. It is not a system that automatically approves retrained models or exempts all changes.

Premarket Pathways and Modifications for Medical Devices

PCCP is not the name of an AI model. It is a plan submitted by a manufacturer to manage device modifications. The FDA's relevant final guidance was first issued on December 4, 2024, and the current document's issue date is August 18, 2025. One must distinguish between the statutory mechanism in the law, the guidance explaining that mechanism, and the plan authorized for an individual device. The mere fact that a manufacturer read the guidance and drafted a plan does not grant permission to modify a device.

The premarket pathways for medical devices in the United States include the 510(k), which reviews substantial equivalence to a predicate device; De Novo, for the classification of novel low-to-moderate-risk devices; and Premarket Approval (PMA) for high-risk devices. Depending on the pathway, the FDA terms clearance, authorization, and approval are used.

When a device is modified, the impact of that change on safety, effectiveness, and intended use must be assessed. Significant modifications may require a new premarket submission or a supplement. The objective of PCCP is to accommodate both the demand for frequent improvements to AI models and the necessity of ensuring the reliability of modified devices.

Statutory Basis and Scope of Application

The statutory basis for PCCP is Section 515C of the Federal Food, Drug, and Cosmetic Act (FD&C Act). When modifications conform to an FDA-cleared or approved change control plan and satisfy the relevant conditions, it reduces the burden of resubmitting for every individual change.

The manufacturer presents the planned modifications and the methods to verify safety. The regulatory agency reviews the plan as part of the device evaluation process. The mere creation of a plan does not grant an exemption; actual modifications must conform to the established scope and conditions.

The FDA's guidance on software functions in AI devices explains how to present and evaluate a PCCP across premarket pathways such as 510(k), De Novo, and PMA. The required evidence varies depending on the individual device's intended use, risk profile, and nature of modifications.

Three Components of a Modification Plan

For example, consider a hypothetical device that analyzes chest images. If a manufacturer intends to train the model on additional images from different imaging equipment, it must define what is being added and for which patients it will be used. It must also be possible to compare whether performance is maintained on images from existing equipment. This example is intended to illustrate the concept of change management and is not an authorized case of a specific product.

There is also a reason for establishing acceptance criteria before reviewing test results. If criteria are lowered after results become known, it becomes difficult to determine whether actual improvement has occurred. Even if an average score increases, lesions may be missed more frequently in specific age groups. Examining both overall performance and subgroup results before and after modifications helps identify such issues.

The Description of Modifications specifies what will be altered. A generalized goal of improving performance is not sufficient. The data to be used, the functions, and the expected range of performance changes must be explicitly defined.

The Modification Protocol is the procedure for developing, validating, and implementing modifications. It includes data collection and segregation, retraining, performance evaluation, acceptance criteria, and user notification. Even for the same modification, the actual validation process must be appropriate for implementation to proceed under the plan.

The Impact Assessment evaluates the benefits and risks of the modifications. It considers not only the performance of individual functions, but also the cumulative impact when multiple modifications are implemented together, as well as the safety of the entire system.

These three components connect the scope of modifications, the methodology for execution, and the impact of the outcomes. Rather than merely evaluating a device once, this approach determines in advance how subsequent improvements will be governed.

Retraining and Continuous Performance Management

Training data is used to adjust the internal parameters of the model. Validation data is used to verify whether those adjustments hold true for unseen cases. If similar images from the same patient are mixed across both sets, testing may become artificially easy. Whether the model performs well on data from entirely new patients must be verified independently.

Retraining and deployment are also distinct stages. Building a new model does not mean it is immediately deployed to clinical interfaces. A process is required to confirm that it has passed testing and to record which version was implemented. Whether the system can roll back to a previous state if problems arise can also be inspected. Such practical measures are detailed in accordance with the device's risks and the authorized change plan.

The PCCP is a framework capable of reviewing both human-controlled retraining and automated modifications. However, whichever method is used, it must follow the pre-approved scope and verification conditions. Modifications outside the plan may require separate regulatory review.

In model improvement, maintaining existing performance is also important. As new data is learned, catastrophic forgetting

may occur, where performance on types previously handled deteriorates. Regression testing is a method to identify such changes, but it does not encompass all situations.

Therefore, training data and verification data must be distinguished, and evaluation must be conducted under diverse patient and equipment conditions. Even after modification, real-world use performance must be monitored, and procedures must be established to respond when issues arise.

Reasons for Clarifying the Scope of Modification

Adding data from new equipment or improving performance for specific patient populations also requires review depending on the impact on the intended use and performance. Modifications that alter the disease, target patients, or the role of medical staff must be judged even more carefully to determine whether they are included in the existing plan.

Modifications outside the scope permitted through the PCCP must not be arbitrarily applied. Management is required to document the actual deployed version and validation results, and to notify users of changes. The convenience of modifications and regulatory responsibilities must be addressed together.

Relationship with International Principles

Just because multiple agencies have issued joint principles does not mean that authorization in one country is automatically recognized in another country. The laws and review procedures of each region must be verified separately. The joint principles present a direction of specifically limiting modifications, maintaining evidence, and allowing users to be aware of changes. When applied to actual products, the plan approved in that specific region serves as the standard.

The FDA, Health Canada, and the UK MHRA have jointly discussed principles related to Good Machine Learning Practice (GMLP) and PCCP. The direction is that the scope of modifications must be clear, grounded in risk and evidence, and take into account the product lifecycle and transparency.

Rather than uniformly restricting improvements in medical AI, this approach aims to manage them within a predictable process. Manufacturers must specify the necessary modifications and evaluation plans, and healthcare institutions must be able to understand the product's current functions and change history.

The core of the PCCP lies in designing modifications in advance and agreeing upon verification conditions. To simultaneously consider the pace of technological improvement and patient

safety, not only the review of the plan but also the execution of actual modifications and continuous performance management are important.

7.2 FDA Competency Assessment and Risk Assessment

The FDA's generative AI discussion paper states from the very beginning that it is not finalized guidance. Although one can read proposals for categorizing risks and testing actual task competencies, this should not be taken as an enforced clearance requirement.

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

Generative AI can answer a variety of questions in sentences and suggest follow-up actions depending on the context of the conversation. Therefore, evaluating only fixed classification results makes it difficult to account for the full range of capabilities. In addition to answer accuracy, one must examine information gathering, risk identification, communication, and adherence to the scope of use together.

In its discussion paper dealing with generative AI medical devices, the FDA discussed risk assessment and a competency-based approach. This document is a resource for gathering feedback, not a finalized guideline or mandatory authorization standard. The contents of this section should also be understood by distinguishing between currently implemented systems and evaluation directions to be reviewed in the future.

Roles of Functions and Consequences of Errors

For example, one can compare a function that explains difficult terminology in medication guides with a function that alters prescription details. Both output sentences, but their impact on the user's subsequent actions differs. Even for the same error, the risk varies depending on whether medical staff have the opportunity to cross-check against source data. This is why the risk level cannot be determined solely by the length or tone of the response displayed on the screen.

The discussion paper presents an approach that considers both device activity and the consequences of erroneous output. Providing information, recommending user action, and taking direct action have different impacts on decision-making.

Even among recommendation functions, the harm caused by an error may differ. General health information and specific insulin dose adjustments do not carry the same risk when misused. The level of healthcare professional oversight and the likelihood of detecting or correcting an error must also be taken into account.

This framework does not determine risk solely by the degree of autonomy. It examines whether the user is a patient or a healthcare professional, whether the accuracy of the output can be directly verified, and whether providing information

turns into specific instructions as the conversation progresses. The actual impact of a function does not disappear merely because disclaimer or guidance statements have been added.

However, the discussion paper did not uniformly assign medical device classes or premarket pathways to each combination. One should not conflate the classification systems of the United States and Europe, nor describe the framework under discussion as already being applied as an authorization standard.

Purpose of Competency-Based Assessment

Competency refers to the ability required to perform assigned tasks. For a medical agent, both the ability to recognize missing information and the ability to answer based solely on verified data are necessary. For instance, if a patient's allergy information is blank, whether to interpret it as non-existent or to request additional verification becomes an object of evaluation. The ability to correctly answer many knowledge questions with definitive answers is not the same as this operational judgment.

A foundation model serves as the AI backbone shared by multiple products. When hospital data retrieval, tool execution, and user interfaces are added to this, it becomes a final product. Even when using the same foundation model, one

product may retrieve outdated data while another retrieves up-to-date data. For this reason, evaluation results cannot simply be substituted for one another merely because the model name is the same.

The competency-based approach is a method aimed at evaluating the abilities required for tasks that must actually be performed. Because it is difficult to individually test every answer that generative AI might generate, the required functions are systematically categorized and examined across representative scenarios.

The object of evaluation is not solely the foundation model itself, but the entire device as actually deployed. This is because the same model may behave differently depending on retrieval data, user interface, tool connections, and supervisory procedures.

Safety competencies include recognizing dangerous situations, adhering to the scope of use, communicating uncertainty, and, when necessary, handing over judgment to a human. Clinical competencies include knowledge and analysis, information gathering, interpretation of measurement results, and quality of communication.

Generalization capability is also critical. It must be evaluated whether performance is maintained across different

institutions, patient populations, and input conditions. Agents executing multi-step processes require additional competencies, such as tool use and adherence to authority boundaries. The importance of each item varies depending on the product's intended use and risk level.

Benchmarking and Clinical Confirmation

Shadow deployment is an approach that receives actual clinical workflow inputs while observing performance in a way that prevents AI outputs from affecting clinical judgment. For example, outputs generated separately from routine care can be reviewed later. While this method provides an opportunity to test real-world data, it cannot demonstrate changes that occur when medical staff act upon seeing the AI. Performance when humans and AI work together requires a separate evaluation.

Benchmarking is an evaluation that compares capabilities using predefined datasets and criteria. Establishing evaluation scope, scoring methods, expert judgments, and acceptance criteria in advance aids in consistent review. Synthetic data and virtual patients can also be used to test specific capabilities.

However, performance in test environments alone cannot predict all operations in real-world clinical care. Clinical confirmation is a stage aimed at examining functionality and

safety in actual clinical care or under conditions representative of it.

Methods vary, including retrospective evaluations of historical patient data, shadow deployments that do not expose outputs in actual workflows, conversations with standardized patients, adjudications by independent clinicians, and prospective studies. This does not mean requiring identical clinical trials for all products; evidence must be prepared in accordance with the risks and mode of use.

Standalone AI performance must also be distinguished from the collaborative performance of humans and AI. Even if the AI provides accurate information, users may misunderstand it or overlook important warnings. Comparators and evaluation metrics must be determined depending on which tasks of the medical staff are being supported.

Independent Evaluation and Postmarket Management

Drift is not used solely to mean that the AI has broken down. Performance can also change if the patient demographics alter, or if testing equipment and documentation formats change. For example, if the format for writing medication names changes, a model accustomed to the previous format may miss necessary information. The modified environment must be documented so that model issues and data issues can be examined separately.

Evaluation data and evaluators independent of the manufacturer can contribute to increasing the reliability of results. This involves managing datasets tailored to the target of evaluation, reviewing real-world use cases, and ensuring the reproducibility of tests. However, the specific operational methods for independent certification remain an area requiring further discussion.

Postmarket, foundation models, retrieval data, prompts, and operational environments can change. Procedures are needed to re-evaluate whether functions are maintained after modifications and to have clinicians review input-output samples. Drift, which is performance degradation resulting from changes in patient populations or data environments, must also be monitored.

While the PCCP in Section 7.1 is a framework for planning modifications in advance, how to apply it to the diverse variations of generative AI requires separate review. Management that links the scope of modifications, performance monitoring, and response when problems arise is essential.

Additional Risks to Consider

A prompt is a request or instruction given to an AI. In prompt injection, a statement directing the AI to change its behavior is embedded within data that is supposed to be read. For example, an instruction to send records elsewhere might be mixed into an external document. The contents of the document are material to be reviewed, not commands to alter the system's execution privileges. This is why read privileges and modification/transmission privileges must be separated.

Beyond the severity of an error, its likelihood of occurrence, reversibility, the presence of alternative safeguards, and traceability of outcomes are also important. Feedback submitted by hospital associations and other organizations on the discussion paper also highlighted these elements.

Prompt injection is an attack that attempts to alter the behavior of an AI through instructions hidden within user inputs, retrieved data, or tool results. In medical agents that

process external information and take action, it must be verified that they do not follow unauthorized instructions.

Regulatory evaluation of generative AI is broadening its scope from high scores on specific tests to safety in actual workflows. The direction is toward defining competencies appropriate to risks, evaluating interactions between the final product and humans, and managing performance even after adoption. However, proposals at the discussion stage and finalized regulatory requirements must continue to be distinguished.

7.3 UK AI Airlock and EU AI Act

The MHRA's AI Airlock Phase 2 report introduces participating cases and regulatory issues. In the European Commission's AI Act guidance, one can directly check high-risk classifications and timeline of application by article.

<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>

Regulation of medical AI is evolving through two approaches: supplementing frameworks while testing new technologies, and establishing legal obligations to verify compliance. The UK's AI Airlock is a program where developers and regulatory agencies review the challenges of actual technologies, while the European Union's AI Act is a regulation establishing legal requirements applicable to AI systems.

Although the two approaches differ in nature, both attach great importance to risk-proportionate evaluation, product lifecycle management, and human oversight. Institutions developing or adopting medical AI must understand not only the technical capabilities but also the regulatory structure of the region of application.

UK Regulatory Sandbox

AI Airlock is also not the name of a product or a diagnostic model. It is a regulatory review program operated by the MHRA. Phase 2 was conducted from April 2025 to May 2026. The official report was released on June 9, 2026, and updated on July 27. It also includes an explanation that the report itself is not formal MHRA guidance.

AI Airlock is the medical AI regulatory sandbox of the UK Medicines and Healthcare products Regulatory Agency (MHRA). A regulatory sandbox is a program that examines the regulatory challenges of new technologies within a defined scope. Developers and regulatory agencies jointly examine issues regarding intended use, evidence generation, and change management.

The program conducted discussions covering large language models, voice tools, and diagnostic assistance technologies. Key topics included how to apply existing regulations to new functionalities, how to structure change management through PCCP, and what to evaluate when expanding intended uses.

Participation itself does not signify market authorization or approval for clinical use of a product. The program's case reports must also be distinguished from official guidelines. The purpose of Airlock is not to immediately authorize individual

products, but to accumulate regulatory knowledge from issues uncovered during the actual development process.

Post-Adoption Management

Subsequent phases of AI Airlock are also broadening their focus to postmarket surveillance and lifecycle management. This is because data, models, and operating environments can change even after the adoption of medical AI.

It is difficult to predict all real-world changes solely through evaluation during the development process. Systems for collecting errors that emerge during use, verifying modifications, and sharing necessary information with medical staff are crucial. These discussions are also connected to change management and competency assessments in the United States.

EU AI Act and Medical Devices

The European AI Act is a comprehensive AI law enacted as Regulation (EU) 2024/1689. It establishes requirements by distinguishing between prohibited practices, high-risk systems, transparency obligations, and general-purpose AI (GPAI) models.

In medical device AI, the relationship with product legislation is critical. If a product or safety component is subject to the

Medical Device Regulation (MDR) or In Vitro Diagnostic Medical Devices Regulation (IVDR) and requires a third-party conformity assessment, the product-related high-risk requirements of the AI Act must be reviewed.

Therefore, not all health-related software falls under the high-risk category in the same manner. Whether it qualifies as a medical device, its intended purpose, product classification, and evaluation procedures must be assessed together. Nor does the AI Act automatically alter the existing classification of a medical device.

Key Requirements for High-Risk AI

A conformity assessment is a procedure that verifies whether a product satisfies applicable requirements. Third-party assessment refers to cases where an assessment body outside the manufacturer participates. The designation of medical device and the classification of high-risk AI address questions under different laws. After confirming what medical purpose it serves, AI-related requirements must also be reviewed. Conflating the two assessments into one can lead to overlooking necessary documentation and responsible entities.

High-risk systems require risk management, data quality and bias checks, technical documentation, operational logging,

and provision of information to users. Human oversight must be considered so that healthcare professionals can understand the outputs and, if necessary, modify them or discontinue their use.

Accuracy, robustness, cybersecurity, and quality management systems are also essential. These requirements evaluate not only model performance scores but also how the actual product is developed and operated. Conformity assessments under relevant product legislation and the review of AI requirements are also linked.

The roles of manufacturers and deploying organizations can differ. Developers are responsible for design, documentation, performance, and risk management, while medical institutions must operate the system in accordance with established conditions of use and oversight procedures. For specific record-keeping and reporting obligations, relevant legal articles and applicability conditions must be verified.

Implementation Timeline and Institutional Changes

According to the Commission's consolidated guidance on July 27, 2026, for product-related high-risk AI under Article 6(1) and Annex I, the relevant requirements of Chapter III, Sections 1 to 3 apply from August 2, 2028. The corresponding application date for high-risk AI under Annex III is December 2, 2027. However, because there are article-specific exceptions, these dates should not be taken as the start date for all AI obligations. Existing medical device-related requirements and other obligations already in force must also be verified separately.

The obligations of the AI Act have different dates of application depending on the sector and specific provision. Because implementation schedules and related frameworks may be influenced by amendment discussions, actual business operations must verify the latest consolidated legislation and confirmed transitional provisions.

Separate from the timing of application for AI Act-related obligations, existing requirements under the MDR and IVDR apply to medical devices. Adjustments to AI regulation do not mean that overall safety and performance evaluations for medical devices are exempted.

Sanctions for violations are also specified in the law, but a distinction must be made between statutory ceilings and actual penalties imposed. Companies should prepare by focusing not only on anticipated fines, but also on product classification, applicable obligations, data management, and oversight systems.

Comparison of the Two Approaches

The AI Airlock is a participatory program that examines regulatory uncertainties using real products as case studies. While it offers flexibility in discussing evaluation methods for new features, the outcome itself does not substitute for regulatory approval.

The AI Act is a framework designed to ensure compliance with legally mandated obligations. Its purpose is to clarify responsibilities throughout the development and deployment process, and its practical application requires harmonization with standards, assessment bodies, and product legislation.

The advancement of medical AI requires both the process of examining novel features and the process of verifying established safety requirements. Early consultation with regulatory authorities, clarification of the intended purpose, and continuous performance management are common tasks across both approaches.

Chapter 8. Patient Safety and the Economics of Medical AI



8.1 Interpreter Error Propagation and Speech Recognition Bias

The study on interpreter errors involving UCLA, AltaMed, and others presents simulated clinical utterances alongside AI documentation side by side. How patient statements change through interpretation and summarization can be seen in the case table.

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

AI scribes convert clinical conversations into draft documentation. In consultations involving interpretation, utterances from the patient, interpreter, and healthcare provider must be distinguished, and their respective meanings must be accurately connected. If errors during the interpretation process are reflected in the record, healthcare providers may misunderstand the patient's condition.

Interpreter errors include omission of content, substitution with different expressions, and addition of content not originally present. If the AI accepts interpreted content as fact, these errors can be carried over into drafts of official records. This is known as Interpreter Error Propagation.

Issues Arising in Multilingual Consultations

The preprint of the study was released on December 1, 2025, and the peer-reviewed journal article was published on July 28, 2026. The names of the two commercial products evaluated were not disclosed. Therefore, the performance or release timeline of any specific company's product cannot be inferred from this experiment alone. The research institutions that authored the paper and the companies that developed the products evaluated must also be distinguished.

The UCLA research team introduced interpretation errors into simulated clinical dialogues in Spanish and English and evaluated how commercial AI scribes documented them. Both products reflected some interpretation errors in their final notes. Errors remained frequently when the patient's statements were conveyed inaccurately.

These results demonstrate that AI, while transcribing dialogues, does not automatically guarantee the accuracy of interpretation. However, because this was a small-scale scripted experiment, it cannot be generalized to real-world clinical error rates.

Errors may also appear in the documentation with an opposite meaning. In a study case, the patient mentioned hearing loss, but when it was omitted in interpretation, the AI record

included a statement to the effect that there were no other hearing problems. A symptom the patient actually reported was thus negated in the record.

There were also instances where the physician inquired about multiple symptoms, but the interpreter conveyed only some of them. While the patient answered only the questions that were conveyed, the AI documented that the patient had answered all of the original questions. This is an example of failing to distinguish between "items not asked" and "items the patient reported as absent."

Such documentation can be regarded as an unsupported negative finding. Here, a negative finding means the absence of a specific symptom or abnormality. When reviewing draft notes, one must examine not only whether sentences read naturally, but also whether the content was actually verified.

Stages Where Errors Occur

The input consists of spoken dialogue exchanged among multiple individuals. The system converts speech to text, separates who spoke, and then selects the information needed for clinical records. Finally, it outputs sentences structured according to specific categories. If a subsequent stage fills in missing information from a preceding stage without verification, a record differing from the actual conversation can be produced.

For example, suppose a patient states they felt dizzy after taking medication, and subsequently adds that they are fine now. If the two sentences are combined and summarized as having no dizziness, the past reaction is lost. Conversely, if it is documented that the patient is still dizzy now, the current status is altered. This example illustrates why preserving the timeline is essential. Even if all spoken words are recognized, the relationships between sentences can still be linked incorrectly.

The processing pipeline of an AI scribe includes automatic speech recognition (ASR) and organization via language models. Misrecognizing spoken sounds can affect subsequent content organization. Even if speech is recognized accurately, speakers can be confused, or relationships between statements can be misinterpreted.

In multilingual consultations, it is also important which content—the source language or the interpretation—serves as the basis for documentation, and how conflicts between disparate statements are handled. Differences in training data volume across specific languages and speakers may have an effect, but it is difficult to identify a single cause of error based on current research alone.

Therefore, not only the final documentation but also each stage—speech recognition, speaker diarization, and content summarization—must be evaluated. When a product is updated, it must also be verified whether previous evaluation results are maintained.

Differences in Speech Recognition Across Speaker Groups

When examining word error rates, one must also look at where the errors occurred. For instance, misidentifying a single word in a greeting has a vastly different impact on the record than missing a negative expression indicating that medication was not taken. Even if the overall error rate decreases, errors in safety-critical expressions can persist. This is why whether drug names and numbers are recognized accurately, and whether negations and temporal markers are preserved, must be reviewed separately.

Speech recognition accuracy varies depending on language, accent, dialect, recording environment, and other factors. Studies comparing general speech recognition systems have reported differences in error rates between Black and white speakers' utterances. Medical studies analyzing home healthcare conversations have also demonstrated performance disparities across speaker groups.

Word Error Rate (WER) and other metrics are used for evaluation. This metric evaluates substitution, omission, and insertion errors in recognition results relative to the reference word count. However, because it counts words of differing medical significance as equivalent units, the numerical score alone cannot account for the entirety of clinical risk. Errors in drug names, dosages, allergies, and the presence or absence of symptoms must be examined separately.

Performance disparities between groups may be linked to insufficient representativeness of training data and generalization limits of acoustic models. Speech characteristics, equipment, and ambient noise also exert influence. While evaluating observed disparities, specific causes and improvement methods must be identified.

Evaluations and Reviews Required During Deployment

Upon discovering an error, one may not stop at simply modifying the final sentence. Identifying which speaker's utterance the problem originated from and whether the interpretation or speech recognition stage had an effect makes it easier to pinpoint targets for improvement. However, access permissions and retention standards for patient information must be observed throughout the process of reviewing audio recordings and source texts. Materials necessary for review cannot be equated with the indefinite retention of recordings.

Healthcare institutions must conduct pre-deployment evaluations that reflect the actual users' languages and clinical contexts. In addition to clear, single-language speech, interpretation, multi-speaker dialogues, accent variations, and ambient noise must also be considered.

Documentation must clearly reflect who made each statement and whether it was verified. Healthcare providers must verify whether symptoms, medication information, and negative expressions match the actual conversation and modify content where necessary. The fact that AI output is a draft remains unchanged even when interpretation is involved.

Developers must evaluate performance using datasets that include diverse languages and speaker groups, and inspect group-level differences. The process of observing error types and editing burdens in actual clinical practice is also necessary.

Interpreter errors and speech recognition bias jointly affect the accessibility and safety of medical AI. For diverse patients to benefit from the technology, systems must not only provide functionality across multiple languages, but must also be structured so that meanings are documented accurately and healthcare providers can verify errors.

8.2 Automation Bias and Human-in-the-Loop Review

A mammography study presented by the Radiological Society of North America shows the impact of incorrect AI suggestions on readers. Stanford researchers' MedAgentBench data describes a method for testing medical agents' task performance in simulated medical records.

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

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

Automation bias is the tendency of users to place excessive trust in a system's suggestions, thereby failing to sufficiently examine their own judgment or other evidence. Even if medical AI presents natural sentences and justifications, there is no guarantee that the content is accurate. As tasks become more complex and difficult to verify, this problem becomes increasingly significant.

Automation complacency refers to a state in which monitoring and verification diminish after observing a system perform reliably over repeated occasions. Neither phenomenon can be completely prevented through clinical experience alone. User education must be paired with reviews of clinical workflows and interface design.

Cases of Following Incorrect Suggestions or Missing Omissions

Automation bias is not a label attached to a specific company's AI. It is a behavioral problem that can arise when humans and automated tools work together. For example, if healthcare providers view the AI's conclusion before reading the primary data, they may begin by looking for information that conforms to that conclusion. The process of verifying whether the system is mistaken is also subject to the user's attention and available time.

A 2023 mammography experiment intentionally introduced errors into classification values presented as AI suggestions. It differs from studies that measure the natural error rate of a specific commercial product. Such experiments are well-suited for examining how much humans are influenced when confronted with erroneous suggestions. To evaluate real-world deployment outcomes, both the frequency of product errors and how healthcare providers interact with the tool must be understood.

A commission error occurs when an inappropriate action is taken following an incorrect AI suggestion. An omission error occurs when necessary checks or actions are not taken because the AI failed to issue an alert.

Image reading studies have reported that incorrect AI suggestions can influence healthcare providers' judgment. Highly experienced physicians were also affected, and a pattern of decreased attention to areas missed by AI was observed. However, the findings of controlled interpretation experiments should not be generalized to error rates in actual clinical practice.

Even in studies where physicians trained in AI evaluated clinical cases, incorrect suggestions influenced their judgment. This does not mean training is unnecessary, but rather that training alone is insufficient to adequately prevent the impact of errors. The system's method of use and review conditions must be designed together.

Presentation of Alternatives and Uncertainty

Presenting two answers does not always ensure independent verification. Relying on similar data and training methods can lead to the same errors. Increasing the number of answers is different from contrasting different rationales. Displaying why an alternative was generated and what information is missing enables clinical staff to compare them. The mere fact that opinions agreed does not guarantee accuracy.

Approaches that present multiple suggestions and alternatives instead of providing a single answer are being studied. Early

research comparing opinions from two models showed a reduced impact of incorrect suggestions under certain conditions. However, the same effect has not been confirmed across all clinical care situations.

Being able to verify alternative hypotheses, grounds for judgment, and uncertainty can help medical staff reexamine results. Conversely, if information is excessive or explanations are complex, the review burden increases. In practice, which screens and procedures are appropriate must be evaluated according to the users and tasks.

It must be easy to verify the source data and reasons for recommendations needed for review, and to modify or reject them when in disagreement. When AI is not confident, the capability to clearly communicate this and request human judgment is also important.

The Meaning of Human Review

An example demonstrating a narrow-scope autonomous device is IDx-DR, developed by IDx. The FDA classified this device through the De Novo pathway on April 11, 2018. The approved scope at the time was to evaluate retinal images of certain adult diabetic patients using a specified camera. This date is the FDA decision date, not the initial development date of the technology. It should not be interpreted as a case permitting autonomous medical care for all ophthalmic diseases.

For human review to function substantively, one must be able to state reasons or modify results when opinions differ. If a downstream system executes the original proposal even after an incorrect suggestion was rejected, supervision is broken. A framework is needed to verify whether modified content has been reflected in records and subsequent tasks.

Human-in-the-loop (HITL) is a framework in which humans participate in the AI processing workflow to review and make decisions. For medical staff to make substantive judgments, they must have the time, information, and authority to modify results. A procedure of merely pressing an approval button pro forma is not sufficient.

The method of supervision varies depending on the product's intended use and risks. This does not mean that a physician must pre-approve every output of all medical AI. There are screening devices authorized to produce results without physician interpretation for narrow, well-defined tasks. In contrast, general-purpose agents that assess multiple diseases and even execute prescriptions handle more complex risks.

Relevant laws and guidelines treat human oversight as an important design element. However, specific obligations must be confirmed according to product classification, authorization conditions, and the timing of implementation. An operational structure is required where individuals with clinical responsibility can review results and suspend use if necessary.

Agent Execution and Safety Procedures

MedAgentBench is an evaluation benchmark created by Stanford researchers. Its initial preprint release was on January 24, 2025, and the journal paper was published in August of the same year. It is not a commercial AI product for treating patients. It provides simulated electronic medical records and task requests, evaluating whether the agent completes the necessary queries and tasks.

A plausible answer is different from a completed task. For example, if an agent replies that it ordered a test but the actual

order entry was not created, the task is incomplete. If it operated on another patient's record, an even greater problem arises. When evaluating agents, one must verify not only the sentences, but also the target patient, execution results, and any omitted steps.

Agents can not only explain information but also perform actions such as record retrieval, document generation, and order entry. Even in evaluations using simulated electronic medical records, they did not successfully handle all tasks. However, task performance failure and actual patient harm are different metrics and must be distinguished.

Execution privileges can be restricted according to risk. Information retrieval, draft creation, and high-risk prescribing do not need to be handled with the same privileges. Emergency situations or cases with insufficient evidence can be escalated to attending clinical staff, and actions requiring approval can be configured to execute only after review.

Logging the rationale for suggestions, modification details, and approval and execution history helps trace the process when errors occur. Such an architecture is an example of applying various safety principles to operations, rather than a standardized procedure that all institutions must follow in the exact same sequence.

Review Burden and Continuous Evaluation

Requiring medical staff to re-review all AI outputs from scratch diminishes the effect of workload reduction. Conversely, skipping reviews can lead to undetected errors. Necessary verification items must be determined by considering the risks and error types associated with each function.

Medical institutions must monitor the acceptance and modification of suggestions, error detection, and alert burden throughout the utilization process. The issue of using general-purpose chatbots as if they were clinical products must also be distinguished and managed. Verifying whether design and evaluation appropriate for medical purposes have been conducted takes priority.

The goal of mitigating automation bias is not to use AI less. It is to assist clinical staff in understanding results and appropriately accepting or rejecting them. Technical accuracy, reviewable design, user education, and adequate working conditions must all be established together.

8.3 The Healthcare Agentic AI Market and CPT Reimbursement

The American Medical Association's CPT Appendix S guidance explains how to categorize AI outputs based on their clinical role. The Bipartisan Policy Center report is a resource that examines why coding, insurance payment, and actual value assessment are separate processes.

<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/>

The healthcare agentic AI market is expanding into administrative automation, clinical decision support for medical staff, drug discovery, and data analysis. Workforce shortages, cost burdens, and repetitive information processing demands form the background of adoption. However, market growth projections, actual purchases by medical institutions, and the effects observed in patients must be viewed distinctly.

The Meaning of Market Forecasts

Grand View Research projected that the healthcare agentic AI market will grow rapidly. The forecast estimated the market at approximately 760 million dollars in 2025 and projected it to reach approximately 17.3 billion dollars in 2033.

These figures are market research firm forecasts, not finalized revenue statistics. Projections also vary across institutions depending on the definition of agentic AI and the products included. While the compound annual growth rate (CAGR) summarizes the anticipated rate of growth over the period, it does not mean that the same growth will actually occur every year.

Adoption trends can be observed in cases of administrative automation, clinical support, and research platforms. However, an increase in partnership announcements or product deployments alone does not prove economic viability and clinical value. Sustainable operation requires addressing cost burdens and establishing reimbursement structures.

CPT Codes and Insurance Reimbursement

A code is akin to a common language describing what service was provided. Insurance coverage is a matter of determining for which patients and under what conditions that service will be reimbursed. The payment rate is a matter of how much will be paid for the service agreed to be reimbursed. Distinguishing these three stages avoids the error of assuming that the news of a code being established automatically generates revenue for hospitals.

To understand medical AI reimbursement in the United States, one must distinguish between the American Medical Association's (AMA) Current Procedural Terminology (CPT) codes and payers' payment systems. CPT is a coding system that classifies medical procedures and services. The existence of a code does not mean that all payers will reimburse the cost.

CPT Category I codes are for established services, Category II for performance measurement, and Category III for tracking the utilization of emerging technologies and services. Category III has the character of temporary codes and does not guarantee reimbursement. The Centers for Medicare & Medicaid Services (CMS) and payers apply separate coverage and pricing criteria.

Currently, reimbursement cases for medical AI are primarily concentrated in defined functionalities such as imaging, electrocardiograms, and retinal examinations. Relevant examples include CT-based fractional flow reserve analysis (FFR-CT) and autonomous diabetic retinopathy screening. General diagnostic test codes and codes for AI functionality must be distinguished.

Assistive, Augmentative, and Autonomous

CPT's AI taxonomy also does not use the model's name or size as a criterion. Even when using the same technology, the role differs depending on whether the output summarizes source data, provides clinical metrics, or renders an independent interpretation and conclusion. The AMA refined the taxonomy descriptions in May 2026 and updated its official guidance in September of the same year. It is necessary to verify whether the latest definitions align with actual services.

The AMA's AI taxonomy examines what role the output plays clinically. Assistive AI provides information requiring interpretation and reporting by medical staff. Augmentative AI derives clinically meaningful metrics from input data. Autonomous AI connects analysis results to conclusions regarding diagnosis or management.

This distinction helps evaluate the functionality and accountability of actual services rather than grouping AI into a single technology. However, classification, code assignment, and insurance reimbursement determination are distinct processes. This is also why agentic platforms combining conversation, information retrieval, and multiple actions are difficult to describe with a single existing code.

Inpatient and Outpatient Reimbursement Structures

In the Medicare Severity Diagnosis-Related Group (MS-DRG) bundled payment for inpatient care, related costs are handled within a fixed bundle. AI costs can also be included within this scope. For new technologies meeting certain requirements, transitional reimbursement programs such as the New Technology Add-on Payment (NTAP) may apply.

In outpatient care, temporary classifications and separate pricing determinations are also utilized for new technologies. However, not all AI services receive adequate reimbursement. In practice, billing volume may be limited, or disparities in utilization may appear across institutions.

Even in the absence of a separate fee schedule, hospitals can adopt tools through their own budgets or subscription contracts. In doing so, the gains in documentation time, staffing burden, and patient flow must be weighed against the costs of acquisition, integration, training, and review.

Fee-for-Service and Value-Based Payment

Fee-for-service reimburses based on the performance of individual services. If AI performs the same task faster or reduces certain services, those benefits may not be adequately reflected in the current reimbursement structure. Conversely, continuously adding AI-related services could potentially increase overall expenditures.

Value-based payment is an approach that seeks to consider patient outcomes, healthcare quality, and costs together. If AI contributes to reducing hospitalizations or unnecessary tests, it can align with this model. However, it is difficult to isolate which factor among various systems and tools generated the performance, and contracts vary by payer.

The Bipartisan Policy Center (BPC), a policy think tank, proposed reimbursing for proven value, differentiating by AI type, ensuring access in underserved areas, using representative data, and conducting continuous reevaluation. These are recommendations regarding policy directions, not finalized payment criteria.

Criteria for Economic Evaluation

For example, suppose the time spent drafting documentation has decreased. A portion of the saved time is spent again verifying the draft and correcting errors. The hospital must also examine whether the remaining time is allocated to patient communication, rest, or additional patient visits. Time savings do not automatically translate into reduced labor costs or increased revenue.

The party bearing the cost and the party receiving the benefit may also differ. A hospital may purchase the system while patients reduce travel and waiting burdens. Payers can experience cost-saving effects when unnecessary services decrease. Unless specified from whose perspective the calculation is made, evaluations of the same technology as economically viable or not can diverge. Separately documenting patient health outcomes and cost burdens makes the differences easier to understand.

The economic viability of medical AI is difficult to determine solely by purchase price. The costs required for system integration and maintenance, clinical staff review, data updates, and error response must be calculated together. Even if time is saved, if other operational burdens increase, the overall effect may change.

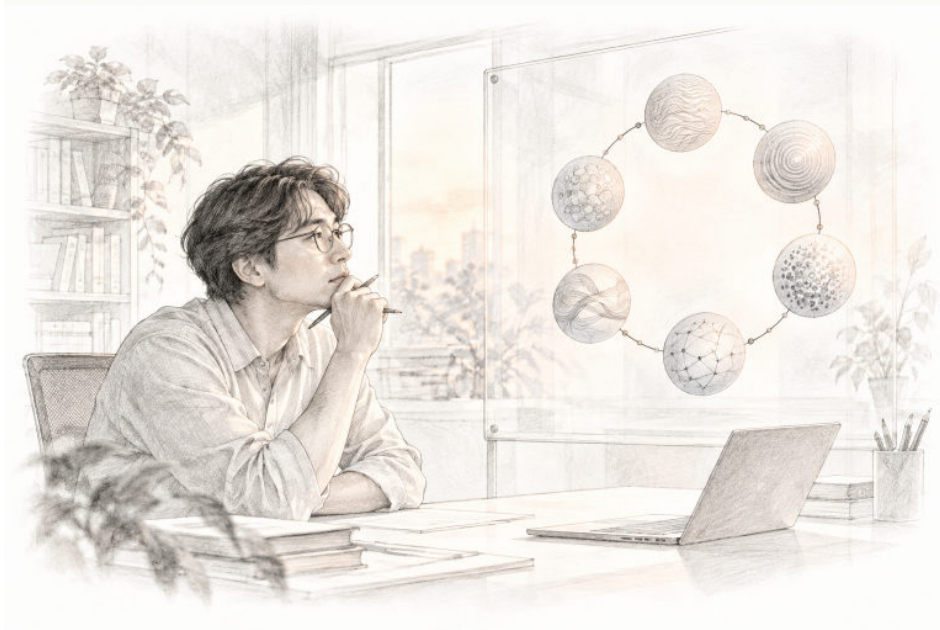
Changes observed in patients are also critical. Outcomes appropriate to the intended use—such as the appropriateness of testing and treatment, waiting times and accessibility, adverse events, and readmissions—must be examined. Increases in revenue and utilization volume cannot directly substitute for these values.

Because insurance and new health technology assessment systems differ across countries, the U.S. CPT case cannot be directly applied to Korea or other nations as is. One must verify what services are reimbursed and what evidence is required in each respective system.

For medical agentic AI to be used sustainably, technical feasibility, cost burden, and reimbursement must be connected. Rather than using market forecasts as a rationale for adoption, it is essential to verify what improvements have occurred in actual workflows and for patients, and to establish payment models appropriate for that value.

Part 5. Medical AI Agents and Collaborative Care

Chapter 9. Healthcare Agentic AI



9.1 Limitations of Single Prediction Models

Reading the paper published by the AgentClinic research team reveals an evaluation methodology that arrives at a diagnosis by questioning simulated patients. It is a resource examining the difference between a test where all information is provided upfront and a situation where it must be actively gathered. <https://arxiv.org/abs/2405.07960>

Medical AI has advanced in well-defined tasks such as finding lesions in images or calculating the risk of specific diseases. Examples include detecting pulmonary nodules on chest X-rays and classifying melanoma on skin images. These models analyze designated inputs and present outputs.

The single predictive model discussed in this section refers to a system with limited capability to continuously monitor a patient's progress or select the next course of action. Although not all predictive models share the same architecture, a distinction must be made between performance on a specific task and the ability to support comprehensive clinical care.

The Patient's Clinical Context

For example, suppose an imaging model flags a small abnormal lesion in the lung. The mere presence of a flag cannot determine the diagnosis. A physician checks whether the same area was present on previous imaging, whether new symptoms have developed, and other factors. The model's output is merely one piece of data that feeds into this judgment. One must not assume that the model inherently knows circumstances outside the input image.

In clinical practice, not only current symptoms but also medical history, current medications, allergies, and changes in laboratory results are all equally important. The significance of a specific finding on an electrocardiogram (ECG) can also vary depending on the patient's symptoms, organ function, and previous tests.

If necessary information is not included in the input, it is difficult for the model to take it into account. This is why accurate analysis of a specific image or test does not equate to an appropriate judgment regarding the patient as a whole. Capabilities to connect diverse data and interpret changes over time are necessary.

Large language models (LLMs) can process extensive records, but they cannot utilize all information in long inputs at the

same level. The 'Lost in the Middle' research addressed the issue of missing relevant information located in the middle of an input. Even as context windows expand, whether models accurately locate and apply necessary information must be evaluated separately.

Characteristics of Real-World Clinical Records

For instance, past prescriptions may remain on a medication list, while a recent clinical note states that the medication was discontinued. Even if both sentences are retrieved together, the answer will be incorrect if the system cannot discern which point in time the query refers to. The ability to read extensive records and the ability to organize them according to the current situation must be verified separately. Even when condensing long records, conditions that alter clinical judgment—such as allergies and reasons for discontinuation—must be preserved.

Electronic health records (EHRs) feature diverse document formats and abbreviations, and information may be missing or contradictory. The same content may be repeated, and there are cases where the time of documentation must be distinguished from the time of the actual event.

While curated examination questions provide all necessary information, in real-world clinical practice, crucial information

must be identified first. A systematic review examining medical LLM evaluations also pointed out that studies utilizing real patient data remain limited. Test scores must not be directly interpreted as the ability to support real-world clinical care.

Information Gathering and Revision of Judgment

Diagnosis is a process of formulating hypotheses, selecting necessary questions and examinations, and revising judgments based on the results. Even after treatment, patient response is monitored and plans are adjusted. A model that produces an output from a single input cannot easily perform this process on its own.

Capabilities to utilize external tools such as calculators, search, and clinical guidelines, request missing information, and re-examine prior judgments may be necessary. Architectures that maintain state and iterate through multiple steps represent an approach aimed at addressing these demands.

MedLA is a reasoning framework developed by a joint research team from Westlake University, Ant Group, and others. The preprint was uploaded on September 28, 2025, and it compares premises and conclusions across multiple agents.

However, simply increasing the number of steps does not guarantee accuracy. Prior premises may be incorrectly

maintained, or contradictions may arise between steps. MedLA investigated a method in which multiple agents compare syllogistic reasoning. This represents an attempt to examine the consistency between premises and conclusions in differential diagnosis.

Differences Revealed by Interactive Evaluation

AgentClinic is the name of a research evaluation environment, not a commercial product name for a clinical service. A research team led by Samuel Schmidgall and colleagues first released the paper on May 13, 2024. Researchers from Johns Hopkins University, Stanford University, ETH Zurich, and other institutions participated in the paper. The revised edition of May 2025 must be distinguished from the initial release date. When a model acting as a physician inputs a question, another role provides patient information or test data. Evaluators then examine the resulting conversation and the final judgment.

AgentClinic is a benchmark designed for AI to converse with virtual patients, request additional information, and make diagnoses. It demonstrates that model performance can vary depending on whether necessary information is provided all at once or whether the model must collect it autonomously.

The results are influenced not only by medical knowledge but also by dialogue strategy, the permitted number of steps, tool

usage, refusal to respond, and other factors. Therefore, a low score should not be directly interpreted as a lack of knowledge, nor should virtual patient diagnostic results be equated with real-world clinical accuracy.

Even if an agent connected to tools shows improvement on certain tasks, processing costs and time may increase. Token usage, latency, and hallucinations must also be evaluated together. A more complex architecture is not always the better choice.

Evaluation When Used in Collaboration with Humans

Retrospective studies evaluate performance using pre-existing historical records. Prospective studies observe events as they unfold according to a predefined plan. In evaluations where the diagnosis in historical records is provided as ground truth, what the physician did not know at the time may not be adequately reflected. This is why, when reading study results, one must examine both the sequence in which data were collected and the scope of information accessible to the AI.

In meta-analyses synthesizing multiple studies, the efficacy of LLM assistance varied depending on the type of case. This is because well-curated case reports and retrospective records of actual patients present differing evaluation conditions.

Even in experiments where laypersons assessed medical situations, the standalone performance of a model did not align with the outcomes when humans utilized the same model. The process through which users communicate necessary information and understand responses influences real-world effectiveness.

Therefore, beyond standalone performance, medical AI must be evaluated on how it changes the judgment and behavior of users. Evaluating outcomes on real patients requires fit-for-purpose evidence, such as prospective studies.

Expansion into Agentic Systems

An agentic system is designed to observe necessary information, plan subsequent actions, and re-examine the results of tool use. Through memory and iterative processing, it represents an approach to bridging the outputs of single tasks into the clinical workflow.

Single predictive models continue to play an important role in executing well-defined tasks. While agents can connect these models, data, and tools, they also introduce new modes of error and operational overhead. Which features should be added must be judged based on the demands and risks of real-world workflows.

9.2 Sense-Think-Act and Memory

The ReAct paper by researchers from Princeton University and Google explains how planning and tool use are interconnected. The MedAgentBench paper from Stanford University demonstrates the testing of such capabilities within virtual medical records. <https://arxiv.org/abs/2210.03629>
<https://arxiv.org/abs/2501.14654>

An agent is a system that receives information from its environment, selects a subsequent action, and verifies the outcome. In medical agents, this can be implemented as a process of reviewing medical records and test results, retrieving necessary data, and modifying the plan based on new information. This is described as a loop—an iterative cycle of Sense-Think-Act.

This structure differs from a capability that generates a response in a single step. To execute multi-step tasks continuously, the system must manage the current state, previous actions, and newly acquired information together. Memory supports the retention and retrieval of such information.

Sense: Collection and Organization of Patient Information

In the sensing stage, information is gathered from electronic health records (EHRs), vital signs, medical imaging, conversations, and other sources. Current values are compared with prior records, and the patient's condition is organized by verifying the timestamps and sources of the data.

The mere presence of data does not make all of it reliable. There may be outdated records, omissions, input errors, and conflicting information. The agent must distinguish between verified facts, assumptions, and items that require further confirmation.

Think: Planning and Re-examination

The initial preprint of ReAct was released on October 6, 2022. It was subsequently presented at the ICLR 2023 conference. It is a general agent design methodology researched jointly by Princeton University and Google Research, and the date does not signify the launch date of a dedicated medical model. Reflexion was proposed by researchers from Northeastern University, MIT, and Princeton University. The paper was first made public on March 20, 2023. It is an approach that incorporates verbal feedback summarizing previous execution into subsequent attempts. The names of these two studies must not be interpreted as hospital-approved product names.

In the planning stage, the objective is broken down into subtasks. For example, discharge support may include reviewing lab results, reviewing medications, providing medication guidance, and formulating follow-up care plans. It is the process of determining which data and tools are required and organizing the sequence of execution.

The ReAct study presented an approach that alternates between reasoning and tool use, adjusting the plan based on observation results. Reflexion investigated a method of recording the results of previous attempts and referencing them during subsequent execution. However, because models do not always accurately identify their own errors, a distinction

must be made between self-reflection and external verification.

Act: Utilization of Authorized Tools

For example, upon receiving a request to retrieve test results, the agent verifies the patient and invokes a lookup tool. It then reads the values and dates returned by the tool to formulate a response. In this context, displaying the test name on the screen and actually registering a test order are actions carrying different authority levels. Even if the same text-generation technology is utilized, the boundary between query and order entry must be separately defined within the system.

In the action stage, records are queried through an API (Application Programming Interface), or calculators and search tools are employed. When results are returned, the patient state and plan are updated. Executable functions must be restricted according to the product's authorization level and intended purpose.

MedAgentBench, which evaluated task execution within virtual electronic health records, addresses tasks such as record lookup and order entry.

It is an evaluation environment created by Stanford University researchers including Yingshi Zhang and Jonathan Chen. The preprint paper was released on January 24, 2025, and the NEJM AI paper was published on August 14 of the same year. One must check whether it is the initial or subsequent version alongside the dates to understand differences in test configuration. Such evaluations are helpful for examining the potential and failure modes of tool-integrated agents. However, the success rate on virtual tasks cannot be directly interpreted as the safety of actual patient care.

Types of Memory

Here, memory refers to data retained by the computer and the rules for retrieving that data. It does not imply equivalence to the human process of understanding and remembering experiences. For example, saving a conversation summary and pasting it into a subsequent request is a form of external memory. Even if storage succeeds, retrieving another patient's data cannot be considered safe memory management.

Working memory consists of immediately required information, such as the current dialogue and recent test results. It can be managed within the model's context window or through summarization. Care must be taken to ensure that critical conditions are not omitted during the summarization process.

Episodic memory consists of records of events that occurred to a specific patient and previous tasks. It may include adverse drug reactions or past test results. Necessary information can be retrieved from external storage, such as vector databases and relational databases. Retrieval-Augmented Generation (RAG) is a method of connecting such information to answers and clinical judgments.

Semantic memory is knowledge regarding medical concepts and relationships. It relates to standardized terminologies and knowledge graphs. Procedural memory represents execution methods, such as clinical guidelines, hospital operational rules, and tool-use procedures.

This categorization is addressed in language agent architecture research such as CoALA (Cognitive Architectures for Language Agents). In practical systems, each memory type may be stored separately or managed collectively in part. The crucial factor is accurately utilizing the necessary information at the appropriate moment.

Memory Management and Updating

CoALA is a design framework for language agents released on September 5, 2023, by researchers from Princeton University, including Theodore Sumers and Shunyu Yao. It is a study that categorizes and explains various types of memory and action, rather than an independent medical product. MemGPT was released on October 12, 2023, by researchers from the University of California, Berkeley, including Charles Packer. Although 'GPT' is in its name, it does not refer to a model released by OpenAI. It is a system that manages information by dividing what the model reads immediately from what is stored externally.

MemGPT investigated an approach to managing information between a constrained working space and external storage. Instead of inputting long records all at once, it retrieves and utilizes necessary information as needed. Although performance improvements using notes have also been observed in AgentClinic and other settings, the results of a specific evaluation cannot be generalized to all clinical workflows.

Maintaining memory allows previous information to be used continuously, but errors can also be repeated. Referring back to an incorrectly saved diagnosis or outdated guidelines impacts subsequent judgments. The source, creation time, and

revision history of data must be managed, and updates should be made according to authorization.

Updating external memory is different from retraining the model itself. Retraining a neural network can cause catastrophic forgetting, where previously learned capabilities decline. In the case of changing external memory, model weights are not directly modified, but the accuracy and appropriateness of the retrieved content must be verified.

Application to Laboratory Work

The Latent Labs research team first released the Latent-Y technical report on March 31, 2026. The agent connects scientific literature with biological tools, and uses Latent-X2 from the same company to generate antibody candidates. Latent-Y is the agent that carries out tasks, while Latent-X2 is the generative model that produces candidates. The initial commercial sale date of each product cannot be determined from this report alone. <https://arxiv.org/abs/2603.29727>

Latent Labs' Latent-Y is a research case that connects literature review, target analysis, antibody candidate design, and computational validation. It was introduced as a method that performs multiple design tasks and generates candidates according to user instructions.

Although experimental binding of some candidates has been reported, this is a preprint technical report released by a company. It cannot be considered that the entire process of iteratively improving the next design by automatically reflecting experimental results is complete. Connecting multiple tasks must be distinguished from continuously learning based on real-world results.

Control of Iterative Processing

Perception-reasoning-action structures and memory support tasks that extend over long periods. However, as the number of iterations increases, errors, costs, and processing time can also increase. Task completion criteria, stopping conditions, and situations requiring human handoff must be clearly defined.

Medical agents must be configured so that they do not automatically continue to make judgments when evidence is insufficient or risks are high. Continuous information management is a means of expanding autonomy, yet simultaneously an object requiring oversight. Designing accurate memory, limited authorization, and auditable records together is crucial.

9.3 Taxonomy and Technological Maturity of Medical Agentic AI

The seven-domain taxonomy paper by Vatsal et al. compares medical agent research by functionality. It is helpful when discerning which functions have been reported and what has not been evaluated. <https://arxiv.org/abs/2602.04813>

Medical agents vary widely in functionality, configuration, and scope of autonomy. It is difficult to evaluate systems that retrieve information to answer queries and systems that maintain patient records while executing multiple tools using the same criteria. A taxonomy helps organize these differences and identify functions that require further development.

The literature review of medical LLM agents by Vatsal et al. proposed a seven-domain taxonomy. This framework is a method for comparing designs reported in research, rather than a measurement of the level of all commercial products.

Seven Taxonomy Domains

This taxonomy itself is not an AI product. It is a study authored by Shubham Vatsal, Harsh Dubey, and Aditi Singh, released on arXiv on February 4, 2026. Because it is the result of categorizing capabilities reported as implemented in the literature, it cannot be used as certification that a system is safe for patients. Even for the same memory function, its actual role differs depending on whether it merely retains summaries during consultation or updates data over the long term.

Cognitive capabilities encompass planning, multi-step reasoning, and uncertainty handling. Knowledge management addresses linking external literature and clinical guidelines, as well as memory storage and retrieval. Interaction modalities include not only dialogues with users but also functions that respond to specific events.

Adaptation and learning refer to the ability to review prior performance and respond to changes. Safety and ethics examine privacy protection, bias, regulatory compliance, and safeguards. Framework types distinguish between single- and multi-agent configurations and central coordination methods.

Core tasks refer to the medical duties actually performed. These include record summarization, question answering,

differential diagnosis (鑑別診斷), treatment planning, study design, and patient monitoring. Even when using the same architecture, task risks and required performance levels may differ.

Frequently Studied Functions and Underrepresented Capabilities

In the literature review, external knowledge integration and multi-agent designs were relatively widely implemented. Typical approaches include referencing literature via Retrieval-Augmented Generation (RAG) and dividing multiple roles to review answers. Question answering, decision support, and virtual environment evaluations were also major tasks.

On the other hand, event-driven activation, which autonomously operates when an event occurs, and drift detection and mitigation, which detects performance changes, were rare. Central orchestration, which coordinates the work of multiple agents, was also insufficiently implemented in some cases.

This gap suggests that medical agent research has addressed continuous operations and adaptation to change less than information-processing functions. However, because these findings are limited to the analyzed papers and timeframe,

they should not be interpreted as statistics representing the capabilities of all current products.

Expansion of Research and Practical Adoption

Other literature reviews have also shown a trend where medical agent research expands into clinical support, patient engagement, oncology, and radiology. However, an increase in the number of papers differs from the adoption rate in actual hospitals.

For functions proposed in research to be used in the field, issues of data access, system integration, review accountability, costs, and training must be resolved. Without standardized evaluations, it is also difficult to compare scores and effectiveness across different studies.

Evidence for Evaluating Technological Maturity

When evaluating maturity, one must examine the environment in which functions were validated. Benchmarks solving curated problems, virtual patient dialogues, retrospective analyses of real patient records, and prospective (前向的) use studies provide different types of evidence. Evaluations confirming patient outcomes and operational sustainability after implementation in actual clinical practice are also necessary.

This distinction is not an official rating that uniformly ranks research. It is a benchmark for judging what questions the evidence can answer. Virtual environments are suitable for repetitive experiments and error analysis, but they cannot fully reproduce the behaviors of actual clinicians and patients.

AgentClinic and MedAgentBench are evaluation environments that allow comparison of information gathering and task performance. However, performance in such environments alone does not prove safety and efficacy in actual clinical care. Follow-up evaluations tailored to the intended use are necessary.

Operational Capabilities and Maturity

For example, a system that signals an error when a lab inquiry fails and a system that presents previous values instead may look similar on screen. However, the latter carries the risk that users might mistake outdated data for new results. Beyond scores on successful requests, evaluations must assess what information is displayed and who is notified upon failure. The evidence hospitals require when considering adoption encompasses these operational conditions.

In continuous operations, processing time and computational costs are also critical. Repeatedly using tools or holding discussions among multiple agents can increase token usage

and response latency. It must be verified whether additional computation genuinely enhances accuracy and safety.

Monitoring to observe errors and performance degradation, logs to verify execution processes, and procedures for handoff to personnel in charge are also necessary. If a system cannot respond to new data or model changes, maintaining the same level of performance over the long term is difficult, even if initial evaluations are favorable.

The technological maturity of medical agents is not determined solely by the number of functions or degree of autonomy. What matters is whether they reliably perform necessary tasks, detect errors, and permit human review. A taxonomy can serve as a tool to comprehensively examine these conditions.

Chapter 10. Multi-Agent Systems and AI Hospitals



10.1 Multi-Agent Collaborative Design

The MedAgents paper by researchers from Yale University and Shanghai Jiao Tong University illustrates the process in which multiple expert roles review answers. It allows one to examine the sequence of dividing roles and the methods for synthesizing opinions. <https://arxiv.org/abs/2311.10537>

Multi-agent collaborative design is a method where multiple agents divide roles and share results to execute tasks. In healthcare, areas such as information retrieval, reviewing diagnostic candidates, verifying medications, and documenting records can be distinguished. The objective lies in managing complex tasks through role-specific processing and cross-verifying results.

Different roles can be assigned to the same base model, or multiple models and tools can be combined. Therefore, it is difficult to judge system diversity or reliability based solely on the number of agents. Roles and permissions, information transfer, and result integration methods must be examined together.

Utilization in Research

In the literature review of medical LLM agents, multi-agent architectures were found to be widely used. This demonstrates that researchers are actively exploring designs that divide complex tasks into multiple functions.

However, the proportion of analyzed papers does not imply the adoption rate in real-world hospital products or clinical superiority. The fact that multi-agent systems were used and the fact that safer outcomes were delivered to patients must be evaluated separately.

Division of Roles

MedAgents was first released on November 16, 2023, by Xiangru Tang and fellow researchers. It is a research collaborative framework published in ACL-related proceedings in 2024. Rather than training a separate medical model from scratch, it focuses on assigning roles and discussion sequences to existing models. It analyzes input medical problems from multiple perspectives and synthesizes reports to provide answers.

The first collaboration method is the division of roles by function. It differentiates planning, literature retrieval, clinical analysis, verification, and explanation, providing each role with the necessary information and tools. When inputs and outputs

for each role are clear, tracking tasks and resolving problems become straightforward.

MedAgents proposed an approach where agents acting in expert roles produce analyses and integrate them. Although it demonstrated performance improvements in evaluating certain medical problems, virtual expert roles do not possess the qualifications and responsibilities of real-world medical specialists.

Debate and Cross-Verification

For example, suppose one role organizes test results while another checks medication lists. If the first role mismatches the patient's name, the remaining roles will review another person's data despite working conscientiously. Rather than the mere fact that multiple answers match, the ability to trace which source data each answer originated from is crucial. Creating distinct roles does not inherently increase independent evidence.

The second method involves cross-examining candidates and evidence. Regarding a diagnosis proposed by one agent, another agent can present missing information and counter-evidence. When results do not align, the system can be configured to request additional searches or human review.

However, there is also the potential to adopt the same error after reviewing each other's answers. Because agents based on the same model or similar training data may share common biases, consensus among multiple opinions does not guarantee accuracy.

Central Orchestration and Task-Specific Configuration

MDAgents was proposed by Yubin Kim and researchers at MIT. Researchers from Google Research and Seoul National University Hospital also participated in this paper. It was preprinted on April 22, 2024, and presented at the NeurIPS conference the same year. Although its name is similar to MedAgents introduced earlier, it is an independent study. While MedAgents focused on role-based discussions, MDAgents extends to the process of selecting collaborative configurations tailored to the problem.

The orchestrator manages task sequences, information flow, and termination conditions. It is the capability that determines which roles to invoke, how to handle conflicting results, and when to hand over to a human.

MDAgents investigated an approach that selects single- or multi-agent configurations based on task difficulty. Instead of employing the same number of agents for every task, it scales up computation only when additional collaboration is

necessary. In practical deployment, the accuracy of difficulty assessment and the associated costs must also be evaluated.

Sharing Reasoning Structures

MedLA is a research framework developed by Xiqi Ma and researchers. The paper lists affiliations including Westlake University, Ant Group, the Centre for Artificial Intelligence and Robotics of the Hong Kong Institute of Science & Innovation at the Chinese Academy of Sciences, and China-Japan Friendship Hospital. It was preprinted on September 28, 2025, and accepted to AAAI 2026. Major premise refers to a general rule, minor premise refers to the conditions of the current case, and conclusion refers to the judgment linking the two. Verifying whether premises are true must be distinguished from verifying whether the conclusion is properly derived from the premises.

In addition to dividing roles, there is also the method of aligning the structure of the judgment process. MedLA represents premises and conclusions in the form of syllogisms (三段論法), enabling agents to compare them and inspect contradictions.

This approach seeks to verify what premises led to the conclusion rather than merely comparing the phrasing of lengthy explanations. However, because logical consistency

differs from medical factual accuracy, verification of source data and clinical evidence is also required.

Virtual Hospital Research

Agent Hospital was developed by the Institute for AI Industry Research at Tsinghua University and associated researchers. The paper was first released on May 5, 2024. A separate study, AI Hospital, involved researchers from Alibaba, Huazhong University of Science and Technology, and Fudan University, among others. It was preprinted on February 15, 2024, and presented at the COLING conference in 2025. Although both names include 'hospital', the papers introduced here deal with clinical simulations inside computers. They are not announcements of having admitted real patients and treated them autonomously. <https://arxiv.org/abs/2405.02957>
<https://aclanthology.org/2025.coling-main.680/>

Agent Hospital constructed a virtual environment where agents in the roles of patients and medical staff interact. It studied methods for doctor agents to iterate through cases and improve performance by referencing previous results.

Other studies, such as AI Hospital, also use medical care simulators involving multiple roles as evaluation environments. While such environments are well-suited for repeatedly experimenting with information transmission and

collaboration, generated patients and virtual care cannot reproduce all the complexities of an actual hospital. Achievements in virtual evaluations must not be interpreted as real clinical efficacy.

Cost and Error Management

Without termination conditions, agents may iterate through debates, merely paraphrasing the same arguments. In the meantime, computational costs mount and the delivery of results is delayed. A prolonged discussion alone cannot satisfy hospital demands. Actionable and reviewable information must be delivered within a designated timeframe. When an answer cannot be finalized, items requiring additional verification should be documented and handed over to the person in charge.

As the number of agents increases, token usage, response times, and the complexity of result integration can grow. It is essential to verify how much the additional computation resulting from collaboration actually improves task accuracy and safety.

If hallucinations are shared, subsequent judgments may be based on flawed premises. Mitigating this requires source verification, external validation, and discrepancy resolution procedures. Although failure studies of tool-based agents

illustrate these risks, observations made in a specific architecture cannot be generalized to the error rates of all multi-agent systems.

Beyond model performance, failures can stem from unclear roles and instructions, missing information, or insufficient verification. Execution records across the entire system must be examined to distinguish causes and ensure that erroneous consensus does not lead to automated execution.

Evaluation Criteria for Collaboration Design

The value of a multi-agent design lies not in the number of roles, but in how reliably it performs required tasks. One must evaluate how accuracy, review burden, cost, and latency change compared to a single model.

In healthcare applications, the intended purpose, authority, medical staff review, and error response must be clearly defined. Actual implementation cannot be justified solely on the grounds that a structure is widely used in research. It is necessary to establish evidence tailored to patients and clinical tasks before determining the scope of application.

10.2 MAI-DxO's Division of Roles and Complex Diagnosis

Microsoft's official October 2026 announcement outlines the evaluation and clinical study plans for MAI-DxO. One can read it while distinguishing between research scores and the actual clinical effectiveness to be confirmed in the future.
<https://microsoft.ai/news/building-medical-superintelligence/>

In complex diagnoses, it is difficult to reach a conclusion based on initial information alone.

The full name of MAI-DxO is Microsoft AI Diagnostic Orchestrator. Microsoft researchers preprinted a paper on sequential diagnosis on June 27, 2025, and published research in Science in October 2026. The timing of a paper's publication differs from the commercial release date of a product that anyone can use in a hospital. As an orchestration framework connecting multiple models, it is also distinct from the name of a single, fixed neural network. Clinicians must select necessary questions and tests, and re-examine likely diseases depending on the results. Microsoft's MAI-DxO is research aimed at supporting such sequential diagnosis through the division of roles across multiple agents.

This system separately handles diagnostic candidates, test selection, counter-evidence, costs, and result checking. It

examines how performance is affected not only by medical knowledge, but also by the sequence of information acquisition and the orchestration of multiple judgments.

Sequential Diagnosis Evaluation Environment

For example, if the initial description lists only fever and fatigue, the agent can either select a disease name right away or request additional information. Depending on the requested items, answers are retrieved from the records and returned, and the agent determines its next action after reading the new data. This explanation is a hypothetical example to help understand the evaluation structure. The act of receiving test results on a screen does not encapsulate the entire process of drawing blood and waiting for results in an actual hospital.

SDBench is a benchmark designed to review complex clinicopathological conferences (CPC) and other cases in a step-by-step manner. Initially, only a limited case summary is provided, and additional information is presented in response to questions or test requests.

The Gatekeeper, an information-providing role, delivers the content corresponding to requests. The final diagnosis is compared against criteria by a model in the Judge role, which also estimates the cost of requested tests. Unlike tests that

provide all necessary data upfront, this structure evaluates which information to select.

Review by Role

In test selection, the system considers what information is needed to distinguish diseases. In cost review, it re-examines whether a judgment can be made using existing information. When the demands of both roles conflict, the orchestration framework determines the next action. Translating these computational results into actual medical cost savings requires investigating not only test prices, but also delays, re-testing, and medical staff verification time.

The hypothesis review role maintains possible diagnoses and adjusts priorities based on new information. The test selection role suggests tests that will help differentiate candidates. The counter-argument role examines evidence inconsistent with current hypotheses and addresses over-reliance on initial judgments.

The cost management role evaluates the benefits of additional information against test costs, while the inspection role verifies test names, reasoning consistency, and other elements. By orchestrating results across multiple roles, it decides on the next question, test, or final diagnosis.

This structure is described as a model-agnostic orchestrator that is not tied to any specific foundation model. While different models can be connected to the same role configuration, actual results may vary depending on the models and evaluation conditions.

Possibilities Shown in Evaluations

Published studies reported high diagnostic accuracy in tasks based on complex NEJM cases and emergency department records. Results were higher than those of general practitioners under identical evaluation conditions, and configurations factoring in test costs were also tested.

These results demonstrate the potential for role division and information gathering planning to contribute to complex diagnoses. However, scores obtained from models and settings at different points in time should not be conflated as a single result. Costs are also estimates within the evaluation environment, not actual billed amounts.

Differences from Clinical Application

SDBench is a controlled virtual evaluation. It does not encompass actual patient communication, physical examinations, or real-world working conditions. The comparison group of physicians and the conditions for data usage may also differ from actual specialist consultations in reality.

Therefore, one cannot conclude based on evaluation results that it can replace an actual team of medical specialists or that patient prognosis has improved. The term "medical superintelligence" also represents a corporate development goal rather than a current clinical achievement.

Microsoft announced that it is pursuing research in actual clinical environments in collaboration with Beth Israel Deaconess Medical Center (BIDMC). Prospective studies can compare patient interactions with independent judgments by medical staff. However, study plans must be distinguished from completed results.

Collaboration Design with Medical Staff

In real-world deployment, beyond the standalone accuracy of AI, its effectiveness when used alongside medical staff is crucial. Methods in which medical staff make independent judgments first and the AI summarizes points of agreement and discrepancy are also being studied.

Case-based studies have shown the potential of such structures to enhance diagnostic accuracy. However, effects can vary depending on the backgrounds of medical staff and users, as well as the presentation of information. Medical staff should not be steered into uncritically following AI suggestions; rather, they must be enabled to compare the necessary evidence.

The significance of MAI-DxO lies in connecting information gathering, diagnosis, and cost review by orchestrating multiple roles. Going forward, its usability, safety, and review burden must be verified in actual clinical practice. To translate research findings into beneficial support for patients, the roles of medical staff and the authority of the system must be designed in tandem.

10.3 ClinVerse and the Clinical Care Process of AI Hospitals

The ClinVerse paper by joint researchers from Fudan University and other institutions presents the interactions and evaluation criteria across six roles. One can read it while distinguishing information taken from real records from information supplemented for simulation purposes. <https://arxiv.org/abs/2610.34878>

Evaluating medical agents requires examining not only the correct diagnostic answer, but also the process of gathering information and selecting tests.

ClinVerse is collaborative research by Fudan University, Shanghai Innovation Institute, Zhejiang University, and Huazhong University of Science and Technology. Sirry Chen and colleagues posted a preprint on October 26, 2026. ClinPatient is a collection of cases used for evaluation, and ClinLab is a component that predicts missing portions of laboratory test information. These names do not refer to separately approved medical products. ClinVerse is a research environment that simulates these interactions based on real electronic health records (EHR). It is not a product operating a real hospital, but a simulator for training and evaluating agents.

In this environment, an agent starts from initial symptoms, asks the patient questions, requests tests, and then presents a diagnosis. Unlike multiple-choice evaluations that provide all necessary information from the start, it evaluates the ability to choose which information must be obtained.

Virtual Cases Based on Real Records

The ClinPatient dataset of ClinVerse contains approximately 150,000 diagnostic cases constructed from public clinical records. It is based on de-identified data such as MIMIC-IV, MIMIC-IV-Note, and SICdb. Because it primarily consists of inpatient and intensive care records from the United States and Austria, it does not represent outpatient populations or all countries.

Lifestyles and social histories lacking in the records were supplemented using Synthea, a synthetic patient generation tool.

Synthea is synthetic patient generation software led by MITRE, a non-profit research organization in the United States. The tool paper by Jason Walonoski and colleagues was released online on August 30, 2017, and published in a journal volume in 2018. It creates records by simulating a patient's life and disease progression through rules. It is a different type of tool from language models themselves. Therefore, the phrase

"based on real records" does not mean that all information was actually observed. One must distinguish real data from synthetic information when evaluating.

Division of Roles in the Clinical Process

The doctor agent performs medical interviewing, test requests, and diagnosis. The patient simulator separately manages information that the patient knows and can explain versus information to be confirmed through tests and other procedures. The clinical pathology role conveys laboratory test data, and the environment controller orchestrates the overall progress. The diagnostic judge evaluates the results and the process without intervening in the conversation.

This division of roles is designed to provide only information that matches questions or tests. It restricts information that could reveal the diagnosis in advance from being inappropriately delivered. However, complete prevention of data leakage cannot be guaranteed.

For missing tests, the ClinLab module can predict and provide values based on other data. These values are components of the simulation, not actual measurement results. Because prediction errors can influence subsequent diagnoses, they must be examined separately.

Evaluation of the Final Diagnosis and Process

The evaluation utilizes structured clinical evidence and International Classification of Diseases (ICD-10) diagnostic codes. It examines how much necessary medical history was gathered, how thoroughly relevant tests and abnormal findings were verified, and whether diagnostic categories and specific details match.

Even if the final accuracy rates are similar, the ways information is collected can differ. An agent might request many tests but fail to link the results to a diagnosis, or it might reach a conclusion relying primarily on initial information. Process-specific metrics help identify these differences.

Evaluation scores can also be used as reward signals for learning. However, whether actions that maximize designated scores truly align with sound medical care must be continuously scrutinized.

Iterative Learning and Modification of Memory

Dynamic Cheatsheet was proposed by researchers including Mirac Suzgun from Stanford University and Together AI. The initial paper was published on April 10, 2025. It summarizes useful methods in text from previously solved problems and feeds them into subsequent inputs. It should not be understood as a proprietary model name created by ClinVerse. Rather, an existing memory management method was adopted and tested within a diagnostic simulation environment.

The ClinVerse study compared methods for extracting skills and lessons from existing conversations, updating memory during interactions, and orchestrating the update process itself. It describes this as a self-evolution or self-improvement approach.

Rather than directly changing model weights, several methods modify external memory and execution rules. Methods such as Dynamic Cheatsheet organize previous experiences to reference during subsequent execution. In some evaluations, information gathering strategies and diagnostic performance improved.

However, these results represent changes observed in a virtual environment. The very fact that test information is missing

might serve as a diagnostic clue, and predicted test values could distort evaluations. One must analyze not only model performance, but also the simulator's assumptions and limitations.

Capabilities Connectable to Real-World Clinical Work

ClinVerse is not a system deployed in actual hospitals, but it enables research on structures that connect medical interviewing, record retrieval, test candidate recommendations, and differential diagnosis preparation. To apply it to clinical support in the future, it is necessary to verify how each function assists medical staff in their workflow.

For example, even if AI organizes medical history, it must be possible to additionally confirm important symptoms. When proposing candidate tests, existing results and the patient's condition must be reviewed together. Candidate diagnoses and rationales must be provided in a format that medical staff can verify and modify. This is an explanation regarding the direction of application, not an outcome proven by ClinVerse in an actual hospital.

Roles and Limitations of Virtual Environment Evaluation

For example, if responses from the patient role always follow the same format, the model may become accustomed to that format rather than gaining medical understanding, potentially inflating scores. To distinguish this, performance must also be examined in cases where sentences are altered or created using different data. Because both the agent under evaluation and the simulator providing information jointly influence the results, it is also necessary to record and distinguish the change history of both.

ClinVerse is meaningful in repeatedly evaluating diverse agents under identical conditions and analyzing issues in information gathering. However, the study discussed in this manuscript is a preprint prior to peer review and did not evaluate real patient outcomes.

Actual clinical practice involves diverse variables, including noise, unclear statements, documentation errors, system connectivity issues, and interruptions to medical staff's work. Virtual patients and predicted test values cannot fully replicate all of these.

The significance of AI hospital research does not lie in immediately automating all hospital judgments. It lies in testing methods for connecting multiple roles and

information, and analyzing the process through which failures occur. Real-world application must undergo function-by-function verification, medical staff supervision, and the clarification of authority and responsibility.

'Medical superintelligence' is an expression representing a future goal rather than a currently achieved level of technology.

Part 6. The Future of Medicine and AI

Chapter 11. Collaboration Between Physicians and AI Agents



11.1 AI Screening, Human Review, and Comprehensive Diagnosis

The tele-ECG paper by researchers affiliated with Sanya People's Hospital and Sichuan University explains the roles of regional institutions and interpretation centers. One can confirm the process where results flagged by AI are delivered after undergoing human review. <https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2026.1869716/full>

When applying medical AI to clinical practice, one must determine which functions to assign and who will verify the results. Even for a model with high accuracy, its effectiveness is limited if there are too many alerts or if they are not delivered to responsible personnel in a timely manner. Therefore, the workflow must be designed alongside the model's performance.

Connecting AI screening, review by trained personnel, and comprehensive judgment by medical staff is one approach to implementation. The AI prioritizes and classifies information to be reviewed, the personnel confirm the appropriateness of the source data and outputs, and the physician decides on diagnosis and follow-up actions while considering the patient's

condition. Specific roles vary depending on the product and institution.

Regional ECG Collaborative System

This case addresses an operational system connecting regional medical institutions rather than a single specific AI model. The study published in 2026 analyzed operations following its introduction in the second half of 2025. The paper describes the core AI as a commercial solution developed by a medical technology company. However, the paper does not disclose the company name or the initial release date of the product. The publication date of the hospital's platform study must not be recorded as the development date of that AI.

The tele-ECG cloud platform in Sanya (三亚), China, is an example connecting interpretation systems between regional medical institutions and higher-level hospitals. When a 12-lead ECG is recorded at a regional facility, data is shared and reviewed stepwise by AI, interpretation personnel, and specialists.

The cloud serves as the infrastructure through which multiple institutions transfer and process data. When a higher-level institution verifies an ECG obtained regionally, it can support necessary judgments and transfer preparations. The achievements of this structure should be understood not

merely as the interpretation competence of individual regional medical staff, but as the result of connecting information and expertise.

Stage 1: AI Screening and Quality Verification

A 12-lead ECG is a test that records the electrical signals of the heart from multiple directions. The input consists of these waveforms, and the AI outputs risk probabilities by comparing them with learned features. Noise caused by hand or body movement can also be mixed into the waveforms. This is why a process is required to verify whether the signal is suitable for analysis before reviewing risk indications.

AI flags potential risks from the waveforms and also verifies electrode attachment and signal quality. In conditions such as acute myocardial infarction, it is important to establish screening criteria so as not to miss risks. ST-segment elevation myocardial infarction (STEMI) is also a major subject of review.

However, in settings that increase sensitivity, normal or low-significance findings may be flagged as abnormal.

Sensitivity refers to how well individuals with actual abnormalities are detected. A false negative occurs when an abnormality is present but missed, while a false positive occurs when an alert is triggered for a person without an abnormality.

Positive predictive value is the proportion of individuals with actual abnormalities among those who received an alert. Even for the same model, the meaning of an alert varies depending on the disease prevalence in the tested population. The meaning of an alert must be evaluated by examining the false negative rate, false positive rate, and positive predictive value (PPV) together.

Stage 2: Raw Data Review

For example, if an AI alert arrives but parts of the waveform are disconnected, personnel can check whether it is a measurement error and communicate the need for re-measurement. If alerts are continuously sent to specialists without this process, judgments may not become faster even as information to be reviewed increases. The function of recording who received and confirmed notifications is also part of collaboration.

Personnel at the central interpretation center verify the original ECG waveforms and review the content flagged by AI. Their role is to reduce signal errors and excessive alerts, organizing information for specialists to review.

When unnecessary notifications recur, alarm fatigue can arise, blunting responses even to critical warnings. The intermediate review stage is a method to manage such burdens, but proper

training and workload management for personnel, as well as criteria for transmitting results, must be in place.

Stage 3: Comprehensive Judgment by Specialists

Specialists review not only the ECG, but also chest pain (胸痛) and other symptoms, alongside test results such as troponin. This is because all diagnoses cannot be confirmed by abnormalities in a single signal alone.

The timing of presenting AI outputs can also influence judgment. To mitigate anchoring—over-relying on the first proposed answer—and automation bias, independent review and comparison of evidence must be possible. Blinding evaluation methods in research and routine procedures in actual hospitals must be designed distinctively.

Changes Reported in Operations

The Sanya study reported improvements in diagnostic processes, decision-making time, and the delivery of interpretation results at regional institutions. Sharing examination data has the potential to reduce the burden of repeating identical tests at higher-level hospitals and to support preparations prior to patient arrival.

However, because these are observational findings before and after implementation, the entirety of changes cannot be

attributed solely to the effects of AI. Limitations in sampling and documentation, seasonal patient variations, and network connectivity conditions must also be considered. Survival outcomes in a small number of patients should not be extrapolated into a survival improvement effect of the platform.

Collaboration in Other Clinical Fields

The Taiwan case evaluated an AI ECG alert system created by researchers from the National Defense Medical Center and Tri-Service General Hospital. The clinical trial paper was published on April 29, 2024. This date does not signify the initial release date of the overall related technology. The system calculates risk from the ECG, notifies medical staff when it exceeds a set threshold, and medical staff verify the patient's condition. The need for real clinical research and follow-up verification can also be confirmed in the guidance published by the hospital. <https://www.tsgh.ndmutsgh.edu.tw/news/191/10010/24887/10864>

Taiwan's AI ECG alert study was structured to deliver risk signals to medical staff and facilitate responses. This case is meaningful in that it evaluated whether screening results translate into actual clinical actions.

In Sweden's MASAI breast cancer screening study, reading workflows were adjusted according to AI risk classification.

MASAI is the name of a study by Lund University and others, not the name of an AI product. Transpara, used in the study, is an image analysis software from ScreenPoint Medical in the Netherlands. It receives mammography images as input and provides suspicious regions and scores. The US FDA 510(k) decision date for the initial product was November 21, 2018. This is the date of the corresponding US review record, and is not identical to the world's first release date or the start date of the MASAI study. The structure is one where AI supports part of the review sequence and task allocation, while medical staff are responsible for interpretation. Patient-related outcomes, such as interval cancer—cancers detected between regular screenings—were also evaluated.

Case studies combining human and AI diagnoses have also demonstrated the potential for complementarity. However, accuracy in written case presentations differs from real-world clinical performance. Clinical evidence appropriate for each field must be verified.

Scope of Autonomous Functions and Supervision

IDx-DR is a medical device developed by IDx in the United States, later known under the name LumineticsCore by Digital Diagnostics. The FDA De Novo clearance decision date was April 11, 2018. It analyzes retinal photographs of the inside of the eye to produce screening results for a defined range of diabetic retinopathy. The fact that reading results does not require separate image interpretation by an ophthalmologist is different from saying that follow-up care for the patient is unnecessary. FDA documents clearly specify referral criteria based on results and testing limitation conditions. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf

There are also devices authorized to provide results without separate interpretation by medical staff for narrow and well-defined tasks. The diabetic retinopathy screening device IDx-DR, currently LumineticsCore, is a relevant example. Therefore, one should not generalize that the identical form of final review is required for all AI outputs.

A pilot project by Doctronic, which handles renewals of existing prescriptions under restricted conditions, was also introduced.

This is a pilot project of a prescription renewal tool operated by Doctronic in the United States. In the status report published

by the responsible agency in Utah, it is stated that approval from a licensed healthcare provider is required for all prescription requests in Phase 1. Therefore, when introducing the pilot project, the phase and supervisory conditions at that time must be stated together. The initial release date of the company's model cannot be determined solely from this guidance. <https://commerce.utah.gov/ai/regulatory-relief-4/authorized-pilots/doctronic/> However, repetition of existing prescriptions, new prescriptions, and dosage modifications are different tasks. Pilot authorization in a limited region must not be understood as approval for general autonomous clinical care.

The level of supervision must be determined in accordance with the intended use, risk, and regulatory clearance conditions of the product. Criteria for escalating cases requiring specialized judgment to medical staff, along with alternative contingency procedures when the system fails to operate, are also necessary.

Design of Collaborative Systems

The structure of AI screening, human review, and comprehensive diagnosis is an approach that clarifies roles by function. Its purpose is to transmit information without omissions, detect errors at each stage, and secure the clinical context necessary for the final judgment.

In practical implementation, an increase in tests and alerts alone should not be viewed as an achievement. One must verify whether appropriate review and responses take place. Turnaround time, repeat testing, operator burden, and patient outcomes should be evaluated together, while also preparing for connectivity failures. The efficacy of medical AI is determined not only by individual models, but by this entire operational framework.

11.2 The Role of Physicians and Comprehensive Clinical Judgment

A randomized trial paper comparing modes of collaboration between medical staff and AI examines differences between independent judgment and opinion comparison. Results derived from case evaluations can be read separately from actual patient treatment outcomes.

<https://www.nature.com/articles/s41746-026-02545-1>

When AI supports documentation and information retrieval, the structure of a physician's work can also change. Rather than bearing the burden of organizing test results, compiling medication lists, and entering clinical notes, the role of reviewing presented information and making appropriate decisions for the patient becomes paramount.

However, automating information organization does not restrict the physician's role solely to judgment. History-taking, physical examination, communication, and treatment coordination are interrelated tasks. Actual changes also differ depending on where AI is applied.

Documentation Work and Clinical Consultation Time

Ambient AI scribe refers to a product category that listens to surrounding conversation during clinical consultations and generates draft notes. Because it is not a single model name produced by one company, a common release date cannot be defined. It converts speech into text and categorizes symptoms, care plans, and other items. Because it may add details absent from the conversation or miss negative expressions, medical staff must verify the draft.

Ambient AI scribes can reduce the documentation burden on medical staff by drafting clinical notes. Multiple studies have reported potential improvements in documentation time and burnout, but effects varied depending on the product and usage environment. In some studies, after-hours work time was not distinctly reduced.

Therefore, whether the time freed by AI can be utilized for dialogue with patients or clinical review depends on the hospital's workflow operations. If scheduling, staffing, and other administrative tasks remain unchanged, a reduction in documentation time alone will not resolve the overall burden.

Even in systems that link AI screening and staff review, such as regional ECG collaborations, the role of the specialist remains. Specialists interpret organized findings by linking them to

symptoms and other diagnostic results. The division of information processing and the responsibility for clinical judgment must be designed together.

The Role of Understanding Patient Context

For example, even if a plan recommending regular outpatient visits is proposed, it is difficult to implement if the patient lives far away or lacks someone to assist with transportation. If such conditions cannot be found in the medical record, one must ask directly. Beyond merely verifying whether the phrasing of an AI-generated plan sounds plausible, one must evaluate whether the patient can realistically adhere to it.

Although electronic health records (EHRs) contain a vast amount of information necessary for clinical care, they do not represent the entirety of a patient's situation. Factors such as living conditions that enable continuation of treatment, caregiver support, and the patient's own understanding and concerns often need to be verified through conversation.

Physicians must verify whether the information referenced by AI is sufficient and up to date, and also examine conditions not present in the record. It is also an important role to review the priority of possible diagnoses, explain uncertainties, and determine the need for additional verification.

The Role of Choosing Treatment Together with Patients

The appropriateness of treatment is not determined solely by average efficacy. Side effects, treatment burden, lifestyle, and goals valued by the patient must be considered together.

Shared decision-making is a process of explaining the pros and cons of options and making decisions that reflect patient preferences. While AI can organize data and assist in explanations, what the patient understands and which choice they desire must be identified through interactions with clinical staff.

The Role of Reviewing AI Outputs

To reduce automation bias, recommendations must be reviewed in light of one's own clinical knowledge and the patient's condition. This is a process of verifying whether the evidence aligns with actual data, whether critical conditions were omitted, and whether alternative options exist.

The review method and accountability vary depending on product features, regulatory clearance conditions, and task risks. While this does not mean physicians must repeat every calculation, they must be able to verify information necessary for judgment and modify or reject recommendations.

A role is also needed to reconcile differences in opinion across multiple medical specialties and AI tools. Priorities must be established taking into account the patient's condition, treatment goals, resources, schedule, and other factors. Even a consensus presented by multi-agent systems cannot automatically substitute for this clinical review.

Effectiveness of Collaboration and Usage Methods

Studies analyzing human-AI collaboration have shown inconsistent efficacy. Even when benefits emerge in curated case studies, advantages may be small in complex real-world records. Even if an AI's standalone score is high, users do not necessarily achieve the same level of outcomes.

Methods in which clinicians form independent judgments before comparing points of agreement and discrepancy with AI have also been studied. Although potential for improvement appeared in some case evaluations, this should not be regarded as proven clinical outcomes in actual patients.

Designs presenting opinions from two models or alternative hypotheses can also support review. On the other hand, excessive information increases burden, so user experience (UX) and actual workflows must be evaluated together. Granting access to tools alone is not sufficient.

Maintaining Professional Competence

Responding when AI fails or stops is also part of training. Clinicians must possess the ability to read source data in order to detect errors in automated analysis and continue workflow through alternative procedures. In training, having trainees explain why an incorrect recommendation must be revised, rather than merely showing correctly answered cases, can be helpful. This is a proposal for curriculum design and not an efficacy metric proven by specific studies.

When AI takes over portions of detection and analysis, the frequency with which medical personnel directly perform those tasks may decrease. Observational studies in endoscopy reported changes in performance during procedures conducted without AI following the use of AI assistance. However, because causation cannot be conclusively established, one should not generalize that automation inevitably causes degradation of competence.

Healthcare institutions must maintain training and audits so that necessary clinical judgments can be executed even without AI. Residency and training curricula must teach not only how to verify outputs, but also how to interpret source data and detect errors.

Evaluating Changes in Roles

Changes in the physician's role should be judged by how actual clinical tasks are restructured rather than changes in job titles. Time spent organizing and reviewing information, communication with patients, appropriateness of decision-making, education, and fatigue must be examined collectively.

When AI reduces repetitive tasks and provides necessary evidence, it can support clinicians' decision-making. To turn this into beneficial changes for patients, adequate conditions for review, clarification of accountability, and continuous education are required. Verifying workflows and patient outcomes following technology adoption is the criterion for assessing the value of role transformations.

Chapter 12. Technical and Ethical Challenges of Future Medical AI



The U.S. FDA's change control guidance explains how modification plans and verification of AI medical devices are linked. It is a resource to be read from the perspective that safety and performance must be confirmed even after initial authorization. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

Medical AI is expanding its scope of application from documentation and imaging analysis to clinical workflow support, drug candidate discovery, and clinical trial design. Connecting multiple tools and agents enables handling complex tasks, but the impact of errors and the management burden also increase.

Medical superintelligence is a term that describes the ultimate goal of such advancements. Demonstrating high performance on specific tests differs from operating safely across healthcare in real-world practice. Current research findings must be distinguished from future prospects, and evidence tailored to the intended use must be secured.

Hallucinations and Cascading Errors

For example, suppose a clinical summary mistakenly records a discontinued drug as an active medication. If a subsequent agent reads only that summary, it inherits the incorrect premise directly. Even if the same explanation is reviewed multiple times, errors may not be detected without re-examining the source data. This is why sources and dates must be transmitted together at every stage.

Large language models can generate counterfactual statements in a natural manner. Hallucinations manifest as incorrect drug information, unsubstantiated diagnostic explanations, nonexistent literature, and more. Across multi-step tasks, preceding errors can propagate into subsequent searches and judgments.

The ability to provide citations does not resolve the issue on its own. Clinicians must confirm whether cited content matches the actual original source and is applicable to the current patient. Verification procedures are also required during the agent's tool usage and outcome integration processes.

Performance Variations and Data Representativeness

Even if overall averages are maintained, errors in certain sub-populations can increase. For example, after imaging hardware changes, performance might degrade exclusively on images from a specific institution. Before modifying the model, one must examine whether input data have shifted or whether the actual disease distribution has altered, allowing interventions targeted to the root cause.

Drift refers to problems where model performance shifts due to changes in patient demographics, equipment, data generation protocols, and related factors. Even if favorable results were achieved in the initially evaluated setting, there is no guarantee that identical performance will be maintained thereafter.

Training datasets biased toward specific regions and populations can create performance disparities for other patients. The prediction target itself must also be scrutinized, because proxy metrics of health needs, such as healthcare spending, can reflect disparities in healthcare access across different groups.

Performance must be evaluated not only on overall averages but also across patient demographics, institutions, and disease categories. When shifts are detected, systems must be capable

of responding through re-evaluation, modification, and adjustments to the intended scope of use.

Biosecurity

While protein- and DNA-design AI supports drug discovery and life science research, there is also potential for misuse in generating sequences with dangerous biological functions. Biosecurity must be considered not only during model development but also throughout actual gene-synthesis ordering and laboratory research.

Related studies have shown that existing sequence screening tools may fail to adequately recognize newly generated variants. Therefore, detection systems must be continuously improved, and procedures verifying the purpose of synthesis orders and customer identities must be operated concurrently. To mitigate risks, verification procedures among model providers, synthesis services, and research institutions must be interconnected.

Liability and Privacy Protection

The U.S. Department of Health and Human Services (HHS) HIPAA guidance explains covered entities and principles for protecting health information. Because identical laws do not apply to all health apps or in all countries, the operating entity and jurisdiction of the service must first be verified. <https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>

When errors occur, liability varies depending on the roles and actual causes among model developers, product manufacturers, healthcare institutions, and end users. It cannot be resolved simply by assigning all liability to the final approver. Contracts and operational guidelines must clearly delineate authority, review duties, reporting requirements, and response procedures.

Privacy protection is necessary throughout the entire lifecycle of data collection, model training, retrieval, and archiving. Legal requirements by jurisdiction, such as the U.S. HIPAA and European GDPR, must be verified, and access and utilization must be restricted to what is necessary.

Federated learning is a method of training a shared model without centralizing data in a single location.

When each institution transmits training updates learned from its own data, these are aggregated to update the global model. Although it has the advantage of not transferring raw data, traces of information can still remain in the transmitted training updates. Federated learning is the name of a technological methodology and cannot be described by a single company or release date. Nevertheless, it does not completely eliminate the risk of information leakage. Potential attacks against models and training processes must be evaluated, and technical safeguards alongside institutional controls must be established.

Clinical Competence and Oversight

As reliance on AI increases, hands-on experience performing certain tasks directly may decrease. Although competence changes observed in studies must be interpreted with caution regarding causation, they demonstrate the necessity of maintaining education and periodic auditing.

To reduce automation bias and alarm fatigue, clinicians must be provided with sufficient time to verify results, access to raw data, and authority to make modifications. Designs enabling clinicians to form independent judgments before comparing them with AI recommendations can be beneficial. However, the appropriate approach must be assessed according to task risks and clinical practice settings.

Autonomy must also be differentiated by function. A device that independently yields results for a narrow screening task differs from a general-purpose agent that diagnoses multiple conditions and decides on novel therapies. One must not describe autonomy as having been granted beyond cleared indications and local pilot program conditions.

Modification and Lifecycle Management

The relevant final guidance currently released by the FDA is the August 2025 version. It recommends that modification details, development, verification, implementation methodologies, and the impact of changes be incorporated into the plan. PCCP denotes a Predetermined Change Control Plan addressed in regulatory review, distinct from the name of an AI model. It does not imply automatic approval for functional modifications that extend beyond the pre-specified scope.

For AI, outputs can vary whenever models, retrieved data, or prompts change. A Predetermined Change Control Plan (PCCP) is an approach that establishes the scope and verification methods of planned modifications in advance. It is not a mechanism that freely permits arbitrary changes.

Pre- and post-modification evaluations, user instructions, and post-market performance monitoring must be linked in accordance with the product's intended use and risk profile.

Because related laws and guidelines differ in finalization status and effective dates, the latest requirements must be verified upon actual deployment.

Hospital Governance Systems

For example, an AI's risk flags can be saved separately without displaying them on the active clinical screen, and compared retrospectively with clinicians' judgments. Because such trials still process patient data, access privileges and retention periods must be specified. Hiding the outputs does not exempt a study from research ethics and security procedures.

Healthcare institutions cannot leave AI adoption solely to technical procurement departments. Clinicians, nurses, information security officers, and legal and ethics experts must collectively review functionalities, risks, and operational parameters. Patient perspectives and accessibility must also be taken into account.

Before impacting actual patient care, silent trials can be considered to evaluate performance using real-world clinical data. This approach confirms operation under actual workflow conditions without utilizing outputs in treatment decisions. It must be conducted after defining data processing protocols, privacy safeguards, and evaluation parameters.

Following deployment, errors and edge cases must be collected, and staff response actions and modification logs must be maintained. Criteria to restrict or suspend use upon discovering safety issues are also essential.

Patient Outcome-Oriented Evaluation

Findings claiming time savings must also be scrutinized regarding whose time was measured and over what scope. While data entry time may decrease, time spent reviewing and correcting can increase. Comparing total workflows, including tasks after clinical encounters conclude, allows for a more accurate understanding of the impact on patients and clinicians.

Benchmarks facilitate reproducible comparisons but do not capture the full scope of real-world patient impact. The processes through which users convey information and interpret responses, clinical staff reactions, and hospital system integrations all influence outcomes.

In areas such as AI-based electrocardiogram alerts and mammography screening, studies evaluating actual clinical workflows and patient-relevant outcomes have been conducted. As in these examples, outcome metrics tailored to each intended use must be defined. For high-risk functionalities, corresponding clinical evidence is essential.

Time savings, cost, diagnostic accuracy, patient access, and safety must be examined together. While increasing the diversity of research data and expanding multi-institutional collaboration, privacy protection and evaluation independence must also be maintained.

Directions for the Advancement of Medical AI

The value of medical AI cannot be judged solely by its autonomy or model scale. What matters is whether it accurately processes essential information, supports clinicians' judgment, and ensures that patients receive appropriate clinical care.

Alongside technological progress, data quality, validation methodologies, liability and compensation, training, and operational frameworks must be enhanced. Translating research-stage potential into practical healthcare services requires ongoing verification of these conditions. Setting the scope of application based on patient benefits and risks constitutes the fundamental trajectory for the advancement of medical AI.

References

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.
Company press release, 2025-10-15.

<https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZFOBI> (Corporate announcement: "Up to 70% reduction (in-house survey)")

SIND (Year unknown, accessed 2026-10-28). Caretomo

看護記録作成AI支援システム product website. <https://caretomo.jp/> (Corporate announcement: trial results by hospital)

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> (Observational study, real-world operation)

Lukac, P. J., Turner, W., Vangala, S., et al. (2025). Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI*, 2(12). Online 2025-11-26.

<https://doi.org/10.1056/AIoa2501000> (Randomized controlled trial, single institution)

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> (Observational study, 5 institutions)

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> (Institutional self-report)

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*, online

advance publication 2026-10-27. <https://doi.org/10.7326/ANNALS-25-05294> (Observational study, single system)

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> (Media report, for auxiliary verification of figures)

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> (Observational study, single institution)

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> (Simulated clinical encounter experiment)

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> (Observational study, 6 healthcare systems, pre- and post-survey)

Rotenstein, L. S., Holmgren, A. J., Thombly, 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> (Observational study, 5 academic medical centers)

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/> (Institutional self-report)

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> (Real-world clinical operations report, no control group, some co-authors affiliated with the company)

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> (Media report, citing institutional announcement)

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> (Commentary from a specialty society medium)

Abridge (2026). The Evolution of Enterprise AI at Sutter: A Timeline. 2026-08-27. <https://www.abridge.com/blog/sutter-health-partnership-timeline> (Corporate announcement)

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>) (Observational study, pre-post comparison without control group, self-reported. Verification of Tables 2 and 3: 263 analysis participants, burnout model 186 participants, adjusted multivariable model 184 participants)

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/AIoa2500945> (ClinicalTrials.gov NCT06517082) (Randomized controlled trial, University of Wisconsin outpatient, 66 participants, 24 weeks)

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> (Observational study)

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*, online advance publication 2026-10-27. <https://doi.org/10.7326/ANNALS-25-05294> (Observational study)

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> (Media report)

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> (Observational study, repeated cross-sectional survey, baseline 1,600 participants, 6-month 101 participants)

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> (Benchmark, case-based comparison)

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. Online 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> (Benchmark, repeated query experiment)

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> (Observational study, retrospective across 2 institutions, prior to real-world clinical implementation)

(Background knowledge) Shortliffe, E. H. (1976). Computer-Based Medical Consultations: MYCIN. New York: Elsevier. (Historical background, foundational 1976 source)

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. Official announcement, 2026-10-21.

<https://www.unisound.com/news/165> (English edition)

<https://www.unisound.com/en/news/165> (Corporate announcement, self-reported benchmark)

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> (Benchmark design, 2025 foundational source)

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> (Randomized study, confirmed through additional investigation) (Publisher correction notice on April 17, 2026: Nature Medicine 32(5):1937, <https://doi.org/10.1038/s41591-026-04404-8>. unable to confirm whether the correction affected figures in the main text)

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> (Benchmark, bibliography and abstract confirmed through additional investigation)

Microsoft AI (2026). Building Medical Superintelligence. 2026-10-22.

<https://microsoft.ai/news/building-medical-superintelligence> (Corporate announcement)

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> (Simulation study, preprint)

Liévin, V., Palepu, A., Weng, W.-H., Saab, K., Stutz, D., Cheng, Y., Kulkarni, K., Mahdavi, S. S., et al., Karthikesalingam, A., & Schaeckermann, M. (2026). Towards conversational artificial intelligence for disease management. *Nature*, 655, 1292-1299. <https://doi.org/10.1038/s41586-026-10764-5> (Simulation study, patient actors, randomized and blinded)

Tu, T., Schaeckermann, M., Palepu, A., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 642, 442-450. <https://doi.org/10.1038/s41586-025-08866-7> (Simulation study, 2025 foundational source)

Google (2026). Google advances its AMIE research medical AI from diagnosis to treatment. Google Blog, 2026-06-17. <https://blog.google/innovation-and-ai/mode-ls-and-research/google-research/amie-for-disease-management-in-nature> (Corporate announcement)

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> (Simulation study, based on MIMIC-IV real patient records)

Brodeur, P., Koshy, J. M., Palepu, A., Saab, K., et al., Schaeckermann, 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 online. [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) (Preprint: A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. arXiv:2603.08448) (Prospective study, real-world clinical practice, supervised)

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> (Commentary by participating research institution)

Nature Medicine (2026). Agents AMIE and MIRA advance medical AI capabilities. 2026-06-30. <https://www.nature.com/articles/d41591-026-00034-2> (Journal news, title and existence confirmed only)

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> (Institutional press release)

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> (Peer-reviewed paper)

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> (Peer-reviewed paper)

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> (Preprint, posted June 14, 2026)

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> (Preprint, 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> (Peer-reviewed paper)

Chaves, J. M. Z., et al. (2024). Tx-LLM: A large language model for therapeutics. *arXiv:2406.06316*. <https://arxiv.org/abs/2406.06316> (Preprint)

Wang, E., Schmidgall, S., et al. (2025). TxGemma: Efficient and agentic LLMs for therapeutics. arXiv:2504.06196. <https://arxiv.org/abs/2504.06196> (Preprint)

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> (Peer-reviewed paper)

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> (Peer-reviewed paper, detailed AUC figures not cross-checked)

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> (Preprint)

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> (Peer-reviewed paper)

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> (Peer-reviewed paper)

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> (Peer-reviewed conference paper, older foundational literature)

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> (Preprint, v3 June 10, 2026)

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> (Peer-reviewed paper, older foundational literature)

Waddington, C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology*. George Allen & Unwin. (Monograph, older foundational literature)

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> (Preprint, May 27, 2026)

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> (Peer-reviewed paper, foundational dataset)

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> (Peer-reviewed paper, foundational dataset)

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> (Peer-reviewed paper)

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> (Preprint, v3 August 31, 2026)

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> (Preprint)

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> (Preprint, v3 October 9, 2026. Full co-authors not cross-checked in this book's research)

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) (Peer-reviewed commentary paper, volume and pages not cross-checked)

Arc Institute. (2025, December). Virtual Cell Challenge 2025 wrap-up: Winners and reflections. <https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up> (Institutional announcement)

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> (Institutional announcement). Related paper: Virtual Cell Challenge 2026: Benchmarking zero-shot generalization across cellular contexts. Cell (2026), PMID 42648290 (Author information not cross-checked)

Virtual Cell Challenge. (2026). Virtual Cell Challenge 2026: Timeline and rules. <https://virtualcellchallenge.org/> (Final test set release on October 22, 2026, deadline confirmed for November 5)

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> (Peer-reviewed paper, older foundational literature)

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> (Peer-reviewed paper, older foundational literature)

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> (Preprint, May 14, 2026)

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> (Preprint, July 13, 2026)

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> (Peer-reviewed policy paper)

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> (Institutional report, detailed wording not cross-checked)

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> (Peer-reviewed paper)

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> (Peer-reviewed review, August 24, 2026)

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> (Peer-reviewed paper, feasibility trial)

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> (Clinical trial registry information)

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> (Regulatory document, final guidance November 2023)

Lim, Il-gon. (2026, July 23). OPRIMED showcases digital twin platform at dementia conference. *Business Watch*. <https://news.bizwatch.co.kr/article/healthcare/2026/07/23/0014> (News report)

Kwon, Sang-hee. (2026, July 28). OPRIMED pursues digital twin clinical trial hub in Chuncheon Enterprise Innovation Park. *Electronic Times*. <https://www.etnews.com/20260728000025> (News report)

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 (Regulatory document)

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> (Regulatory document)

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> (Regulatory document, draft. Commentary noting draft status as of August 2026: 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> (Peer-reviewed paper, older key literature)

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> (Industry report, February 2021, older foundational material. Analysis of 12,728 phase transitions, confirming 7.9% likelihood of approval from Phase 1)

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> (Corporate announcement)

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> (Regulatory document)

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> (Explanatory material)

International Council for Harmonisation. (2000). ICH E10: Choice of control group and related issues in clinical trials. <https://www.ich.org/page/efficacy-guidelines> (Regulatory document, older foundational literature)

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> (Peer-reviewed review, analysis of FDA approval decisions from 2000-2019)

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> (Regulatory document, draft)

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> (Regulatory document)

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> (Peer-reviewed paper)

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> (Peer-reviewed paper)

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> (Peer-reviewed review)

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> (Peer-reviewed paper, older foundational literature)

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> (Regulatory agency material)

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/> (Statute, section numbers were not cross-checked during research for this book)

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> (Congressional record)

U.S. Food and Drug Administration. (2025, April). Roadmap to reducing animal testing in preclinical safety studies. <https://www.fda.gov/media/186092/download> (Regulatory document)

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> (Regulatory document)

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> (Regulatory document)

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> (Regulatory document, final guidance)

U.S. Food and Drug Administration. (2018-). Model-Informed Drug Development Paired Meeting Program. <https://www.fda.gov/> (FDA regulatory program information. Exact web address and detailed phrasing were not cross-checked during research for this book)

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> (Regulatory agency notice. Step 4 final version: 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> (Regulatory agency notice)

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> (Peer-reviewed paper, online 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> (Peer-reviewed

paper. The DOI 10.1038/s41591-025-03120-w appearing in some secondary sources is not the DOI for this paper)

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> (Corporate announcement)

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> (Clinical trial registration information, verified updated version as of 2026-10-13)

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> (Peer-reviewed paper)

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/NEJMra2204673

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

✂ Materials not covered: For Collaco et al. (2026, npj Digital Medicine) and Alhazba et al. (2026) in some secondary sources, the original texts could not be verified; the latter was not cited because the journal name and DOI did not match each other.

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/news-release-details/tempus-acquire-personalis-more-tightly-integrating-molecular> / Supplementary: 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 (Conference abstract, technical before-and-after comparison. Tempus publicly released presentation materials: <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, DEF14A). U.S. Securities and Exchange Commission. <https://www.sec.gov/Archives/edgar/data/1527753/000119312526427872/d141436ddefm14a.htm> / Supplementary: Personalis, Inc. Form 8-K (date of report 2026-10-02). <https://www.sec.gov/Archives/edgar/data/1527753/000119312526415547/d155320d8k.htm>

※ Materials whose original texts could not be verified: NVIDIA-Tempus joint technical white paper (2025), primary materials related to the PRECISION Challenge.

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. Published June 2026. 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> (Preprint, before peer review)

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 (earnings presentation investor relations activity record). 巨潮资讯网(cninfo) disclosure. <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> (English edition: <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) disclosure.

<http://static.cninfo.com.cn/finalpage/2026-04-27/1225212083.PDF> /

Supplementary: 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 republication. <https://finance.sina.com.cn/stock/relnews/cn/2026-02-03/doc-inhkpvfi7241242.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, preprint)

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/guihuaxxs/c100132/202511/25d0e47241b04a6c85214aa1351e7ff7.shtml>

Ant Group (2025, December 15). Press release on AQ upgrade and naming of '蚂蚁阿福'.

<https://www.antgroup.com/en/news-media/press-releases/1765779300000>

(Primary source recorded in supplementary research materials; re-verification of page content failed during the research for this book)

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). Published on the Chinese Government Website (中国政府网).

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

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). Cninfo (巨潮资讯网) disclosure.

<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/> / Shenzhen Municipal Health Commission (深圳市卫生健康委员会) post (2025-03-18) https://wjw.sz.gov.cn/wzx/content/post_12075823.html (B~C) / Xinhua Net (新华网) (2025-12-11) <http://www.news.cn/tech/20251211/5daf09a9b5034f1eb1e6b6688293cb32/c.html>

※ Materials not covered: The original texts of the '15th Five-Year Plan for National Health (July 2026, target life expectancy of 80 years)' found in some secondary sources, the 2024 84 Scenarios Guidelines (国卫办规划函〔2024〕420号), and the Sidebench (2026) report were not included in the bibliography because their original texts could not be verified during the research for this book.

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 (initial final version December 4, 2024; current posted version issued August 18, 2025). 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 page updated 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> ; See also: 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 (updated as of September 2026). 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 (posted 2025-10-20). 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> ; Consolidated version (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 (Note: Describes application dates prior to revision)

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 (secondary source). <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 (secondary source). <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. Online 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/AIoA2501001>

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* (pre-peer review).

<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/AIdbp2500144>

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> ; Report: 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 (secondary market data for comparison). <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) and 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 (Key prior study from 2024)

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> (Extended version 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> (Preprint)

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> (Publisher's correction notice on April 17, 2026: *Nature Medicine* 32(5):1937, <https://doi.org/10.1038/s41591-026-04404-8>. ; whether the correction affected the figures in the main text could not be verified)

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>

※ Bibliographic note: "AgentClinic, npj Digital Medicine 8:112 (DOI 10.1038/s41746-025-01452-1)" and "meta-analysis 8:245 (DOI 10.1038/s41746-025-01588-0)" found in some secondary sources were confirmed to be different papers upon checking their DOIs and were not used. The actual source of the 1.4% and 1.0% figures is the 2026 paper "Benchmarking large language model-based agent systems for clinical decision tasks" by Liu et al.

Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson. (Classical definition of agent, textbook)

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/AIdbp2500144>

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> (Corporate technical report, preprint)

[Supplementary Reference] 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> (Defines agentic AI as 'perception-reasoning-action cycle, persistent memory, tool use')

※ Bibliographic note: Regarding "Latent-Y / DBTL (Nature 2026)" in some secondary sources, publication in Nature could not be confirmed, and the actual source is the arXiv technical report by the Latent Labs Team. "ProtoPilot" (arXiv:2606.31763, an agent for automated generation of biological experiment protocols) is an existing preprint paper, but was not covered in the main text as it is not directly related to the topic of this section.

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> (Preprint 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>

※ Bibliographic note: The Sidebench industry report, the McKinsey figure of approximately 6%, and the MIT Project NANDA 95% figure found in some secondary sources were not used because their primary sources could not be directly verified during the research for this book. The portion in some secondary sources that stated the scope of Vatsal's study as "2021 to early 2026" was corrected to "October 2023 to June 2025" after verifying the full text of the paper.

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. (Citation re-verified against the references of [7])

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> (Citation verified; detailed body text not viewed)

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>

※ Reference Note: The "MAGDA" case from some secondary sources was not covered because the primary source could not be verified. "ProtoPilot" (arXiv:2606.31763) is an existing preprint paper, but was not addressed as it falls outside the scope of this section.

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> (Published 2026-10-22; all authors affiliated with Microsoft)

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> (Preprint version; figures differ from the Science edition)

Microsoft AI (2026, October 22). Building Medical Superintelligence.

<https://microsoft.ai/news/building-medical-superintelligence> (Official corporate announcement)

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-healthcare/> (Corporate press release; link verified in [3])

Microsoft AI Research Team (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)

※ Reference Note: The Science volume and issue citation "394(6821), eaeh7930" found in some secondary sources was corrected to "394(6822), 423–428" based on a Crossref query. The statement regarding "programmatic innovation in EHRs by 2027" in some secondary sources (citing Chinese media) was not used because the primary source could not be verified.

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> (Preprint, before peer review)

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> (Primary data source; key 2023 material) / SICdb: Rodemund, N., et al. (2023). Salzburg Intensive Care database (SICdb). Re-cited from the references of [1], provided by PhysioNet.

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> (Included in 2018 volume; published online 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> (Source of the MedAgent-Zero method)

Nori, H., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930> (Comparison target SDBench)

※ Reference Note: Claims in some secondary sources such as "ClinPatient focuses on specific Chinese hospitals," "guardian agents," "nine techniques including DPO/PPO," and "50% reduction in charting time" did not match the original paper or could not be verified, and were therefore corrected or omitted.

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> (Correction notice on July 16, 2026: <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). Doctronic regulatory mitigation agreement: AI prescription renewal pilot (Posted 2026-01; contract term indicated as October 2025–October 2026).

<https://commerce.utah.gov/ai/regulatory-relief/authorized-ai-pilots/doctronic/>

Utah Office of Artificial Intelligence Policy. (2026, May 19). Key statistics on the Doctronic pilot program (Public Report 1). <https://commerce.utah.gov/wp-content/uploads/2026/05/Doctronic-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/AIoa2501000>

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/AIoa2500945>

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 [Corporate announcement]. <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 [Corporate announcement]. 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 [Corporate announcement]. <https://microsoft.ai/news/building-medical-superintelligence> (Related paper introduced in the same article: The building blocks of medical superintelligence. *BMJ Digital Health & AI*, 2(1), e000395. <https://bmjdigitalhealth.bmj.com/content/2/1/e000395> — authors and DOI unconfirmed in this book's investigation)

ECRI. (2026). Top 10 health technology hazards for 2026: Misuse of AI chatbots tops annual list [Press release]. <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) [Including COM(2022)496, AI Liability Directive]. Official Journal of the European Union, C series. <http://data.europa.eu/eli/C/2025/5423/oj>

Utah Office of Artificial Intelligence Policy. (2026). Doctrionic regulatory mitigation agreement and Public Report 1 (May 19, 2026).

<https://commerce.utah.gov/ai/regulatory-relief/authorized-ai-pilots/doctrionic/> ;
<https://commerce.utah.gov/wp-content/uploads/2026/05/Doctrionic-Outcomes-May-2026.pdf>

U.S. Food and Drug Administration. (Initial final version 2024-12-04; current posted version 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>

Additional References for the Supplementary Edition

TXP Medical. SpeechER Award and Product Announcement. <https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZF0BI/>

SIND. Caretomo Product Description. <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 et al. 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>

Capital Medical University, Tsinghua University, et al. Antibiotic selection CDSS. <https://medinform.jmir.org/2026/1/e98207>

Unisound. U2-Med Launch. <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, et al. 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 Research Team. Tx-LLM. <https://arxiv.org/abs/2406.06316>

Google Research Team. TxGemma. <https://arxiv.org/abs/2504.06196>

Wang, Tian, et al. 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 et al. CGA-Net. <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/full>

Christopher Baker et al. Octopus. <https://arxiv.org/abs/2607.16262>

Christopher Baker et al. Octopus Revised Manuscript. <https://www.biorxiv.org/content/10.64898/2026.07.05.736565v2>

Christopher Baker et al. CIWM. <https://arxiv.org/abs/2603.02274>

UNC·UT Austin·UPenn. Chreode. <https://arxiv.org/html/2605.28111v1>

Shanghai AI Lab et al. SCALE. <https://arxiv.org/html/2603.17380v3>

Di Francesco et al. 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, et al. 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 et al. 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 Product Description. <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

EMA. Original Copy of EMA Qualification Opinion. 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/xmjnsn4l091-insilico-initiates-phase-iii-clinical-tr>

K-MELLODDY. Project Consortium Introduction.
<https://www.kmelloddy.org/aboutus>

Insilico Medicine. Insilico official Technology introduction.
<https://insilico.com/mediakit>

VoiceCare AI. VoiceCare AI Launch and Mayo Pilot Project. <https://www.prnewswire.com/news-releases/agentic-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 Institutional Overview. <https://newsnetwork.mayoclinic.org/discussion/new-mayo-clinic-study-advances-personalized-prostate-cancer-education-with-an-ehr-integrated-ai-agent/>

Tempus. Tempus Next Official Product Introduction. <https://www.tempus.com/solutions/next/>

Tempus. Lens Initial Release. <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. Next-Generation Lens Unveiling. <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 Paper PMC Full Text. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12714722/>

First Hospital of Hebei Medical University Research Team. Clinical Artificial Intelligence Implementation in Routine Care. <https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

Winning Health Research Team. 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 Launch. <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>

Shenzhen Municipal Health Commission. Introduction of ICU Critical Care Model to Shenzhen. https://wjw.sz.gov.cn/wzx/content/post_12075823.html

National Health Commission of China. Policy Interpretation on Artificial Intelligence + Healthcare. <https://www.nhc.gov.cn/guihuaxxs/c100132/202511/25d0e47241b04a6c85214aa1351e7ff7.shtml>

U.S. FDA. PCCP Final Guidance. <https://www.fda.gov/media/166704/download>

FDA/Health Canada/MHRA. PCCP International Guiding Principles. <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 Consolidated Legislation. <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>

UCLA, AltaMed, et al. Research Team/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>

Research Team/Radiology. Automation Bias in Mammography. <https://pubs.rsna.org/doi/10.1148/radiol.222176>

Stanford Research Team. MedAgentBench Project. <https://stanfordmlgroup.github.io/projects/medagentbench/>

Stanford Research Team. MedAgentBench Preprint.

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

Stanford University, Yixing Jiang, Jonathan H. Chen, et al. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents.

<https://doi.org/10.1056/AIdbp2500144>

U.S. FDA. IDx-DR De Novo Decision.

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, et al. Joint Research Team. 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, et al. Joint Research Team. 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, et al. MedAgentBench: Dataset for Benchmarking LLMs as Agents in Medical Applications. <https://arxiv.org/html/2501.14654v1>

Princeton University, Theodore R. Sumers, Shunyu Yao, et al. Cognitive Architectures for Language Agents. <https://arxiv.org/abs/2309.02427>

Princeton University, Theodore R. Sumers, Shunyu Yao, et al. 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 joint research team. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making.

<https://arxiv.org/abs/2404.15155>

MIT, Google Research, Seoul National University Hospital joint research team. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making.

<https://arxiv.org/html/2404.15155v3>

Tsinghua University AIR et al. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. <https://arxiv.org/abs/2405.02957>

Tsinghua University AIR et al. 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/>

No single development entity; expressed as a goal by Microsoft et al. 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-led and HIKER Group et al. joint research team. 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>

Paper research team / 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>

Platform research: research team affiliated with Sanya People's Hospital and Sichuan University; commercial AI manufacturer undisclosed in the paper. 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 research team. 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 research team. 人工智慧心電圖之落地運用-三軍總醫院經驗 (2024-05-31 Grand Round announcement). <https://www.tsgh.ndmutsggh.edu.tw/news/191/10010/24887/10864>

MASAI: Lund University et al.; Transpara: ScreenPoint Medical BV. Mammography Screening with Artificial Intelligence. <https://portal.research.lu.se/en/projects/mammography-screening-with-artificial-intelligence/>

MASAI: Lund University et al.; Transpara: ScreenPoint Medical BV. Transpara, 510(k) K181704 decision and summary. https://www.accessdata.fda.gov/cdrh_docs/pdf18/K181704.pdf

MASAI: Lund University et al.; Transpara: ScreenPoint Medical BV. Breast AI Research Portfolio. <https://screenpoint-medical.com/evidence>

IDx LLC, currently Digital Diagnostics. De Novo classification request for IDx-DR, DEN180001. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf

IDx LLC, currently 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/>

US FDA regulatory guidance. 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

Wittmann et al.; Microsoft, IBBIS, et al. research team. 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>

Book Information

The Great Transformation of AI and Healthcare

Author | Attorney Kim Kyung-jin

Publisher | Kim Kyung-jin

Publishing House | Attorney Kim Kyung-jin Publishing House

Publisher Registration | 2025. 3. 10. (No. 2025-000015)

Address | Room 304, Baekil Building, 91 Jeonnong-ro, Dongdaemun-gu, Seoul

Telephone | 02-6338-1905

Email | kimkj008@gmail.com

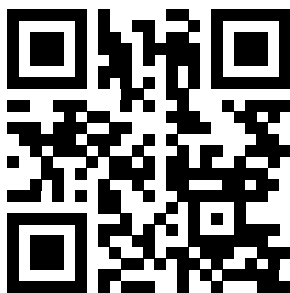
© Kim Kyung-jin 2026

This book is the intellectual property of the author, and unauthorized reproduction or copying is prohibited.

Note) The cover and chapter illustrations are conceptual representations generated by artificial intelligence and are not photographs taken of actual clinical settings. Portions of this book's content and editing process were assisted by artificial intelligence tools.

Support This AI Library Book

If this book was useful, you can support future translations, public editions, and open AI knowledge projects through PayPal.



PayPal

<https://paypal.me/kimkjj>

Scan the QR code or open the link above.