Print ISSN: 2434-9186 Online ISSN: 2434-9194

GHM

Global Health & Medicine

Volume 8, Number 3
June 2026

www.globalhealthmedicine.com

Print ISSN: 2434-9186
Online ISSN: 2434-9194
Issues/Year: 6
Language: English

Global Health & Medicine

Global Health & Medicine

Global Health & Medicine (Print ISSN 2434-9186, Online ISSN 2434-9194) is an international, open-access, peer-reviewed journal, published by the Japan Institute for Health Security (JIHS), which is a national research and development agency in Japan that covers advanced general medicine, basic science, clinical science, and international medical collaboration.

1. Mission and Scope

Global Health & Medicine is dedicated to publishing high-quality original research that contributes to advancing global health and medicine, with the goal of creating a global information network for global health, basic science as well as clinical science oriented for clinical application.

The articles cover the fields of global health, public health, and health care delivery as well as the seminal and latest research on the intersection of biomedical science and clinical practice in order to encourage cooperation and exchange among scientists and healthcare professionals in the world.

2. Manuscript Types

Global Health & Medicine publishes Original Articles, Brief Reports, Reviews, Policy Forum articles, Communications, Editorials, Letters, and News on all aspects of the field of global health and medicine.

3. Editorial Policies

Global Health & Medicine will perform an especially prompt review to encourage submissions of innovative work. All original research manuscripts are to be subjected to an expeditious but rigorous standard of peer review, and are to be edited by experienced copy editors to the highest standards.

We aspire to identify, attract, and publish original research that supports advances of knowledge in critical areas of global health and medicine.

Editorial and Head Office:

Global Health & Medicine
Japan Institute for Health Security,
1-21-1 Toyama Shinjuku-ku,
Tokyo 162-8655, Japan
URL: www.globalhealthmedicine.com
E-mail: office@globalhealthmedicine.com

Print ISSN: 2434-9186
Online ISSN: 2434-9194
Issues/Year: 6
Language: English

Global Health & Medicine

CONTENTS

Volume 8, Number 3, 2026

CORRESPONDENCE

COVER FIGURE

Global Health & Medicine. 2026; 8(3):140-147.
POLICY FORUM
DOI: 10.35772/ghm.2026.01052

Japanese regulation and approval process for medical artificial intelligence (AI) as software as a medical device (SaMD): Current status and emerging challenges

Sara Takahashi*, Tomohiko Makino, Reiko Mizutani, Takanori Hirano, Yumiko Nomura

Medical Device Evaluation Division, Pharmaceutical Safety Bureau, Ministry of Health, Labour and Welfare, Tokyo, Japan.

Abstract: The rapid expansion of artificial intelligence (AI) in healthcare has led to increasing adoption of AI-based software as a medical device (SaMD). This paper reviews the current regulatory and approval framework for AI-based SaMD in Japan and discusses emerging challenges associated with generative and adaptive AI technologies. Under the Pharmaceuticals and Medical Devices Act (PMD Act), software intended for diagnosis, treatment, or prevention is regulated as a medical device when classified as Class II or higher, and its clinical utility, performance, and safety are evaluated. While the number of approved AI-based SaMDs has increased, most existing products are task-specific systems supporting clinical decision-making within defined scopes. Recent advances in generative AI introduce novel regulatory issues, including difficulties in defining intended use, evaluating reliability of natural language outputs, and managing continuously evolving performance after market entry. These characteristics challenge conventional regulatory paradigms based on fixed product specifications. In light of ongoing international regulatory developments, key issues include clarifying scope of regulated functions, strengthening lifecycle and change management approaches, enhancing transparency, and improving user literacy. Developing adaptive regulatory frameworks that balance innovation, patient safety, and regulatory clarity will be essential for responsible integration of generative AI into healthcare.

***Keywords*:** medical device, Software as a Medical Device (SaMD), AI, regulation, Japan

1. Introduction

The use of artificial intelligence (AI) in the medical field has been advancing rapidly. Medical AI encompasses a wide variety of products, including software aimed at reducing the workload of healthcare professionals by assisting with in-hospital tasks such as managing consultation schedules and maintaining a medical device (MD), as well as software that supports physicians in deciding on treatment and diagnosis strategies by analyzing patient data. Furthermore, recent advancements in AI technology have been remarkable, and its potential seems limitless. However, from a MD regulation perspective, there are currently no special frameworks in place for evaluating or managing risks associated with generative AI technology.

This paper aims to summarize current Japanese regulatory and approval framework for AI-based software as a medical device (SaMD), identify emerging regulatory challenges posed by generative and adaptive AI, and discuss future directions for balancing innovation, patient safety, and regulatory clarity.

2. Current status of regulations for medical devices in Japan

In Japan, the Pharmaceuticals and Medical Devices Act (PMD Act) defines MDs as "machinery, equipment, *etc.* that are intended for use in diagnosis, treatment, or prevention of diseases in humans, or intended to affect structure or functioning of the bodies of humans". Furthermore, in the 2014 amendment to the PMD Act, it was clarified that standalone software is included in the scope of MDs.

As shown in Figure 1, hardware MDs are divided into four classes according to the risks they pose. Class I (lowest class) products, which pose little risk to humans even in the event of failures or malfunctions, do not require review by a regulatory body and can be manufactured by the manufacturer simply by notifying the Pharmaceuticals and Medical Devices Agency (PMDA). However, Class II and above products are required to undergo review by the PMDA or a third-party certification body. On the other hand, regarding SaMDs, Class I software is exempt from regulation under the

PMD Act, and Class II to IV software is subject to regulation under the PMD Act. Therefore, not all AI technology used in the medical fields qualifies as a MD; rather, only software used for diagnosis, treatment, or prevention, and possessing functionality equivalent to Class II to IV, qualifies as a MD.

The Ministry of Health, Labour and Welfare (MHLW) has issued guidelines and case examples outlining the criteria for determining whether software qualifies as a MD (*1,2*). Based on the documents issued by the MHLW, Figure 2 shows the fundamental concept used in Japan when determining whether software falls under the definition of a MD. When considering whether software qualifies as a MD or how to classify it, it is necessary

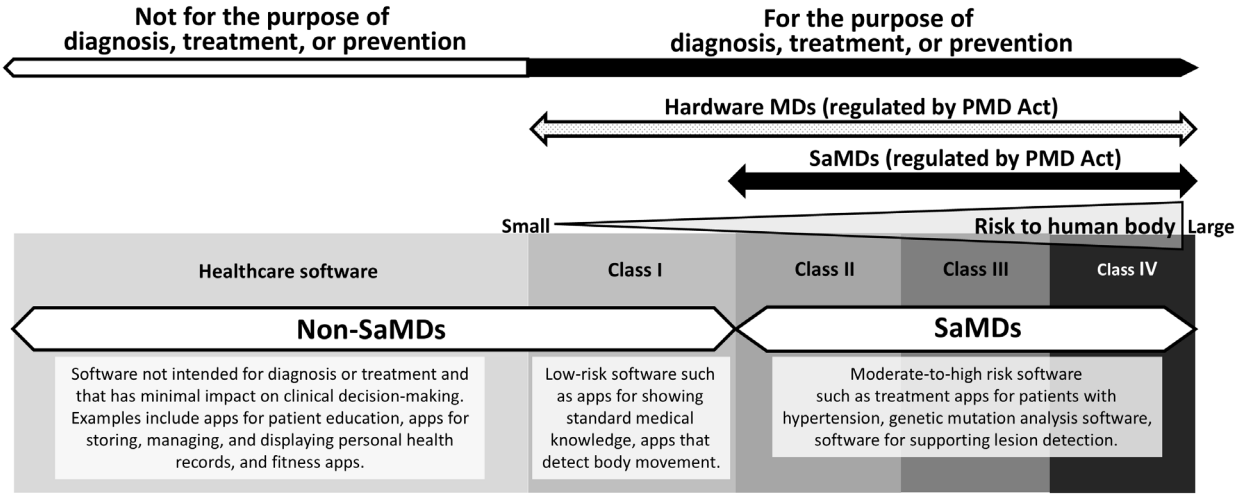


Figure 1. Japanese regulation of medical software. Software intended for diagnosis, treatment, or prevention that presents a moderate or higher risk to the human body (Class II or above) is regulated as SaMD. In contrast, low-risk software intended for purposes such as health management or information provision is categorized as non-SaMD. Thus, the scope of software subject to regulation is systematically defined based on the presence of a medical purpose and the level of risk. *Abbreviations*: MD, medical device; PMD Act, Pharmaceuticals and Medical Devices Act; SaMD, software as a medical device.

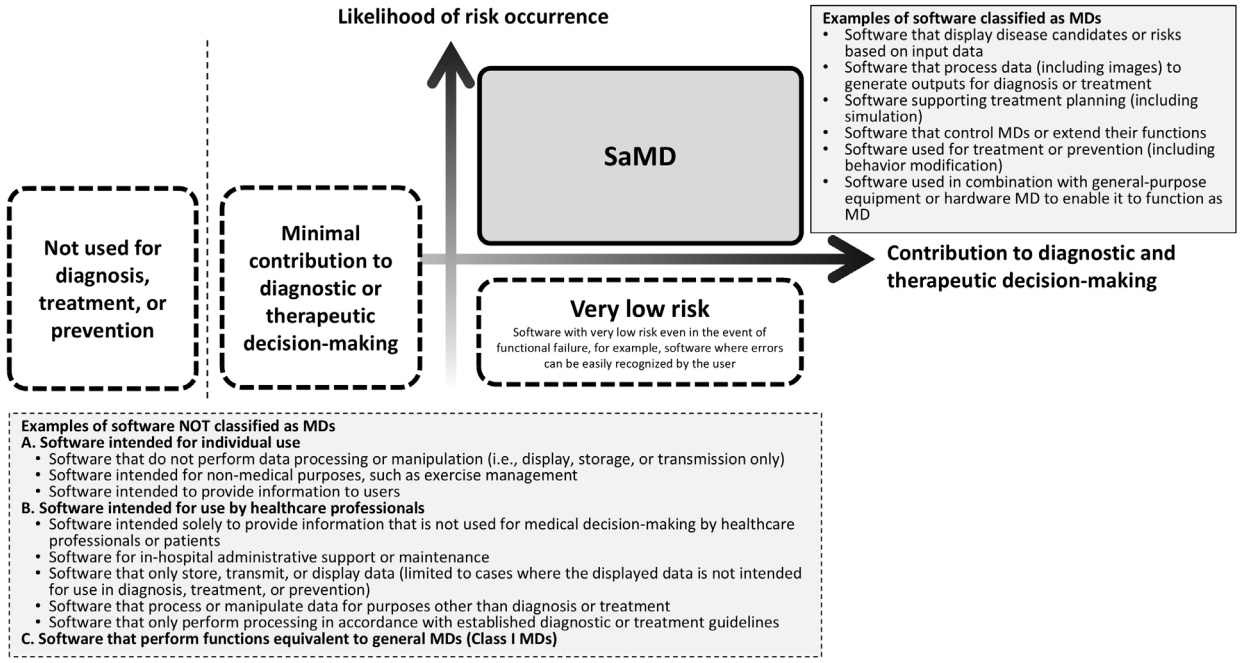


Figure 2. Fundamental concept for classification of SaMDs and non-SaMDs in Japan. Whether software is classified as SaMD is determined from two perspectives: the degree to which it contributes to diagnosis, treatment, or prevention, and likelihood of risk to the human body. Software with a low level of involvement in clinical practice—such as simple information display or storage functions—or software used for non-medical purposes is considered non-SaMD. In contrast, software that has functions influencing clinical decision-making, such as presenting diagnostic candidates or supporting treatment planning, is subject to regulation as SaMD. *Abbreviations*: MD, medical device; SaMD, software as a medical device.

to consider its contribution to the diagnosis, treatment, or prevention of diseases and the probability of overall risk to humans. This paper primarily focuses on software that is regulated as SaMD classified as Class II or higher under the PMD Act.

Figure 3 illustrates the process that software vendors must follow when introducing software intended for medical use in Japan. It is essential to first determine whether the product falls within the scope of MDs. If software qualifies as a MD, it must undergo review by a third party certification body or the PMDA.

Figure 4 shows the number of approved or certified SaMDs in Japan. This figure illustrates the number of SaMDs approved by the PMDA or certified by third-party certification bodies as of September 30, 2025. These counts were derived from the lists of approved and certified MDs on the PMDA website. Because use of AI in certified MDs is not publicly disclosed, this figure presents the total number of all approved or certified MDs, irrespective of whether they incorporate AI. While diagnostic SaMDs have been prevalent in the past, development of therapeutic SaMDs is progressing, expanding the scope of treatment beyond hypertension, alcoholism, depression, attention deficit hyperactivity disorder (ADHD), insomnia, and other conditions. Furthermore, there is an increasing number of over-the-counter (OTC) SaMDs (home-use SaMDs) that detect signs of diseases and are primarily intended for general consumer use. Variations of home-use SaMDs are also increasing, including SaMDs that notify users of signs of Sleep Apnea Syndrome (SAS) and hearing-related diseases. Development of gene mutation analysis SaMDs is also becoming more active within the field of diagnostic SaMDs.

3. Regulation of AI-based SaMDs in Japan

In the Japanese conventional regulation of MD, its principle, structure, materials, specifications, usage methods, and manufacturing methods for each MD should be stipulated in its regulatory application, and the regulatory body reviews whether its effectiveness, safety, and quality are guaranteed. In other words, the evaluation is conducted after defining what the MD is and how the MD is used. Furthermore, in evaluating MDs, it is necessary to demonstrate that the MD has clinically significant effectiveness and safety for the specific disease that it is intended for. Similarly, regulation for SaMDs also specifies the specifications and usage methods of the SaMDs and evaluates effectiveness and safety of the SaMDs for the intended use.

The number of approved SaMDs that utilize AI technology as their functions is increasing. As of September 2025, there are 51 approved SaMDs reviewed

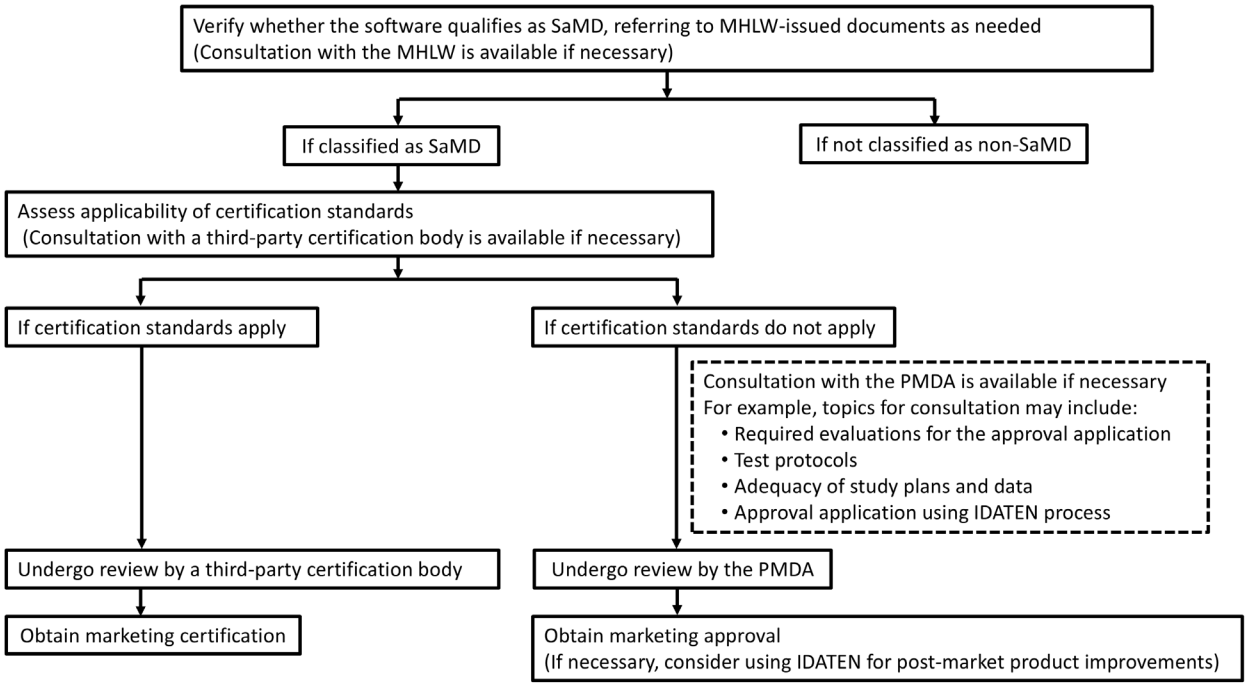


Figure 3. Process for introducing SaMD expected to be used in clinical practice into Japanese market. For software intended for clinical use, the first step is to determine whether it qualifies as SaMD. If it does, applicability of certification standards is then assessed. Where certification standards apply, the product undergoes review by a registered third-party certification body; where they do not apply, approval by the PMDA is required. In addition, the overall regulatory pathway to market entry is structured through a series of processes, including pre-submission consultations, consideration of study protocols, and use of mechanisms such as the IDATEN process. *Abbreviations*: IDATEN, Improvement Design within Approval for Timely Evaluation and Notice; PMDA, Pharmaceuticals and Medical Devices Agency; SaMD, software as a medical device.

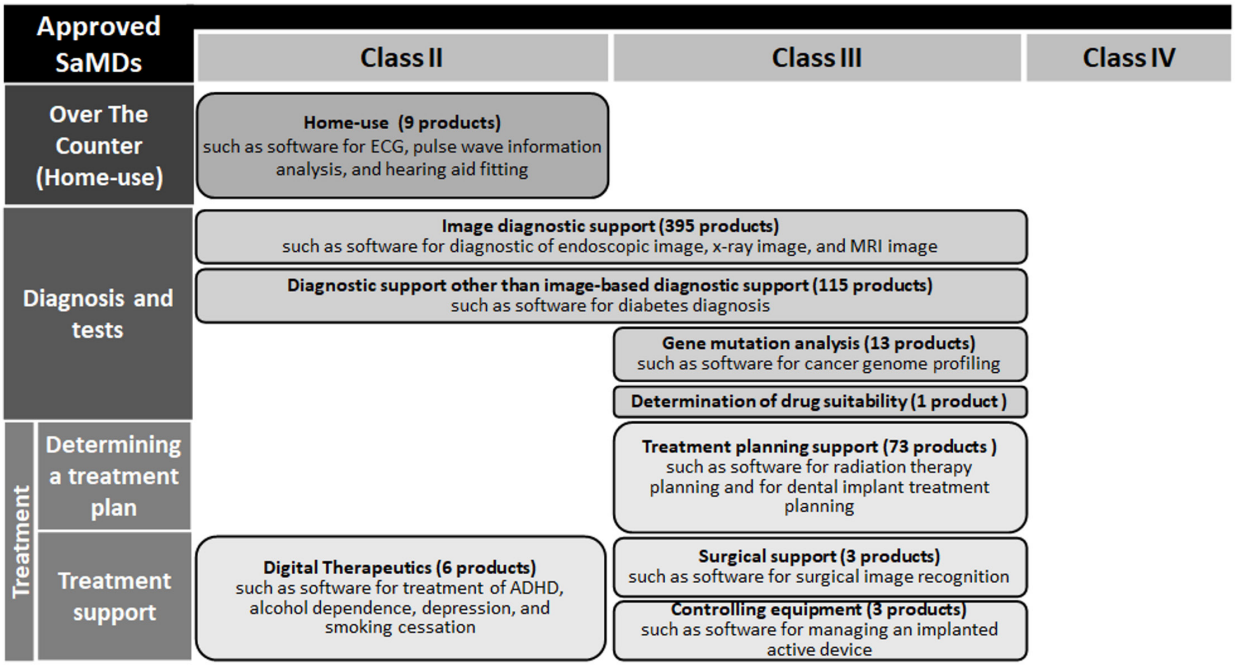


Figure 4. The number of approved SaMDs reviewed by PMDA and certified by third-party certification bodies in Japan as of September 30, 2025. Among SaMDs that have been approved or certified, image analysis and diagnostic support constitute the majority. At the same time, their applications have expanded to a wide range of purposes, including diagnostic support, treatment planning support, and therapeutic software. In terms of classification, most products fall into Class II and Class III, and range of application areas is also broadening to include lifestyle-related diseases, psychiatric disorders, and genomic analysis. In addition, a certain number of consumer-oriented products, such as home-use SaMD, have emerged, suggesting an expanding scope of application. *Abbreviations*: ADHD, attention-deficit/hyperactivity disorder; ECG, electrocardiogram; PMDA, Pharmaceuticals and Medical Devices Agency; SaMD, software as a medical device.

by the PMDA and published on the PMDA website as AI-based SaMD (*3*), which have functions using AI technology, regardless of whether they are initial approvals or approvals with partial changes to approved items. In addition, functions not intended to improve patient outcomes—such as those aimed solely at reducing workload of healthcare providers—are excluded from this list; therefore, it does not encompass all SaMDs that utilize AI.

Examples of AI-based SaMD include SaMDs that detect neoplastic lesions from endoscopic images (*4*), cerebral aneurysms, pulmonary nodules, and pneumonia, *etc*. from X-ray and MRI images, and SaMDs that indicate location and area of anatomical structures on surgical images during surgery. There are differences among countries in evaluation and regulation of AI-based MDs (*5*), but in Japan, regardless of whether AI technology is used, the general review criteria required for regulatory approval of SaMDs remain the same, and evaluation focuses on clinical utility (medical value that SaMD brings to diagnosis), clinical performance (SaMD's output information accuracy and processing capabilities), basic performance (whether it works as designed), and basic safety. When evaluating AI-powered functions, it is necessary to pay attention to bias in the evaluation dataset and relationship between the training dataset and evaluation dataset, in order to determine whether results can be generalized to the target population in actual clinical practice.

As shown in Table 1, several discussions have taken place in Japan regarding the regulatory approach to AI-based SaMD (*6-10*). To date, AI-based SaMD has generally been a product that supports physicians' diagnosis and treatment within a limited scope, such as specific diseases or medical fields, and discussions have focused on task specific identification AI specialized for specific tasks.

Generative AI technology capable of generating text, images, programs, and more has rapidly advanced. It is expected that Generative AI technology will greatly broaden the possibility of software functions supporting treatment and diagnosis, as it will be possible to generate diverse answers and creative content not dependent on specific tasks, in response to various questions from users in natural language or voice.

Until now, AI-based SaMDs that have been put into practical use have allowed for the clear definition of target functions, and accuracy of the output could be quantitatively evaluated using rule-based methods by outputting inference results as labels, and other structured outputs. However, with SaMDs that utilize generative AI technology, it may be difficult to limit scope of use due to the high versatility and multipurpose nature of generative AI, and there is greater ambiguity in determining accuracy of natural language output. Additionally, with the emergence of generative AI, it is anticipated

that SaMDs that utilize generative AI technology will continue to learn after marketing, and their functions and performance will change continuously. Considering these characteristics, it is necessary to clarify which functions should be regulated as MDs, the appropriate timeline for such regulation, and specific risks that should be evaluated in comparison with existing MDs.

Furthermore, since generative AI built on extremely large datasets may utilize AI models provided by external vendors, it is important to specify, on a product-by-product basis, the extent to which manufacturers of MDs should acquire information about the underlying AI models and extent of their responsibility for explanation, oversight, and management.

A hallucination, meaning AI's production of plausible but completely wrong or false information, is one of emerging challenges unique to generative AI. Some countermeasures for hallucination can be considered, such as Retrieval-Augmented Generation (RAG) techniques. In addition, users' literacy is critical so that users understand risks of hallucination, can interpret accuracy and appropriateness of information presented by AI-based SaMD, and can select information that they deem useful for diagnosis, treatment, or prevention of diseases. When formulating regulations, it is necessary to consider how to protect effectiveness and safety of patients while not hindering development of AI-based SaMDs which are valuable to patients and healthcare professionals.

4. Global context: Regulatory evolutions regarding MDs utilizing AI overseas

The adoption of medical AI is also thriving overseas, and its use in diagnosis and treatment is progressing, with medical AI such as Open Evidence's Deep Consult clinical decision support platform being widely put into practical use (*11*). As shown in Table 2, various countries have published guidance and are conducting demonstration projects on how to regulate MDs utilizing AI (*12-16,18-24*). In the United States, for example, discussion papers and guidance on regulation of AI/ML-based SaMD have been published since around 2020, and more recently, guidance on lifecycle management

Table 1. Main initiatives focused on MDs utilizing AI in Japan

Year	Main initiatives focused on MDs utilizing AI (*Ref.*)
2017–2018	The Subcommittee on Artificial Intelligence and its Applications in the Medical Field of the Science Board was held in 2017. The subcommittee discussed the characteristics, handling, and challenges of AI medical systems that utilize machine learning including deep learning. The report was published in 2018 (*6*).
2019	"Guidance for Evaluation of Artificial Intelligence-Assisted Medical Imaging Systems for Clinical Diagnosis" was published (*7*).
2020	With the amendment of the PMD Act, the IDATEN process has been introduced, which serves as a system to confirm change plans for MDs using AI technology, and other software based functions that are expected to be frequently improved after-market release (*8*).
2021	Based on the "Review of Approval and Review Processes Regarding SaMD" in the Regulatory Reform Council's "Implementation Items for Regulatory Reform in the Immediate Future", the PMDA has begun building a flexible and speedy review system and improving review standards to cope with frequently updated AI-based SaMD and other technologies (*9*).
2023	The Subcommittee on SaMD Utilizing AI and Machine Learning of the Science Board was held in 2023. The subcommittee considered how to reuse data, the conditions required for evaluation data, and how to review Adaptive AI intended to change performance after marketing and published the "Report on SaMDs Utilizing AI" (*10*).

Abbreviations: AI, artificial intelligence; IDATEN, Improvement Design within Approval for Timely Evaluation and Notice; MD, medical device; PMD Act, Pharmaceuticals and Medical Devices Act; PMDA, Pharmaceuticals and Medical Devices Agency; SaMD, software as a medical device.

Table 2. Examples of initiatives focused on MDs utilizing AI overseas

Region / Regulator body	Recent main initiatives focused on MDs utilizing AI (*Ref.*)
United States / U.S. Food and Drug Administration (FDA)	• Publication of guidance for AI-enabled device software functions and clinical decision support software (*12-14*) • Publication of AI Enabled MDs List (*18*)
United Kingdom / Medicines and Healthcare Products Regulatory Agency (MHRA)	• Publication of guidance for software and AI as a MD (*19,20*) • AI Airlock sandbox (*16*)
European Union (EU)	• Publication of document for MDAI (*21*)
Republic of Korea / Ministry of Food and Drug Safety (MFDS)	• Publication of guidance for AI/generative AI-based MDs (*15,22-24*)

Abbreviations: AI, artificial intelligence; MD, medical device; MDAI, medical device artificial intelligence.

and marketing submission of AI-enabled device software functions was published in 2025 (*12*). Also, guidance on change management plans for AI software was published in August 2025, a development that is presumed to consider the adoption of Adaptive AI (*13*) , and guidance on Clinical Decision Support Software was published in January 2026 (*14*), indicating that proactive regulatory reforms are underway. Guidance on Clinical Decision Support Software provides developers of medical AI with guidance on scope of regulation and points to consider when introducing such systems to the market, and it is expected to promote their practical application. Efforts in various countries show a trend toward adopting a management approach that encompasses the entire product lifecycle from pre-marketing to post-marketing, including change management. With regard to guidance on MDs using generative AI, the Ministry of Food and Drug Safety (MFDS) of the Republic of Korea published guidance in January 2025 (*15*).

Furthermore, sandbox projects are being implemented in various countries. In the United Kingdom (UK), for example, the AI Airlock Sandbox Program involves regulatory authorities, healthcare professionals, and development companies collaborating to identify regulatory gaps and conduct demonstration projects to resolve issues to enable safe and rapid introduction of AI technology into healthcare settings (*16*). In Utah, the United States (USA), based on the "Artificial Intelligence Regulatory Sandbox Act" enacted in 2024, a demonstration program is being implemented to introduce autonomous AI for prescription updates for chronic diseases, advancing efforts to enable AI-driven prescription update decisions for patients with chronic illnesses (*17*). These sandbox strategies tell us that countries intend, when regulating MDs using generative AI, to first accumulate a certain level of usage experience in controlled environments and then appropriately identify challenges and examine corresponding countermeasures.

5. Discussion

As described above, discussions are ongoing in various countries regarding the appropriate regulatory frameworks for SaMD incorporating AI into their functionality. In Japan, deliberations on generative AI have also recently begun within a research project commissioned by the MHLW. Taking into account international developments as well as discussions within the research group, four major issues can be identified as requiring further consideration in Japan.

First, there is a need to clarify the scope of AI-based software that should be regulated as SaMD. As noted above, the MHLW has already published several documents outlining the scope of software subject to MD regulation. It is considered that fundamental principles applied to conventional software-based MDs are also applicable to AI-based MDs. However, providing more detailed and specific guidance, particularly for MDs utilizing AI—especially generative AI—is expected to facilitate practical development and implementation for software developers.

Second, there is the issue of the regulatory approach under the PMD Act. For SaMD that may change after market entry, regulatory frameworks such as IDATEN, which manage post-market modification plans, have already been introduced. In addition to expanding regulations that take such change management into account, it is also necessary to re-examine the conventional regulatory paradigm, which assumes fixed product specifications. This includes considering new approaches to premarket review that anticipate product modifications, as well as lifecycle-oriented regulatory oversight spanning both premarket and post-market phases.

Third, attention should be given to initiatives aimed at improving information provision to users and enhancing user literacy. Currently, public disclosure regarding approved or certified MDs mainly consists of package inserts and lists of approved or certified MDs. However, there is an increasing need to consider providing more detailed information on device functionality and to promote initiatives that enhance literacy of users. The importance of transparency in the disclosure of evidence for SaMD has also been highlighted, particularly by the clinical community (*25*), and further discussion among industry, government, and academia will be necessary moving forward.

Fourth, there is a need to establish mechanisms that enable empirical validation. In Japan, there is a "regulatory sandbox system" under jurisdiction of the Cabinet Office for the social implementation of new technologies and business models, but there is no sandbox system specific to MDs utilizing AI. The overseas initiatives can serve as a reference when considering regulations on AI-based SaMD utilizing generative AI technology in Japan, and it is hoped that practical considerations will progress in Japan in the future, through use of sandbox systems and other similar mechanisms.

6. Expectations for the future introduction of medical AI

To facilitate the digital transformation of healthcare through medical AI as a MD in Japan, it is necessary to clarify the scope of regulation and scope of responsibility of manufacturers and distributors of MD. So as not to fall behind other countries, it is critical to promptly organize regulatory frameworks that take into account product changes after marketing. The MHLW is seriously working on facilitating AI in medical practice. Placing excess responsibility for ensuring effectiveness and safety of AI-based SaMDs on manufacturers and distributors of MDs would hinder development and deployment of

such devices. It is also necessary to nurture literacy of users regarding use of AI-based SaMDs. Versatile AI that utilize adaptive learning may urge MHLW and regulators to reform and renovate regulatory algorithms. Carefully designed regulatory frameworks for generative AI-based SaMDs are essential to balance between innovation and patient safety. In this context, Japan's experience in SaMD review may provide valuable insights for future regulatory discussions through multistakeholder collaboration.

Funding: None.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Ministry of Health, Labour and Welfare. Regarding the partial revision of the guidelines concerning whether software is classified as a medical device. *https://www.mhlw.go.jp/content/11120000/001082227.pdf* (accessed April 30, 2026). (in Japanese)
2. Ministry of Health, Labour and Welfare. Regarding the determination of whether software qualifies as a medical device. Administrative Notice. March 31, 2020. *https://www.mhlw.go.jp/content/11120000/001082229.pdf* (accessed April 30, 2026). (in Japanese)
3. Pharmaceuticals and Medical Devices Agency. List of approved SaMDs. *https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/devices/0052.html* (accessed April 30, 2026). (in Japanese)
4. Uchida D, Kawarasaki S, Ikuma M. Software as a medical device (SAMD) based on artificial intelligence and machine learning: The Pharmaceutical and Medical Devices Agency (PMDA) Perspective. Gastroenterol Endosc. 2021; 63:2297-2307. (in Japanese)
5. Yuba M, Iwasaki K. Systematic analysis of the test design and performance of AI/ML-based medical devices approved for triage/detection/diagnosis in the USA and Japan. Sci Rep. 2022; 12:16874.
6. Chinzei K, Shimizu A, Mori K, *et al.* Regulatory science on AI-based medical devices and systems. Adv Biomed Eng. 2018; 7:118-123.
7. Ministry of Health, Labour and Welfare. Guidance for evaluation of artificial intelligence-assisted medical imaging systems for clinical diagnosis. *https://www.mhlw.go.jp/content/10601000/000515843.pdf* (accessed April 30, 2026). (in Japanese)
8. Ministry of Health, Labour and Welfare. Handling application for confirming change plans for medical devices. *https://www.mhlw.go.jp/content/11120000/000665757.pdf* (accessed April 30, 2026). (in Japanese)
9. Regulatory Reform Promotion Council. Review of approval and review processes regarding SaMD. *https://www8.cao.go.jp/kisei-kaikaku/kisei/publication/opinion/211222.pdf* (accessed April 30, 2026). (in Japanese)
10. Subcommittee on Software as a Medical Device Utilizing AI and Machine Learning of the Science Board. Report on AI-based Software as a Medical Device (SaMD). *https://www.pmda.go.jp/files/000266099.pdf* (accessed April 30, 2026).
11. OpenEvidence. *https://www.openevidence.com/* (accessed April 30, 2026).
12. U.S. Food and Drug Administration. Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations (Draft Guidance). *https://www.fda.gov/media/184856/download* (accessed April 30, 2026).
13. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. *https://www.fda.gov/media/166704/download* (accessed April 30, 2026).
14. U.S. Food and Drug Administration. Clinical decision support software: Guidance for industry and Food and Drug Administration staff. *https://www.fda.gov/media/109618/download* (accessed April 30, 2026).
15. Ministry of Food and Drug Safety, Republic of Korea. Guidance on licensing and review of generative AI medical devices. *https://www.mfds.go.kr/brd/m_1060/view.do?seq=15628&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=6* (accessed April 30, 2026). (in Korean)
16. Medicines and Healthcare products Regulatory Agency. AI Airlock: The regulatory sandbox for AIaMD. *https://www.gov.uk/government/collections/ai-airlock-the-regulatory-sandbox-for-aiamd* (accessed April 30, 2026).
17. Utah Department of Commerce. News release: Utah and Doctronic announce groundbreaking partnership for AI prescription medication renewals. *https://commerce.utah.gov/2026/01/06/news-release-utah-and-doctronic-announce-groundbreaking-partnership-for-ai-prescription-medication-renewals/* (accessed April 30, 2026).
18. U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices list. *https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices* (accessed April 30, 2026).
19. Medicines & Healthcare products Regulatory Agency. Guidance: Software and AI as a Medical Device Change Programme roadmap. *https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap* (accessed April 30, 2026).
20. Medicines & Healthcare products Regulatory Agency. Guidance: Software and artificial intelligence (AI) as a medical device. *https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device* (accessed April 30, 2026).
21. Joint Artificial Intelligence Board and Medical Device Coordination Group. Interplay between the Medical Devices Regulation (MDR) & *In vitro* Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). *https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en?filename=mdcg_2025-6_en.pdf* (accessed April 30, 2026).
22. Ministry of Food and Drug Safety, Republic of Korea. Guidance on the review and approval of artificial intelligence (AI)-based medical devices. *https://www.mfds.go.kr/eng/brd/m_40/view.do?seq=72627* (accessed April 30, 2026).
23. Ministry of Food and Drug Safety, Republic of Korea. Guidance on clinical trials design of artificial intelligence (AI)-based medical devices. September 20, 2023. *https://*

www.mfds.go.kr/eng/brd/m_40/view.do?seq=72628&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=2 (accessed April 30, 2026).

24. Ministry of Food and Drug Safety, Republic of Korea. Guidance on the review and approval of digital therapeutics (DTx) (Revision). *https://www.mfds.go.kr/eng/brd/m_40/view.do?seq=72624&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=2* (accessed April 30, 2026).
25. Kikuchi T, Walston SL, Takita H, *et al*. Scoping review of regulatory transparency in AI-based radiology software: Analysis of PMDA-approved SaMD products. Jpn J Radiol. 2026; 44:1095-1111.

Received May 11, 2026; Revised June 10, 2026; Accepted June 14, 2026.

Released online in J-STAGE as advance publication June 19, 2026.

**Address correspondence to*:
Sara Takahashi, Medical Device Evaluation Division, Pharmaceutical Safety Bureau, Ministry of Health, Labour and Welfare (Current Affiliation: Office of Software as a Medical Device, Pharmaceuticals and Medical Devices Agency (PMDA), Shin-Kasumigaseki Building, 3-3-2, Kasumigaseki, Chiyoda-ku, Tokyo 100-0013, Japan).
E-mail: takahashi-sara@pmda.go.jp

Global Health & Medicine. 2026; 8(3):148-153.

DOI: 10.35772/ghm.2026.01046

Proactive adoption of generative artificial intelligence (AI) in the operations of Japan's Pharmaceuticals and Medical Devices Agency (PMDA): Current initiatives, governance, and future perspectives

Kohei Amakasu[1], Junichi Kawana[1], Osamu Kotera[1], Tomoharu Numanyu[2], Akihiro Nakajima[3], Koichi Ishikawa[3], Yuka Kobayashi[4], Yoshiaki Uyama[5,*]

[1] Office of Regulatory Science Coordination, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan;
[2] Office of New Drug I, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan;
[3] Office of Information Technology Promotion, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan;
[4] Office of Planning and Operations, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan;
[5] Center for Regulatory Science, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan.

Abstract: The Pharmaceuticals and Medical Devices Agency (PMDA) continues to face increasing operational demands stemming from growing regulatory complexity, expanding data volumes, and evolving scientific and societal expectations. In this context, the appropriate adoption of generative artificial intelligence has emerged as a potential approach for enhancing operational efficiency while reinforcing scientific rigor and accountability. This article describes the current status of generative artificial intelligence utilization at PMDA, outlines its governance framework, and discusses future perspectives for its sustainable application based on institutional experience, internal policy development, and planned/ongoing proof-of-concept activities conducted within PMDA. We summarize a phased implementation strategy that combines commercially available generative artificial intelligence tools for administrative support with the exploration of large language models in secure internal environments for scientifically specialized tasks. Central to this approach is a governance framework that emphasizes human-in-the-loop decision-making, staged evaluation, information governance, and staff capacity building. We also present practical use cases across information collection, analysis and evaluation, and dissemination activities to illustrate how generative artificial intelligence may support regulatory work without replacing human judgment. In conclusion, PMDA's experience suggests that proactive yet cautious adoption of generative artificial intelligence, grounded in robust governance and organizational learning, can improve productivity and enhance scientific capacity within regulatory authorities while maintaining public trust and institutional accountability.

***Keywords*:** generative artificial intelligence, regulatory science, operational efficiency

1. Introduction

The Pharmaceuticals and Medical Devices Agency (PMDA) conducts its daily operations with the objective of ensuring the quality, efficacy, and safety of pharmaceuticals, medical devices, and related products in Japan through its core functions of regulatory reviews, safety measures, and relief services for adverse health effects, supported by administrative and management activities (*1*).

Recently, with the continued evolution of the environment surrounding PMDA, various challenges, such as drug loss and ensuring a stable supply of pharmaceuticals, have emerged (*2*). Alongside advances in science and technology, regulatory responses to digital transformation, including modeling and simulation, real-world data, adaptive design, and decentralized clinical trials, have become increasingly important. To continue fulfilling its expected role, PMDA needs to further strengthen its scientific capacity and its ability to address emerging challenges.

In particular, artificial intelligence (AI) technologies have attracted growing attention as tools capable of dramatically improving productivity, including through their accelerated application in drug development and related areas (*3*). Among AI technologies, generative AI has recently emerged as a practical tool for supporting knowledge-intensive tasks in drug development and regulation (*4,5*). This article therefore focuses on the current status of generative AI utilization at PMDA,

including practical use cases, fundamental principles guiding its application, and future perspectives for its use in PMDA.

2. Implementation of generative AI and governance framework at PMDA

PMDA has long promoted initiatives aimed at improving operational quality and efficiency. Considering the changes in its operating environment, PMDA has worked toward introducing generative AI to advance its operational activities.

As outlined in Figure 1, the strategic utilization of generative AI across PMDA's core functions (regulatory reviews, safety measures, and relief services) enhances the productivity of PMDA and improves quality while expanding human resource capacity for matters requiring advanced scientific judgment. Through these efforts, PMDA can further contribute to its organizational purpose of "making everyone's lives brighter together".

To facilitate the implementation of generative AI, PMDA established the Action Plan for the Use of AI in Operations on September 26, 2025 (*6*). This action plan consists of three pillars. The first and second pillars address technical introduction and implementation strategies, while the third pillar focuses on establishing a governance framework. The first pillar represents the initial phase of generative AI adoption, which involves the introduction of commercially available and widely implemented generative AI technologies, such as Microsoft Copilot, to improve the efficiency of PMDA's operations, particularly processes that involve administrative tasks. The second pillar represents the subsequent phase that involves the introduction of other generative AI models into a secure internal environment to support highly specialized and scientific operations in PMDA. In this phase, tasks may require a high level of medical and pharmaceutical expertise required for regulatory assessments, as well as careful consideration of the complexity and context-dependence of scientific judgment in regulatory decision-making. To assess the technical feasibility and regulatory applicability of generative AI and to inform the development of PMDA-specific generative AI tailored to our requirements, a series of proof-of-concept (PoC) studies is conducted. With the aim of enabling practical, reliable, and secure use in real-world settings, domain-specific generative AI systems tailored to medical applications have been developed in Japan (*e.g.*, the SIP-JMED initiative) (*7,8*). The third pillar focuses on strengthening the governance framework, including the establishment of internal rules and security measures, as well as enhancing information technology and AI literacy among PMDA staff through training programs that support appropriate AI utilization and informed decision-making. As part of this governance structure,

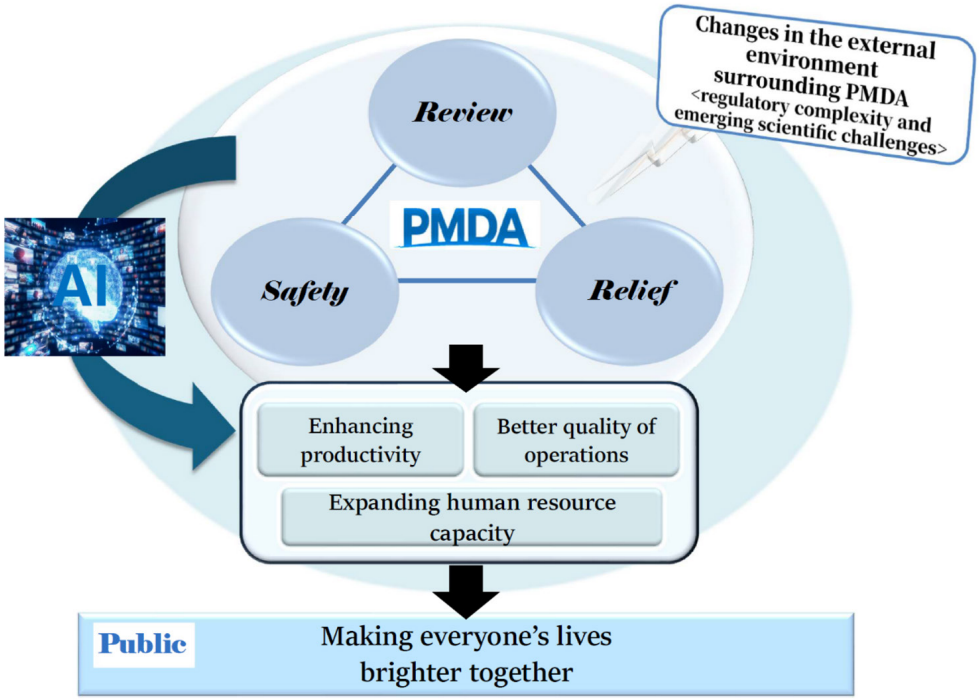


Figure 1. Use of generative artificial intelligence (GenAI) across PMDA's core functions to achieve its mission. Generative AI is applied across PMDA's core functions—review, safety, and relief—in response to increasing regulatory complexity and evolving scientific challenges. Its integration enhances operational quality, productivity, and effective allocation of human resources. This enables staff to focus on tasks requiring advanced scientific judgment while maintaining accountability and supporting PMDA's mission.

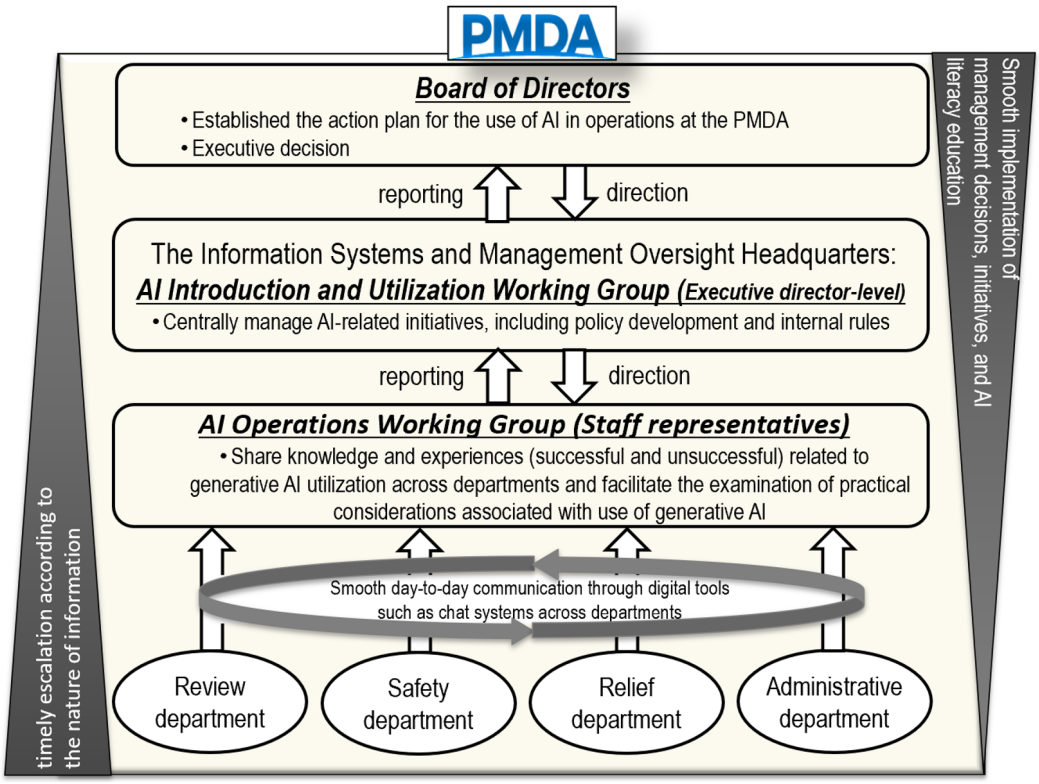


Figure 2. Governance framework for generative AI (GenAI) utilization at PMDA. A multi-layered structure is led by executive decision-making and supported by dedicated working groups for strategy and operations. This framework enables coordinated implementation, knowledge sharing across departments, and organization-wide AI literacy. Arrows indicate the reporting lines from operational units to governance bodies and the strategic direction from executive-level committees to operational teams.

PMDA has appointed a Chief AI Officer and established an AI Introduction and Utilization Working Group composed of executive director-level members under the Information Systems and Management Oversight Headquarters (Figure 2). Furthermore, an AI Operations Working Group consisting of staff representatives has been created to share knowledge and experiences, both successful and unsuccessful, related to generative AI utilization across departments and facilitate the examination of practical considerations associated with generative AI use. This framework enables rapid decision-making by top management while ensuring seamless sharing of knowledge, experiences, and challenges related to generative AI across departments. PMDA centrally coordinates discussions and necessary adjustments related to generative AI implementation through timely escalation from individual departments, depending on the nature of the information. Through prompt examination of issues related to operational requirements, security, internal regulations, operations, education, and evaluation, PMDA can smoothly and systematically implement management decisions while simultaneously promoting agency-wide AI-related initiatives and strengthening AI literacy among staff.

3. Practical use cases of generative AI at PMDA

PMDA carries out its responsibilities by communicating with various stakeholders, including pharmaceutical and medical device companies, healthcare institutions, academia, and patients. To ensure the quality, efficacy, and safety of medical products, information collection, analysis and evaluation, and dissemination processes must be executed promptly and appropriately. PMDA believes that generative AI has the potential to support many stages of these workflows (Figure 3).

In the information collection phase, generative AI may be applied to tasks such as drafting meeting minutes, translation, information retrieval, and technical checks of data/information. In the analysis and evaluation phase, generative AI may be utilized for data cleaning, development of analytical programs, data analysis, and drafting evaluation or review reports. In the dissemination phase, generative AI may support the preparation of public communication and outreach materials. Across these phases, PMDA, as a regulatory agency, focuses on applying generative AI to scientific and regulatory data/information collected from publicly available sources and/or submitted by stakeholders, including healthcare professionals, patients, and marketing authorization holders (*e.g.*, pharmaceutical companies). By embedding generative AI into these workflows, PMDA aims to enhance the sophistication of

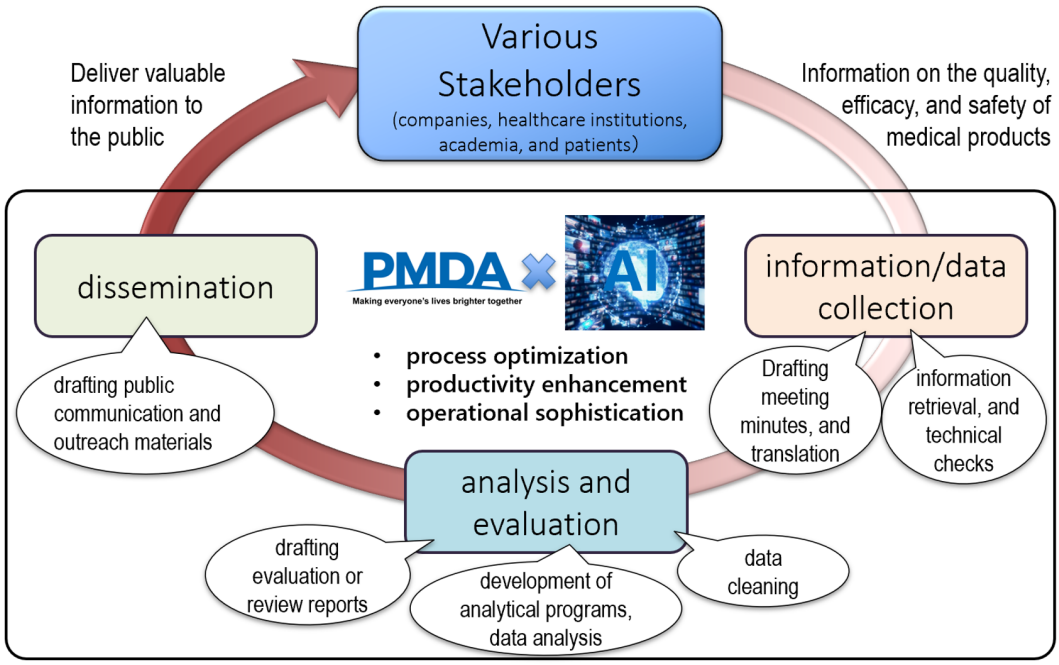


Figure 3. The information-to-value cycle at PMDA enabled by generative artificial intelligence (GenAI). Generative AI supports processes from information collection to analysis, evaluation, and dissemination. It assists with tasks such as data processing, document drafting, and analytical support across the workflow. Its integration enhances efficiency and contributes to the timely delivery of high-quality regulatory information.

scientific assessment and continuously deliver valuable information to the public, thereby ensuring the quality, efficacy, and safety of medical products.

PMDA has advanced the first pillar of the action plan using Microsoft Copilot, which has already been implemented in certain activities, such as drafting meeting minutes and translation. For the second and third pillars, PoC studies have been initiated, with close inter–working group communication to efficiently advance AI-related initiatives. We have explored its applications across all pillars through ongoing initiatives and pilot activities. The governance-related components in the third pillar have also progressed to support both current implementation and future development, reflecting PMDA's phased and cautious approach.

As an example under the second pillar, PMDA has been conducting PoC studies related to new drug review to assess whether domain- and task-specific generative AI can assist with summarizing clinical trial results during the preparation of review reports, thereby enabling staff to devote more time toward advanced scientific evaluation (*9*). Other potential applications in the first and/or second pillars include support for developing analytical programs for electronic clinical trial data, retrieving relevant materials from past scientific discussions, evaluating complex clinical and nonclinical data, and detecting drug safety signals, although concrete implementation strategies remain under consideration. Key challenges in using generative AI at PMDA may include ensuring data security and confidentiality, maintaining transparency and accountability of AI-supported decisions, and establishing appropriate human oversight (*10*). Specifically, PMDA-specific AI will require not only language generation capabilities but also a deep understanding of regulatory context, document structures, and accumulated institutional knowledge, which may not be fully addressed by general large language models.

4. Key considerations in use of generative AI

Utilization of generative AI involves numerous considerations that must be carefully addressed. Recently, the US Food and Drug Administration and the European Medicines Agency have jointly published guiding principles for good AI practice (*11*). Likewise, PMDA emphasizes the importance of conducting AI-related initiatives with appropriate risk management based on a thorough understanding of both characteristics and limitations of generative AI.

A fundamental principle guiding AI utilization at PMDA is that generative AI should not replace human judgment. Presently, generative AI should be regarded as a tool to support our work. As such, final regulatory decisions and associated responsibilities remain with human staff, in accordance with the human-in-the-loop principle (*12,13*). PMDA also stresses the need to clearly define appropriate use cases, noting that applications substituting for human judgment or relying on insufficiently verified outputs are inappropriate. In addition, use of generative AI in handling highly

sensitive or confidential information requires appropriate safeguards. These considerations are reflected in development of internal rules and guidance to ensure responsible and appropriate use of generative AI within PMDA. Given the known risks of AI utilization, such as hallucinations, PMDA staff should ensure reliability of AI-generated outputs, as PMDA continues to fulfill its institutional accountability.

It is also essential to confirm whether generative AI systems demonstrate appropriate performance consistent with their intended purposes, as well as key characteristics such as explainability, robustness, and generalizability. PMDA considers the potential introduction of AI technologies tailored to specific operational requirements by leveraging accumulated knowledge and document resources related to review, safety, and relief activities. For each individual application, repeated proof-of-concept studies and staged evaluations that account for model performance and limitations, implementation and operational costs, and security requirements need to be conducted. In this context, establishing appropriate evaluation framework tailored to each specific use case is considered important. Applicability of generative AI will be examined based on multiple dimensions, including task time reduction, accuracy of generated summaries, hallucination frequency, human correction rate, traceability to source documents, user satisfaction, and information security. Although these efforts remain at an early stage, PMDA emphasizes integrated assessment of both performance and feasibility, particularly given the need to handle sensitive personal and confidential data within regulatory workflows.

5. Future perspectives: Continuously creating "tomorrow's normal" through generative AI

PMDA will continue to promote utilization of generative AI across its operations and aims to improve productivity and optimization through business process reengineering that incorporates AI technologies. Through these initiatives, PMDA will strengthen its scientific capacity and its ability to respond to new and emerging challenges, thereby enhancing overall organizational performance.

By appropriately responding to advances in science and technology and facilitating the timely and appropriate availability of medical products required in clinical practice, PMDA fulfills its mission as a life platform that works together with society to continuously create "tomorrow's normal" while actively contributing to improvement of public health and safety.

Moreover, PMDA believes that its experience with utilization of generative AI under an appropriate governance framework will provide valuable insights for industry stakeholders considering application of generative AI across pre- and post-marketing stages. By sharing these experiences, PMDA will play an important role in promoting appropriate use of generative AI in medical product development and fostering international harmonization.

Acknowledgements

Views expressed in this article are those of the authors and do not necessarily reflect official views of the Pharmaceuticals and Medical Devices Agency.

Funding: None.

Conflict of Interest: One of the authors (Y. K.) has a family member employed by a pharmaceutical company. This employment is unrelated to the subject matter of this study. The other authors have no conflicts of interest to disclose.

References

1. Pharmaceuticals and Medical Devices Agency. Profile of services. *https://www.pmda.go.jp/files/000271450.pdf* (accessed May 29, 2026).
2. Kohno Y, Ishiguro A, Yasukawa T, Yasuda N, Tanaka D, Araki Y, Takahashi Y, Uyama Y, Fujiwara Y. Expediting drug development in Japan: A PMDA perspective. Clin Pharmacol Ther. 2025; 118:1262-1264.
3. Zhang K, Yang X, Wang Y, Yu Y, Huang N, Li G, Li X, Wu JC, Yang S. Artificial intelligence in drug development. Nature Med. 2025; 31:45-59.
4. Williams J, Boyce D, Collu G, *et al.* Generative AI: A generation-defining shift for biopharma regulatory affairs. Nat Rev Drug Discov. 2025; 24:651-652.
5. DISRUPT-DS Industry Roundtable. Generative AI in pharmaceutical R&D: From large language models to AI agents to regulation. Drug Discov Today. 2026; 31:104593.
6. Pharmaceuticals and Medical Devices Agency. Action plan for the use of AI in operations at the PMDA. *https://www.pmda.go.jp/files/000277511.pdf* (accessed May 29, 2026).
7. National Institute of Informatics. LLM-jp: Establishing the baseline for Japan's LLM development. *https://llm-jp.nii.ac.jp/en/home-en/* (accessed May 29, 2026).
8. Yahata S, Wan Z, Cheng F, Kurohashi S, Sato H, Nagai R. Causal tree extraction from medical case reports: A novel task for experts-like text comprehension. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25849-25867, Suzhou, China. *https://aclanthology.org/2025.emnlp-main.1313/* (accessed May 29, 2026).
9. Public research group on promoting the use of generative AI for the preparation of regulatory submission and review documents, Database of public research groups supported by the Ministry of Health, Labour and Welfare. *https://mhlw-grants.niph.go.jp/project/180404* (accessed May 29, 2026).
10. Ning Y, Teixayavong S, Shang Y, *et al.* Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist. Lancet Digit Health. 2024; 6:e848-e856.

11. US Food and Drug Administration, European Medicines Agency. Guiding principles of good AI practice in drug development. *https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf* (accessed May 29, 2026).
12. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nature Med. 2019; 25:44-56.
13. Wang A, Freeman S, Magrabi F. Governance for safe and responsible AI in healthcare organisations: A scoping review of frameworks. NPJ Digit Med. 2026 May 1. doi: 10.1038/s41746-026-02679-2. Epub ahead of print.

Received April 28, 2026; Revised May 30, 2026; Accepted June 8, 2026.

Released online in J-STAGE as advance publication June 13, 2026.

**Address correspondence to*:
Yoshiaki Uyama, Center for Regulatory Science, Pharmaceuticals and Medical Devices Agency, Shin-Kasumigaseki Building, 3-3-2 Kasumigaseki, Chiyoda-ku, Tokyo 100-0013, Japan.
E-mail: uyama-yoshiaki@pmda.go.jp

Global Health & Medicine. 2026; 8(3):154-160.
DOI: 10.35772/ghm.2026.01047

Beyond consent: Reconstructing ethical justification in medical adaptive machine learning systems

Keiichiro Yamamoto[1,*], Makoto Udagawa[2], Eisuke Nakazawa[3]

[1] Department of Clinical Research Management, Center for Clinical Sciences, Japan Institute for Health Security, Tokyo, Japan;
[2] Section of Bioethics, Department of Clinical Research Support, National Center of Neurology and Psychiatry, Tokyo, Japan;
[3] Department of Biomedical Ethics, Faculty of Medicine, The University of Tokyo, Tokyo, Japan.

Abstract: Medical Adaptive Machine Learning Systems (MAMLS) that continuously update their models using clinical data blur the conventional boundary between therapy and research, prompting the argument that their use should be classified as research and governed by informed consent requirements. Although informed consent remains normatively and legally important, this paper contends that consent-centered ethics faces two structural limitations in the context of MAMLS. First, the irreversibility inherent in deep learning models substantially undermines withdrawability—an important ancillary right of consent—thereby suggesting that consent may be transformed from an instrument of ongoing self-determination into a form of delegation to institutions. Second, the problem of data representativeness and bias shifts the unit of ethical analysis from the individual to the population, creating an "autonomy dilemma" in which respect for individual consent can paradoxically undermine the protection of autonomy at the collective level. Under these conditions, ethical justification must be complemented by, and in some contexts repositioned toward, public trust in institutions. The paper concludes that the ethical challenges surrounding MAMLS cannot be adequately addressed within the framework of research ethics alone, but must instead be taken up within the broader framework of public health ethics, with particular attention to transparency, accountability, and participatory governance.

***Keywords*:** medical artificial intelligence, informed consent, public trust, research ethics, public health ethics

1. Introduction

In recent years, there has been growing interest in artificial intelligence (AI) in healthcare—particularly Medical Adaptive Machine Learning Systems (MAMLS)—with increasing attention to their potential clinical deployment (*1,2*). Unlike conventional medical AI systems that rely on fixed, pre-trained models, MAMLS continuously update their models by incorporating patient data generated in clinical settings after initial deployment (*1,3*). In this sense, MAMLS are not merely tools for diagnosis or treatment support; they are technologies in which the act of use itself encompasses a process of knowledge generation. Since the clinical environment thus doubles as a research environment, the distinction between "therapy" and "research"—a distinction the Belmont Report (1979) deemed essential to maintain—becomes increasingly blurred, both institutionally and conceptually (*4,5*).

On this point, Sparrow and colleagues argue that, given the continual-learning nature of MAMLS, their use should be classified as research (*6*). Their argument can be summarized as follows: since MAMLS continuously update their models by incorporating patient data in clinical settings, each patient's data contributes to a process that produces generalizable knowledge affecting future patients, thereby structurally satisfying the Belmont Report's definition of "research" (*4*). Furthermore, patient data is used not only to optimize individual treatment but also to improve the model, meaning that patients are not necessarily involved solely for their own benefit. Given that manufacturers have financial interests in system improvement, they argue that continual learning without treating patients as research subjects gives rise to a risk of moral hazard. Consequently, they conclude that IRB oversight and prospective written opt-in informed consent are ethically required (*6*).

This argument is consistent with the mainstream tradition of research ethics shaped by the Belmont Report. However, the consent-centered regulatory vision appears to be in tension with actual institutional

trends. In Japan, the "three-yearly review" of the Act on the Protection of Personal Information has examined legislative designs that would permit AI development and statistical use without individual consent in certain circumstances (*7*). The Personal Information Protection Commission's reform policy has raised as an issue the appropriate role of individual consent in facilitating AI development understood as a form of statistical processing, and the public consultation process generated numerous responses concerning "data utilization that does not require individual consent" (*7,8*). Among these responses, some characterized the relaxation of consent requirements as a shift from ex ante regulation to ex post governance (*8,9*). These reform discussions are still ongoing, and the final legislative design has not yet been determined. Although these developments are specific to Japan, they illustrate a broader international challenge in medical AI governance: how to reconcile large-scale, socially valuable data use with the continuing normative importance of individual consent, accountability, and public oversight. They therefore provide a useful institutional example of the wider shift from ex ante consent-centered regulation toward ex post governance and trust-based accountability.

What this institutional context reveals is a marked gap between normative principles and practical realities: while consent is positioned as central to AI development from the perspective of research ethics, it is simultaneously recognized as being in tension with large-scale data utilization in practice (*10-12*). Particularly in MAMLS, it is necessary to continuously acquire broad and minimally biased data in order to adequately cover the relevant data distribution and thereby ensure model performance and safety (*13*). In other words, data-related bias must be minimized as far as possible. This creates a marked divergence between norm and practice: on the one hand, stronger consent requirements are advocated from the standpoint of research ethics, while on the other hand, consent requirements are being relaxed in the name of AI utilization policy.

This paper seeks to reframe this tension not merely as a clash of policy choices, but as a manifestation of the structural limits of consent as conventionally understood in research ethics. First, it argues that MAMLS inherently involve irreversibility, which substantially undermines withdrawability as an important ancillary right of consent. Second, it contends that problems of data representativeness and bias shift the unit of ethical analysis from the individual to the population. Third, it argues that, as a consequence, the focus of ethical justification should be repositioned from individual consent to public trust in institutions. On this basis, the paper concludes that the ethical challenges posed by MAMLS cannot be adequately addressed within the framework of research ethics alone, but must also be taken up within the broader framework of public health ethics.

2. Irreversibility and the breakdown of consent

To understand the characteristics of MAMLS from the perspective of research ethics, it is first necessary to examine their underlying technical features. Many contemporary medical AI systems are based on deep learning models, including convolutional neural networks (*14,15*). In such models, training data is not preserved as discrete records, but is incorporated into the model through changes to its internal parameters (*16*). Because data is embedded throughout the model's internal structure, it is difficult to determine how any particular data point has influenced the model.

Under such a structure, selectively deleting data once incorporated into training and restoring the model to a state in which that data had not been incorporated is technically and operationally difficult under current conditions (*17*). While methods such as retraining and machine unlearning are theoretically conceivable and increasingly explored, they may require substantial computational resources, time, and system-level reconstruction in practice (*18*). Retroactively removing only the contribution of specific data may therefore be difficult to implement reliably in deployed MAMLS. Data utilization in MAMLS thus appears to take on a practically irreversible character under realistic technical and operational conditions. Sparrow and colleagues note that MAMLS carry the risk of "catastrophic forgetting"—a phenomenon in which the model overwrites previously acquired information during the update process (*6*)—but the implications of such practical irreversibility for the concept of consent itself do not appear to have been sufficiently addressed.

This irreversibility stands in tension with the philosophical premises underlying informed consent. In both research ethics and clinical ethics, informed consent has generally been understood to include withdrawability as an important ancillary right, alongside informedness, understanding, and voluntariness (*19*). Withdrawability is a crucial condition for understanding consent not as a one-time act of permission, but as a temporally extended process of self-determination (*20,21*). Within research ethics in particular, ensuring the voluntariness and withdrawability of consent to participation is a central implication of the principle of respect for persons and an important safeguard against unjust exploitation of research participants (*22,23*).

In MAMLS, however, this premise is significantly destabilized. First, since data is integrated into the model in ways that may be difficult to reverse, even if the formal right to withdraw consent is recognized, it is difficult to give substantive effect to that withdrawal. Second, given the nature of continuous learning, the

scope and impact of data use change over time, and the full picture cannot be grasped in advance. Under these conditions, consent necessarily becomes something given comprehensively in relation to indeterminate future uses—less a reversible choice than a commitment whose practical consequences may not be fully undone.

Conventionally, consent has been understood as a means by which individuals exercise ongoing control over matters concerning themselves (*20,21*). In MAMLS, however, such control is severely constrained by technical and operational limitations. Consequently, consent takes on a character closer to "delegation"—an act of entrusting discretion regarding future uses and their consequences to institutions and professionals (*24*). The question, therefore, is under what conditions such delegation can be justified in a setting that presupposes data use that is difficult to reverse in practice. This, in turn, opens onto broader questions about how ethical justification is to be secured within institutional and social arrangements.

3. From individuals to populations

The characteristics of MAMLS examined above are not limited to irreversibility: through the ways in which data is collected and used, they also potentially reconfigure the very unit of ethical analysis. The focus of the problem thus appears to shift from the choices of individual patients to the nature of the populations constituted through data.

In MAMLS, the quantity and diversity of training data are critical determinants of model performance and safety. Particularly in the medical domain, data reflecting diverse attributes—such as age, sex, genetic background, and socioeconomic status—are required (*13*). However, opt-in data collection necessarily depends on participants' choices and therefore tends to result in certain groups being over- or underrepresented—the well-known problem of selection and sampling bias (*1*). This problem has already been observed in real-world systems. In dermatological image-recognition AI, for example, diagnostic accuracy for patients with darker skin tones is significantly lower owing to the overrepresentation of White patients in the training data (*25,26*). Consequently, biases in training data can reduce the model's predictive accuracy for particular groups.

Such data bias is not merely a technical problem. Through reduced diagnostic accuracy and increased misdiagnoses for certain patient groups, it gives rise to substantive disadvantage and thereby raises concerns about both scientific and ethical validity. Sparrow and colleagues also recognize the problem of bias: they argue that, in MAMLS that employ collective learning, models may not necessarily improve—and may even worsen—for particular subpopulations, thereby raising the ethical concern that such research subjects may be treated as "mere means" (*6*). However, they seek to address this problem within the framework of risk-benefit assessment in research ethics. What this paper suggests, by contrast, is that the problem may involve a more fundamental transformation—one that shifts the very unit of ethical analysis from the individual to the population—and therefore cannot be adequately contained within such a framework.

Here, an ethical tension becomes manifest. The more thoroughly individual autonomy is respected and consent-based data provision is pursued, the more likely it is that data bias will arise, potentially leading to a decline in model performance. Conversely, efforts to ensure the comprehensiveness and representativeness of data tend to reduce the scope for individual choice and weaken the substantive meaning of consent (*27*). This situation represents what might be called an "autonomy dilemma": efforts within the traditional research ethics framework to secure respect for persons through consent-centered mechanisms may paradoxically undermine the effective protection of data providers across the population, including their autonomy-related interests.

Under this dilemma, the very unit of ethical analysis is reconfigured. Conventional research ethics has been built primarily around individual patients and research participants, with a primary focus on their rights and interests (*4,22*). In MAMLS, however, individual data items have no meaning in isolation; they fulfill their function only as part of an aggregate. The focus of the problem thus shifts from the individual question of "who consented" to the structural question of "what kind of population is constituted through data". This shift demands a change in ethical framework: from protecting individual research participants to protecting the broader population that provides clinical data, with greater emphasis on collective interests, distributional fairness, and the distribution of risks—thereby suggesting the need for a shift toward public health ethics (*28,29*).

4. From consent to trust

As we have seen, irreversibility and collectivity fundamentally undermine the conventional consent-centered ethical framework in MAMLS. What is at issue is not merely that consent procedures are inadequate; rather, the question is whether consent itself can function as an adequate principle of justification in this context. This argument should not be read as denying the normative or legal importance of consent. Rather, consent remains an important expression of respect for persons and may retain legal significance even when it cannot, by itself, bear the full justificatory burden of MAMLS governance.

One source of this difficulty lies in the limits of understanding. The internal structure of deep learning

models is highly complex, and it is difficult for individual patients to adequately understand either the specific processes by which decisions are made or the system's future behavior. The black-box problem that pervades AI systems means that only a limited number of specialists can understand such systems sufficiently (*30*). Furthermore, in systems like MAMLS that engage in continual learning, the scope and impact of data use change over time, and the full picture cannot be grasped in advance (*1*). Under these conditions, the ideal of "fully informed consent" is difficult to realize in practice.

An anticipated counterargument holds that "complete understanding is not a necessary condition for consent; understanding information that is substantially relevant to decision-making is sufficient" (*31*). Certainly, informed consent theory does not require complete understanding of technical details. However, in the case of MAMLS, the problem is not limited to the degree of understanding; rather, it lies in the fact that the very object of understanding changes over time. The mode of use understood by the patient at the time of consent may subsequently change through continual learning, and that understanding does not necessarily extend to future uses. Consent in MAMLS is therefore marked by a structural gap between understanding and actual use.

In such circumstances, there is no doubt that efforts to explain technical matters in accessible language and to facilitate patients' understanding are important. However, such "translation" is bound to remain inherently incomplete. A structural asymmetry exists between expert knowledge and ordinary understanding, and it is difficult to overcome this asymmetry entirely. In this respect, consent may not always function as a choice grounded in adequate understanding.

Sparrow and colleagues themselves recognize this point to some degree, noting that IRBs must determine on a case-by-case basis whether consent is required, and that replacing written consent with verbal consent or waiving consent may in some cases be permissible (*6*). From the perspective of this paper, however, this limitation should not be understood as a problem that can be resolved through procedural adjustments alone; rather, in the context of AI systems such as MAMLS, it calls for a reconsideration of consent as a principle of justification.

Under these circumstances, the character of consent is changing. Conventionally understood as an expression of autonomous choice grounded in understanding, consent in MAMLS increasingly takes the form of an act that presupposes trust in institutions and professionals. Patients do not choose on the basis of having fully grasped the details of data use; rather, they accept such use on the premise that it will be appropriately managed in the future. Consent is thus shifting from "choice based on understanding" to "trust-based delegation". Biobanks present a structurally analogous case: participants must consent on the basis of trust in the biobank system without knowing in advance the details of future research uses. This structural similarity has already been noted in international discussions, and the accumulated insights from biobank ethics are highly relevant here (*32-34*).

Importantly, this trust is not limited to trust in individual healthcare providers. Rather, it is directed toward the broader institutional framework that establishes norms for data use and monitors and enforces those norms—regulatory authorities, ethics review bodies, academic institutions, scholarly societies, academic journals, and wider structures of social governance. The focus of ethical justification is thus shifting from the presence or absence of individual consent to the legitimacy of the institutional system as a whole. Such institutional trust requires more than general reassurance. It must be supported by concrete arrangements, including transparency about data use and model updating, accountability mechanisms for harms and biases, continuous oversight after deployment, regular evaluation of model performance across subpopulations, patient and public involvement in governance, and clearly defined roles for regulatory bodies, ethics committees, healthcare institutions, professional societies, and academic journals. These arrangements can help transform trust from a merely psychological attitude into a justified confidence in institutional practices.

Accordingly, while the distinction in consent form—opt-in versus opt-out—remains an important topic of discussion, it does not constitute the core of the problem. This is not to deny the normative significance of opt-in consent: consent that is procedurally valid but substantively thin can still play a symbolic role in protecting patient autonomy and helping to sustain trust in institutions. However, this is not a sufficient condition, and properly structuring consent procedures does not in itself amount to the establishment of institutional trust.

More important is the question of through what processes rules governing data use are formed, and how those processes can acquire public trust. In other words, the ethical challenges surrounding MAMLS need to be repositioned not as a matter of individual choice, but as a problem of building institutional trust. In this respect, the ethics of medical AI extends beyond the conventional research ethics framework centered on consent and connects to a broader theory of social justification—namely, public health ethics and the domain of political philosophy that provides its theoretical foundations. What is at stake there is how to reconcile respect for individual choice with the realization of collective goods, and the resolution of this problem depends to a considerable extent on the construction of a trustworthy institutional framework.

To clarify how the technical and institutional features of MAMLS give rise to ethical challenges and corresponding governance responses, Table 1 summarizes the central structure of the argument.

5. Conclusion

This paper has clarified, through two pathways, that in the context of MAMLS the framework of consent has become difficult to sustain as a viable premise. First, the irreversibility inherent in deep learning models substantially undermines withdrawability, an important ancillary right of consent, thereby suggesting that consent may be transformed from a guarantee of ongoing self-determination into something closer to delegation to institutions. Second, the problem of data representativeness and bias shifts the unit of ethical analysis from the individual to the population, bringing to light an "autonomy dilemma" in which efforts within the traditional research ethics framework to respect individual autonomy may paradoxically undermine the effective protection of data providers throughout the population, including their autonomy-related interests.

Under these conditions, the foundation of ethical justification appears to be shifting from the presence or absence of individual consent to trust in the institutional system as a whole. Consent-centered ethics continues to play important normative and legal roles, but under the technological and institutional conditions of MAMLS, it remains open to question whether consent alone can suffice as a principle of justification.

In this respect, while taking the argument of Sparrow and colleagues as a point of departure, this paper seeks to extend its scope one step further. The heart of the problem lies not in the thoroughgoing implementation of opt-in consent, but in the question of what framework of justification may still retain its validity under technological and institutional conditions in which the consent-based interpretation of respect for persons within the traditional research ethics framework is proving insufficient. This question cannot be adequately examined within the framework of research ethics alone, but must instead be pursued in connection with public health ethics and political philosophy.

Recent developments in the review and amendment of the Act on the Protection of Personal Information in Japan also suggest the practical implications of this problem. The fact that the consent principle itself is being institutionally reconfigured in the context of AI development appears to suggest that the conventional framework, which places consent at the center of ethical justification, is being compelled to reconsider its practical sustainability. These developments may be rooted in a deeper institutional transformation that cannot be adequately captured by the binary of strengthening versus relaxing consent.

Therefore, the tasks ahead are not limited to improving methods for obtaining consent. Rather, the broader question of governance must be confronted: how to institutionally secure the legitimacy of data use and public trust among data providers, who should be involved in this process, and how deliberation and collective decision-making should be organized. In this context, alongside transparency, accountability, and fairness, institutional designs that enable the participation of diverse stakeholders will be essential.

From this perspective, what is being asked in AI medicine is not the technical or procedural question of how consent should be obtained. Rather, under institutional conditions in which consent-centered

Table 1. Key ethical challenges of Medical Adaptive Machine Learning Systems (MAMLS) and corresponding governance responses

Technical / institutional feature	Ethical / scientific challenge	Limit of consent-centered ethics	Governance response
Continual learning and data integration	Irreversibility and weakened withdrawability	Withdrawal cannot always be fully implemented after data have influenced model parameters	Transparency about limits of withdrawal; accountability for data use and downstream effects
Model updating over time	Uncertain and evolving future uses	Consent at one time point cannot fully cover later model behavior or new uses	Continuous oversight after deployment; periodic review of model updates and risks
Large-scale data requirements	Selection bias and underrepresentation	Opt-in participation may reduce representativeness and reinforce bias	Fairness monitoring; population-level evaluation; attention to underrepresented groups
Population-level impact	Distribution of risks and benefits across groups	Individual consent alone does not address collective harms	Public health ethics; participatory governance; patient and public involvement
Institutional dependence	Trust-based delegation	Consent alone cannot justify data use when understanding and control are limited	Institutional transparency; regulatory and ethics oversight; clear responsibility structures

approaches to ethical justification are being called into question, we are confronted with a more fundamental set of questions: what do we entrust, to whom, how can that delegation be justified, and how can public trust in healthcare—including medical research—be secured?

Funding: This work was supported by AMED under Brain/MINDS 2.0 (Multidisciplinary Frontier Brain and Neuroscience Discoveries), Grant Numbers JP24wm0625001 and JP24wm0625012, and by the Japan Society for the Promotion of Science (JSPS) through a Grant-in-Aid for Scientific Research (A), "A Comprehensive Study on Moral Distress", Grant Number 23H00005.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. Lancet Digit Health. 2021; 3:e337-e338.
2. Yuan H. Toward real-world deployment of machine learning for health care: External validation, continual monitoring, and randomized clinical trials. Health Care Sci. 2024; 3:360-364.
3. Bruno P, Quarta A, Calimeri F. Continual learning in medicine: a systematic literature review. Neural Process Lett. 2025; 57:Article 2.
4. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (NCPHSBBR). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. *https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html* (accessed May 21, 2026).
5. Levine RJ. Ethics and regulation of clinical research. 2nd ed. New Haven (CT): Yale University Press; 1986.
6. Sparrow R, Hatherley J, Oakley J, Bain C. Should the use of adaptive machine learning systems in medicine be classified as research? Am J Bioeth. 2024; 24:58-69.
7. Personal Information Protection Commission, Japan. Interim report on considerations for the triennial review of the Act on the Protection of Personal Information. *https://www.ppc.go.jp/files/pdf/240627_02_houdou_betten1.pdf* (accessed April 23, 2026).
8. Personal Information Protection Commission, Japan. Results of public comments on the interim report concerning the triennial review of the Act on the Protection of Personal Information. *https://www.ppc.go.jp/files/pdf/240904_shiryou-1-2.pdf* (accessed April 23, 2026). (in Japanese)
9. Personal Information Protection Commission, Japan. System Reform Policy under the Triennial Review of the Act on the Protection of Personal Information. *https://www.ppc.go.jp/en/topix/triennial_review_2026_02/* (accessed May 28, 2026).
10. Balch JA, Evans BJ, Shickel B, Bihorac A, Upchurch GR Jr, Loftus TJ. The dilemma of consent for AI in healthcare. Surgery. 2024; 175:1456-1457.
11. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. PLoS Med. 2018; 15:e1002689.
12. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019; 25:37-43.
13. Guan H, Bates D, Zhou L. Keeping medical AI healthy and trustworthy: A review of detection and correction methods for system degradation. IEEE Trans Biomed Eng. 2025; PP:10.1109/TBME.2025.3642706.
14. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019; 25:44-56.
15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017; 42:60-88.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521:436-444.
17. Bourtoule L, Chandrasekaran V, Choquette-Choo CA, Jia H, Travers A, Zhang B, Lie D, Papernot N. Machine unlearning. In: 2021 IEEE Symposium on Security and Privacy (SP). Piscataway (NJ): IEEE; 2021. p. 141-159.
18. Xu H, Zhu T, Zhang L, Zhou W, Yu PS. Machine unlearning: A survey. ACM Computing Surveys. 2023; 56: Article 9.
19. Beauchamp TL, Childress JF. Principles of biomedical ethics. 8th ed. New York (NY): Oxford University Press; 2019.
20. Manson NC, O'Neill O. Rethinking informed consent in bioethics. Cambridge: Cambridge University Press; 2007.
21. Miller FG, Wertheimer A, editors. The ethics of consent: Theory and practice. New York (NY): Oxford University Press; 2009.
22. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human participants. *https://www.wma.net/policies-post/wma-declaration-of-helsinki/* (accessed April 23, 2026).
23. Council for International Organizations of Medical Sciences. International ethical Guidelines for Health-related research involving humans. Geneva: Council for International Organizations of Medical Sciences; 2016.
24. Wiertz S, Boldt J. Evaluating models of consent in changing health research environments. Med Health Care Philos. 2022; 25:269-280.
25. Daneshjou R, Vodrahalli K, Novoa RA, *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv. 2022; 8:eabq6147.
26. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. J Am Acad Dermatol. 2022; 87:157-159.
27. Yamamoto K, Ibuki T, Nakazawa E. The fine balance between complete data integrity in medical adaptive machine learning systems and the protection of research participants. Am J Bioeth. 2024; 24:101-103.
28. Childress JF, Faden RR, Gaare RD, Gostin LO, Kahn J, Bonnie RJ, Kass NE, Mastroianni AC, Moreno JD, Nieburg P. Public health ethics: Mapping the terrain. J Law Med Ethics. 2002; 30:170-178.
29. Kass NE. An ethics framework for public health. Am J Public Health. 2001; 91:1776-1782.
30. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018; 16:31-57.
31. Faden RR, Beauchamp TL. A history and theory of

informed consent. Oxford University Press, New York (NY): Oxford University Press; 1986.
32. Caulfield T, Kaye J. Broad consent in biobanking: Reflections on seemingly insurmountable dilemmas. Med Law Int. 2009; 10:85-100.
33. Steinsbekk KS, Myskja BK, Solberg B. Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem? Eur J Hum Genet. 2013; 21:897-902.
34. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: A patient interface for twenty-first century research networks. Eur J Hum Genet. 2015; 23:141-146.

Received April 28, 2026; Revised May 28, 2026; Accepted June 2, 2026.

Released online in J-STAGE as advance publication June 13, 2026.

Address correspondence to:
Keiichiro Yamamoto, Department of Clinical Research Management, Center for Clinical Sciences, Japan Institute for Health Security, 1-21-1 Toyama, Shinjuku-ku, Tokyo 162-8655, Japan.
E-mail: yamamoto.kei@jihs.go.jp

Global Health & Medicine. 2026; 8(3):161-165.
POLICY FORUM
DOI: 10.35772/ghm.2026.01067

Human-in-the-loop reconsidered: Shadow use and reliance management in drug development

Yusuke Inoue*

Department of Healthcare Ethics, Kyoto University School of Public Health, Kyoto, Japan.

Abstract: This article examines the ethical governance of artificial intelligence (AI) use in drug development through joint principles of good AI practice issued by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). It argues that the significance of the principles lies in moving beyond AI exceptionalism: AI should neither be uniformly prohibited nor uniformly permitted but assessed in a risk-based manner according to context, purpose, and potential impact across the drug lifecycle. Among the ethical and governance risks associated with AI, this study focuses on two organizational risks that are particularly relevant to implementation. The first is shadow use, in which AI involvement remains insufficiently visible, documented, or reviewed. The second is reliance management. Once AI is integrated into research and regulatory workflows, some degree of reliance is inevitable; however, such reliance must remain conscious, proportionate, reviewable, and supported by meaningful human oversight. Overreliance and deskilling are risks associated with poorly managed reliance. Ethical governance should therefore make AI use visible and reviewable while preserving the practical ability to question, verify, escalate, or set aside AI-assisted outputs.

***Keywords*:** regulatory science, drug development, shadow use, reliance management, human oversight, risk-based governance

1. Introduction

Use of artificial intelligence (AI) in medicine and research has expanded beyond diagnostic support to include document drafting, translation, clinical trial documentation management, data analysis, pharmacovigilance, and manufacturing control. The American Medical Association has framed medical AI not as "artificial intelligence" intended to replace human judgment but as "augmented intelligence" designed to support physicians' judgment and work (*1,2*). Although this framework was developed primarily for clinical medicine, it is also relevant to drug development, where AI increasingly supports document preparation, evidence generation, data interpretation, and regulatory process. This framing indicates a shift in the ethical focus of AI from the abstract question of whether AI should be used to more practical questions of implementation: in what contexts, by whom, to what extent, and under what governance and accountability arrangements AI should be used.

2. The U.S. Food and Drug Administration (FDA)/ European Medicines Agency (EMA) joint principles as an implementation framework

As a basis for examining ethical governance of AI in drug development, this article focuses on the "Guiding Principles of Good AI Practice in Drug Development" issued in 2026 by the FDA and EMA (*3,4*). These principles are based on recognition that AI can generate evidence relevant to regulatory decision-making across the entire drug lifecycle, including nonclinical research, clinical trials, manufacturing, and post-marketing safety surveillance. Accordingly, such evidence may affect assessments of quality, efficacy, and safety, and inappropriate use of AI may have implications for product approval.

Significance of the FDA/EMA joint principles lies not in treating AI as an exceptional technology requiring uniform regulation but in translating AI governance into concepts already familiar in pharmaceutical development and regulation. These concepts include human-centered design, a risk-based approach, adherence to existing standards, a clear definition of the context of use, multidisciplinary expertise, data governance, model design, risk-proportionate performance assessment, lifecycle management, and clear communication (*3*). In this respect, the principles move ethical discussion away from the binary question of whether AI should be

allowed, and toward more practical questions: whether a given AI use is fit for its intended purpose, whether it has been evaluated in proportion to its risk, whether it remains reliable after deployment, and whether assignment of responsibility is clearly defined.

This implementation-oriented framing is particularly important in drug development, which involves complex, multistage, and often transnational processes. AI may be introduced at different points in the lifecycle, used by different actors, and connected to decisions of varying regulatory significance. Therefore, a uniformly permissive or prohibitive rule is poorly suited to practice. By emphasizing risk-based governance, alignment with existing standards, performance monitoring, and life-cycle management, the FDA/EMA principles provide a practical framework for calibrating oversight according to intended use of AI, its potential impact, and its degree of influence on decision-making.

Joint issuance of these principles by the FDA and EMA is also significant. Substantial divergence in regulatory approaches to AI use across regions increases uncertainty for sponsors, research institutions, and developers involved in international drug development. By articulating a shared regulatory direction for the United States and Europe, the principles may provide a basis for future international regulatory convergence and discussion in forums such as the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use. Indeed, these principles have been perceived as a step toward facilitating international regulatory harmonization and global implementation (*3,5*).

However, the FDA/EMA joint principles should be understood as high-level guiding principles that provide a shared vocabulary for assessing AI use in a risk-based manner rather than as a complete operational manual or legally binding regulatory framework (*3,6*). This article does not attempt to provide a comprehensive taxonomy of ethical issues raised by AI in drug development, such as data quality, bias, privacy, transparency, validation, and security. Instead, this article focuses on two related but distinct implementation challenges: shadow use and reliance management. These challenges test whether human-in-the-loop is understood as a substantive governance requirement rather than merely as a procedural label. Shadow use raises the question of whether organizations can identify where AI is being used, including uses embedded in ordinary tools and uses not fully recognized by users themselves. Reliance management raises a different question: once AI is integrated into drug development workflows and some reliance on AI-assisted outputs becomes unavoidable, can organizations ensure that such reliance remains conscious, proportionate, reviewable, and resilient? Overreliance and deskilling are risks that may arise when reliance becomes unexamined, excessive, or disconnected from meaningful human judgment.

3. Shadow use as a visibility problem

The first question concerns visibility: whether organizations can identify, document, and review how and where AI is used. The logic of the FDA/EMA principles draws attention to the problem of shadow use, although they do not provide an operational definition of the term. If organizations do not know where and how AI is used, they cannot meaningfully assess risk, allocate responsibility, document decisions, or manage AI systems throughout their lifecycle. In this sense, shadow use is not a peripheral compliance issue but a central governance problem.

Shadow use refers to a situation in which analysts, researchers, or other staff use large language models or similar AI tools in routine work, even though the organization has not explicitly positioned, documented, or governed such use, and management lacks sufficient visibility in actual practice (*5,7*). Shadow use may occur in routine tasks such as document drafting, translation, summarization, searching, code generation, data formatting, and issue framing. It may also occur through embedded AI functions in applications, such as autocomplete, translation assistance, search support, or text revision, even when users do not clearly recognize that they have "used AI".

Therefore, shadow use should not be treated solely as a matter of rule violation or individual negligence. Intentional concealment or inappropriate AI use should not be tolerated, particularly when such use involves confidential information, personal data, proprietary data, or regulatory submissions. However, as AI functions become increasingly embedded in routine work environments, some forms of unrecognized or undocumented AI involvement may arise even without deliberate misconduct. Moreover, AI may influence upstream stages of reasoning, such as organizing issues, considering counterarguments, and discussing analytical strategies, even when no AI-generated text remains in the final document. Therefore, a governance system that merely asks individuals to declare whether they have used AI is unlikely to capture the full extent of their involvement with AI.

One implication is that organizations may need to consider how to make AI involvement more visible within routine workflows rather than relying solely on retrospective self-reporting. This does not require all uses of AI to be treated as posing the same level of risk. A more feasible approach may be to distinguish between low-risk routine uses and uses that may require documentation, human review, or restrictions on entering sensitive or submission-relevant information, with the level of oversight proportionate to associated risk.

Therefore, practical governance measures should be designed at the workflow level. Possible measures include an internal AI-use registry covering approved tools and use cases; context-of-use documentation

incorporated into analysis plans, validation reports, or submission-relevant documents; risk-based thresholds that determine when AI use requires disclosure, validation, or independent review; restrictions on entering confidential information, personal data, clinical trial data, proprietary information, or submission-relevant data into external AI tools; audit trails for AI-assisted codes, summaries, translations, and regulatory text; and internal consultation pathways for borderline cases. The purpose is not merely to detect and punish hidden use but to convert otherwise invisible AI involvement into reviewable, accountable, and proportionately governed use.

Risk-based governance also requires distinguishing among contexts of AI use. Low-risk uses may include language editing, formatting, and preliminary summarization, provided that no confidential or submission-relevant data are entered into unapproved external AI systems. Moderate-risk uses may include internal literature reviews, coding assistance, data formatting, and drafting internal working documents. High-risk uses may include statistical programming, data cleaning that affects dataset analysis, safety signal detection, manufacturing controls, and regulatory submission documents. The highest-risk uses are those in which AI outputs directly influence evidence generation, benefit-risk assessment, quality evaluation, or regulatory decision-making. Although these categories should be adapted to each organization's workflows, they illustrate why governance should be proportionate to context of use and potential regulatory impact.

Such an approach may also help avoid framing shadow use only as individual misconduct. If AI use is treated solely as an exceptional or prohibited act, researchers may be less willing to seek advice or disclose borderline cases. Clearer expectations regarding when AI use should be discussed, documented, or reviewed can make AI involvement easier to identify. The aim is not to eliminate every form of routine AI assistance but to make AI use that is relevant to governance more transparent, explainable, and reviewable.

4. Reliance management and human oversight

The second implementation challenge concerns reliance management, which involves human oversight. In AI-assisted drug development, this goal cannot eliminate reliance on AI outputs. Once AI tools are used for literature reviews, data cleaning, statistical programming, safety signal detection, manufacturing control, and regulatory writing, some degree of reliance is expected and operationally necessary. The governance question is therefore not whether professionals rely on AI but whether such reliance is conscious, proportionate to risk, and supported by conditions that enable critical review.

Accordingly, human-in-the-loop should not be treated as an ethical guarantee. Although the FDA/EMA principles emphasize human-centered design, simply designating a human as final decision-maker is insufficient. If researchers, analysts, or reviewers are expected to use AI-assisted outputs in fast-moving workflows but lack the time, information, expertise, or authority to question those outputs, human oversight may become a procedural label rather than a substantive safeguard.

Distinguishing three related failure modes of poorly managed reliance helps clarify these concerns. Overreliance refers to excessive trust in AI outputs for a particular task, such as accepting AI-generated code, summaries, analyses, or regulatory texts without sufficient verification. Deskilling refers to the gradual erosion of professional judgments or domain-specific expertise through repeated dependence on AI. Merely formal or nominal human oversight refers to situations in which a human decision-maker remains formally responsible but lacks the time, expertise, authority, or information required for critical evaluation of AI-assisted outputs. These concepts are analytically distinguishable but not mutually exclusive. Formal oversight may facilitate task-specific overreliance, while repeated overreliance may contribute to deskilling over time.

A review of AI-induced deskilling noted that AI-based decision support may contribute to erosion of professional skills and reduce opportunities to acquire them (*8*). However, direct empirical evidence of AI-induced deskilling in drug development remains limited. Deskilling in this field should therefore be framed as a plausible organizational risk that requires monitoring and governance, rather than as an already established outcome across all AI-assisted workflows. Likewise, the literature on automation bias helps explain how reliance may become excessive in particular tasks (*9,10*), whereas critiques of human oversight caution that merely placing a human reviewer in the workflow may create only a formal safeguard unless practical conditions for meaningful review are present (*11*).

This framing is consistent with the practical distinction between confidence in an AI tool and appropriate reliance on its output in a specific clinical or research context. A tool's having undergone a certain level of evaluation does not by itself determine how much weight should be placed on its output in a particular situation (*12*). Appropriate reliance should vary with the intended use, data quality, uncertainty, reversibility, regulatory significance, and consequences of errors. Therefore, the same AI output may warrant different levels of scrutiny, depending on whether it is used for preliminary exploration, internal drafting, formal evidence generation, or support for regulatory submission.

Similarly, the World Medical Association emphasized that AI should augment human judgment and that systems should be in place to ensure availability of alternative procedures in the event of AI system failure,

Table 1. Two implementation challenges of AI governance in drug development and possible organizational responses

Implementation challenge	Potential ethical concerns	Possible governance responses: directions and examples
Shadow use	AI involvement may remain invisible to organizations or reviewers.	Make AI involvement visible and traceable: approved-tool registry; context-of-use records.
	Responsibility, documentation, and accountability may become unclear.	Set proportionate documentation and review rules: documentation thresholds; audit trails.
	A structural visibility problem may be treated only as individual misconduct.	Enable safe discussion of borderline cases: consultation pathway; guidance on sensitive or submission-relevant data.
Reliance management	AI outputs may be accepted without scrutiny appropriate to the task's risk.	Calibrate reliance to risk: verification triggers; escalation criteria.
	Human-in-the-loop may become a sign-off exercise rather than critical oversight.	Secure meaningful human oversight: reviewer time, expertise, authority, and information.
	Repeated reliance may weaken skills needed to detect errors and question assumptions.	Maintain professional judgment over time: training; task rotation; periodic competency checks.

Note: The measures listed are illustrative organizational responses and are not intended as formal regulatory requirements.

critical evaluation of AI outputs, and incident reporting (*13*). These requirements are important because meaningful human oversight depends on more than mere presence of a human reviewer. Reviewers must have sufficient information, time, expertise, and authority to question or override AI-assisted output. Without these conditions, human-centered design may become a procedural label rather than an operational safeguard.

Related evidence from radiology suggests that AI may not always reduce professional burden and may instead create additional interpretive, post-processing, or psychological demands in certain settings (*14*). Although this evidence cannot be directly generalized to drug development, it raises a relevant possibility: AI-assisted workflows may shift professional work away from producing outputs and toward verifying, documenting, and explaining them. In this sense, reliance management is not only a matter of individual professional training but also a governance concern for maintaining the credibility of AI-assisted drug development.

Table 1 summarizes the main ethical concerns and illustrative governance responses to the two implementation challenges discussed in this study.

5. Conclusions

The ethical challenge of AI use in drug development is not simply whether AI should be used or whether reliance on it can be avoided. Once AI is integrated into research and regulatory workflows, a certain degree of reliance becomes inevitable. The central question is whether AI involvement can be made visible and reviewable and whether reliance on AI-assisted outputs can remain conscious, proportionate, and resilient. The FDA/EMA joint principles provide a practical starting point by framing good AI practices around human-centered design, risk-based governance, data governance, performance assessment, lifecycle management, and clear communication.

These principles have limited impact unless organizations translate them into workflow-specific governance practices. Shadow use illustrates that AI involvement may remain insufficiently visible, including when AI functions are embedded in ordinary tools and not fully recognized by users. Reliance management illustrates a second challenge: the human-in-the-loop concept is meaningful only when reviewers have sufficient time, expertise, authority, and information to question AI-assisted outputs, and when organizations preserve professional judgment needed to decide when AI should be accepted, verified, escalated, or set aside. The aim is not to discourage responsible AI use but to govern inevitable reliance in ways that remain compatible with scientific and regulatory integrity.

Funding: This study was supported by the Japan Agency for Medical Research and Development (JP 26oa0439001, JP223fa627001).

Conflict of Interest: The author has no conflicts of interest to disclose.

References

1. American Medical Association. 2026 physician survey on augmented intelligence. *https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf* (accessed May 30, 2026).
2. American Medical Association. Augmented intelligence in medicine. *https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine* (accessed May 30, 2026).
3. U.S. Food and Drug Administration; European Medicines

Agency. Guiding principles of good AI practice in drug development. *https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development* (accessed May 30, 2026).
4. European Medicines Agency. EMA and FDA set common principles for AI in medicine development. *https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0* (accessed May 30, 2026).
5. Health Policy Watch. EU and US regulators reach landmark accord on AI principles in drug development. *https://healthpolicy-watch.news/eu-and-us-ai-principles/* (accessed May 30, 2026).
6. Oualikene-Gonin W, Jaulent MC, Thierry JP, Oliveira-Martins S, Belgodère L, Maison P, Ankri J; Scientific Advisory Board of ANSM. Artificial intelligence integration in the drug lifecycle and in regulatory science: Policy implications, challenges and opportunities. Front Pharmacol. 2024; 15:1437167.
7. Klotz S, Kopper A, Westner M, Strahringer S. Causing factors, outcomes, and governance of shadow IT and business-managed IT: A systematic literature review. International Journal of Information Systems and Project Management. 2019; 7:15-43.
8. Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. Artificial Intelligence Review. 2025; 58:356.
9. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012; 19:121-127.
10. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: An attentional integration. Hum Factors. 2010; 52:381-410.
11. Green B. The flaws of policies requiring human oversight of government algorithms. Comput Law Secur Rev. 2022; 45:105681.
12. NHS AI Lab; Health Education England. Understanding healthcare workers' confidence in AI: Report 1 of 2. *https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/dart-ed/understandingconfidenceinai-may22.pdf* (accessed May 30, 2026).
13. World Medical Association. WMA statement on artificial and augmented intelligence in medical care. *https://www.wma.net/policies-post/wma-statement-on-artificial-and-augmented-intelligence-in-medical-care/* (accessed May 30, 2026).
14. Liu H, Ding N, Li X, Chen Y, Sun H, Huang Y, Liu C, Ye P, Jin Z, Bao H, Xue H. Artificial intelligence and radiologist burnout. JAMA Netw Open. 2024; 7:e2448714.

Received June 5, 2026; Revised June 17, 2026; Accepted June 18, 2026.

Released online in J-STAGE as advance publication June 20, 2026.

**Address correspondence to*:
Yusuke Inoue, Department of Healthcare Ethics, Kyoto University School of Public Health, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan.
E-mail: inoue.yusuke.6m@kyoto-u.ac.jp

Global Health & Medicine. 2026; 8(3):166-181. REVIEW
DOI: 10.35772/ghm.2026.01048

Clinical artificial intelligence (AI) in Japan: Regulatory pathways, domain-specific evidence, and its data infrastructure from an international perspective

Kenji Karako[1,*], Wei Tang[1,2]

[1] Department of Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan;
[2] National Center for Global Health and Medicine, Japan Institute for Health Security, Tokyo, Japan.

Abstract: Artificial intelligence (AI) has advanced rapidly across clinical domains, generating both a growing evidence base and dedicated regulatory frameworks for AI-based software as a medical device (SaMD). This review provides a comprehensive assessment of clinical AI across five major domains—diagnostic imaging, gastrointestinal endoscopy, cardiology and remote patient monitoring, diagnosis of infectious diseases, and an AI-ready data infrastructure—examining Japan's regulatory framework, approved device portfolio, and research contributions in an international context. We reviewed literature published between 2019 and 2026, using Japan's regulatory trajectory, approved device portfolio, and domain-specific research output as the primary lens for international comparison and prioritizing prospective studies, multicenter trials, and real-world implementation reports. The state of evidence varies markedly across the domains examined: endoscopy AI has the strongest randomized trial base, while diagnostic imaging AI has seen a systematic decline in real-world performance despite large-scale regulatory approval. Across the three dimensions examined, Japan has a distinctive profile: its strengths are a regulatory and clinical deployment infrastructure—evince by an established program medical device pathway and among the world's highest densities of diagnostic imaging systems and endoscopy volumes—while the data infrastructure lags, constrained by limited open-access resources relative to programs such as The Cancer Imaging Archive and the European Health Data Space. Large language models and generative AI, falling largely outside existing SaMD frameworks, carry the risk of hallucinations and gaps in oversight that healthcare systems in Japan and abroad are only beginning to address. Japan's established program medical device regulatory pathway, high-volume clinical deployment infrastructure, and proven regulatory-approval-to-reimbursement pathway provide a strong foundation for clinical AI adoption; post-approval change management frameworks and clinical accountability mechanisms need to be strengthened, AI-ready data accessibility needs to be expanded, and validated tools need to be embedded within reimbursed clinical workflows to translate this foundation into internationally competitive AI development and deployment.

***Keywords*:** artificial intelligence (AI), machine learning, Japan, software as a medical device, data infrastructure

1. Introduction

Artificial intelligence (AI), and in particular machine learning and deep learning, has rapidly transformed clinical medicine over the past decade. Applications now span diagnostic imaging, electrocardiogram (ECG) interpretation, gastrointestinal endoscopy, diagnosis of infectious diseases, radiation treatment planning, surgical assistance, and remote patient monitoring (*1*). In several narrowly defined tasks—such as detecting diabetic retinopathy from fundus photographs or adenomas during colonoscopy—AI systems have demonstrated diagnostic accuracy comparable to or exceeding that of expert clinicians (*2*). These developments have generated substantial clinical interest and have accelerated the regulatory and health technology assessment processes needed to translate research findings into safe, effective, and reimbursable clinical tools (*3*).

The governance of AI-enabled clinical software has converged on the concept of software as a medical device (SaMD), defined by the International Medical Device Regulators Forum (IMDRF) as software that fulfils a medical purpose independently of hardware (*4*). Major regulatory entities—including the United States Food and Drug Administration (FDA), the European Union (EU), the United Kingdom (UK), Japan, China, Singapore, and South Korea—have each established or are actively developing SaMD-

specific frameworks that address risk classification, clinical evaluation, post-marketing surveillance, and the particular challenge of managing AI model updates over the product lifecycle (*5-10*). In parallel, the research community has developed standardized reporting guidelines—including CONSORT-AI for clinical trials (*11*), TRIPOD+AI for prediction model studies (*12*), and STARD-AI for diagnostic accuracy research (*13*)—to improve the transparency and reproducibility of AI-based clinical evidence (*14*). Despite this regulatory and methodological progress, systematic comparative analysis of clinical AI development across major healthcare systems—jointly examining regulatory maturity, clinical deployment infrastructure, and data governance—remains limited.

Meaningful international comparison of medical AI development requires examining not only regulatory frameworks but also three complementary dimensions: the maturity and transparency of regulatory and reimbursement pathways, the clinical infrastructure available to support deployment, and the data resources required for model training and validation. Japan has classified AI-based software as "program medical devices" under the Pharmaceuticals and Medical Devices (PMD) Act; as of September 2025, 51 of the 172 approved program medical devices carried an AI-utilization designation, spanning imaging, endoscopy, cardiology, and infectious disease applications (Figure 1) (*8,15*). The United States FDA maintains a non-exhaustive public list of over 1,400 AI-enabled medical device authorizations spanning three decades, reflecting a markedly higher absolute volume of cleared or approved AI tools (*16*). With respect to clinical infrastructure, Japan's density of CT, MRI, and PET scanners—184 units per million population, compared to 86 in the United States, 74 in Germany, and 19 in the United Kingdom—indicates a high-volume imaging environment that represents a substantial deployment base for AI-assisted clinical applications (Figure 2) (*17,18*). The third dimension—data infrastructure for AI model development and validation—varies substantially across countries in scale, accessibility, and governance, and is examined in detail in Section 7.

This review examines the current state of clinical AI across five major domains, positioning Japan's regulatory maturity, clinical deployment infrastructure, and data resources within an international comparative framework to assess where Japan stands and what developments are required to sustain progress. We examine evidence published from 2019 through 2026, with an emphasis on studies published since 2021, prioritizing prospective studies, multicenter trials, and real-world implementation reports, supplemented by key regulatory and guideline documents. Following a summary of the global regulatory landscape (Section 2), we address AI in diagnostic imaging (Section 3), endoscopy (Section 4), cardiology and remote monitoring (Section 5), diagnosis of infectious diseases (Section 6), and an AI-ready data infrastructure (Section 7). Cross-cutting challenges—including the methodological and regulatory implications of large language models and generative AI, which represent an emerging category of clinical tools with a distinct evidence and oversight profile—are discussed in Section 8, followed by conclusions in Section 9. Table 1 provides an overview of the landmark clinical studies discussed across these domains.

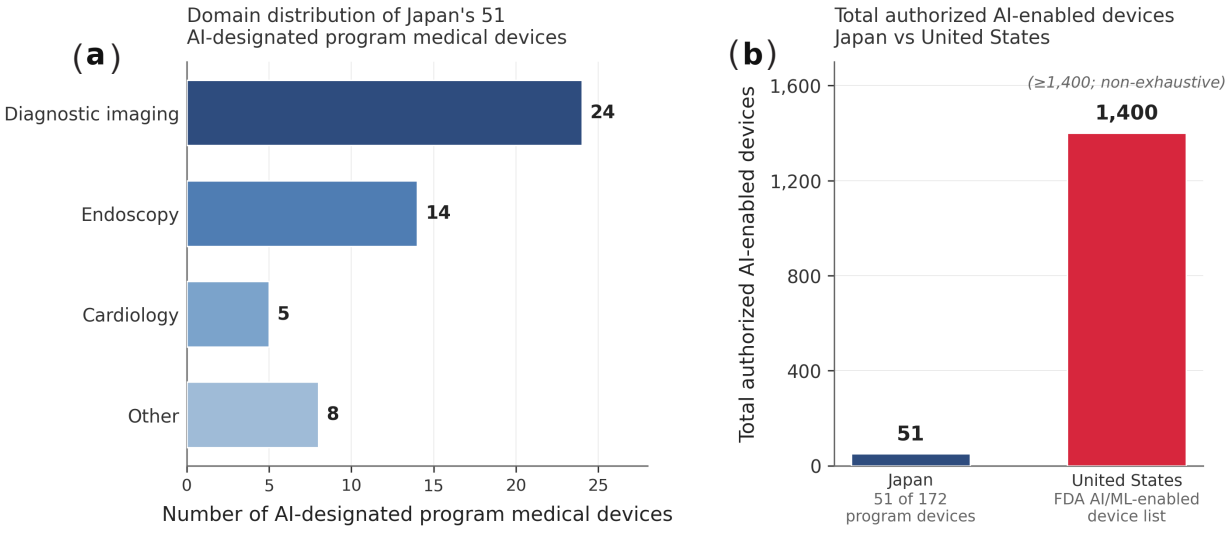


Figure 1. AI-enabled medical device authorizations as of September 2025. (a) Distribution of Japan's 51 designated program medical devices using AI; "Other" encompasses the diagnosis of infectious disease and additional applications not individually enumerated. **(b)** Total AI-enabled devices authorized in Japan (51 of 172 program medical devices) versus the United States FDA AI/ML-enabled device list (≥ 1,400 authorizations; non-exhaustive public list). Data sources: PMDA program medical device list (*15*); MHLW approval status (*33*); FDA AI/ML-enabled device list (*16*).

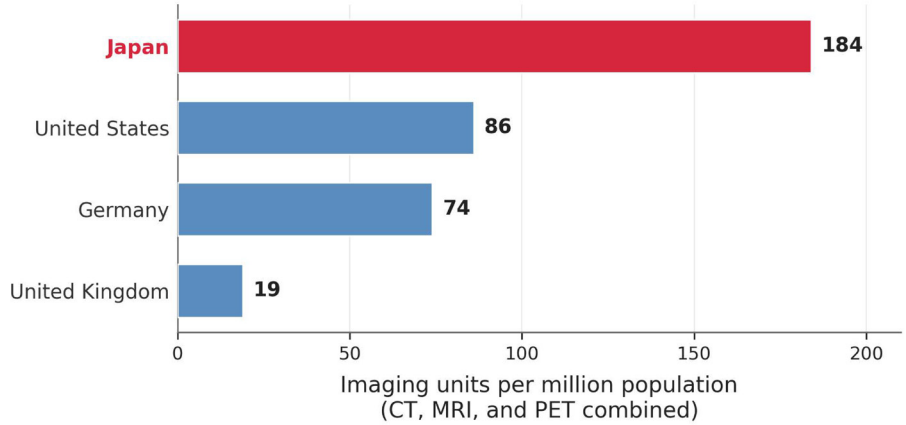


Figure 2. Density of CT, MRI, and PET scanners per million population in selected countries (2023). Japan's substantially higher scanner density compared to other major economies provides a high-volume imaging environment that constitutes a large deployment base for AI-assisted clinical applications. Data source: OECD Health Statistics 2025 (*17,18*).

Table 1. Landmark clinical AI studies by domain discussed in this review

Domain	Study	Design	Endpoint	Key finding
Radiology	McKinney *et al*. 2020 (*20*)	Retrospective validation	Breast cancer detection	Outperformed clinical reads on UK+US test sets; surpassed all 6 radiologists in US reader study
Endoscopy	Wang *et al*. 2019 (*21*)	RCT	ADR	34% *vs* 28% (control)
Endoscopy	Repici *et al*. 2020 *(22)*	RCT	ADR	54.8% *vs* 40.4% (control)
Cardiology (ECG)	Attia *et al*. 2019 (*23*)	Retrospective	AF detection in sinus rhythm	AUC 0.87 (454,789 ECGs)
Cardiology (wearable)	Perez *et al*. 2019 (*24*)	Prospective cohort	PPV for AF notification	84% concurrent AF confirmation (419,297 participants)
Infectious diseases	Okiyama *et al*. 2022 *(25)*	Prospective validation	Accuracy of influenza diagnosis	Comparable to or exceeding antigen rapid test, especially in early illness

Note: ADR, adenoma detection rate; AF, atrial fibrillation; AUC, area under the receiver operating characteristic curve; ECG, electrocardiogram; PPV, positive predictive value; RCT, randomized controlled trial.

2. The global regulatory landscape for medical AI

While major regulatory jurisdictions have converged on the SaMD concept established by the IMDRF, they differ considerably in how they classify, evaluate, and govern AI-based software over its clinical lifecycle—differences that directly determine the pace of and conditions under which AI tools reach clinical deployment.

2.1. The SaMD framework

The IMDRF classifies SaMD risk along two axes—the significance of information provided for clinical decision-making and the healthcare situation in which it is used—yielding four risk categories with progressively stringent requirements; clinical evaluation must demonstrate analytical validity, clinical validity, and clinical utility (*4,26*). The IMDRF N88 document (2025) formalizes shared Good Machine Learning Practice (GMLP) expectations—covering data quality, model transparency, and post-marketing monitoring—that underpin lifecycle management requirements across jurisdictions (*27,28*).

2.2. Regulatory approaches across major jurisdictions

Despite adopting the SaMD concept, major jurisdictions differ considerably in their classification systems, authorization pathways, and AI-specific guidance (Table 2).

The FDA's Predetermined Change Control Plan (PCCP) allows pre-specified model updates without a new submission; the EU AI Act (August 2024) classifies medical AI as high-risk under a dual MDR/AI Act compliance regime (*5,6,29*). The UK's MHRA publishes dedicated software and AI lifecycle guidance, and the National Institute for Health and Care Excellence (NICE) maintains an Evidence Standards Framework that stratifies evidence requirements for digital health technologies

Table 2. Regulatory frameworks for AI-based medical devices across major countries and regions

Country/Region	Regulatory framework	AI-specific characteristics
United States	Class I–III; 510(k)/De Novo/PMA (FDA)	PCCP for pre-specified model updates; Good Machine Learning Practice principles
European Union	MDR/IVDR; Notified Body conformity assessment	AI Act (high-risk classification); dual MDR–AI Act compliance requirement
United Kingdom	UK MDR 2002; MHRA authorization	MHRA SaMD/AI lifecycle guidance; NICE Evidence Standards Framework for NHS adoption
China	NMPA Class II/III registration	AI classification principles (2021); PIPL data governance requirements
Singapore	HSA SaMD lifecycle regulation	Change management guidance; AI in Healthcare Guidelines (AIHGle 2.0)
South Korea	MFDS; Digital Medical Products Act	Dedicated AI device evaluation guidelines covering trial design and validation
Japan	PMD Act; Program Medical Device	Machine learning review guidance from the PMDA; post-approval change management is considered

Note: FDA, Food and Drug Administration; GMLP, Good Machine Learning Practice; HSA, Health Sciences Authority; IVDR, In Vitro Diagnostic Regulation; MDR, Medical Device Regulation; MFDS, Ministry of Food and Drug Safety; MHRA, Medicines and Healthcare products Regulatory Agency; NHS, National Health Service; NICE, National Institute for Health and Care Excellence; NMPA, National Medical Products Administration; PCCP, Predetermined Change Control Plan; PIPL, Personal Information Protection Law; PMD Act, Pharmaceuticals and Medical Devices Act; PMDA, Pharmaceuticals and Medical Devices Agency; PMA, Premarketing Approval; SaMD, software as a medical device; UK MDR, UK Medical Device Regulations.

according to their function and risk, providing the basis for NHS procurement decisions (*7,30*).

In Asia, South Korea enacted the Digital Medical Products Act to provide a dedicated framework for AI-enabled digital health, Singapore's Health Sciences Authority has adopted a lifecycle approach with explicit change management guidance and has issued its AI in Healthcare Guidelines, and China's National Medical Products Administration published classification principles for AI-based medical software in 2021 (*9,10,31*). Japan regulates AI-based software as "program medical devices" under the PMD Act, with the Pharmaceuticals and Medical Devices Agency (PMDA) having published review considerations specific to machine learning-based devices covering risk–benefit assessment, analytical and clinical validation, and post-market change management (*8,32*). As of September 2025, 172 program medical devices held valid marketing authorization in Japan under the PMD Act framework, 51 of which carried an AI-utilization designation spanning imaging, endoscopy, cardiology, and infectious disease applications (Figure 1) (*15*). Regulatory authorization under the PMD Act is a prerequisite for, but distinct from, reimbursement under Japan's national health insurance system; coverage classification is determined separately by the Central Social Insurance Medical Council under the C1 and C2 new-technology categories—a pathway examined in the context of the nodoca system (Section 6.2) and discussed as a policy challenge in Section 8.3.

3. AI in diagnostic imaging

Of the 51 AI-utilization–designated program medical devices approved by the PMDA as of September 2025, 24 fall within the imaging category—the largest single subcategory—with approved indications spanning chest CT and plain radiography (detection of lung nodules, quantification of interstitial lung disease, and identification pneumonia), brain MRI (detection of an intracranial aneurysm), breast ultrasound (characterization of lesions), and musculoskeletal applications (detection of a fracture) (*15,33*). This concentration of approvals reflects both the maturity of the PMDA regulatory pathway for image analysis software and Japan's exceptional imaging infrastructure: with 184 CT, MRI, and PET units per million population—more than twice the density in the United States or Germany—Japan has one of the highest-volume deployment environments for AI-assisted clinical applications worldwide (Figure 2) (*17,18*).

In breast cancer screening, a deep learning model outperformed clinical reads on held-out UK and US test sets—reducing false-positive rates by 1.2% (UK) and 5.7% (US)—surpassing all six radiologists in an independent US reader study, and it reduced second-reader workload by 88% in a simulated UK double-reading workflow (*20*). In chest radiology and CT, AI systems have displayed improved performance across detection of lung nodules, characterization of interstitial lung disease, and triage of acute findings (*34*). In time-critical settings, automated large vessel occlusion (LVO) detection software significantly reduced door-to-groin (DTG) times and the time from CT initiation to the start of endovascular therapy (EVT) in acute stroke care (*35*).

Despite these proof-of-concept demonstrations, systematic analyses have consistently identified a

critical gap between single-facility performance and generalization to external populations: AI models validated at one site frequently exhibit marked performance degradation when applied to images acquired with different scanners, acquisition protocols, or patient demographics (*36*). Japan's imaging devices span a diverse mix of domestic manufacturers alongside international platforms operating under varying acquisition protocols, and training datasets compiled from a single center are unlikely to represent the full range of image characteristics encountered in clinical practice; differences in population-level prevalence and temporal drift from scanner upgrades further erode post-deployment calibration. These challenges underscore the importance of prospective multi-site validation and post-marketing performance monitoring (*12,13*).

Published real-world deployment evidence in Japan remains limited relative to the volume of regulatory authorizations. A prospective evaluation of AI approved for the detection of an intracranial aneurysm by Ito *et al*. highlights this gap: while diagnostic sensitivity was maintained in routine clinical use, the false-positive burden in unselected cases substantially exceeded that reported in the regulatory validation study (*37,38*). Comparable post-deployment evaluations of AI systems for chest imaging—the most numerous category within Japan's approved imaging portfolio—have not yet been widely published, reflecting a broader structural gap between pre-marketing validation requirements and the post-marketing performance evidence base.

4. AI in endoscopy

Gastrointestinal endoscopy is among the most evidence-rich domains for clinical AI, with a substantial body of randomized controlled trial (RCT) data supporting the integration of AI-based detection and characterization tools into colonoscopy practice. As of September 2025, 14 of the 51 AI-utilization–designated program medical devices approved by the PMDA fall within the endoscopy category—the second largest subcategory—reflecting high procedural volumes and an exceptional gastrointestinal malignancy burden (*15*). The endoscopy setting offers several features that facilitate AI development and evaluation: procedures are video-based and therefore generate large volumes of labelled training data; the primary clinical endpoint—the adenoma detection rate (ADR)—is a measurable surrogate for colorectal cancer prevention; and the procedural context allows direct real-time display of AI output to the endoscopist without requiring a separate reporting infrastructure.

4.1. Computer-aided detection and diagnosis: CADe and CADx

Computer-aided detection (CADe) systems analyze the live video stream and generate real-time alerts when the AI model identifies a region with a suspected polyp or other mucosal lesion, prompting the endoscopist to inspect the flagged area. Computer-aided diagnosis (CADx) systems go further, characterizing a detected lesion—for example, distinguishing a hyperplastic polyp from an adenoma, or predicting the depth of submucosal invasion—to aid in deciding on a strategy to "resect and discard" or "diagnose and leave," thus reducing unnecessary polypectomies and associated costs.

The most extensively studied clinical application is CADe-assisted colonoscopy for detection of colorectal adenoma. Wang *et al*. conducted an RCT in which patients undergoing colonoscopy were allocated to AI-assisted detection or conventional colonoscopy, and they found that the group with AI-assistance had a significantly higher ADR (34% versus 28%), with the system detecting additional adenomas that were missed in the withdrawal phase (*21*). A subsequent multicenter RCT by Repici *et al*. confirmed these findings in an Italian population, demonstrating that real-time AI-aided colonoscopy increased the ADR compared to standard colonoscopy (54.8% versus 40.4%) (*22*). Multiple meta-analyses of RCTs have since confirmed that CADe-assisted colonoscopy improves the ADR as a pooled outcome, with consistent findings across different AI systems and endoscopic settings (*39*).

Despite these efficacy data, CADe implementation raises important questions: the ADR is a surrogate endpoint for interval cancer prevention with no prospective proof that AI-assisted ADR gains reduce the long-term incidence of interval cancer; false-positive alerts add inspection time and risk operator fatigue; and the benefit is heterogeneous across endoscopists, suggesting AI assistance functions primarily as a quality levelling tool rather than a universal performance enhancer (*40*).

High-confidence optical characterization of diminutive polyps (≤ 5 mm) is the prerequisite for the "resect and discard" strategy, in which diminutive adenomas are removed without pathological examination and hyperplastic polyps are left in situ. Achieving the negative predictive value threshold (≥ 90% for high-confidence diagnoses) required by international guidelines requires not only high AI accuracy but also rigorous operator training and standardized image capturing conditions—requirements that have proved difficult to satisfy in routine endoscopy practice outside expert centers (*41*).

4.2. Japan's AI endoscopy portfolio and multicenter trial contributions

The age-standardized rate (ASR) for the incidence of colorectal cancer in Japan is 36.6 per 100,000 population—substantially above rates in the United States (27.0), United Kingdom (30.9), and Germany

(25.7)—while the ASR for the incidence of gastric cancer is 27.6 per 100,000, compared to 4.1 in the United States (*42,43*). This dual malignancy burden translates into high endoscopic procedure volumes, an established endoscopic training culture, and a correspondingly large supply of annotated training data for AI development.

EndoBRAIN (Olympus Corporation / Cybernet Systems) is a CADx system that uses a convolutional neural network to characterize colorectal lesions in real time during a colonoscopy, outputting a probability score for neoplastic versus non-neoplastic tissue (*40,44*). The device received PMDA approval and has been deployed in clinical practice at multiple Japanese facilities. The EndoBRAIN-EYE variant, which integrates CADe functionality with the existing CADx platform, has been evaluated in prospective clinical trials registered in the University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR), including the EYE-OPENER trial (*45*). These prospective studies are designed to assess whether the combined CADe/CADx platform improves clinically meaningful outcomes in routine endoscopy practice, going beyond the ADR to examine resection rates, procedure duration, and operator workload. Japan's high burden of gastric malignancy has similarly driven the development of AI for gastric endoscopy: PMDA-approved program medical devices in this domain include tools for detection of gastric lesions and for assessment of the depth of early gastric cancer invasion—the latter is a clinically critical distinction that determines eligibility for endoscopic versus surgical resection (*15*). The existence of approved tools across both colorectal and gastric applications reflects how Japan's epidemiological profile has shaped the scope of its regulatory portfolio in AI-assisted endoscopy.

Recognizing that AI systems trained and validated predominantly in single-country populations may not be generalizable across the diverse ethnicities, dietary patterns, and bowel preparation practices of Asian populations, a multicenter trial initiative coordinated by the National Cancer Center Japan was announced in 2024 (*46*). This program aims to evaluate the efficacy of AI-assisted colonoscopy across multiple Asian countries in a prospectively registered, multicenter design, with the explicit goal of generating evidence of external validity applicable to the broader Asia-Pacific region. The initiative is particularly significant because the majority of high-quality RCTs underpinning AI-assisted colonoscopy have been conducted in predominantly East Asian or European populations with limited cross-population generalizability; Japan's capacity to coordinate multinational Asian trials positions it as a key contributor to closing this gap in external validity.

5. AI in cardiology and remote patient monitoring

Cardiovascular disease remains the leading cause of death globally, and the cardiological domain offers particularly rich sources of structured longitudinal data—electrocardiograms (ECGs), echocardiograms, implantable device telemetry, and wearable biosignals—that are well-suited to machine learning. Of the 51 AI-utilization–designated program medical devices approved in Japan as of September 2025, five fall within cardiology—the smallest subcategory by volume, reflecting the complexity of cardiac AI development relative to image analysis tasks—and yet Japan's high cardiovascular disease burden, aging population, and extensive cardiac registry infrastructure represent a deployment environment with considerable potential (*15,18*).

5.1. AI for ECG interpretation and cardiac risk stratification

The standard 12-lead ECG is the most widely performed cardiac examination worldwide, and its digital format makes it immediately tractable for deep learning analysis. A foundational study by Attia *et al*. at the Mayo Clinic trained a convolutional neural network on 454,789 ECGs from 180,922 patients and demonstrated that the model could identify patients who had AF but were currently in sinus rhythm at the time of recording, with an area under the receiver operating characteristic curve (AUC) of 0.87 in a validation set of 36,280 patients (*23*). This finding implied that the AI system detected subtle ECG features reflecting atrial remodeling that precede the clinical manifestation of atrial fibrillation (AF)—a latent clue not evident in conventional ECG interpretation. A companion study by the same group used a similar architecture to detect left ventricular (LV) systolic dysfunction, achieving an AUC of 0.93 at identifying patients with an ejection fraction of 35% or less, suggesting that population-level screening for LV dysfunction using routine ECGs may be feasible (*47*).

These results have stimulated substantial interest in research, but several barriers to clinical translation remain. Most published models are trained and validated on retrospective data from single facilities, and external validation across different ECG recording systems, electrode placement standards, and demographic populations has revealed a significant heterogeneity in performance (*48*). The clinical actionability of a positive AI alert also depends critically on the availability of a downstream cardiological assessment: a model that correctly identifies patients at risk of future AF is of clinical value only if those patients can be promptly referred for confirmatory testing and, where indicated, started on anticoagulation therapy. Without a structured care pathway linked to AI output, the population-level benefit of screening remains theoretical. Early prospective evidence from a Japanese cohort study suggests that AI-ECG screening can identify AF with a higher discriminative ability than conventional risk scoring, offering a practical approach to population-based

screening in older adults (*49*). Prospective multicenter validation studies of AI-ECG tools specifically in Japanese populations remain limited relative to the volume of approved products, and the existing cardiac registry infrastructure—notably the Japan Registry of All Cardiac and Vascular Diseases (JROAD, covering 1,500 facilities as of 2023) and the Japan Percutaneous Coronary Intervention Registry (J-PCI, with over 2.4 million cumulative cases)—represents an underutilized platform with which to close this gap in external validity (*50,51*) (discussed further in Section 7).

5.2. Wearable devices and AF screening

The Apple Heart Study, conducted in the United States between 2017 and 2018, enrolled 419,297 participants who wore an Apple Watch and consented to receive irregular pulse notifications triggered by an algorithm detecting inter-beat interval irregularity (*24*). Among participants who received a notification, 84% had AF confirmed with a simultaneous ECG patch worn at the time of notification, and 34% had AF identified in a subsequent 90-day monitoring period. While the positive predictive value for concurrent AF was high, the study highlighted a fundamental challenge of population screening: persistent AF was not ultimately confirmed in the majority of notification recipients during follow-up, requiring medical evaluation for a large number of individuals whose downstream management trajectory remained uncertain.

The Fitbit Heart Study subsequently evaluated PPG-based irregular rhythm detection on a different consumer platform in a US cohort and reported a high positive predictive value (98.2%) for AF when an irregular rhythm notification was issued during concurrent ECG monitoring (*52*). Taken together, these landmark studies established that consumer wearables can achieve clinically relevant sensitivity in AF detection, but they also demonstrated that the translation of a wearable alert into a meaningful clinical outcome requires a carefully designed confirmation and care pathway. High notification rates in low-prevalence populations generate large numbers of notifications requiring medical follow-up, potentially overwhelming cardiology services and exposing patients to the anxiety and risks associated with unnecessary examination.

Calibrating the detection algorithm threshold—balancing sensitivity against the false-positive notification rate—is a design decision with direct implications for healthcare system workload and patient experience that must be addressed before population-scale wearable AF screening programs can be implemented responsibly (*53*). In Japan, home-use program medical devices incorporating AF detection notifications and sleep apnea detection alerts have received regulatory approval and are commercially available (*15,33*), but systematic evaluation of their real-world impact on clinical outcomes and healthcare utilization in the Japanese context has not yet been widely reported. The contrast with the Apple Heart Study—which enrolled 419,297 participants in a single prospective study—illustrates the scale at which the leading wearable AF screening programs have been evaluated internationally; prospective wearable studies of comparable scale have not yet been reported in Japanese cohorts, despite the direct relevance of AF screening to Japan's aging population. Japan's demographic profile—the highest proportion of adults age 65 or older among OECD countries—creates a strong demand for AI-assisted home monitoring (*18,19*). CureApp HT (a smartphone-based hypertension management device) began to be covered by national insurance in 2022 after demonstrating cost-effectiveness, illustrating the pathway through which evidence-based digital health tools can qualify for reimbursement in Japan (*54*).

6. AI in the diagnosis of infectious diseases

The diagnosis of infectious diseases presents a distinctive set of challenges for AI: the pathogen in question, disease prevalence, and the clinical presentation of infection vary with season, geographic region, circulating strain, and host immunity—factors that shift the data distribution that a model encounters after deployment relative to the distribution on which it was trained. AI applications in this domain span three broad functional categories: image-based diagnosis (interpreting chest radiographs or CT scans for pneumonia and related conditions), augmentation of a rapid test (automating or checking the quality of the reading of lateral flow or other point-of-care tests), and multimodal clinical integration (combining symptom profiles, vital signs, and laboratory or imaging results to generate a differential diagnosis probability—as exemplified by the nodoca system for influenza diagnosis discussed in Section 6.2). Japan holds a distinctive position in this domain: despite a smaller portfolio of AI tools approved for diagnosis of approved infectious diseases compared to imaging or endoscopy, it has produced one of the most operationally complete regulatory-approval-to-reimbursement trajectories for an AI-based clinical decision-making support system in the Asia-Pacific region, through the nodoca system for diagnosis of influenza, and it has since expanded approved AI applications to include culture-free identification of species causing bacterial infections (*15*).

6.1. The COVID-19 experience and methodological lessons

Studies proliferated rapidly during the COVID-19 pandemic, with many reporting a high diagnostic performance in detecting pneumonia due to SARS-CoV-2 on chest CT or plain radiography using deep learning models. A systematic methodological review

by Roberts *et al*. evaluated 415 published machine learning studies involving COVID-19 detection and prognosis based on chest imaging, and it concluded that none met the requirements for clinical use (*55*). The most commonly identified problems included dataset contamination (patients appearing in both training and test sets), use of retrospective convenience samples with unrepresentative negative controls, a lack of external validation, and failure to account for the clinical context in which the model would be deployed. This analysis provided a cautionary benchmark for the field, illustrating how the urgency of a public health crisis can accelerate publication volume while simultaneously eroding methodological standards.

A key challenge specific to AI diagnosis of infectious diseases is the distributional shift driven by changing prior probability: a model trained during a COVID-19 wave will encounter a very different disease prevalence in an inter-epidemic period, substantially altering positive and negative predictive values even if model sensitivity and specificity are unchanged (*3*). This structural vulnerability means that AI systems to diagnose infectious diseases require more frequent performance re-evaluation and retraining than systems applied to stable disease populations, with post-marketing monitoring designed around seasonal and epidemic periodicity.

6.2. nodoca: AI-assisted diagnosis of influenza in Japan

nodoca (Aillis Inc., Tokyo) is a program medical device that acquires images of the posterior pharynx using a dedicated imaging device and it integrates this visual information with structured clinical data—including patient age, symptom onset, body temperature, and vaccination history—to estimate the probability of influenza and generate a differential diagnosis to assist the clinician in his or her assessment. Okiyama *et al*. reported that the deep learning model trained on pharyngeal images and structured clinical records achieved an influenza detection accuracy comparable to, and in some subgroups exceeding, conventional rapid antigen tests—particularly in the early phase of illness when viral loads may fall below standard kit sensitivity (*25*).

The PMDA reviewed and approved nodoca as a program medical device, with a regulatory review report addressing its intended clinical use—an aid to diagnose influenza—together with evidence of its analytical and clinical validation, the performance boundaries of the model, and post-marketing surveillance requirements (*56*). In December 2022, the Central Social Insurance Medical Council granted nodoca a C2 reimbursement classification, evaluating it as a new technology fee rather than as a standard device covered by insurance (*57*). This sequence—prospective clinical validation, PMDA regulatory authorization, and Chuikyo's approval of its reimbursement—constitutes the most complete and transparent example of the regulatory–reimbursement chain for an AI-based clinical decision-making support system in Japan to date, and its serves as a reference pathway for developers of subsequent AI diagnostic tools.

Implementation challenges include image quality consistency in busy primary care settings, the need for annual recalibration as dominant circulating strains and vaccine components change across influenza seasons, and unresolved assignment of liability when an AI-assisted diagnostic decision leads to an adverse outcome (*3*).

6.3. Broader applications and future directions

Beyond influenza and COVID-19, AI's applications in the diagnosis of infectious diseases are expanding into detection of tuberculosis on chest radiography (*58*) and prediction of sepsis risk based on clinical and laboratory data. Sepsis represents a disease burden on a scale that warrants AI-based detection tools: global estimates indicate approximately 48.9 million cases and 11 million deaths annually, accounting for roughly 20% of all deaths worldwide (*59,60*). In 2024, the Sepsis ImmunoScore—an AI tool predicting sepsis severity based on host immune biomarkers—received FDA De Novo authorization, establishing a regulatory reference point for AI to evaluate sepsis that Japan has not yet matched under the PMDA framework (*61*). In Japan, the portfolio of AI approved for diagnosis of infectious diseases does extend beyond influenza: PMDA-authorized program medical devices include AI software for estimating bacterial species causing urinary tract infections without culture and a microbial classification support program—tools that complement nodoca's multimodal approach and signal an expanding scope for culture-free AI-based diagnosis of infectious diseases in Japan (*15*). Systematic clinical validation data regarding these newer approved tools have not yet been widely published, representing an evidence gap comparable to that observed in the imaging domain.

Adherence to STARD-AI and TRIPOD+AI reporting standards (*12,13*) is especially critical to AI for diagnosis of infectious diseases given the domain's structural susceptibility to distributional shift—a challenge that post-marketing surveillance frameworks in Japan and abroad are only beginning to systematically address.

7. An AI-ready data infrastructure: International initiatives and Japan's position

The performance of AI models in clinical medicine is bounded not only by algorithmic design but by the quality, scale, and interoperability of the data on which they are trained and validated. As evidence has accumulated that models trained at single facilities regularly fail to generalize across sites, the field has

converged on a consensus that a large-scale, multimodal, and externally accessible data infrastructure is a necessary precondition for robust AI development. This section surveys the principal international initiatives for building an AI-ready medical data infrastructure, characterizes Japan's current position, and comparatively analyzes strengths, gaps, and trajectories. Japan's position in this landscape exemplifies a structural paradox that recurs across the domains reviewed in this paper: the country generates some of the world's largest volumes of clinical and administrative health-related data through its universal insurance system, and yet the infrastructure for converting this data into AI-ready, externally accessible research platforms lags materially behind that of comparable OECD nations—a gap that is the central analytical focus of this section.

7.1. International initiatives for AI-ready medical data

Maintained by the National Cancer Institute, the Cancer Genome Atlas (TCGA) provides multi-omic and clinical data for over 20,000 samples across 33 cancer types and has underpinned a substantial fraction of published computational oncology and pathology AI research (*62*). The Cancer Imaging Archive (TCIA) complements the TCGA with over 30.9 million medical images from 37,568 subjects across a wide range of modalities and clinical contexts, making it the world's largest openly accessible repository of annotated medical imaging data (*63*). The NIH Bridge2AI program and the All of Us Research Program extend this infrastructure with explicit AI-readiness mandates and large-scale multimodal cohorts designed to support prospective validation on a population scale (*64,65*).

In Europe, the European Health Data Space (EHDS), which entered into force in March 2025, establishes a cross-member harmonized framework for the secondary use of electronic health data, including provisions for federated analysis without requiring data to leave national borders (*66*). The EHDS is expected to create a governance infrastructure for large-scale multicenter AI training and validation across EU member states—a capability structurally absent from European medical AI research to date. The United Kingdom's NHS AI Lab developed shared medical imaging platforms, though independent evaluation has highlighted implementation challenges in translating infrastructure investments into active use in research (*67*).

In Asia, South Korea has adopted a centralized curation strategy: the National Information Society Agency (NIA) manages the AI Hub, a nationally maintained repository of annotated datasets—including medical imaging, clinical records, and biosignal collections—made available to domestic and international researchers under defined access conditions (*68*). The South Korean model, in which government-funded agencies take direct responsibility for dataset quality and annotation, offers an instructive counterpoint to the federated governance designs favored in Europe and the market-facilitated repository model in the United States. In the Asian context, South Korea also represents the most direct structural comparator to Japan: both countries operate universal health insurance systems that generate large-scale administrative claims data, and yet South Korea has invested in a nationally coordinated annotation infrastructure for AI research that Japan has not yet replicated at a comparable scale—a contrast that frames the analysis of Japan's data position in Section 7.2. Key attributes of these international initiatives as well as Japan's principal infrastructure programs are compared in Table 3.

7.2. Japan's medical AI data infrastructure

Japan possesses several large-scale clinical databases that provide a foundation for medical AI research, though these have developed primarily for disease surveillance and pharmacovigilance rather than as AI-ready research platforms. The National Database of Health Insurance Claims and Specific Health Checkups (NDB) covers virtually the entire Japanese population through the universal health insurance system; while the scale of the claims data is unmatched, a systematic review of NDB-based research has identified access difficulties, issues with patient identification, and the lack of links to imaging or biomarkers as constraints on its utility for AI model development (*69*). The Medical Information Database Network (MID-NET), operated by the PMDA and linking hospital information systems at participating facilities to cover approximately 8.7 million patients—roughly 7% of Japan's population, in contrast to the NDB's near-universal coverage—was designed principally for post-marketing drug safety surveillance but represents a resource for AI-based research on pharmacovigilance and clinical outcomes (*70*).

Domain-specific registries further strengthen Japan's data position. The National Clinical Database (NCD), a surgical outcomes registry operated jointly by relevant surgical societies, had accumulated 28.48 million cumulative surgical cases as of 2023 (*71*). The JROAD, covering 1,500 participating facilities as of 2023, and the Japan Percutaneous Coronary Intervention Registry (J-PCI), with over 2.4 million cumulative cases, collectively represent one of the most complete national procedure-level cardiovascular datasets among OECD countries (*50,51*). These registries have supported observational and epidemiological research but have not yet been systematically capitalized upon as training or validation platforms for clinical AI models—a gap attributable to access governance designed for epidemiological use cases rather than AI development, the lack of links to imaging and molecular data, and the lack of any regulatory or institutional mandate requiring AI model validation to be conducted using these

Table 3. Principal international and Japanese initiatives on a data infrastructure for medical AI

Country	Initiative	Data type	Scale	Access model
United States	TCIA	Medical imaging	30.9M images, 37,568 subjects	Open access
United States	All of Us	EHR + genomic + wearable	> 873,000 participants	Controlled access
United States	Bridge2AI	Multi-modal (AI-ready)	$130M program	Controlled access
Europe	EHDS (2025)	EHR (federated)	Cross-border EU	Federated
United Kingdom	NHS AI Lab	Imaging + pathology	National platforms	Institutional
South Korea	AI Hub (NIA)	Annotated datasets	National repository	Controlled access
Japan	NDB	Claims + checkups	Near-universal coverage	Restricted
Japan	J-MID (JRS/AMED)	Medical imaging	543M images, 10 univ. hospitals	Controlled access
Japan	NCD/ ROAD/J-PCI	Surgical/cardiovascular	28.48M/1,500 inst./2.4M cases	Research access

Note: AMED, Japan Agency for Medical Research and Development; EHDS, European Health Data Space; EHR, electronic health record; J-MID, Japan Medical Imaging Database; J-PCI, Japan Percutaneous Coronary Intervention Registry; JROAD, Japan Registry of All Cardiac and Vascular Diseases; JRS, Japan Radiological Society; NCD, National Clinical Database; NDB, National Database of Health Insurance Claims and Specific Health Checkups; NHS, National Health Service; NIA, National Information Society Agency; TCIA, The Cancer Imaging Archive.

resources.

In the medical imaging domain, a high imaging volume and fragmented institutional data ownership means that Japan generates large quantities of diagnostic images and yet lacks the access infrastructure for competitive AI model development and multi-institutional validation—a gap identified by Ueda *et al*. and which prompted the J-MID initiative (*72*). The Japan-Medical Image Database (J-MID), a Japan Radiological Society (JRS) initiative supported by the Japan Agency for Medical Research and Development (AMED), addresses this gap by aggregating anonymized CT and MRI images with diagnostic reports from 10 major university hospitals *via* the SINET academic network; as of May 2024, J-MID had assembled over 543 million images (1.68 million cases), constituting what the project describes as an unparalleled repository of real-world radiological data in Japan (*73*).

The Next-generation Medical Infrastructure Act, promulgated in May 2023 and later amended, provides a new legal category of pseudonymized medical information, enabling certified data operators to link and process health records from multiple sources for research purposes under a unified governance regime (*74*). Progress on interoperability standards has been slower: the adoption of Japan-specific HL7 FHIR profiles for the exchange of medication and clinical data is still in at an early stage of implementation, which constrains the technical basis for cross-institutional data linkage that more mature federated research infrastructures require (*75*).

7.3. Comparative analysis: Strengths, gaps, and convergence

The NDB, MID-NET, and domain-specific registries represent population-wide or procedure-level datasets of a scale uncommon outside the United States, but access for research use is administratively difficult and the datasets lack the multimodal linkage—integrating imaging, genomic, and longitudinal clinical records—that characterizes the most capable international AI platforms.

The design philosophies of these initiatives differ materially: the United States combines open-access repositories with a federated infrastructure for sensitive records; the EHDS prioritizes federated-first data sovereignty; and Japan's framework under the Next-generation Medical Infrastructure Act is architecturally closer to the federated model, though a certified data operator ecosystem is still being established. Federated learning can approach centralized training performance when data distributions are similar across sites, but heterogeneous data quality and non-IID distributions remain active research challenges (*76*).

The competitive trajectory for Japan's medical AI will be materially shaped by three converging developments: the pace at which the Next-generation Medical Infrastructure Act framework generates research-accessible pseudonymized data pools, with only a handful of certified data operators having been designated as of early 2026; the extent to which the J-MID expands access beyond its current restricted institutional-partner model to support externally validated research comparable to that enabled by the TCIA, given that access remains limited to institutional partners despite the database having assembled over 543 million images from 10 major university hospitals; and the breadth of HL7 FHIR adoption enabling cross-institutional data linkage, currently in an early stage of implementation across Japan's hospital information systems. The clinical registries described above—the NCD, JROAD, and J-PCI— already provide a procedure-level evidence base that few countries can match and could become platforms for prospective AI validation studies if linked to imaging and molecular data under the governance framework the Next-generation Medical Infrastructure Act enables. Realizing this potential will require coordinated investment in data standardization, access infrastructure, and governance that goes beyond individual institutional initiatives and that matches the ambition of the Bridge2AI and EHDS programs described earlier.

8. Challenges and future directions

This section synthesizes the cross-cutting challenges that recur across the clinical domains reviewed earlier—other than data infrastructure and governance, addressed in Section 7—and it identifies the policy, methodological, and research developments needed to facilitate the responsible advancement of clinical AI.

8.1. Bridging the gap between performance metrics and clinical outcomes

A fundamental tension in clinical AI research is the reliance on intermediate performance metrics—the AUC, sensitivity, Dice similarity coefficient, and adenoma detection rate—as proxies for the patient-level outcomes that determine clinical value. The literature on AI-assisted imaging related to COVID-19, in which hundreds of studies reported high AUC values for retrospective data while none met the requirements for clinical deployment, illustrates the extent of this gap (*55*).

External validity is a related and persistent problem. Models trained at a single facility frequently exhibit a degradation in performance when applied to different scanners, patient populations, or clinical settings—as documented across multiple domains including the Japanese literature on AI detection of aneurysms (*38*). Prospective, multi-site evaluation studies are the methodological standard required to support regulatory authorization and health technology assessment (*12*).

International reporting frameworks—CONSORT-AI and SPIRIT-AI for clinical trials (*11,77*), TRIPOD+AI for prediction models (*12*), and STARD-AI for diagnostic accuracy studies (*13*) —provide the infrastructure for consistent, auditable evidence generation. Consistent application will raise the evidentiary floor for regulatory submissions and health technology assessments, though journal enforcement remains uneven.

8.2. Model updates, distribution shift, and post-marketing surveillance

AI systems deployed in clinical practice are subject to a distributional shift: changes in the patient population, clinical workflow, imaging equipment, or disease prevalence alter the input distribution relative to training data, gradually or abruptly diminishing model performance. AI to diagnose infectious diseases exemplifies this vulnerability through seasonal drift: the nodoca system to diagnose influenza requires re-evaluation as circulating strains and vaccine components change annually. A temporal drift in radiology, driven by scanner upgrades, acquisition protocol changes, or evolving diagnostic criteria, poses analogous challenges in that domain.

The United States FDA has addressed this challenge through the PCCP framework, which allows manufacturers to specify in advance the types of model updates—retraining on expanded data, threshold adjustments, or feature additions—that may be implemented without requiring a new regulatory submission, provided that the changes fall within pre-agreed performance bounds (*5*). In the European Union, the AI Act creates a "double compliance" burden for medical AI manufacturers—simultaneous MDR/IVDR conformity assessment alongside high-risk transparency and monitoring obligations pursuant to the AI Act (*6,29*).

In Japan, the regulatory review process for program medical devices that utilize machine learning has explicitly identified the management of continuous or post-approval performance changes as a key area requiring further policy development (*8,32*). Articulating a domestic framework for change management that is compatible with PCCP principles and enables seamless post-marketing surveillance through Japan's single-payer claims data infrastructure represents a priority policy task for the PMDA and the Ministry of Health, Labour, and Welfare (*78*). Accountability for adverse outcomes attributable to AI-assisted decisions also yet to be resolved: Japan's program medical device framework defines pre-marketing performance requirements but does not yet specify post-marketing liability norms for AI-assisted clinical decisions.

8.3. Reimbursement design and healthcare system integration

Regulatory authorization is a necessary but not sufficient condition for clinical AI adoption: the design of the reimbursement mechanism—whether AI-assisted procedures incur an additional fee, are assigned a separate billing code, or are subsumed into existing diagnostic fees—directly determines the economic incentive for healthcare providers to adopt and maintain AI systems. Japan's C1/C2 reimbursement classification system for new medical devices and technologies, exemplified by nodoca's C2 classification as a new technology fee, provides a formal pathway for health technology assessment of AI-based clinical decision-making support tools (*57*). However, the criteria for classification decisions, the evidence requirements for reclassification as standard care under insurance, and the mechanisms for incorporating evidence of real-world effectiveness into reimbursement still need to be fleshed out further for AI-specific applications.

Beyond reimbursement classification, healthcare system integration presents practical challenges that Japan has yet to systematically address. Alert fatigue from high false-positive rates and the workforce implications of AI integration represent cross-cutting policy challenges that the existing C1/C2 classification framework does not yet address (*1*).

8.4. Large language models and generative AI in clinical practice

Large language models (LLMs) and generative AI systems present a capability and risk profile distinct from the task-specific diagnostic models reviewed in the preceding sections. Benchmark evaluations have demonstrated that general-purpose LLMs—including GPT-4, which exceeded the passing threshold for the United States Medical Licensing Examination (USMLE) by over 20 points without domain-specific fine-tuning—possess substantial medical knowledge and reasoning ability (*79*). These results have prompted clinical interest in LLM applications including structured drafting of radiology reports and assisting with a differential diagnosis in electronic health records (*80*).

A 2025 systematic review of 15 studies that evaluated LLM-generated radiology reports found that while automated metrics and assessments by radiologists generally indicated acceptable linguistic quality, evidence from prospective implementation studies indicated that there were limited measurable improvements in clinical workflow efficiency or diagnostic accuracy (*81*). The principal risk specific to generative AI in clinical documentation is hallucination—the generation of plausible but factually incorrect content—which in clinical contexts can introduce erroneous findings, incorrect drug names, or fabricated examination results into a medical record in ways that are difficult to detect without systematic quality assurance processes. Unlike task-specific diagnostic AI, which typically produces a structured output (a probability score or bounding box) amenable to threshold-based review, LLM outputs are free-form text that have subtle and context-dependent errors.

A further challenge is regulatory classification. Many clinical LLM applications—and particularly those supporting documentation, summarization, and administrative workflow—do not meet the SaMD definition, as their primary function is not the direct generation of diagnostic or treatment outputs, placing them outside the PMDA program medical device framework in Japan and analogous regulatory perimeters in other jurisdictions. This regulatory gap means that LLM-based clinical tools can be deployed without pre-marketing clinical evaluation, post-marketing surveillance requirements, or change management obligations that apply to approved AI medical devices, creating asymmetric oversight that may underestimate population-level risk as adoption scales.

In Japan, the additional challenge of linguistic specificity is acute. Clinical documentation in Japanese involves a mix of kanji, kana, and medical abbreviations that are tokenized and semantically represented in models trained predominantly on English-language corpora, leading to substantial degradation relative to their native-language performance. Recent work has demonstrated that fine-tuned LLMs can achieve cross-hospital generalizability at recognition of diseases in clinical notes in Japanese, but deployment at scale across the heterogeneous documentation practices of Japan's hospital information systems remains an active research challenge (*82*). Developing and validating Japanese-language medical LLMs on pseudonymized data pools enabled by the Next-generation Medical Infrastructure Act represents a natural extension of the data infrastructure agenda discussed in Section 7. Progress in this area could have direct implications for clinical productivity and documentation quality across Japan's hospital system.

9. Conclusion

The evidence base across the five clinical domains examined in this review differs markedly. AI to assist diagnostic radiology has received regulatory approval at scale—accounting for 24 of Japan's 51 AI-designated program medical devices—but real-world deployment has revealed systematic gaps between controlled validation performance and operational accuracy, compounded in Japan by fragmented institutional data ownership that constrains the multi-institutional external validation needed to close that gap. AI to assist endoscopy has accumulated the most robust clinical trial evidence, driven in part by Japan's dual burden of colorectal and gastric malignancy that has produced 14 PMDA-approved devices and prompted Asian initiatives to conduct multicenter validation to expand the generalizability of existing RCT evidence—derived from predominantly East Asian and European populations—across the broader Asia-Pacific region. AI to assist cardiology has demonstrated the potential for ECG analysis and consumer wearables to facilitate population-level screening for arrhythmia while also exposing care pathway and reimbursement challenges that remain largely unaddressed in Japan despite a strong demographic demand from a population with the highest proportion of adults age 65 or older among OECD member countries. The infectious disease domain has produced Japan's most complete regulatory-approval-to-reimbursement pathway for a clinical AI tool, as exemplified by the nodoca system for diagnosis of influenza; it offers a reference for subsequent submissions involving AI-based clinical decision-making support. Comparative analysis of the AI data infrastructure has revealed that Japan's substantial clinical registry base—spanning cardiovascular procedures, surgical outcomes, and pharmacovigilance—is matched by few OECD countries at the procedure level, and yet a structural gap in accessible medical imaging data and multimodal linkage leaves Japan behind the data accessibility that programs such as TCIA and the EHDS framework afford international researchers. LLMs and generative AI systems, meanwhile, have introduced a category of clinical tool that largely falls outside existing SaMD regulatory frameworks, creating oversight asymmetries that healthcare systems in Japan and abroad

are only beginning to address.

Japan has a distinct position across all three comparative dimensions. In terms of its regulatory infrastructure, Japan's early legislative recognition of SaMD, the lists of approved program medical devices published by PMDA, and the complete regulatory-approval-to-reimbursement pathway demonstrated by nodoca represent a framework that is more operationally developed than that in many comparator countries. In clinical terms of the deployment infrastructure, Japan's density of diagnostic imaging systems and high endoscopy volume—reflected in the concentration of AI approvals within the imaging and endoscopy categories—provide a high-volume deployment base that few healthcare systems can match. In terms of the data infrastructure, however, a structural asymmetry persists: Japan's substantial clinical registries clash with limited data accessibility for external researchers and a medical imaging database in its early stages, in contrast to the open access of the TCIA or the cross-border federated governance of the EHDS. Sustained progress needs to be made across all three dimensions examined in this review—strengthening regulatory frameworks for post-approval model change management and clinical accountability, expanding the accessibility and scale of the AI-ready data infrastructure, and embedding validated AI tools into reimbursed clinical workflows—in order to realize the potential of clinical AI as a durable contributor to healthcare quality, efficiency, and equity in Japan and abroad.

Funding: This work was supported by JSPS KAKENHI Grant Number JP26K21655.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Organisation for Economic Co-operation and Development. Artificial Intelligence and the health workforce. *https://www.oecd.org/en/publications/artificial-intelligence-and-the-health-workforce_9a31d8af-en.html* (accessed April 13, 2026).
2. Liu X, Faes L, Kale AU, *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit Health. 2019; 1:e271-e297.
3. World Health Organization. Regulatory considerations on artificial intelligence for health. *https://www.who.int/publications/i/item/9789240078871* (accessed April 13, 2026).
4. International Medical Device Regulators Forum. Software as a medical device: Possible framework for risk categorization and corresponding considerations. *https://www.imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations* (accessed April 13, 2026).
5. U.S. Food and Drug Administration. 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* (accessed April 13, 2026).
6. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. *https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng* (accessed April 13, 2026).
7. Medicines and Healthcare products Regulatory Agency. Software and artificial intelligence (AI) as a medical device. *https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device* (accessed April 13, 2026).
8. Pharmaceuticals and Medical Devices Agency. Regarding the review of medical devices utilizing machine learning. *https://www.pmda.go.jp/files/000265866.pdf* (accessed April 13, 2026). (in Japanese)
9. National Medical Products Administration. NMPA Announcement on guidance for the classification defining of AI-based medical software products. *https://english.nmpa.gov.cn/2021-07/08/c_660267.htm* (accessed April 13, 2026).
10. Health Sciences Authority. Regulatory guidelines for software medical devices including machine learning-enabled medical devices – A Life Cycle Approach. *https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/gl-04-r4-regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach-(2025-dec)-pub.pdf* (accessed April 13, 2026).
11. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. BMJ. 2020; 370:m3164.
12. Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024; 385:e078378.
13. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Golub RM, Denniston AK, Shetty S, Moher D, Bossuyt PM, Darzi A, Ashrafian H, STARD-AI Steering Committee. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025; 31:3283-3289.
14. Karako K, Gao J. Artificial intelligence (AI)-assisted ultrasound in clinical trials: Endpoint automation, decentralized monitoring, and regulatory readiness. Drug Discov Ther. 2026; 20:91-103.
15. Pharmaceuticals and Medical Devices Agency. List of approved program medical devices. *https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/devices/0052.html* (accessed April 13, 2026). (in Japanese)
16. U.S. Food and Drug Administration. Artificial intelligence-Enabled Medical Devices. *https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices* (accessed April 13, 2026).
17. Organisation for Economic Co-operation and Development. Health at a Glance 2025: OECD indicators. *https://www.oecd.org/en/publications/health-at-a-glance-2025_8f9e3f98-en.html* (accessed April 13, 2026).
18. Organisation for Economic Co-operation and

Development. Health at a Glance 2025: Japan. *https://www.oecd.org/en/publications/health-at-a-glance-2025_15a55280-en/japan_319bfc39-en.html* (accessed April 13, 2026).
19. Karako K. Integration of wearable devices and deep learning: New possibilities for health management and disease prevention. Biosci Trends. 2024; 18:201-205.
20. McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. Nature. 2020; 577:89-94.
21. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, Lei L, Li L, Guo Z, Lei S, Xiong F, Wang H, Song Y, Pan Y, Zhou G. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): A double-blind randomised study. Lancet Gastroenterol Hepatol. 2020; 5:343-351.
22. Repici A, Badalamenti M, Maselli R, *et al*. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020; 159:512-520.
23. Attia Z, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. Lancet. 2019; 394:861-867.
24. Perez MV, Mahaffey KW, Hedlin H, *et al*. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med. 2019; 381:1909-1917.
25. Okiyama S, Fukuda M, Sode M, Takahashi W, Ikeda M, Kato H, Tsugawa Y, Iwagami M. Examining the use of an artificial intelligence model to diagnose influenza: Development and validation study. J Med Internet Res. 2022; 24:e38751.
26. International Medical Device Regulators Forum. Software as a Medical Device (SaMD): Clinical evaluation. *https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation* (accessed April 13, 2026).
27. U.S. Food and Drug Administration and Health Canada and Medicines and Healthcare Products Regulatory Agency. 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* (accessed April 13, 2026).
28. International Medical Device Regulators Forum. Good machine learning practice for medical device development: Guiding principles (IMDRF/AIML WG/N88 FINAL:2025). *https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles* (accessed April 13, 2026).
29. Medical Device Coordination Group. MDCG 2025-6 Interplay between the Medical Devices Regulation (MDR) & In vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). *https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en?filename=mdcg_2025-6_en.pdf* (accessed April 13, 2026).
30. National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies. *https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies* (accessed April 13, 2026).
31. Ministry of Food and Drug Safety, Republic of Korea. Regulatory update on medical devices in the Republic of Korea. *https://www.imdrf.org/sites/default/files/2025-09/Korea%20%28MFDS%20MC%29.pdf* (accessed April 13, 2026).
32. Ministry of Health, Labour, and Welfare, Japan. Responses to the regulatory treatment of medical devices using AI under the PMD Act. *https://www.mhlw.go.jp/content/10601000/000361102.pdf* (accessed April 13, 2026). (in Japanese)
33. Ministry of Health, Labour, and Welfare, Japan. Approval status of AI-enabled medical device programs. p.12. *https://www.mhlw.go.jp/content/11124500/001329696.pdf* (accessed April 13, 2026). (in Japanese)
34. Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, Chartrand G, Kadoury S. Deep learning: An update for radiologists. Radiographics. 2021; 41:1427-1445.
35. Martinez-Gutierrez JC, Kim Y, Salazar-Marioni S, *et al*. Automated large vessel occlusion detection software and thrombectomy treatment times: A cluster randomized clinical trial. JAMA Neurol. 2023; 80:1182-1190.
36. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. Radiol Artif Intell. 2022; 4:e210064.
37. Ishihara M, Shiiba M, Maruno H, Kato M, Ohmoto-Sekine Y, Antoine C, Ouchi Y. Detection of intracranial aneurysms using deep learning-based CAD system: Usefulness of the scores of CNN's final layer for distinguishing between aneurysm and infundibular dilatation. Jpn J Radiol. 2023; 41:131-141.
38. Ito R, Asai R, Nakamichi R, Nakane T, Taoka T, Naganawa S. Evaluation of approved AI-based brain aneurysm detection software in clinical practice: Comparison with radiologist assessment and image re-review. Magn Reson Med Sci. 2026; 25.
39. Makar J, Abdelmalak J, Con D, Hafeez B, Garg M. Use of artificial intelligence improves colonoscopy performance in adenoma detection: A systematic review and meta-analysis. Gastrointest Endosc. 2025; 101:68-81.
40. Misawa M, Kudo S, Mori Y. Implementation of artificial intelligence in colonoscopy practice in Japan. JMA J. 2025; 8:60-63.
41. Houwen BBSL, Hazewinkel Y, Giotis I, Vleugels JLA, Mostafavi NS, van Putten P, Fockens P, Dekker E. Computer-aided diagnosis for optical diagnosis of diminutive colorectal polyps including sessile serrated lesions: A real-time comparison with screening endoscopists. Endoscopy. 2023; 55:756-765.
42. World Cancer Research Fund. Colorectal cancer statistics. *https://www.wcrf.org/cancer-trends/colorectal-cancer-statistics/* (accessed April 13, 2026).
43. World Cancer Research Fund. Stomach cancer statistics. *https://www.wcrf.org/cancer-trends/stomach-cancer-statistics/* (accessed April 13, 2026).
44. Kudo SE, Misawa M, Mori Y, *et al*. Artificial intelligence-assisted system improves endoscopic identification of colorectal neoplasms. Clin Gastroenterol Hepatol. 2020; 18:1874-1881.
45. University Hospital Medical Information Network Clinical Trials Registry. Clinical trial registration: EndoBrain multicenter study (UMIN000047839). *https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000055293* (accessed April 13, 2026). (in

Japanese)
46. National Cancer Center Japan. An Asian multicenter clinical trial evaluating efficacy of computer-aided detection for colonoscopy in colorectal cancer screening. *https://www.ncc.go.jp/en/information/press_release/2024/0111/index.html* (accessed April 13, 2026).
47. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019; 25:70-74.
48. Yagi R, Goto S, Katsumata Y, MacRae CA, Deo RC. Importance of external validation and subgroup analysis of artificial intelligence in the detection of low ejection fraction from electrocardiograms. Eur Heart J Digit Health. 2022; 3:654-657.
49. Masumura M, Ohno A, Yoshinaga H, Sasaki T, Yamauchi Y, Hachiya H, Takahashi A, Imai Y, Fujita H, Ihara K, Ebana Y, Tanaka T, Furukawa T, Sasano T. AI-ECG for early detection of atrial fibrillation: First-year results from a stroke prevention study in Shimizu, Japan. J Arrhythm. 2025; 41:e70132.
50. Japanese Circulation Society. JROAD (Japan Registry of All Cardiac and Vascular Diseases) 2024 Annual Report. *https://www.j-circ.or.jp/jittai_chosa/about/report/* (accessed April 13, 2026).
51. Japanese Association for Cardiovascular Intervention and Therapeutics. J-PCI Registry 2024 Annual Report. *https://www.cvit.jp/docs/registry/annual-report/j-pci/2024.pdf* (accessed April 13, 2026).
52. Lubitz SA, Faranesh AZ, Selvaggi C, Atlas SJ, McManus DD, Singer DE, Pagoto S, McConnell MV, Pantelopoulos A, Foulkes AS. Detection of atrial fibrillation in a large population using wearable devices: The Fitbit Heart Study. Circulation. 2022; 146:1415-1424.
53. Simonson JK, Anderson M, Polacek C, Klump E, Haque SN. Characterizing real-world implementation of consumer wearables for the detection of undiagnosed atrial fibrillation in clinical practice: Targeted literature review. JMIR Cardio. 2023; 7:e47292.
54. Nomura A. Digital health, digital medicine, and digital therapeutics in cardiology: Current evidence and future perspective in Japan. Hypertens Res. 2023; 46:2126-2134.
55. Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021; 3:199-217.
56. Pharmaceuticals and Medical Devices Agency. Review report: Nodoca (software medical device to aid in the differential diagnosis of influenza). *https://www.pmda.go.jp/files/000248783.pdf* (accessed April 13, 2026). (in Japanese)
57. Central Social Insurance Medical Council. Health insurance coverage for medical devices (December 2022): Nodoca. *https://www.mhlw.go.jp/content/12404000/000989561.pdf* (accessed April 13, 2026). (in Japanese)
58. Han ZL, Zhang YY, Li J, Gao S, Liu W, Yang WJ, Xing ZH. A systematic review and meta-analysis of artificial intelligence software for tuberculosis diagnosis using chest X-ray imaging. J Thorac Dis. 2025; 17:3223-3237.
59. Rudd KE, Johnson SC, Agesa KM, *et al*. Global, regional, and national sepsis incidence and mortality, 1990-2017: Analysis for the Global Burden of Disease Study 2017. Lancet. 2020; 395:200-211.
60. World Health Organization. Sepsis. *https://www.who.int/news-room/fact-sheets/detail/sepsis* (accessed April 13, 2026).
61. Bhargava A, López-Espina C, Schmalz L, *et al*. FDA-authorized AI/ML tool for sepsis prediction: Development and validation. NEJM AI. 2024; 1:AIoa2400867.
62. National Cancer Institute. The Cancer Genome Atlas Program (TCGA). *https://www.cancer.gov/ccg/research/genome-sequencing/tcga* (accessed April 13, 2026).
63. Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, Bennett W, Nolan T, Freymann J. The public cancer radiology imaging collections of The Cancer Imaging Archive. Sci Data. 2017; 4:170124.
64. National Institutes of Health. BRIDGE2AI – Propelling biomedical research with artificial intelligence. *https://bridge2ai.org/* (accessed April 13, 2026).
65. NIH All of Us Research Program. The "All of Us" research program. *https://allofus.nih.gov/* (accessed April 13, 2026).
66. European Commission. European Health Data Space Regulation (EHDS). *https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en* (accessed April 13, 2026).
67. Cresswell K, Williams R, Dungey S, Anderson S, Bernabeu MO, Mozaffar H, Yang X, Sai V, Bea S, Eason S. A mixed methods formative evaluation of the United Kingdom National Health Service Artificial Intelligence Lab. npj Digit Med. 2025; 8:448.
68. National Information Society Agency. AI Hub: Integrated AI training data platform. *https://www.aihub.or.kr/* (accessed April 13, 2026).
69. Suto M, Iba A, Sugiyama T, Kodama T, Takegami M, Taguchi R, Niino M, Koizumi R, Kashiwagi K, Imai K, Ihana-Sugiyama N, Ichinose Y, Takehara K, Iso H. Literature review of studies using the National Database of the Health Insurance Claims of Japan (NDB): Limitations and strategies in using the NDB for research. JMA J. 2024; 7:10-20.
70. Pharmaceuticals and Medical Devices Agency. Understanding MID-NET®: Its features and efforts to promote its use in academia. *https://www.pmda.go.jp/files/000267996.pdf* (accessed April 13, 2026). (in Japanese)
71. Yamamoto T, Takahashi A, Yoshizumi T, *et al*. 2023 National clinical database annual report by the Japan surgical society. Surgery Today. 2025; 55:295-334.
72. Ueda D, Walston S, Takita H, Mitsuyama Y, Miki Y. The critical need for an open medical imaging database in Japan: Implications for global health and AI development. Jpn J Radiol. 2025; 43:537-541.
73. Akashi T, Kumamaru KK, Wada A, Hashimoto M, Hirata K, Hayakawa Y, Sano K, Kamagata K, Hagiwara A, Ikenouchi Y, Aoki S. Japan-Medical Image Database (J-MID): Medical big data supporting data science. Juntendo Med J. 2025; 71:166-172.
74. Ministry of Health, Labour, and Welfare, Japan. Regarding the amended Next-generation Medical Infrastructure Act. *https://www.mhlw.go.jp/content/10808000/001166476.pdf* (accessed April 13, 2026). (in Japanese)
75. Kobayashi S, Kimura M, Kodama Y, Takada A, Nagashima S, Kawazoe Y, Ohe K. Development of Japan-specific HL7 FHIR medication-related profiles. J Med Syst. 2025; 49:46.
76. Teo ZL, Jin L, Li S, Miao D, Zhang X, Ng WY, Tan TF, Lee DM, Chua KJ, Heng J, Liu Y, Goh RSM, Ting DSW.

Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. Cell Rep Med. 2024; 5:101419.
77. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. BMJ. 2020; 370:m3210.
78. Ministry of Health, Labour, and Welfare, Japan. Plan to promote investment in reduced Labour: Medical care. *https://www.mhlw.go.jp/content/10801000/001565769.pdf* (accessed April 13, 2026). (in Japanese)
79. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023).
80. Karako K, Song P, Chen Y, Tang W. New possibilities for medical support systems utilizing artificial intelligence (AI) and data platforms. Biosci Trends. 2023 Jul 11; 17(3):186-189.
81. Artsi Y, Klang E, Collins JD, Glicksberg BS, Nadkarni GN, Korfiatis P, Sorin V. Large language models in radiology reporting — A systematic review of performance, limitations, and clinical implications. Intell Based Med. 2025; 12:100287.
82. Shimizu S, Nishiyama T, Nagai H, Wakamiya S, Aramaki E. Toward cross-hospital deployment of natural language processing systems: Model development and validation of fine-tuned large language models for disease name recognition in Japanese. JMIR Med Inform. 2025; 13:e76773.

Received March 27, 2026; Revised May 4, 2026; Accepted May 9, 2026.

Released online in J-STAGE as advance publication May 13, 2026.

**Address correspondence to*:
Kenji Karako, Department of Surgery, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8655, Japan.
E-mail: tri.leafs@gmail.com

Global Health & Medicine. 2026; 8(3):182-192. REVIEW
DOI: 10.35772/ghm.2026.01051

Artificial intelligence (AI)-aided clinical data management: Applications, human-in-the-loop workflows, and regulatory considerations

Saya Ohi[1], Tomoko Iwamoto[2], Daiki Ikeda[1], Yuichi Kawanishi[1], Koji Kitajima[2], Hajime Ohyanagi[3,*]

[1] Office of Bioinformatics, Department of Joint Center for Researchers, Associates and Clinicians (JCRAC), Center for Clinical Sciences, Japan Institute for Health Security, Tokyo, Japan;
[2] Office of Clinical Data Management, Department of Joint Center for Researchers, Associates and Clinicians (JCRAC), Center for Clinical Sciences, Japan Institute for Health Security, Tokyo, Japan;
[3] Department of Joint Center for Researchers, Associates and Clinicians (JCRAC), Center for Clinical Sciences, Japan Institute for Health Security, Tokyo, Japan.

Abstract: Clinical data management (CDM) is central to the quality of clinical research. In Japan, CDM faces a shortage of qualified personnel, particularly in academic research organizations (AROs), as well as increasing data volume and complexity. Rapid advances in artificial intelligence (AI), especially large language models, have therefore attracted attention as a way to support CDM. This review summarizes domestic and international examples of AI utilization in CDM-related tasks, including data cleaning, medical coding, and query generation. Across the cases reviewed, a common implementation principle emerged: a human-in-the-loop design in which AI performs initial processing or detection, while final judgment remains with human personnel. This design is especially relevant to AROs, where high data quality must be maintained with limited CDM human resources. Regulatory frameworks, including ICH E6 (R3) and the FDA-EMA Guiding Principles, are beginning to address AI use, but how AI-aided processes should be handled under Good Clinical Practice remains under discussion. Comprehensive risk mitigation is therefore essential. AI and data are interdependent: better data improve AI performance, and better AI can further improve data quality. The shift from manual processes to human-AI collaborative workflows is likely to accelerate, and CDM must develop the technical, regulatory, and risk-management frameworks needed to support that transition.

***Keywords*:** artificial intelligence (AI), clinical data management (CDM), human-in-the-loop (HITL), AI-related regulations and guidelines

1. Introduction

Clinical research is human-subjects research that evaluates health, disease, and medical interventions to generate evidence for patient care, regulatory decision-making, and medical practice (*1*). Clinical data management (CDM) underpins the reliability of clinical research data by supporting the design, collection, quality control, and preparation of data for analysis (*2*). Its responsibilities extend from the study planning stage to the finalization of analysis-ready datasets. The International Council for Harmonisation (ICH) E6 (R3) Good Clinical Practice guideline, an internationally harmonized standard for clinical trials, identifies data integrity, traceability, confidentiality, reliability, and fitness for purpose as core principles of data governance (*3*) (Figure 1, bottom line). Because errors introduced at the data level cannot be corrected by subsequent statistical analysis or interpretation, CDM is a foundational function for ensuring the quality of clinical research.

The environment surrounding CDM has changed substantially in recent years. Clinical trials now use diverse data sources beyond conventional case report forms (CRFs), including electronic health records (EHRs), wearable devices, patient-reported outcomes (PROs), and medical images. As a result, data volume has increased, and data structures have become more complex (*4*). These changes have increased the workload of CDM and have made it increasingly necessary to maintain data quality efficiently. At the same time, qualified data managers remain in short supply (*5*),

particularly in organizations without dedicated CDM departments. In this context, rapid advances in artificial intelligence (AI), especially large language models, have attracted attention as a potential means of supporting CDM tasks such as data cleaning, medical coding, and query generation (*6*).

This review summarizes examples of AI use in key CDM-related tasks, including data cleaning, medical coding, and query generation, and discusses the regulatory requirements and risk-management considerations necessary for the responsible implementation of AI in clinical research. A common feature of the cases reviewed is that AI supports initial processing, anomaly detection, coding, or text generation, whereas final judgment remains with qualified human personnel. This operational model is known as human-in-the-loop (HITL) and is based on collaboration between humans and AI rather than full automation (*7*). The HITL concept is particularly important for clinical research organizations that must maintain high data quality with limited CDM resources (*8*). In this review, HITL is positioned as the central framework for discussing AI applications, relevant regulations, and risk management in CDM.

2. Background: CDM challenges, policy background, and AI utilization

2.1. Roles and workflow of CDM

High-quality data are essential for valid analysis and reliable conclusions in clinical research (*3,9*). Even well-designed studies cannot generate robust evidence unless protocol-specified data are collected accurately and completely (*3,10*). Data managers are involved from the planning stage to ensure that primary endpoint data are captured appropriately (Figure 1, Clarification of Endpoints, Protocol Review). Building an electronic data capture (EDC) system is one of their main responsibilities (Figure 1, EDC Construction). Because EDC is only the data-entry infrastructure, data checks, reporting functions, and entry-support must also be implemented to enable accurate and timely data collection (*3*). Accuracy in CDM goes beyond confirming consistency between entered data and source documents; it also requires confirming that protocol-required data have been collected and are available in an analysis-ready format (*3,11*) (Figure 1, Creating a Collection Plan). For example, primary endpoint data should be collected without missing values, and visit-based data should be entered at time points specified in the protocol schedule (*3,10*). Data cleaning systematically refines collected data and is a core CDM activity for ensuring data quality (Figure 1, Data Cleaning). Appropriate recording and management of audit trails for data modifications and updates are also mandatory in clinical research (*3,12,13*) (Figure 1, Recording of Audit Trails).

Data managers support clinical research by building databases that faithfully capture protocol-defined endpoints and by maintaining data-entry quality. Their responsibilities span EDC construction, system operation, data cleaning, and data delivery, making them essential to research quality. In this article, "clinical research" refers broadly to clinical trials in general and is not limited to pharmaceutical trials.

2.2. Current challenges in CDM

The CDM field faces a serious shortage of qualified personnel, particularly in Japanese Academic Research

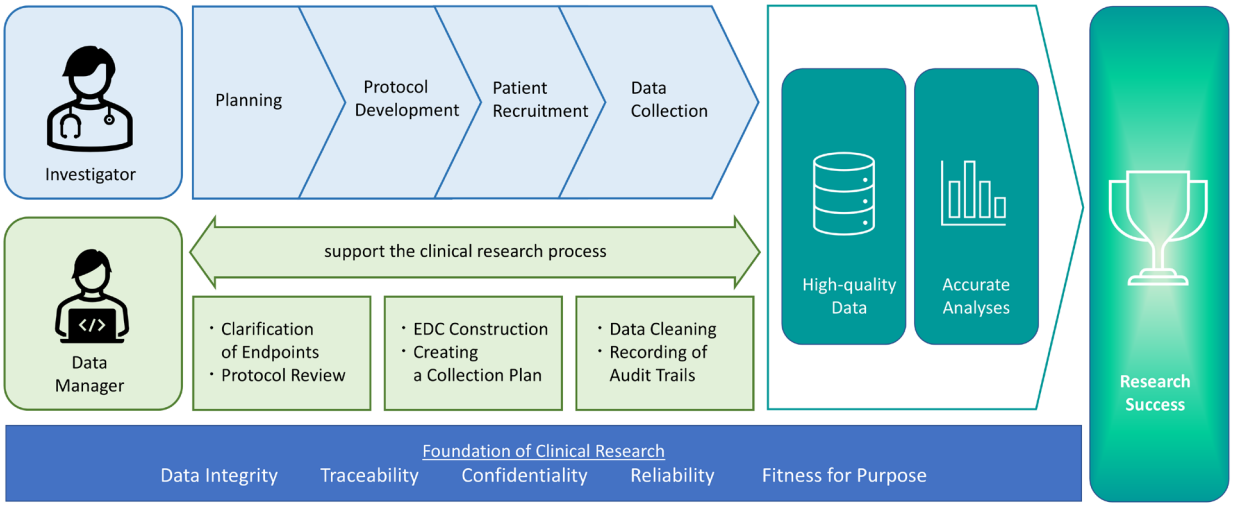


Figure 1. Clinical research process and data management. This simplified schematic highlights clinical research stages and data manager activities relevant to this review. Data managers support investigators across planning, protocol development, patient recruitment, and data collection phases, contributing to high-quality data and accurate analyses that underpin research success. The five principles at the bottom—data integrity, traceability, confidentiality, reliability, and fitness for purpose—represent the foundational CDM requirements specified in ICH E6 (R3).

Organizations (AROs). Unlike pharmaceutical companies, which often have dedicated CDM departments, many AROs rely on a small number of data managers to oversee multiple trials simultaneously. In a report by the All-Japan Conference of Deans of Medical Schools and Hospital Directors, only 16 of 80 universities (20%) reported having three or more full-time data managers (*14*). Including the seven institutions that were developing such capacity, 64 universities had insufficient CDM human resources. Thus, a limited workforce must manage multiple concurrent trials. At the same time, the volume and diversity of data sources are increasing (*15*). Data are becoming more diverse, complex, and voluminous, and decentralized clinical trials (DCTs) have widened the range of sources. Unstructured data, including medical images, are also expected to be incorporated more often (*16,17*). Maintaining data quality under these conditions will become increasingly challenging.

2.3. Policy background

Several policy developments have shaped the CDM environment in Japan. In Japanese clinical research, the Ethical Guidelines for Medical and Biological Research Involving Human Subjects (*18*) and the Enforcement Regulations on the Clinical Research Act (*19*) require research protocols to specify data collection and management methods; monitoring is required under the guidelines, when research involves invasiveness or intervention. These requirements make data quality assurance a planning-stage legal obligation. For pharmaceutical trials, the GCP Ministerial Ordinance establishes equivalent requirements (*20*). In parallel, Japan has adopted the Society 5.0 vision, and the Medical DX Promotion Headquarters, led by the Cabinet Office and the Ministry of Health, Labor and Welfare (MHLW) is accelerating healthcare digitalization (*21*). AI-aided data management is therefore increasingly viewed not as an experimental option, but as a practical necessity. Together with personnel shortages, these policy signals have encouraged active consideration of AI applications in CDM.

2.4. AI in CDM

AI, particularly large language models (LLMs), has advanced rapidly in recent years. Major AI and cloud providers are increasingly developing healthcare-oriented services based on medical data standards such as FHIR and DICOM (Supplementary Table S1, *https://www.globalhealthmedicine.com/site/supplementaldata.html?ID=123*). Among these advances, several features of AI and LLMs are especially relevant to clinical research data (*22*). First, they can process unstructured text, such as physician notes, pathology reports, and adverse event narratives, and extract structured information. This capability addresses an area in which conventional tools have struggled and may create new categories of research data (*23*). Second, LLMs can be adapted to specific domains through fine-tuning on specialized corpora or through careful prompt engineering, reducing the need for large-scale custom development. Third, AI can scale processing volume while improving output consistency. Manual data review is affected by reviewer experience, fatigue, and workload, which can reduce the thoroughness and consistency of checks. AI systems, including LLMs, avoid these sources of variation and can apply uniform criteria repeatedly to large datasets, making them useful complements to quality management (*6*).

Current AI systems also have important limitations and risks, discussed later in this article. The next section uses case examples to examine how AI capabilities can be applied to specific CDM-related tasks.

3. Case examples in AI-aided CDM

A defining feature of AI in professional applications is that non-specialists can operate it through natural language instructions. This has accelerated adoption across many professional domains (*24,25*). The following sections present examples of AI utilization in CDM-related tasks and adjacent data-processing approaches that may be transferable to CDM workflows. Peer-reviewed literature documenting AI applications in routine CDM practice remains limited; therefore, some examples are drawn from publicly accessible online sources. The sources should not be interpreted as official positions of the organizations concerned. The examples should therefore be interpreted according to their maturity: some are already used or evaluated in CDM-related tasks, whereas others illustrate approaches that may require further validation before routine CDM implementation.

3.1. Data cleaning

Among CDM tasks, data cleaning is quality-critical and rule-intensive, making it a strong target for automation (Figure 2A). A notable precedent is the work of Shi *et al.* on Belgian primary care EHR data (*26*). They built a clinical knowledge database (CKD) containing reference ranges, unit-conversion formulas, and outlier-detection criteria for each variable, and implemented an automated cleaning pipeline combining fuzzy matching and outlier detection (*26*). For more than one million records across 52 variables, the pipeline completed in 5.2 minutes a task that would have required 30 to 40 hours of manual work by an experienced statistician. Quality indicators improved for most variables, particularly the proportion of abnormal values caused by digit or unit errors (*26*). A key strength of this approach is that decisions are based on objective clinical knowledge rather than statistical distributions, enabling robust performance even in real-

world datasets with a high proportion of patients with disease.

Bönisch *et al*. proposed a machine learning (ML) approach for predictive data-quality assessment using data from medical data integration centers at German university hospitals (*27*). XGBoost and SVM were applied to echocardiography, laboratory, and medication data, and prediction results were stored as quality metadata in a data warehouse (*27*). Compared with conventional approaches that require item-by-item logical checks, this method may streamline large-scale quality checking while maintaining comparable rigor.

Although these studies were not conducted in routine clinical trial CDM settings, they illustrate data-quality assessment approaches that may be transferable to CDM workflows involving large-scale, heterogeneous clinical data.

We explored AI-aided data checking (Iwamoto *et al.*, The potential of data checking using AI and RPA. In: The 17th Annual Meeting of the Japan Society of Clinical Trials and Research. 2026). The workflow was designed to read raw CSV data downloaded from an EDC system and extract records that met predefined error conditions (Figure 2A). Generative AI produced check scripts from natural language descriptions of error conditions, and robotic process automation (RPA) executed the scripts to extract error data (Figure 2B). More than 94% of the anticipated errors embedded in test data were detected. The study also showed that prompt rules can improve reproducibility in data checking.

Together, these international precedents and the domestic initiative share a common implementation principle: AI performs the initial processing, while humans make the final judgment. This hybrid structure reflects a HITL (see Section 6.1) perspective rather than full automation. Streamlining data checking is central to CDM quality and efficiency, and AI may have a particularly large impact in AROs with lean CDM teams.

3.2. Medical coding

Evidence supporting AI-aided medical coding continues to accumulate. WHODrug Koda, developed by the Uppsala Monitoring Centre (UMC), combines text-processing algorithms, coding rules, and ML to convert free-text drug names in adverse event reports into standard WHODrug Global codes (*28*). In an evaluation of approximately 4.8 million drug entries in VigiBase, Koda increased the automatic coding rate from 61% to 89% compared with a simple direct matching while maintaining 97% coding accuracy (*28*). Its three-tier design automatically selects a code when confidence is high, presents candidate codes when ambiguity remains, and withholds coding when expert judgment is required, making it a rational HITL implementation. The AI-DM Task Force of the Japan Pharmaceutical Manufacturers Association (JPMA) has examined natural language processing-based automatic coding of disease names entered as free text in CRF fields, including medical history and adverse events, to ICD-10, as well as the

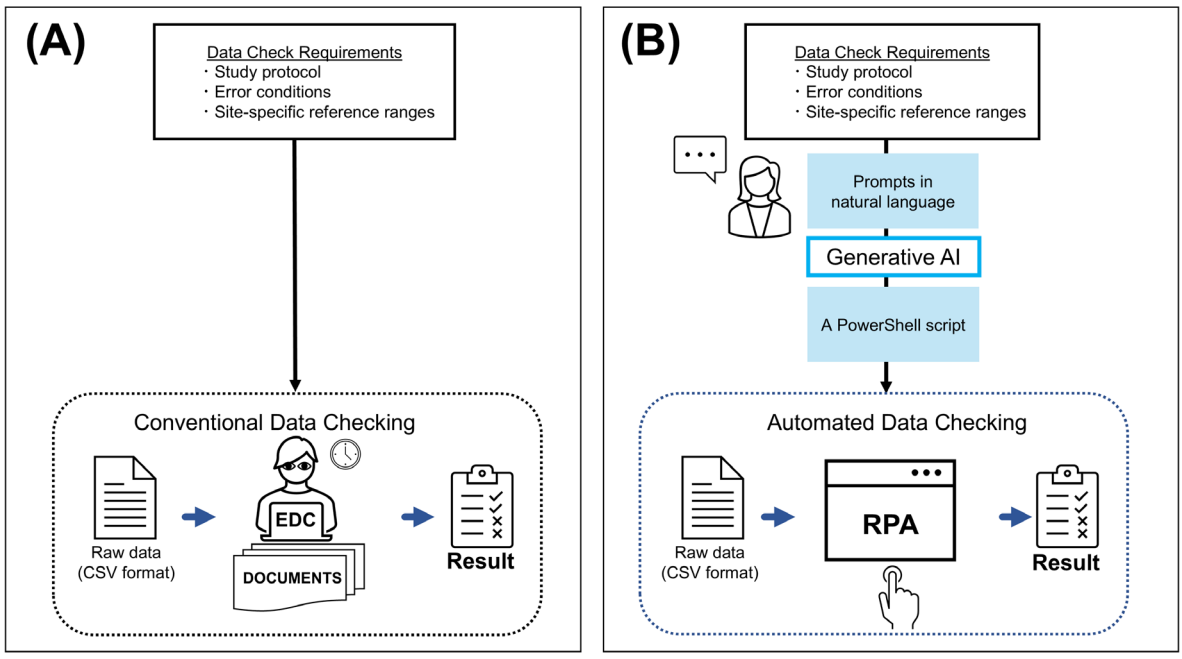


Figure 2. Conventional and AI-aided automated data-checking workflows. In the conventional approach **(A)**, data check requirements are manually translated into EDC-based review by a data manager. In the automated approach **(B)**, generative AI interprets the same requirements and generates executable scripts for robotic process automation (RPA), enabling automated error detection from raw CSV data.

human oversight needed for quality assurance (*29-31*). Similar initiatives have been reported for coding safety information, including adverse events, to MedDRA (*32*). Across drug and adverse event coding, the shared design principle is clear: AI processes high-confidence cases automatically, while human reviewers focus on cases requiring nuanced judgment. This approach helps concentrate limited specialist resources on high-value tasks.

3.3. Query generation

Query management is one of the most labor-intensive components of CDM, yet comprehensive query issuance appears to contribute only modestly to database quality. Stokman analyzed approximately 2 million queries from 20 Phase III trials at seven major pharmaceutical companies and found that the data-correction rate for conventional rule-based automated queries was approximately 1.4% (*33*). This suggests a shift from query quantity to query quality and prioritization, that is, toward risk-based query management. LLMs may offer a new way to address this challenge. In Japan, members of the JPMA task force evaluated two query-generation approaches: fine-tuning a base model and using an existing LLM without additional training. In both approaches, more than 80% of generated queries were judged suitable for direct operational use. The findings suggest that the choice of approach should depend on user-side requirements, including the availability of training data and the desired naturalness of query messages.

AI could support preferential detection of clinically meaningful inconsistencies and generation of corresponding queries by leveraging both structured clinical data and unstructured text, including query messages. Using AI to improve query quality and prioritization may help address a long-standing CDM challenge: maintaining data quality while reducing site burden from large query volumes. Further evidence is expected as real-world validation and quantitative effectiveness evaluations accumulate.

4. Regulations and guidelines for AI-aided CDM

AI use in CDM raises regulatory challenges that must be addressed alongside technical implementation. The current regulatory landscape is fragmented. Relevant documents include established clinical research regulations, AI-specific frameworks with varying degrees of legal force, and medical AI guidelines that often do not directly address clinical research. This section first characterizes the regulatory landscape for AI-aided CDM and then examines expectations across three issues most directly relevant to CDM practice: the integrity of computerized processes, human oversight and accountability, and the privacy and security of clinical research data. Supplementary Table S2 (*https://www.globalhealthmedicine.com/site/supplementaldata.html?ID=123*) summarizes each instrument's coverage of these issues.

4.1. The regulatory landscape

At the international level, three documents are particularly relevant, but they differ in legal force, scope, and direct applicability to CDM. ICH E6 (R3) (January 2025) provides the GCP framework for CDM and becomes binding in each jurisdiction when incorporated into domestic regulations (*3*). Through its standalone data-governance chapter and media-neutral coverage of computerized systems, it extends established CDM expectations to AI-aided workflows. However, it does not specifically target AI. The FDA-EMA Guiding Principles of Good AI Practice in Drug Development (January 2026) provide the most direct regulatory discussion of AI in drug development, including human-AI interactions and traceable documentation, but they are non-binding (*34*). The EU AI Act (Regulation (EU) 2024/1689, in force August 2024) is a legally binding AI-specific framework and classifies AI systems used as medical devices as high-risk, with potential extraterritorial reach where outputs are used in the EU; however, it is not specific to CDM, and AI systems developed solely for research and development are explicitly exempt (*35*). Thus, no single international instrument provides binding requirements and directly targets AI use in CDM.

In Japan, alignment between ICH E6 (R3) and domestic ministerial ordinances under the Pharmaceuticals and Medical Devices Act is underway, but current domestic implementation extends only to E6 (R2) (*36*). In contrast, AI-specific domestic instruments remain voluntary. The MHLW Guidelines for the Utilization of Digital Data in AI Research and Development focus on personal information protection but explicitly exclude clinical research data (*37*). The AI Business Guidelines issued by the Ministry of Economy, Trade and Industry and the Ministry of Internal Affairs and Communications (Version 1.1, METI/MIC, 2025) establish four cross-sector pillars: safety, fairness, transparency, and privacy (*38*). Japan AI Safety Institute (J-AISI/IPA) provides cross-sector references for AI evaluation and data-quality management (*39,40*). The Healthcare AI Platform Collaborative Innovation Partnership (HAIP-CIP) Guidelines for the Use of Generative AI in the Medical and Healthcare Fields (2nd Edition) are the most practically relevant domestic reference for CDM, providing detailed risk classifications across eight use cases, including research data processing (*41*). Thus, in Japan, binding requirements for AI-aided CDM currently come from clinical research regulations for computerized systems, whereas AI-specific domestic documents remain voluntary supplementary guidance and are not tailored to CDM.

4.2. Process integrity: validation, audit trails, and documentation

For AI-aided CDM, the central question is whether the data-management process remains verifiable, reproducible, and auditable when AI is used to support data review, data cleaning, or query generation. Because such outputs may influence how clinical research data are corrected, documented, and judged to be reliable for analysis, AI-aided workflows should be treated as part of the computerized processes subject to CDM quality requirements.

This interpretation is grounded primarily in ICH E6 (R3), which requires risk-proportionate validation and audit-trail maintenance for computerized systems regardless of technology, placing AI-aided data cleaning and query generation within the same quality expectations as conventional EDC tools (*3*). AI-specific frameworks reinforce this direction. The FDA-EMA Guiding Principles emphasize traceable documentation of data provenance, processing steps, and human-AI interactions, while the EU AI Act adds automated logging and lifecycle quality-management requirements for high-risk AI systems (*34,35*).

In Japan, J-AISI/IPA documents provide complementary perspectives on data quality and AI evaluation, although they are not specific to CDM (*39, 40*).

Several CDM-specific issues therefore remain unresolved. Current guidance does not clearly define how prompts, model versions, and model updates should be documented for reproducibility; whether AI-aided data cleaning qualifies as a validated process under existing GCP definitions; or how AI-generated review results should be retained in the Trial Master File (*42-44*). These gaps are important because AI outputs may influence data review, query generation, and final judgments about whether data are sufficiently reliable for analysis.

4.3. Human oversight and accountability

The second practical question is how human oversight should be designed when AI contributes to CDM decisions. AI does not remove accountability from qualified personnel; rather, it changes where human judgment should be placed within the workflow. The key issue is therefore not simply whether a human is present, but whether human review occurs at points where AI outputs may affect data quality, regulatory documentation, or final data interpretation.

Human oversight is consistently treated as a core safeguard across the relevant documents. ICH E6 (R3) requires risk-proportionate human oversight of computerized systems and retains ultimate decision-making with human personnel in clinical trials (*3*). The FDA-EMA Guiding Principles further position HITL as a lifecycle design principle, while the EU AI Act requires high-risk AI systems to allow humans to disregard outputs or intervene in operation (*34,35*).

Japanese guidance follows the same direction: the AI Business Guidelines emphasize a human-centered principle, the MHLW Guidelines for the Utilization of Digital Data in AI Research and Development require human oversight throughout the data lifecycle, and the HAIP-CIP Guidelines require final review by qualified personnel in relevant use cases (*37,38,41*).

Together, these instruments establish HITL as a regulatory expectation rather than an optional design preference. However, they do not specify how much human verification is sufficient for AI outputs to be considered authoritative under GCP (*43,44*). For AI-aided CDM, human review should therefore be defined as an accountable decision point, not as a superficial confirmation of AI-generated results.

4.4. Privacy and secure handling of clinical research data

The third practical question is how patient-level clinical research data can be protected when AI tools are introduced into CDM workflows. AI-aided CDM may involve external vendors, cloud-based processing environments, model improvement processes, or secondary use of submitted data. These features make it necessary to clarify what data are transferred, where they are processed, whether they are retained, and whether they can be used for model training.

Existing clinical research regulations already require protection of participant data and confidentiality, and AI-related frameworks extend this concern to lifecycle governance, cybersecurity, and control over data use. ICH E6 (R3) provides the GCP basis for patient data protection, while the FDA-EMA Guiding Principles and the EU AI Act broaden the focus to AI lifecycle governance and cybersecurity requirements for higher-risk systems (*3,34,35*). In Japan, privacy and secure handling are grounded in the Act on the Protection of Personal Information (*45*). The MHLW Digital Data Guidelines, the AI Business Guidelines, J-AISI/IPA documents, and the HAIP-CIP Guidelines add practical considerations such as privacy-by-design, safety, evaluation of privacy risks, and restrictions on vendor use of submitted data for retraining (*37-41*).

Thus, privacy in AI-aided CDM should not be limited to anonymization alone. It should also include vendor governance, data retention policies, model-training restrictions, and cybersecurity controls.

5. Risks associated with AI utilization

5.1. Risks associated with AI output

Beyond regulatory compliance, four AI-related risks require particular attention: hallucination, bias, privacy and security, and explainability and transparency. These

risks are relevant across domains, but their implications are especially important in clinical research because data accuracy is directly linked to patient safety and regulatory decision-making.

Hallucination, or generation of plausible but factually incorrect outputs, may be the most consequential risk for CDM (*43,46*). Such errors are difficult to detect precisely because they appear credible. In CDM, AI may generate structured data from clinical narratives that appear valid but are wrong, or may misinterpret data relationships when generating queries. These errors differ from random transcription mistakes because they can evade conventional quality-control processes (*47*).

Bias is also a major concern. LLMs trained predominantly on data from specific demographic groups, clinical settings, or disease areas may perform worse when applied to underrepresented populations or rare diseases (*43,48*). These are precisely the settings in which data quality is most critical and often most difficult to ensure. Clinical research involving underrepresented populations or rare diseases therefore requires particularly careful evaluation for AI bias.

Privacy and security add further complexity. LLMs can memorize and reproduce personally identifiable information in their training data, including names and contact details (*49*). Processing patient-level clinical data with AI therefore requires robust anonymization, secure computational environments, and clear policies on data retention and model training (*44,50*).

Explainability and transparency also pose important challenges, particularly for GCP audit-trail requirements and model reproducibility. LLMs may produce different outputs for identical prompts because of model-version updates or infrastructure changes. Standards have not yet been established for recording and preserving the rationale behind AI-generated judgments in audit trails, leaving a gap against fundamental GCP expectations (*51*). Rigorous documentation of prompts and model versions can partially mitigate this issue, but it does not fully solve it.

Addressing these challenges requires multilayered safeguards. Technical safeguards include confidence scoring and anomaly detection (*50*). Procedural safeguards include HITL verification at critical decision points (see Section 6.1) (*51*). Organizational safeguards include governance frameworks, training programs, and audit mechanisms (*52*). Validation protocols designed specifically for AI-aided CDM are an important target for future standardization.

5.2. Risks associated with AI infrastructure

The risks of introducing AI into CDM are not limited to model-intrinsic issues. Supply chain and geopolitical risks must also be considered in real-world operations.

External AI vendors may create continuity risks related to policy, contractual, and geopolitical factors. The 2024 cyberattack on Change Healthcare Inc. exposed the vulnerability of critical infrastructure configurations that depend heavily on a single vendor (*53*). In March 2026, Anthropic was designated as a supply chain risk by the United States government (*54*), illustrating that service continuity can be threatened by non-technical factors. Such service interruptions are a realistic risk for CDM operations that rely heavily on specific vendors.

In Japan, external-vendor risks are being discussed through the lens of data sovereignty in government cloud environments, the MHLW Guidelines for the Secure Management of Medical Information Systems (*55*), and the AI Business Guidelines issued by METI and MIC (*38*). Similar considerations apply to AI introduction in CDM. Institutions should decide from the design stage which operations can be entrusted to which vendors, what alternatives exist if service is interrupted, and how data can be retrieved and migrated.

For CDM, these risks are relevant because interruption, migration failure, or loss of access to AI-supported tools could affect data review timelines, auditability, and continuity of trial operations.

6. Discussion and future perspectives

6.1. Human-in-the-loop: Human-AI collaborative workflows

Most cases reviewed here share the same structural principle: AI performs primary processing and detection, while human personnel retain final judgment. This design, known as human-in-the-loop (HITL) (*7*), may serve as the central implementation framework for integrating AI into CDM.

HITL refers to architectures in which human judgment is embedded in cycles of AI proposal, detection, and generation. Depending on the degree of automation, human involvement can range from "human in the loop", where humans make individual decisions, to "human on the loop", where humans supervise AI (*56*). The appropriate degree of involvement depends on task risk and accuracy requirements (*56*). In CDM, tasks directly linked to patient safety, regulatory decision-making, and final data approval require greater human involvement, whereas routine primary checks can reasonably be delegated to AI. From a quality-assurance perspective, AI-generated confidence scores and supporting rationales can help responsible personnel make decisions. Such risk-based allocation of human oversight is especially practical in AROs, where high data quality must be maintained with limited CDM human resources.

The regulatory basis for HITL is increasingly clear: ICH E6 (R3) requires risk-proportionate human oversight for AI as a computerized system (*3*), and the EU AI Act mandates human oversight for high-risk AI systems (*35*). These provisions provide not only qualitative justification

for human involvement, but also a regulatory basis for incorporating HITL.

HITL design also has challenges. These include human bottlenecks as workload increases, rubber-stamping during verification, and cognitive deskilling caused by over-reliance on AI. Bottlenecks can be reduced through threshold-based designs that automatically process high-confidence cases and reserve human review for cases that genuinely require it (*57*). Rubber-stamping can be mitigated by explainability designs that present AI outputs with supporting rationale (*58*). Cognitive deskilling requires regular training so that CDM personnel maintain an accurate understanding of AI limitations (*59*).

6.2. Advanced data cleaning and quality control

In the near term, integration of EDC and AI may enable real-time quality checking. EDC products vary in their logical-check system, and human personnel currently translate check rules and conditions into system-specific syntax. AI support for logical-check construction could therefore reduce workload. In multisite and multinational trials, adaptive quality management is also emerging: AI can detect site-to-site differences in data quality and language interpretation in real time and feed those findings into monitoring strategies. As shown by the international precedents discussed above (*26,27*), hybrid approaches combining AI with existing databases are already approaching practical implementation, and early adoption in Japanese ARO settings is expected.

In the medium to long term, CDM will need to accommodate diverse data sources, including unstructured data such as images, audio recordings, and free-text narratives. As DCTs spread and wearable devices and electronic patient diaries become new data sources, AI will play a larger role in data-source integration and quality assessment. Automatic generation of check rules is another promising direction. Today, protocol specifications are largely converted into check rules by hand; a workflow in which LLMs interpret protocols and generate draft check rules for CDM personnel to review and refine could reduce trial start-up workload. "Predictive CDM", in which models trained on historical trial data and query records proactively detect potential data issues in new trials, is also attracting interest. Temporal data variation is an important future target for AI. Current checks mainly compare values with reference ranges, for example, whether a laboratory value is within the normal range. Rapid or substantial changes over short intervals can therefore be missed even when values remain within normal limits. Such context-dependent anomalies may signal an adverse event prodrome or measurement errors, but they are difficult to detect with static, rule-based checks. AI-aided time-series analysis could detect clinically meaningful patterns by comparing trends within subjects, within sites, and across the trial as a whole. This would complement the current review processes that rely on CDM personnel's knowledge and experience and could move the field toward quality management that detects meaningful signals earlier. If realized, AI would move beyond retrospective data verification to become a proactive tool for anticipating and managing quality from the trial design stage.

6.3. Interdependent development of AI and data

This article has examined the potential of AI in CDM, including data collection and data cleaning in clinical research. Because AI systems learn from data, improving data quality is essential for improving AI performance. If Japanese medical data are underrepresented in global AI training datasets, AI performance for Japanese patients may be diminished (*60*). Preparing large volumes of high-quality data currently depends on human effort. If AI improves the quality of large datasets, those improved data can then support more capable AI, which can again be applied to improve data quality. AI and data are therefore mutually interdependent. This perspective is recognized internationally. A data-centric challenge led by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) proposed systematic methods for transforming raw clinical research data into AI-ready formats, showing that high-quality clinical datasets form the training foundation for next-generation AI and ML models (*61*). The J-AISI/IPA Data Quality Management Guidebook likewise emphasizes the "garbage in, garbage out" principle, arguing that data quality is the source of AI excellence and that high-quality outputs can fuel a virtuous cycle of further data-quality improvements (*39*). Realizing this interdependent development will require sustained use of AI to improve data quality at each stage of CDM.

7. Conclusions

This review examined AI-aided applications in CDM, with a focus on workflow support, human-in-the-loop implementation, and regulatory and risk-management considerations. Examples in data cleaning, medical coding, and query generation show that institutions are exploring and implementing AI-aided approaches at multiple stages of CDM workflows. In the near term, AI is most likely to contribute to repetitive, rule-based, or text-processing CDM tasks, such as data cleaning support, medical coding assistance, and query generation.

Important challenges remain. The regulatory framework for AI use under Good Clinical Practice continues to evolve, and institutions must remain vigilant about hallucination, bias, privacy and security, and explainability and transparency. Validation of AI-aided processes to a standard comparable to established computerized system validation is a prerequisite for

broader adoption. High-risk decisions and final data judgments should remain under qualified human oversight. In this context, HITL should be regarded as a safeguard that places human judgment at critical points where AI outputs may affect data quality or regulatory accountability. Combining AI-based technical support with human verification and organizational governance will be essential for managing AI-specific risks in CDM.

With policy support and technological maturation, Japan can contribute meaningfully to global discussions on AI-integrated CDM, particularly from the perspective of AROs that must maintain high data quality with limited CDM resources. The transition from predominantly manual, labor-intensive data management to human-AI collaborative workflows is no longer a question of whether, but of when and how.

Funding: None.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. National Institutes of Health (NIH). Clinical research and trials. *https://orwh.od.nih.gov/clinical-research-and-trials* (accessed June 5, 2026).
2. Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol. 2012; 44:168-172.
3. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Guideline for good clinical practice E6(R3). *https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf* (accessed March 24, 2026).
4. Mahadik S, Sen P, Shah EJ. Harnessing digital health technologies and real-world evidence to enhance clinical research and patient outcomes. Digit Health. 2025; 11:20552076251362097.
5. Mitchell EJ, Goodman K, Wakefield N, *et al*. Clinical trial management: A profession in crisis? Trials. 2022; 23:357.
6. Musik S, Sasin-Kurowska J, Panczyk M. Bridging the past and future of clinical data management: The transformative impact of artificial intelligence. Open Access Journal of Clinical Trials. 2025; 17:15-33.
7. Cole S. What is human-in-the-loop? *https://www.ibm.com/think/topics/human-in-the-loop* (accessed April 1, 2026).
8. Olawade DB, Plabon SB, Ojo A, Ogunbona MA, Makanjuola BD, Olasilola OR. Human in the loop artificial intelligence in healthcare: Applications, outcomes, and implementation challenges. Int J Med Inform. 2026; 213:106362.
9. National Institutes of Health (NIH). Data quality management in clinical research. *https://oir.nih.gov/system/files/media/file/2021-08/data_quality_management-2015_05_15.pdf* (accessed March 24, 2026).
10. National Cancer Institute (NCI). NCI guidelines for auditing clinical trials for the NCI national clinical trials network (NCTN) program and NCI community oncology research program (NCORP) including NCORP research bases. *https://dctd.cancer.gov/research/ctep-trials/for-sites/nctn-auditing.pdf* (accessed March 24, 2026).
11. U.S. Food and Drug Administration (FDA). Guidance for Industry - Computerized systems used in clinical trials. *https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/guidance-industry-computerized-systems-used-clinical-trials* (accessed March 24, 2026).
12. eClinical Forum (eCF), the Society for Clinical Data Management (SCDM). Audit trail review: A key tool to ensure data integrity. *https://scdm.org/wp-content/uploads/2024/07/2021-eCF_SCDM-ATR-Industry-Position-Paper-Version-PR1-2.pdf* (accessed May 7, 2026).
13. Prasanna B, Kothapalli P, Vasanthan M. The role of quality assurance in clinical trials: Safeguarding data integrity and compliance. Cureus. 2024; 16:e67573.
14. All-Japan Conference of Deans of Medical Schools and Hospital Directors (AJMC). Summary of questionnaire results regarding the current status of support and management of clinical research. *https://ajmc.jp/wp/wp-content/uploads/2021/04/20151119_1_press.pdf* (accessed May 7, 2026). (in Japanese)
15. Yamaguchi T, Miyaji T, Hayashi Y, Suganami H. Clinical data management in Japan: Past, present, and future. J Soc Clin Data Manag. 2021; 1:1-6.
16. Japan Pharmaceutical Manufacturers Association. The ideal clinical data manager and skill set for the future-Focusing on Vendor Oversight, QMS, and DCT. *https://www.jpma.or.jp/information/evaluation/results/allotment/DS_202305_DM-evol.html* (accessed March 27, 2026). (in Japanese)
17. Harper B, Smith Z, Snowdon J, DiCicco R, Hekmat R, Willis V, Weeraratne D, Getz K. Characterizing pain points in clinical data management and assessing the impact of mid-study updates. Ther Innov Regul Sci. 2021; 55:1006-1012.
18. Ministry of Education, Culture, Sports, Science and Technology, Ministry of Health, Labor and Welfare, Ministry of Economy, Trade and Industry. Ethical guidelines for medical and biological research involving human subjects. *https://www.mext.go.jp/content/20250325-mxt_life-000035486-01.pdf* (accessed April 6, 2026).
19. Ministry of Health, Labor and Welfare. Enforcement regulations on the clinical research act. *https://laws.e-gov.go.jp/law/430M60000100017* (accessed March 31, 2026). (in Japanese)
20. Ministry of Health, Labor and Welfare. Ministerial ordinance concerning standards for conducting clinical trials of pharmaceuticals. *https://laws.e-gov.go.jp/law/409M50000100028?occasion_date=20250401#Mp-Ch_2-Se_1* (accessed March 31, 2026). (in Japanese)
21. Fukuyama M. Society 5.0: Aiming for a new human-centered society. Japan spotlight. 2018; 27:47-50.
22. Morino E, Tokita D. AI in clinical trials: Current status, challenges, and future directions for emergency infectious disease clinical trials-Insights from the 2025 iCROWN Symposium. Glob Health Med. 2026; 8:70-71.
23. Takemura S. The challenge to develop and implement artificial intelligence (AI) technologies in health and medical care in Japan. Journal of the National Institute of Public Health. 2023; 72:2-13.
24. Hartley J, Jolevski F, Melo V, Moore B. The labor market effects of generative artificial intelligence. *https://doi.*

org/10.2139/ssrn.5136877 (accessed May 7, 2026).
25. Morgan Stanley. AI's impact accelerates. *https://www.morganstanley.com/insights/articles/ai-adoption-accelerates-survey-find* (accessed April 1, 2026).
26. Shi X, Prins C, Van Pottelbergh G, Mamouris P, Vaes B, De Moor B. An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. BMC Med Inform Decis Mak. 2021; 21:267.
27. Bönisch C, Schmidt C, Kesztyüs D, Kestler HA, Kesztyüs T. Proposal for using AI to assess clinical data integrity and generate metadata: Algorithm development and validation. JMIR Med Inform. 2025; 13:e60204.
28. Meldau EL, Bista S, Rofors E, Gattepaille LM. Automated drug coding using artificial intelligence: An evaluation of WHODrug Koda on adverse event reports. Drug Saf. 2022; 45:549-561.
29. Utsumi K. Let's consider using AI in data management. *https://www.aro.med.kyushu-u.ac.jp/wp-content/uploads/2024/10/seminar20241018_material-2.pdf* (accessed March 17, 2026). (in Japanese)
30. Japan Pharmaceutical Manufacturers Association. Model prototype for exploring the use of AI in data management operations. *https://www.jpma.or.jp/information/evaluation/results/allotment/q83i5d0000000mtx-att/DS_202406_AI_Prototype.pdf* (accessed March 30, 2026). (in Japanese)
31. Japan Pharmaceutical Manufacturers Association. Application of artificial intelligence in data management - AI to start from now on. *https://www.jpma.or.jp/information/evaluation/results/allotment/DS_202305_2022TF1_1_AI_DM.html* (accessed March 23, 2026). (in Japanese)
32. Jane R. Autocoding adverse events to MedDRA – time to throw out the manual. *https://www.iqvia.com/blogs/2022/03/autocoding-adverse-events-to-meddra-time-to-throw-out-the-manual* (accessed April 1, 2026).
33. Stokman PG, Ensign L, Langeneckhardt D, Mörsch M, Nuyens K, Hochgräber G, Cassan V, Beineke P, Kwock R, Voortman A, Vogelgesang S, Boussetta S, Bitzer B. Risk-based quality management in CDM: An inquiry into the value of generalized query-based data cleaning. Journal of the Society for Clinical Data Management. 2021; 1.
34. U.S. Food & Drug Administration (FDA), European Medicines Agency. Guiding principles of good AI practice in drug development. *https://www.fda.gov/media/189581/download* (accessed March 26, 2026).
35. European Union. Regulation (EU) 2024/1689 of the European parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations. *https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng* (accessed May 7, 2026).
36. Ministry of Health, Labor and Welfare. Regarding the revision of the "Guidance on the ministerial ordinance on standards for conducting clinical trials of pharmaceuticals". *https://www.mhlw.go.jp/web/t_doc?dataId=00tc8160&dataType=1&pageNo=1* (accessed April 7, 2026). (in Japanese)
37. Ministry of Health, Labor and Welfare. Guidelines for the utilization of digital data in AI research and development, *etc*. *https://www.mhlw.go.jp/content/001310044.pdf* (accessed April 7, 2026). (in Japanese)
38. Ministry of Internal Affairs and Communications (MIC), Ministry of Economy, Trade and Industry. AI business guidelines (Version 1.1). *https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20250328_1.pdf* (accessed April 7, 2026). (in Japanese)
39. Japan AI Safety Institute (J-AISI). Data quality management guidebook. *https://aisi.go.jp/assets/pdf/250331_Data_quality_management_guidebook.pdf* (accessed March 19, 2026).
40. Japan AI Safety Institute (J-AISI). Guide to evaluation perspectives on AI safety, Version 1.10. *https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.10_en.pdf* (accessed March 19, 2026).
41. Healthcare AI Platform Collaborative Innovation Partnership (HAIP-CIP). Guidelines for the use of generative AI in the medical and healthcare fields (2nd Edition). *https://haip-cip.org/assets/documents/nr_20241002_02.pdf* (accessed May 7, 2026). (in Japanese)
42. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial intelligence applied to clinical trials: Opportunities and challenges. Health Technol (Berl). 2023; 13:203-213.
43. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023; 6:120.
44. Olawade DB, Fidelis SC, Marinze S, Egbon E, Osunmakinde A, Osborne A. Artificial intelligence in clinical trials: A comprehensive review of opportunities, challenges, and future directions. Int J Med Inform. 2026; 206:106141.
45. Government of Japan. Act on the protection of personal information. *https://www.japaneselawtranslation.go.jp/en/laws/view/4241* (accessed June 8, 2026).
46. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus. 2023; 15:e44720.
47. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, Pimenta D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025; 8:274.
48. Badani A, de Moraes FY, Vollmuth P, Chung C, Mansouri A. AI and innovation in clinical trials. NPJ Digit Med. 2025; 8:683.
49. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A, Raffel C. Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21) Association, 2021; pp. 2633-2650.
50. Chen Y, Esmaeilzadeh P. Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. Journal of Medical Internet Research. 2024; 26:e53008.
51. Zhou J, Li H, Chen S, Chen Z, Han Z, Gao X. Large language models in biomedicine and healthcare. npj Artificial Intelligence. 2025; 1:44.
52. Cao W, Zhang Q, Liu J, Liu S. From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era. J Med Internet Res. 2026; 28:e86550.
53. Arthur F, Daniel S. The cyberattack on change healthcare: Lessons for financial stability. *https://www.financialresearch.gov/briefs/files/OFRBrief-24-05-change-healthcare-cyberattack.pdf* (accessed May 7, 2026).
54. Anthropic. Where things stand with the department of war. *https://www.anthropic.com/news/where-stand-department-war?_bhlid=bfe019c8865f01ef699354be3d20cd037b6c8ff5&utm_source* (accessed March 27, 2026).
55. Ministry of Health, Labour and Welfare. Guidelines for the secure management of medical information systems, Version 6.0. *https://www.mhlw.go.jp/stf/*

shingi/0000516275_00006.html (accessed April 7, 2026). (in Japanese)
56. Michael BC, Rick W. Is Human-on-the-loop the best answer for rapid relevant responses? *https://www.japcc.org/essays/is-human-on-the-loop-the-best-answer-for-rapid-relevant-responses/* (accessed May 7, 2026).
57. Chen M, Wang Y, Wang Q, Shi J, Wang H, Ye Z, Xue P, Qiao Y. Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. NPJ Digit Med. 2024; 7:349.
58. Choudhury A, Chaudhry Z. Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res. 2024; 26:e56764.
59. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012; 19:121-127.
60. Ueda D, Walston S, Takita H, Mitsuyama Y, Miki Y. The critical need for an open medical imaging database in Japan: Implications for global health and AI development. Jpn J Radiol. 2025; 43:537-541.
61. Domagalski MJ, Lu Y, Pilozzi A, Williamson A, Chilappagari P, Luker E, Shelley CD, Dabic A, Keller MA, Rodriguez RM, Lawlor S, Thangudu RR. Preparing clinical research data for artificial intelligence readiness: Insights from the National Institute of Diabetes and Digestive and Kidney Diseases data centric challenge. J Am Med Inform Assoc. 2025; 32:1609-1616.

Received May 7, 2026; Revised June 9, 2026; Accepted June 15, 2026.

Released online in J-STAGE as advance publication June 19, 2026.

**Address correspondence to*:
Hajime Ohyanagi, Department of Joint Center for Researchers, Associates and Clinicians (JCRAC), Center for Clinical Sciences, Japan Institute for Health Security, 1-21-1 Toyama Shinjuku-ku, Tokyo 162-8655, Japan.
E-mail: ohyanagi.h@jihs.go.jp

Global Health & Medicine. 2026; 8(3):193-204.
MINI-REVIEW
DOI: 10.35772/ghm.2026.01058

Artificial intelligence (AI)-assisted diagnosis of skin diseases: From image classification to dermatology-specific multimodal clinical reasoning

Yuhan Cheng[1,2,§], Chu Zhou[3,§], Ping Wang[3], Huanran Liu[4], Yue Han[3,*]

[1] School of Medicine, Tongji University, Shanghai, China;
[2] Department of Nursing, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China;
[3] Department of Dermatology, The Union Hospital, Fujian Medical University, Fuzhou, Fujian, China;
[4] Graduate School of Engineering, Tamagawa University, Machida, Tokyo, Japan.

Abstract: Artificial intelligence (AI) in dermatology has moved beyond the early paradigm of single-image classification. Dermatological diagnosis is achieved based on morphology, distribution, symptoms, tactile findings, temporal evolution, patient history, histopathology, and treatment response. Clinically important differentials, such as eczema versus psoriasis, cutaneous T-cell lymphoma versus chronic dermatitis, drug eruption versus viral exanthem, lupus erythematosus versus dermatomyositis, and melanoma versus atypical nevus, are rarely resolved with one photograph alone. This review therefore frames AI-assisted dermatology around a central argument: the field must progress from lesion recognition to dermatology-specific multimodal clinical reasoning. We summarize major advances in convolutional neural networks, dermoscopic benchmarks, clinical-image datasets, large language models, vision-language systems, and dermatology foundation models. We also analyze challenges that are particularly relevant to dermatology, including morphologic overlap, skin-tone bias, reduced erythema visibility on darker skin, dataset imbalance, variable smartphone imaging, imperfect reference standards, and the gap between benchmark performance and clinical deployment. Special attention is given to fairness, regulatory oversight, software as a medical device, human-AI collaboration, prognosis prediction, biologic-response modeling, longitudinal monitoring, and treatment optimization. Finally, we discuss future directions, including skin-tone-aware foundation models, lesion-level and body-site grounding, pathology-genomics integration, dermatology copilots, post-marketing surveillance, and prospective clinical trials. By prioritizing dermatological reasoning rather than generic AI architecture, this review outlines a clinically grounded pathway for building safe, interpretable, equitable, and useful AI systems for skin disease management.

***Keywords*:** dermatology, artificial intelligence, multimodal reasoning, foundation models

1. Introduction

Skin diseases are among the most common human disorders and encompass inflammatory, infectious, neoplastic, autoimmune, genetic, and drug-related conditions. Dermatology is also an unusually visible specialty: patients, primary care clinicians, dermatologists, pathologists, and digital platforms can all observe the same organ, although at different levels of resolution and with different amounts of context. This visibility has made dermatology an early adopter of computer vision, but it has also encouraged the misleading view that diagnosis is primarily an image classification task. In clinical practice, dermatologists rarely make decisions based on an isolated image. They integrate lesion morphology with body site, distribution, skin tone, symptoms, palpation, tempo of evolution, medication exposure, comorbidities, occupational and environmental triggers, dermoscopy, histopathology, microbiology, serology, and the response to treatment (*1,2*).

The first generation of dermatology artificial intelligence (AI) relied largely on convolutional neural networks (CNNs) trained on clinical or dermoscopic images. Landmark work reported dermatologist-level or near-dermatologist-level performance in selected skin cancer tasks, and public datasets such as HAM10000 and ISIC accelerated algorithm benchmarking (*3-9*). These studies were important because they showed that visual pattern recognition could be scaled, standardized, and compared across groups. At the same time, they revealed

a substantial dataset-to-clinic gap: algorithms usually performed best on curated pigmented-lesion images and often decreased in reliability when confronted with different devices, image quality, disease spectra, clinical settings, and patient populations. This gap is particularly vast in dermatology, where real-world photographs vary in lighting, distance, compression, hair, scale, ulceration, cosmetics, anatomical site, and background pigmentation.

A second generation of systems is now emerging. Large language models (LLMs) and multimodal vision-language models can process patient narratives, electronic health records, pathology reports, and clinical instructions along with images. General medical foundation models, pathology copilots, and dermatology-specific multimodal models offer a route to align morphology with clinical text, histology, and longitudinal data (*10-15*). The goal is not simply to improve top-1 accuracy. Clinically useful systems should generate a ranked differential diagnosis, explain which visual and historical features substantiate each option, indicate when biopsy or referral is needed, express uncertainty, and maintain performance across skin tones and care settings.

This review departs from a generic 'AI in medicine' framework by asking a dermatology-specific question: what must an AI system understand to reason in a clinically dermatological manner? We propose four requirements. First, the system should encode morphology and distribution rather than disease labels alone. Second, it must account for skin color and optical visibility because erythema, scale, pigment network, and vascular structures are not equally visible across skin tones. Third, it should model time, symptoms, and treatment response since many inflammatory and lymphoproliferative diseases declare themselves longitudinally. Fourth, it must be embedded in a supervised workflow with regulatory, ethical, and fairness safeguards.

Dermatology is fundamentally a multimodal specialty, making image-only AI intrinsically incomplete for many inflammatory, lymphoproliferative, drug-related, and longitudinal disorders. The most clinically useful question is therefore not whether more modalities always improve accuracy but which diagnostic decisions are impossible, unsafe, or poorly calibrated without history, distribution, time, pathology, and treatment response.

2. Why dermatology is uniquely amenable to and uniquely difficult for AI

2.1. Morphologic overlap is the central clinical problem

The core challenge for dermatology AI is not merely whether a CNN can detect a border or a pigment network but whether a model can separate diseases that share the same visible vocabulary. Erythematous scaly plaques may represent psoriasis, chronic eczema, tinea corporis, cutaneous lupus erythematosus, pityriasis rubra pilaris, or early mycosis fungoides. Vesicles may indicate allergic contact dermatitis, herpesvirus infection, dyshidrotic eczema, bullous pemphigoid, or a drug eruption depending on distribution, age, symptoms, mucosal involvement, and chronology. A cropped trunk lesion can be photographed and classified, but dermatologists often make a diagnosis by comparing multiple lesions and asking whether they are symmetrical, grouped, photo-distributed, acral, flexural, follicular, dermatomal, annular, targetoid, retiform, or livedoid.

This morphologic overlap helps explain why dermatology AI may involve more difficulty than many standardized imaging tasks. Radiologic images usually follow consistent acquisition protocols and utilize consistent anatomical planes, whereas dermatology photographs may be taken by patients, nurses, primary care physicians, or specialists using different devices, angles, distances, and lighting conditions. The relevant signal may include background skin, hair-bearing status, nail or mucosal involvement, scale texture, or even the absence of a finding. In addition, the diagnostic reference standard is often mixed: some diseases are diagnosed clinically, some require clinicopathologic correlation, and some become evident only after follow-up or treatment failure. AI systems therefore need to learn a diagnostic hierarchy that includes lesion type, color, scale, border, configuration, distribution, temporal behavior, and clinicopathologic consistency. Table 1 summarizes key dermatology-specific problems and the multimodal capabilities required to address them.

2.2. Skin tone is not a secondary fairness issue; it changes the visual signal

Skin-tone bias is one of the defining challenges in dermatology AI. Many training datasets overrepresent lighter skin, while darker skin types, acral lesions, mucosal disease, and conditions common in underserved populations remain comparatively scarce. This imbalance is not only statistical; it changes the visual basis of diagnosis. Erythema may appear bright red on lightly pigmented skin but violaceous, gray, brown, or subtle on deeply pigmented skin. Post-inflammatory hyperpigmentation may mask active inflammation, and pigment networks or vascular patterns may differ in contrast. Psoriasis, atopic dermatitis, lupus erythematosus, dermatomyositis, and drug eruptions can therefore present with skin-tone-dependent cues that a narrow model may not have learned.

Fairness studies have shown that dermatology AI can perform worse on datasets enriched for darker skin tones and uncommon diseases and that post-hoc reweighting cannot fully compensate for models trained on narrow data (*16,17*). Fitzpatrick skin type labels are useful but incomplete because they describe ultraviolet response rather than the full range of skin color, hue, undertone,

Table 1. Dermatology-specific clinical problems that multimodal AI should address

Clinical problem	Typical examples	Why image-only AI is insufficient	Desired multimodal capability
Morphologic overlap	Eczema *vs.* psoriasis, CTCL *vs.* chronic dermatitis, lupus *vs.* dermatomyositis, drug eruption *vs.* viral exanthem	Single images may show the same erythematous plaques, scale, vesicles, or annular patterns without enough context.	Modeling primary and secondary lesion morphology, symptoms, distribution, chronicity, medications, pathology, and treatment response.
Distribution and body-site logic	Extensor psoriasis, flexural eczema, photo-distributed lupus, dermatomal zoster, acral melanoma, nail disease	Cropped lesion photographs remove the body map and may hide symmetry, clustering, or site-specific clues.	Integrating close-up images, distant distribution photos, body maps, anatomical metadata, and prior visits.
Temporal evolution	Changing nevus, flare-remitting eczema, treatment-resistant CTCL, delayed drug eruption	A one-time image cannot show growth, recurrence, latency, dechallenge, or response to therapy.	Using longitudinal images, symptom diaries, medication timelines, and disease trajectory models.
Tactile and bedside signs	Induration, warmth, tenderness, blanching, Nikolsky's sign, scale texture, edema	The model cannot palpate or perform bedside procedures based on pixels alone.	Prompting clinicians or patients for structured signs and flag tests needed before decision-making.
Skin tone and optical visibility	Subtle erythema on darker skin, post-inflammatory hyperpigmentation, low-contrast vascular features	Color cues learned from light-skin datasets may not transfer and may amplify racial disparities.	Reporting skin-tone-stratified performance, calibrating color, enriching darker skin datasets, and using fairness audits.
Weak or mixed gold standards	Clinical diagnosis without biopsy, clinicopathologic discordance, evolving diagnoses	A single disease label can be noisy or misleading.	Using consensus labels, uncertainty-aware training, pathology linkage, and follow-up outcome labels.

Abbreviations: AI, artificial intelligence; CTCL, cutaneous T-cell lymphoma.

and imaging contrast. Future dermatology AI should report performance by skin tone, race and ethnicity when ethically collected, age, sex, anatomical site, image source, and disease prevalence. It should also test erythema-dependent conditions explicitly. Equity must be designed into data collection, annotation, model training, external validation, and post-deployment monitoring rather than added after model development.

2.3. Dermatological diagnosis is longitudinal and often therapeutic

Time is diagnostically important in dermatology. Melanoma screening depends on changes in size, shape, color, and structure. Psoriasis and atopic dermatitis fluctuate with infection, stress, season, adherence, and treatment. Cutaneous T-cell lymphoma (CTCL) may mimic eczema for years before the clinicopathologic pattern reaches a diagnostic threshold. Drug eruptions require attention to medication chronology, latency, dechallenge, rechallenge, and systemic features. Models that rely on a single image at a given time therefore discard information that dermatologists routinely consider essential.

The clinical horizon of dermatology AI is also expanding beyond diagnosis. In chronic inflammatory diseases, the central question may be whether a patient will respond to a biologic, have a flare-up, experience treatment toxicity, or need to be escalated. In melanoma, AI may help integrate histology, molecular markers, and clinical staging for risk assessment. In wounds and ulcers, the task may be to quantify healing trajectory and detect infection or ischemia. These use cases require multimodal and longitudinal modeling rather than isolated image classification (*18-20*).

3. From image classification to multimodal clinical reasoning

3.1. What image-based AI has achieved and where it remains limited

CNNs, EfficientNet, ResNet, Vision Transformers, and ensemble methods have performed well in selected dermoscopic and clinical-image tasks. Early studies demonstrated that deep networks could classify skin cancers at a level comparable to dermatologists under experimental conditions, and ISIC-style challenges created shared benchmarks for lesion segmentation, attribute detection, and classification (*3-9,21-23*). These studies remain foundational because they made dermatology one of the most visible specialties in medical computer vision.

Nevertheless, high area under the curve (AUC) values in curated datasets should not be mistaken for clinical readiness. A pigmented-lesion classifier trained on dermoscopy cannot necessarily diagnose inflammatory rashes, infections, vasculitis, connective tissue disease, or genodermatoses. A model validated on biopsy-confirmed lesions may perform poorly on low-quality smartphone images from primary care. Likewise, a melanoma-versus-nevus classifier may still fail to recommend biopsy when the lesion is amelanotic, the image is incomplete, or the history is concerning. Image AI should therefore be

treated as one component of clinical reasoning and not as the reasoning process itself.

3.2. LLMs may be especially important in dermatology because diagnosis is narrative

LLMs are often described as documentation tools, but in dermatology their deeper value may lie in narrative reasoning. Dermatologists transform a history into diagnostic constraints: acute or chronic, localized or generalized, itchy or painful, febrile or afebrile, drug-associated or spontaneous, recurrent episodes or an initial episode, and treatment-responsive or refractory. LLMs can extract these constraints from consultation notes, referral letters, pathology reports, patient-submitted descriptions, and teledermatology intake forms. They can also translate patient language into dermatological terminology; for example, 'spreading red itchy patches after antibiotics' can be mapped to a differential that includes morbilliform drug eruption, viral exanthem, urticaria, and an early severe cutaneous adverse reaction.

LLMs also create safety risks. They may generate unsupported diagnoses, overstate certainty, suggest an unsafe treatment, or miss emergencies such as Stevens-Johnson syndrome/toxic epidermal necrolysis, meningococcemia, necrotizing infection, or rapidly progressive melanoma. Dermatology-specific deployment therefore requires retrieval from curated guidelines, awareness of local formularies, calibrated uncertainty, escalation rules, and clinician review. A dermatology copilot should not merely provide an answer; it should show which visual signs, historical features, and evidence sources substantiate its differential diagnosis (*12,19*).

3.3. Vision-language and foundation models: Opportunities for dermatology-specific grounding

Vision-language models align images and text in a shared representation space and may enable open-vocabulary recognition, image-text retrieval, and explanations in dermatological terms. General medical foundation models and multimodal systems show how images, text, structured variables, and long clinical records can be interpreted through a unified interface (*10-15*). This architecture is attractive in dermatology because the disease label alone is rarely sufficient. A system should understand that an 'annular scaly plaque with central clearing on the trunk' implies a different differential from a "well-demarcated extensor plaque with silvery scale," even if both appear as erythematous plaques.

Segmentation and multimodal foundation models may aid in lesion boundary detection, body-site mapping, ulcer area monitoring, total-body photography, and clinicopathologic alignment. Dermatology-specific models trained on large collections of clinical photographs, dermoscopy, histopathology, and weakly labeled reports are especially promising (*15*). The next step is not only larger pretraining but clinically grounded pretraining. Models should learn relationships between morphological terms and image regions, lesion distribution and body maps, histopathologic patterns and clinical phenotypes, and longitudinal changes and treatment response.

Foundation models may change dermatology more profoundly than classic CNNs because they are better aligned with the three characteristics of the specialty: label ambiguity, an open vocabulary, and morphology-language coupling. A CNN can learn that a lesion resembles training examples labeled psoriasis, but a vision-language model can represent the phrase "well-demarcated erythematous plaques with silvery scale on extensor surfaces" and relate it to competing diagnoses, body-site logic, and clinical history. This matters because dermatological labels are often provisional, overlapping, or refined after biopsy and follow-up rather than fixed at the moment of imaging (*10,14,15*).

Dermatology may therefore be one of the medical specialties most naturally amenable to multimodal foundation models. Its diagnostic vocabulary is visual but not purely visual: morphology is described in language, distribution is mapped across the body, histopathology provides tissue-level confirmation, and chronic diseases evolve across visits. The field should judge foundation models not only by top-1 disease classification but by whether they can retrieve visually similar cases, ground morphological terms in image regions, recognize missing clinical context, and express uncertainty when a single photograph cannot lead to a safe diagnosis.

3.4. Multimodal fusion should follow the dermatologist's workflow

Early fusion, late fusion, cross-modal attention, contrastive pretraining, and retrieval-augmented generation are useful engineering strategies, but their value depends on whether they reproduce dermatological reasoning. A practical multimodal pipeline should assess image quality and skin tone, segment relevant lesions, identify morphology and distribution, extract a structured history from text, retrieve guidelines or prior images when appropriate, generate a ranked differential diagnosis, calibrate uncertainty, and recommend triage or next steps only when the available information is sufficient.

This workflow differs from generic medical AI because the dermatological examination is partly visual, partly tactile, and partly historical. A model cannot palpate an induration or gauge warmth, tenderness, or blanching, but it can ask the clinician or patient to provide those findings. It cannot directly perform bedside procedures, but it can flag which tests are needed before making a decision. Multimodal dermatology AI should therefore be interactive: it should improve

data acquisition and clinical questioning rather than simply classify what has already been uploaded. The dermatological multimodal reasoning architecture for AI-assisted skin disease diagnosis is shown in Figure 1.

3.5. Which dermatological tasks truly require multimodal AI?

A critical weakness of early dermatology AI was the implicit assumption that every skin task should be solved as an image classification problem. The opposite error would be to assume that every task requires maximal multimodalities. Multimodal AI is valuable when additional context changes the clinical decision, reduces unsafe uncertainty, or distinguishes diseases with overlapping morphology. It is less necessary when the target is a narrowly defined visual triage problem under standardized acquisition conditions.

Melanoma screening illustrates this distinction. For many dermoscopic images of pigmented lesions, visual structure represents a large fraction of the diagnostic signal, and image-based models can provide useful risk stratification. And yet the same model is incomplete for amelanotic lesions, lesions selected from high-risk total-body photography, changing nevi, immunosuppressed patients, or cases in which the relevant clinical decision is biopsy versus short-interval monitoring. In these settings, prior images, lesion evolution, age, site, personal and family history, and a clinician's suspicions become part of the diagnostic signal (*3-9,21-23*).

Inflammatory, drug-related, infectious, and lymphoproliferative dermatoses are different. A photograph of an erythematous scaly plaque may be consistent with eczema, psoriasis, tinea, lupus, dermatomyositis, or early CTCL. The decisive information may be itching, pain, a fever, the timing of medication, mucosal involvement,

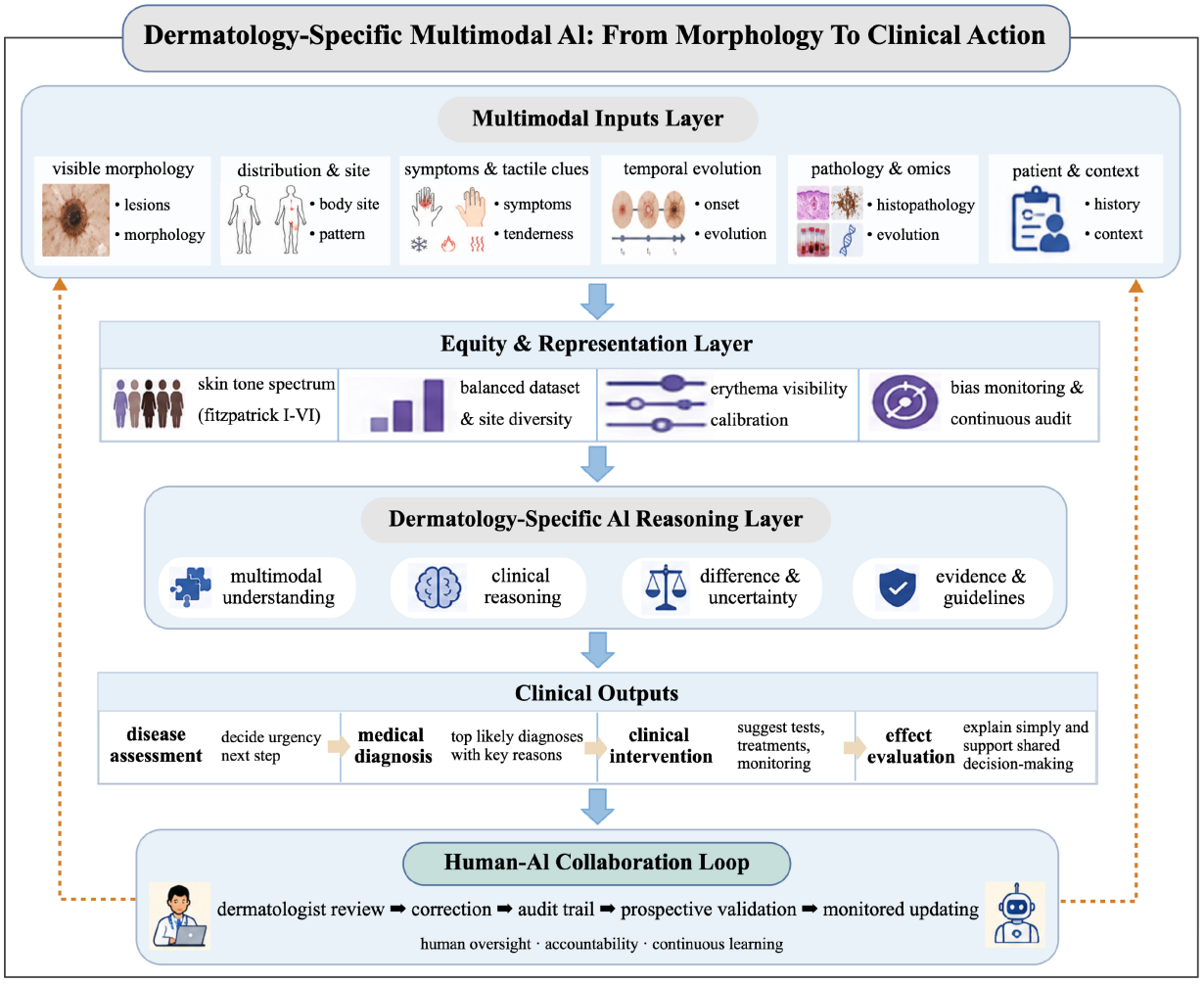


Figure 1. Dermatological multimodal reasoning architecture for AI-assisted skin disease diagnosis. The schematic illustrates a clinically grounded pipeline that integrates multiple data modalities to aid in dermatological decision-making. Inputs include clinical images, dermoscopy, patient history, symptoms, medication timelines, prior images, and histopathology reports. The core reasoning engine performs lesion segmentation, morphology and distribution analysis, detection of temporal changes, and clinicopathologic alignment using vision-language and foundation models. Outputs consist of a ranked differential diagnosis with uncertainty calibration, visual and textual explanations linked to dermatological signs, triage recommendations, and suggestions for missing information. The architecture emphasizes interactive data acquisition, clinician oversight, and fairness across skin tones.

distribution, chronicity, histology, immunophenotype, laboratory findings, or treatment failure. For these tasks, image-only AI is not merely incomplete; it can yield methodologically unsound results because the required clinical variables are absent from the input.

4. Dermatology-specific bottlenecks and challenges

4.1. Data quality and annotation: Labels must depict morphology and not just diagnosis

Many dermatology AI datasets provide disease labels, but dermatologists reason with intermediate descriptors. A label such as 'psoriasis' is less informative than a structured description such as 'well-demarcated erythematous plaque, silvery scale, extensor surface, chronic recurrent course, family history, and no fungal hyphae.' Similarly, a melanoma dataset becomes more clinically useful when it includes body site, lesion size, evolution, dermoscopic structures, histopathology, and whether the lesion was identified during high-risk surveillance or routine care. Disease-only labels may encourage models to learn shortcuts, including device type, background skin, ruler marks, biopsy ink, or center-specific artifacts.

Annotation should therefore move from single labels towards structured dermatological ontologies that include primary lesion type, secondary changes, arrangement, distribution, body site, symptom profile, skin tone, image quality, and diagnostic certainty. Consensus annotation, adjudication of difficult cases, and uncertainty-aware labels are essential because dermatology reference standards are often imperfect. For rare diseases, sharing of registry-based data, active learning, and expert-in-the-loop annotation may be more useful than indiscriminate data scaling.

4.2. Learning of hidden shortcuts by dermatology AI

Dermatology AI can fail even when retrospective accuracy appears high because models may learn features that correlate with labels but do not represent disease biology. Common hidden shortcuts include ruler marks placed beside suspicious lesions, skin-marker ink, biopsy-site framing, surgical drapes, dermoscopy device signatures, clinic-specific backgrounds, compression artifacts, hair removal patterns, and anatomical-site or referral-center cues. A model can therefore appear to detect melanoma while partly detecting the clinical behavior that preceded biopsy or photography.

These failures are especially dangerous because dermatology images are often collected after a clinician has already identified a lesion as concerning. The dataset may encode selection bias, workup bias, and spectrum bias: benign lesions photographed casually, malignant lesions photographed with rulers and dermoscopy, and inflammatory rashes photographed only after referral. If these cues are not controlled, external validation may reveal an abrupt drop in performance in primary care, teledermatology, darker skin tones, or limited-resource settings (*16,17,19,23-25*).

Robust development should therefore include shortcut stress tests. Investigators should perform metadata ablation, background masking, lesion-only versus context-aware comparisons, device-stratified validation, counterfactual removal of rulers and markers, compression robustness testing, and prospective evaluation in the intended workflow. Explanations should be audited for whether they highlight true dermatological signs rather than artifacts. This failure analysis is as important as reporting the aggregate AUC.

4.3. Real-world deployment

Many AI systems perform well on images acquired under standardized conditions but face a different reality in teledermatology and primary care. Patient-submitted photographs may be blurred, underexposed, overcompressed, too close, too far, or taken after the application of emollients, makeup, antiseptics, topical corticosteroids, or dressings. Hair, tattoos, scale, crust, ulceration, nail polish, and anatomical curvature can obscure lesions. Smartphone color calibration and automatic white balance can alter erythema and pigmentation. These factors are not merely technical noise; they directly change the diagnostic signal available to both a clinician and a model (*24,25*).

Deployment-ready systems should provide image-quality feedback, standardized acquisition instructions, color calibration or reference cards when feasible, validation across devices, and the ability to safely reject inadequate images. They should also be tested in common teledermatology scenarios, including multiple-lesion uploads, rashes involving several body sites, darker skin tones, pediatric images, nail and scalp disease, genital or mucosal lesions requiring careful consent, and follow-up photographs taken at different distances or under different lighting conditions.

4.4. Bias, global inequity, and dataset colonialism

Global dermatology AI should not be created based solely on images collected at wealthy, urban, specialist centers and then exported to settings with different diseases, skin tones, pathogens, occupational exposure, climate, access to care, and treatment pathways. Such a pattern, sometimes discussed as dataset colonialism, risks turning underrepresented populations into data sources or target markets without giving them governance authority, clinical benefit, or locally valid tools. It may also miss conditions common in limited-resource settings, including neglected tropical diseases, pigmentary disorders, leprosy, scabies, deep fungal infections, Kaposi sarcoma, and HIV-associated dermatoses.

Equitable dermatology AI requires locally meaningful datasets, partnerships between communities and clinicians, transparent consent, shared governance, subgroup performance reporting, and mechanisms for sharing benefits. Fairness should be assessed not only with an equal AUC but also based on clinically meaningful outcomes: missed malignancies, delayed diagnosis, unnecessary biopsies, referral burden, treatment access, patient trust, and whether the system improves care for groups historically underserved by dermatology (*16,17*).

4.5. Interpretability must be dermatology-specific

Generic heatmaps are insufficient if they simply highlight a lesion without explaining the relevant signs. Dermatologists need explanations that map to clinical language, such as an atypical pigment network, blue-white veil, arborizing vessels, Wickham striae, collarette scale, follicular plugging, Gottron papules, a targetoid morphology, retiform purpura, or palmoplantar involvement. A useful explanation should connect a visible region to a morphology term, the term to a differential diagnosis, and the differential diagnosis to a concrete next step.

Dual-modality explanations are therefore preferable: visual evidence should be paired with a textual rationale and calibrated uncertainty. A system might rank psoriasis as the leading diagnosis because it identifies sharply demarcated scaly plaques on extensor surfaces and a chronic recurrent history while also retaining eczema as a possibility because of pruritus and flexural involvement. Such explanations allow clinicians to challenge the model, identify missing information, and reduce automation bias.

4.6. Regulation and medico-legal responsibility

Clinical dermatology AI may qualify as software as a medical device when it provides diagnostic or triage recommendations. Regulatory evaluation should consider the intended use, target population, user type, autonomy level, risk of harm, and whether the model is locked or adaptive. Earlier devices such as MelaFind demonstrated the feasibility of computerized lesion assessment but also underscored problems related to specificity, workflow fit, and clinical utility; newer primary care systems have renewed interest in regulated AI-assisted skin cancer triage (*26,27*).

Modern multimodal and adaptive systems require representative validation, skin-tone subgroup analysis, human-factor testing, cybersecurity review, data governance, change-control planning, and post-marketing performance monitoring (*28-30*). Responsibility must also be explicit. When an AI output conflicts with a clinician's judgment, the medical record should show who reviewed the output, what evidence was considered, and why the final decision was made. High-risk recommendations should require clinician confirmation and should be linked to visible evidence and current guidance.

5. Future breakthroughs and technical pathways

5.1. Dermatology foundation models should be morphology-aware, skin-tone-aware, and longitudinal

The next generation of dermatology foundation models should not be evaluated only in terms of disease classification accuracy. They should be tested on morphology recognition, body-site grounding, lesion segmentation, image-quality assessment, skin-tone fairness, rare-disease retrieval, detection of temporal changes, and clinicopathologic correlation. Self-supervised and weakly supervised learning can capitalize on large image archives, but clinically meaningful pretraining should pair images with morphology-rich reports, dermoscopic structures, pathology captions, distribution maps, and longitudinal treatment data. Table 2 summarizes landmark evidence and emerging model families.

Longitudinal modeling is especially important. Total-body photography, sequential dermoscopy, patient-reported flare-up diaries, laboratory trends, and treatment changes can all be represented as trajectories. Such models could identify changing melanocytic lesions, quantify psoriasis severity over time, detect atopic dermatitis flare-ups, monitor wound healing, and recognize chronic dermatitis cases that warrant a biopsy for cutaneous lymphoma. They should also report uncertainty and clearly distinguish diagnostic prediction, prognosis, and estimation of the treatment response.

5.2. Clinicopathologic foundation models

The most consequential frontier for dermatology foundation models may be clinicopathologic integration. Dermatologists rarely treat clinical photographs and pathology slides as independent evidence streams; instead, they ask whether the clinical morphology, dermoscopic structures, histopathologic pattern, immunophenotype, molecular alterations, and disease course support the same diagnosis. AI systems should be designed to model this alignment rather than simply concatenate modalities.

Melanoma, CTCL, autoimmune connective-tissue disease, vasculitis, blistering disorders, and complex drug reactions are particularly amenable to this approach. In melanoma, clinical and dermoscopic images can be connected with histologic subtype, Breslow thickness, ulceration, mitotic rate, genomics, and outcomes. In CTCL, persistent patches or plaques, serial photographs, repeated biopsies, T-cell receptor clonality, and treatment response form a longitudinal clinicopathologic pattern. In autoimmune disease, morphology and distribution must

Table 2. Landmark evidence and emerging model families in dermatology AI

Area or model family	Representative evidence or dataset	Dermatology-specific contribution	Remaining gap
Classic skin cancer CNNs	Esteva *et al*. (*3*), Haenssle *et al.* (*4*), Tschandl *et al.* (*5*): human-computer studies.	Established dermatologist-level performance in selected dermoscopic or clinical image tasks.	Often have a narrow disease spectrum, curated images, and limited real-world workflow evaluation.
Public dermoscopy benchmarks	ISIC challenges, HAM10000, BCN20000.	Enabled reproducible lesion segmentation and pigmented-lesion classification.	Underrepresentation of darker skin, non-pigmented lesions, rare disorders, and primary care images.
Clinical image + history systems	Liu *et al.* (*6*): deep learning system for 26 common skin diseases; PAD-UFES-20 smartphone dataset.	Demonstrated that history and metadata improve differential diagnosis beyond lesion pixels.	Require prospective triage evaluation and better handling of incomplete patient contexts.
Broad dermatology classifiers	Han *et al.* (*7*): 134 skin disorders, mobile and few-shot platforms.	Expanded AI from melanoma to inflammatory, infectious, and benign conditions.	May still rely on disease labels rather than morphology and distribution annotations.
Fairness datasets and audits	DDI, Fitzpatrick17k, skin-tone subgroup analysis.	Exhibited clinically significant performance degradation on darker skin tones and uncommon diseases.	Require global, prospective, skin-tone-aware validation and local governance.
Vision-language models	CLIP, BiomedCLIP, GPT-4V dermatology prompting.	Enabled open-vocabulary matching, image-text retrieval, and morphology-grounded explanations.	Prone to prompt sensitivity, hallucinations, and weak spatial grounding without clinical guardrails.
Medical generalist models	Med-PaLM M, Med-Gemini, multimodal pathology copilots.	Supported flexible reasoning over text, images, records, pathology, and long contexts.	Dermatology-specific performance, safety, and fairness remain insufficiently validated.
Dermatology foundation models	PanDerm-like multimodal dermatology pretraining.	Can learn transferable representations across clinical, dermoscopic, and histopathologic modalities.	Need prospective external validation, transparent failure analysis, and equitable deployment.
Regulated or near-regulated systems	MelaFind, DermaSensor, Skin Analytics-type triage tools.	Moved AI from benchmark datasets towards clinical pathways and device regulation.	Need evidence for specificity, workflow impact, liability, and post-marketing monitoring.
Clinicopathologic foundation models	Multimodal pathology copilots, dermatology foundation models, and paired clinicopathologic datasets.	Connected clinical morphology, dermoscopy, histopathology, immunophenotype, genomics, transcriptomics, and outcomes into one reasoning space.	Require lesion-to-biopsy linkage, tissue-level grounding, temporal alignment, external validation, and explicit handling of discordant evidence.

Abbreviations: AI, artificial intelligence; CNNs, convolutional neural networks; DDI, Diverse Dermatology Images; ISIC, International Skin Imaging Collaboration.

be aligned with serology, histology, immunofluorescence, and transcriptomic signatures (*14,15,18*).

Clinicopathologic foundation models would require paired and temporally linked datasets, tissue-level grounding, lesion-level mapping between the clinical image and biopsy site, and outcome labels that distinguish diagnosis, prognosis, and the therapy response. Their value should be judged by whether they detect discordance, recommend a repeat biopsy when morphology and histology conflict, and support cautious reasoning in instances of diseases where a single gold standard is unrealistic.

5.3. Agentic dermatology AI: A copilot, not an autonomous dermatologist

Agentic AI refers to systems that can plan actions, call tools, retrieve information, ask follow-up questions, and coordinate workflow steps. In dermatology, a safe agent might check image quality, request close-up and distant distribution photographs, ask about pain, itching, a fever, medication exposure, mucosal involvement, pregnancy, immunosuppression, and duration, retrieve prior images, generate a differential diagnosis, recommend whether urgent referral is needed, and draft a structured note for clinician review.

Such systems should be designed as copilots rather than autonomous dermatologists. They require escalation rules for red-flag conditions, refusal behavior when data are inadequate, transparent uncertainty, and audit trails. Patient-facing agents should avoid a definitive diagnosis when the risk is high and should direct patients to urgent care when systemic symptoms, a rapidly spreading rash, mucosal involvement, necrosis, purpura, or suspected

melanoma is present. Clinician-facing agents can reduce documentation and the cognitive burden while preserving professional accountability.

5.4. Fairness-by-design and global data collaboration

Federated learning, secure aggregation, differential privacy, and trusted research environments can facilitate collaboration across institutions without centralizing sensitive images or records (*31*). However, privacy-preserving learning does not automatically ensure fairness. A federated network dominated by similar populations can still reproduce bias. Collaboration should therefore be deliberately structured to include diverse skin tones, geographic regions, age groups, anatomical sites, disease categories, and care settings. Synthetic images may help balance rare categories, but they must be reviewed by experts and tested for leakage, artifact amplification, and skin-tone distortion.

A practical global reporting standard should require every dermatology AI study to describe its dataset composition, skin-tone distribution, image source, reference standard, subgroup performance, failure modes, calibration, and external validation. Journals and regulators should also require explicit reporting on darker skin, inflammatory disorders, rare diseases, and real-world smartphone images.

5.5. Prospective clinical trials and post-deployment learning

Most dermatology AI evidence remains retrospective. Prospective studies should determine whether AI improves outcomes that matter clinically: time to melanoma diagnosis, appropriateness of referral, biopsy yield, missed cancer rate, flare-up control, treatment adherence, clinician workload, patient satisfaction, and equity. The SPIRIT-AI, CONSORT-AI, and DECIDE-AI reporting frameworks provide useful principles for AI clinical trials and early-stage evaluations, but dermatology-specific endpoints are still needed (*30,32,33*).

Post-deployment monitoring is equally important because imaging devices, clinical practice, disease prevalence, and treatment options change over time. Drift detection, periodic recalibration, safety audits, and clinician feedback mechanisms should be required. Adaptive updates should remain separate from real-time decision-making support unless a regulated change-control plan is in place. Table 3 provides a practical deployment checklist for dermatology AI.

6. AI's potential applications and future prospects in dermatology

6.1. Teledermatology and primary care triage

Teledermatology is the most immediate setting for multimodal AI because it already depends on asynchronous images and text. AI can assist by checking image quality, collecting a structured history, identifying high-risk lesions, suggesting differentials for common rashes, and prioritizing referrals. In primary care, AI may help clinicians decide whether a lesion requires an urgent dermatology assessment, whether a rash can be treated empirically, or whether biopsy, culture, serology, or emergency evaluation is needed (*24,25*).

The purpose of these systems should not be framed as replacing dermatologists. Rather, AI can help direct the right patient to the right level of care. Low-risk benign lesions may be managed with reassurance and follow-up, whereas suspected melanoma, a rapidly progressing infection, vasculitis, or a severe drug eruption should be escalated. Successful triage depends on a high level of sensitivity to dangerous diseases, transparent uncertainty, skin-tone equity, and workflows that reduce rather than increase the clinician's burden.

6.2. Rare and difficult diseases

Rare and diagnostically difficult diseases are natural targets for multimodal dermatology AI because individual clinicians may encounter only a small number of cases, whereas aggregated data can reveal recognizable patterns. CTCL, autoimmune blistering diseases, genodermatoses, vasculitis, connective tissue diseases, and rare infections often require integration of clinical morphology, distribution, laboratory data, pathology, immunofluorescence, molecular testing, and follow-up. AI systems may aid in these cases by retrieving similar examples, ranking differential diagnoses, and detecting discordance between clinical and histopathologic findings.

Few-shot learning, retrieval-augmented reasoning, and knowledge graphs may be particularly useful. Even when a model cannot confidently classify a rare disease, it may provide safety-oriented reasoning, such as noting that a chronic treatment-resistant patch or plaque dermatitis with poikiloderma and atypical lymphocytes warrants concern about mycosis fungoides and calls for a repeat biopsy. This type of cautious, evidence-linked output may be more clinically valuable than forcing a high-confidence label.

6.3. Treatment decision-making support and efficacy prediction

Dermatology AI is moving from diagnosis towards decision-making support. In psoriasis and atopic dermatitis, multimodal models may combine images, severity scores, itching and sleep measures, comorbidities, laboratory data, prior therapies, adherence, and pharmacogenomic or transcriptomic signals to predict the response to biologics, Janus kinase inhibitors, phototherapy, or systemic immunomodulators. In acne,

Table 3. Dermatology AI deployment checklist: From validation to clinical use

Risk domain	Dermatology-specific manifestation	Required evaluation	Practical mitigation
Skin-tone fairness	Erythema and vascular cues vary by pigmentation; darker skin is underrepresented in many datasets.	Reporting performance by skin tone, disease type, anatomical site, and image source; auditing false negatives and false positives.	Diverse recruitment, skin-tone-aware sampling, color calibration, fairness dashboards, and local validation.
Real-world image quality	Blur, lighting, compression, cosmetics, hair, scale, ulceration, and variable smartphone cameras.	Stress-testing teledermatology and primary care images; measuring the reject rate and clinician override rate.	Image-quality prompts, acquisition guidance, reference cards, and safe rejection when images are inadequate.
Clinical safety	Missed melanoma, severe drug eruption, vasculitis, necrotizing infection, or immunosuppression-related disease.	Evaluate high-risk triage sensitivity, calibration, uncertainty, and red-flag recognition.	Escalation rules, clinician confirmation, conservative thresholds, and emergency warnings.
Workflow integration	Standalone tools add clicks and may not fit EHR or teledermatology pathways.	Ascertaining time, referral quality, biopsy yield, user trust, and documentation burden.	EHR integration, structured output, audit trails, and human-centered interface design.
Regulation and liability	Adaptive multimodal models may change over time and influence biopsy, referral, or therapy decisions.	Defining intended use, autonomy level, locked *vs.* adaptive status, cybersecurity, and change-control plans.	SaMD pathway, predetermined change-control plans, post-marketing surveillance, and clear responsibility.
LLM hallucination	Confident but unsupported diagnosis, unsafe treatment advice, or missed red flags.	Benchmarking factuality, guideline consistency, uncertainty expression, and failure cases.	Retrieval-augmented generation, curated knowledge bases, guardrails, and clinician review.
Longitudinal drift	Disease prevalence, cameras, guidelines, therapies, and patient populations evolve.	Monitoring calibration and subgroup performance over time; comparing with respect to prior model versions.	Drift detection, scheduled recalibration, rollback mechanisms, and prospective auditing.

Abbreviations: AI, artificial intelligence; EHR, electronic health record; LLM, large language model; SaMD, software as a medical device.

hidradenitis suppurativa, and rosacea, longitudinal images and patient-reported outcomes may help monitor severity and treatment response. In melanoma, AI may integrate histology, clinicopathologic variables, and molecular markers for risk stratification and trial matching (*18-20*).

These applications require higher safety standards than classification because they affect treatment exposure, cost, and patient expectations. Recommendations should be linked to guidelines, they should factor in local formulary constraints, and they should be patient-centered and explainable. The model should distinguish diagnostic confidence from therapeutic evidence, and clinicians should remain responsible for final treatment decisions.

6.4. Education and research innovation

Multimodal AI can enhance dermatology education by generating case-based learning materials that include images across skin tones, morphology labels, distribution maps, histopathology, and differential reasoning. It can also give trainees immediate feedback, not only on whether an answer is correct but also on which morphology clues were missed. In research, AI can mine image-text-pathology repositories to identify phenotypes, disease subtypes, biomarker associations, and treatment response patterns. Foundation models may also accelerate clinical trial screening by matching patient phenotypes to eligibility criteria.

Educational and research applications must still ensure privacy, consent, and fairness. Synthetic cases should be clearly labeled, and generated educational material should be reviewed by experts. For trainee education, AI should emphasize uncertainty and differential diagnosis rather than encouraging single-label pattern recognition.

6.5. From diagnosis-centric AI to continuous skin intelligence

The future paradigm of dermatology AI should not be limited to "taking a photograph and receiving a diagnosis". A more ambitious and clinically realistic direction is continuous skin intelligence: systems that track lesions, inflammation, wounds, symptoms, exposure, and treatment response over time. This paradigm fits dermatology because many important decisions depend on trajectory rather than a single visual snapshot.

Such systems could combine patient-submitted images, total-body photography, wearable or smartphone sensing, itching and pain diaries, environmental triggers, medication timelines, sleep and activity data, and clinician-verified outcomes. They could monitor melanoma risk through detection of changes,

quantification of psoriasis and atopic dermatitis flare-ups, follow hidradenitis suppurativa activity, gauge wound healing, and identify patients whose condition needs to be managed before a severe flare-up occurs.

However, continuous monitoring also creates new risks: overdiagnosis, anxiety, privacy leakage, inequitable access, false alarms, and unclear responsibility for unattended alerts. The goal should be clinician-supervised longitudinal decision-making support and not surveillance for its own sake. Safe systems must define what is monitored, who receives alerts, how uncertainty is displayed, and when the model should remain silent.

7. Conclusion

Dermatology AI has progressed from experimental CNN classifiers to multimodal systems that can integrate images, text, structured patient information, and histopathology. The key conceptual advance is that dermatology AI should be built around clinical reasoning rather than generic computer vision. The field must learn morphology, distribution, skin tone, symptoms, temporal evolution, clinicopathologic correlation, and therapeutic context.

The major barriers are also dermatology-specific. Image-only models are vulnerable to morphologic ambiguity, acquisition variability, and missing context. Skin-tone bias can alter the visual signal itself, and particularly in erythema-dependent diseases. Benchmark accuracy may not translate to teledermatology, primary care, rare-disease diagnosis, or treatment decision-making support. LLMs and agentic systems add important capabilities but also risks, including hallucinations, automation bias, privacy leakage, and unclear medico-legal responsibility.

Future progress will depend on foundation models that are morphology-aware, skin-tone-aware, longitudinal, and clinically grounded; datasets that are diverse and responsibly governed; explanations that connect visual evidence to dermatological language; regulatory pathways that address adaptive multimodal software; and prospective studies that measure outcomes valued by patients and clinicians. With these foundations, AI can become a trusted component of digital dermatology - not a replacement for dermatologists, but a tool that expands access, improves triage, improves diagnostic reasoning, and facilitates more personalized care. The central methodological challenge is to decide when AI should recognize a lesion, when it should reason across modalities, and when it should defer because the available inputs do not support a safe dermatological determination.

Funding: This work was supported by the National Natural Science Foundation of China (82203935), the Fujian Province Natural Science Foundation (2026J001590), Joint Funds for the innovation of Science and Technology, Fujian Province (2025Y9299), and Fujian Medical University Union Hospital's Project to Foster Excellent Young Scholars (2022XH027).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Hay RJ, Johns NE, Williams HC, Bolliger IW, Dellavalle RP, Margolis DJ, Marks R, Naldi L, Weinstock MA, Wulf SK, Michaud C, J L Murray C, Naghavi M. The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions. J Invest Dermatol. 2014; 134:1527-1534.
2. Karimkhani C, Dellavalle RP, Coffeng LE, Flohr C, Hay RJ, Langan SM, Nsoesie EO, Ferrari AJ, Erskine HE, Silverberg JI, Vos T, Naghavi M. Global skin disease morbidity and mortality: An update from the Global Burden of Disease Study 2013. JAMA Dermatol. 2017; 153:406-412.
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542:115-118.
4. Haenssle HA, Fink C, Schneiderbauer R, *et al*. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018; 29:1836-1842.
5. Tschandl P, Rinner C, Apalla Z, *et al*. Human-computer collaboration for skin cancer recognition. Nat Med. 2020; 26:1229-1234.
6. Liu Y, Jain A, Eng C, *et al*. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020; 26:900-908.
7. Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, Chae JB, Huh CH, Na JI. Augmented intelligence dermatology: Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J Invest Dermatol. 2020; 140:1753-1761.
8. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018; 5:180161.
9. Marchetti MA, Codella NCF, Dusza SW, *et al*. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol. 2018; 78:270-277.e1.
10. Moor M, Banerjee O, Abad ZSH, *et al*. Foundation models for generalist medical artificial intelligence. Nature. 2023; 616:259-265.
11. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023; 388:1201-1208.
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022; 28:31-38.
13. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019; 25:44-

56.
14. Lu MY, Chen B, Williamson DFK, *et al*. A multimodal generative AI copilot for human pathology. Nature. 2024; 634:466-473.
15. Yan S, Yu Z, Primiero C, *et al*. A multimodal vision foundation model for clinical dermatology. Nat Med. 2025; 31:2691-2702.
16. Daneshjou R, Vodrahalli K, Novoa RA, *et al*. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv. 2022; 8:eabq6147.
17. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018; 154:1247-1248.
18. Choy SP, Kim BJ, Paolino A, Tan WR, Lim SML, Seo J, Tan SP, Francis L, Tsakok T, Simpson M, Barker JNWN, Lynch MD, Corbett MS, Smith CH, Mahil SK. Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. NPJ Digit Med. 2023; 6:180.
19. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019; 17:195.
20. Widiawaty A, Indriatmi W, Jatmiko W, Novianto E, Kekalih A, Gunawan H, Palar PS, Rachmadi MF, Dermawan S, Malahayati TL, Ramadhan AW. Multimodal machine learning approach for diagnosing atopic dermatitis. F1000Res. 2025; 14:952.
21. Tschandl P, Rosendahl C, Akay BN, *et al*. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol. 2019; 155:58-65.
22. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, Schilling B, Haferkamp S, Schadendorf D, Holland-Letz T, Utikal JS, von Kalle C; Collaborators. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer. 2019; 113:47-54.
23. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. Comput Biol Med. 2020; 127:104065.
24. Pacheco AGC, Lima GR, Salomão AS, *et al*. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief. 2020; 32:106221.
25. Freeman K, Dinnes J, Chuchu N, Takwoingi Y, Bayliss SE, Matin RN, Jain A, Walter FM, Williams HC, Deeks JJ. Algorithm based smartphone apps to assess risk of skin cancer in adults: Systematic review of diagnostic accuracy studies. BMJ. 2020; 368:m127.
26. Monheit G, Cognetta AB, Ferris L, Rabinovitz H, Gross K, Martini M, Grichnik JM, Mihm M, Prieto VG, Googe P, King R, Toledano A, Kabelev N, Wojton M, Gutkowicz-Krusin D. The performance of MelaFind: A prospective multicenter study. Arch Dermatol. 2011; 147:188-194.
27. Venkatesh KP, Kvedar JC. Learnings from the first AI-enabled skin cancer device for primary care authorized by FDA. NPJ Digit Med. 2024; 7:167.
28. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): A comparative analysis. Lancet Digit Health. 2021; 3:e195-e203.
29. Benjamens S, Dhunnoo P, Mesko B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. NPJ Digit Med. 2020; 3:118.
30. Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022; 28:924-933.
31. Hossen MN, Panneerselvam V, Koundal D, Ahmed K, Bui FM, Ibrahim SM. Federated machine learning for detection of skin diseases and enhancement of IInternet of Medical Things (IoMT) Security. IEEE J Biomed Health Inform. 2023; 27:835-841.
32. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med. 2020; 26:1351-1363.
33. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020; 26:1364-1374.

Received April 28, 2026; Revised May 21, 2026; Accepted June 9, 2026.

Released online in J-STAGE as advance publication June 13, 2026.

[§]*These authors contributed equally to this work.*
**Address correspondence to*:
Yue Han, Department of Dermatology, The Union Hospital, Fujian Medical University, No. 29 Xinquan Road, Fuzhou 350001, China.
E-mail: dr_hanyue@126.com

Global Health & Medicine. 2026; 8(3):205-214. ORIGINAL ARTICLE

DOI: 10.35772/ghm.2026.01053

Automated radiographic shoulder balance assessment in scoliosis *via* deep learning

Longhao Yang[1,2,§], Fangzheng Xu[1,2,§], Qingzhi Xiang[1,2], Jianwen Fu[3], Xiao Xia[1,2], Fuping Li[1,2], Shaobo Cheng[1,2], Yifei Qin[1,2], Yan Yu[1,2,*]

[1] Division of Spine, Department of Orthopaedics, Tongji Hospital, Tongji University School of Medicine, Tongji University, Shanghai, China;
[2] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Shanghai, China;
[3] Spinextech Medical Technology Co., Ltd., Shanghai, China.

Abstract: The objective of this study was to develop an automated deep learning-based method for the assessment of shoulder balance in adolescent idiopathic scoliosis (AIS) patients using X-ray images in order to provide a reliable and efficient alternative to manual measurements. A total of 940 AIS radiographs were screened; 937 cases were included in the model-development cohort after quality control and were annotated for precise identification and segmentation of the T1 vertebra, both clavicles, and both coracoids. A deep learning neural network was used to segment these structures. Landmarks were extracted based on morphological image processing, and shoulder balance parameters including the clavicle angle (CA), coracoid height difference (CHD), clavicle tilt angle difference (CTAD), radiological shoulder height (RSH), and T1 tilting angle (T1TA) were calculated. The accuracy of the automated measurements was validated using an external dataset (n = 70) assessed by three senior spinal surgeons. The deep learning neural network achieved reliable segmentation performance for foreground anatomical structures, with macro-average intersection over union (IoU) values of 0.77 and 0.73 and Dice coefficients of 0.87 and 0.84 in the internal and external validation datasets, respectively. In the external dataset, the automated measurements displayed a high level of agreement with observer-averaged measurements, with intraclass correlation coefficients ranging from 0.964 to 0.994. Bland–Altman analysis revealed small mean biases across the five shoulder balance parameters, and 90.0 to 98.6% of automated measurements were within the range of interobserver variability. The proposed method provides an efficient and reproducible approach for radiographic shoulder balance assessment and may help reduce observer-dependent measurement variability.

***Keywords*:** adolescent idiopathic scoliosis, automated measurement, neural network, X-ray

1. Introduction

Shoulder balance in adolescent idiopathic scoliosis (AIS) and early-onset scoliosis (EOS) holds immense importance in both diagnostic and treatment contexts, as it influences early detection, treatment planning, prognosis evaluation, and aesthetic outcomes (*1-8*). Precise measurements of shoulder parameters, and particularly the clavicle angle (CA), coracoid height difference (CHD), clavicle tilt angle difference (CTAD), radiological shoulder height (RSH), and T1 tilting angle (T1TA), are crucial for assessing pre- and postoperative shoulder balance (*3,9-12*). However, the assessment of these parameters has been challenging due to the inherent subjectivity and variability associated with manual measurements (*12,13*). This subjectivity not only consumes considerable time but can also lead to overlooked discrepancies, especially during screening processes. A reliable, objective method for their assessment needs to be developed, as it would not only minimize the risk of measurement errors but also enhance the overall management and outcomes of AIS patients.

Achieving automated measurement of shoulder balance parameters requires the semantic segmentation of shoulder structures on X-ray images. This segmentation process should encompass the clavicles, coracoids, and T1 vertebra, as these are the primary anatomical structures commonly used in clinical practice to derive essential shoulder balance parameters. Significant progress has been made in the automated measurement of spinal balance alignment parameters in AIS, and considerable accuracy has been achieved (*14-16*), but identification and segmentation of shoulder

structures have received comparatively less attention in the research setting. Currently, there is a notable absence of established automated methods for measuring shoulder balance parameters. Automatically segmenting shoulder structures in medical images poses challenges due to anatomical overlap, low contrast, potential artifacts, and the need for high-quality annotated data. Moreover, another challenge lies in the recognition of key points within the segmented structures to facilitate parameter calculations.

To address these challenges and the need for automated shoulder balance assessment in AIS, we developed an automated pipeline for segmenting the clavicles, coracoids, and T1 vertebra and quantifying relevant shoulder balance parameters. We validated the proposed method using an external dataset by comparing automated measurements with measurements from three senior spinal surgeons. The evaluation included segmentation performance, landmark detection accuracy, agreement with observer-averaged measurements, and comparison of automated measurement error with interobserver variability.

2. Materials and Methods

2.1. Dataset collection

This retrospective study was approved by the Ethics Committee of Shanghai Tongji Hospital (Approval No. SBKT-2025-319). The Ethics Committee waived the requirement for informed consent because of the retrospective nature of the study. All sensitive patient information was anonymized before data analysis to protect patient privacy. We acquired all X-ray images following established clinical whole spine protocols, utilizing digital X-ray radiography systems including AccE GC85A vision (Samsung Electronics), CXDI-401 (Canon Medical Systems), uDR (United Imaging Healthcare), and Ti-WISH-IL (TaoImage). The images were retrieved and stored in the Digital Imaging and Communications in Medicine (DICOM) 3.0 protocol format from 2017 to 2022. A total of 940 AIS radiographs were initially collected. After quality control, three cases were excluded because of insufficient image quality. Therefore, 937 cases were included in the model-development cohort. The dataset was split at the patient level into a training set of 843 cases and an internal validation set of 94 cases. No patient appeared in more than one subset, thereby avoiding patient-level data leakage. To ensure accuracy, two attending physicians with expertise in musculoskeletal radiology independently annotated the X-ray images using the open-source annotation tool X-Anylabeling (CVHub). Annotations included the precise identification and segmentation of the T1 vertebra and both clavicles and coracoids, with disagreements resolved through consensus.

2.2. Shoulder structure segmentation

We developed a deep learning segmentation model to segment the T1 vertebra, both clavicles, and both coracoids on whole-spine radiographs. The model used RepVGG-A1 (*17*) as the backbone encoder, with the final fully connected classification layer removed. The extracted image features were passed to an upsampling segmentation head to restore the spatial resolution and generate pixel-wise masks for five foreground anatomical classes, including the left clavicle, right clavicle, left coracoid, right coracoid, and T1 vertebra. The network treated the left and right anatomical structures as independent semantic classes to facilitate subsequent landmark extraction and parameter calculation. Figure 1C and 1D show the segmentation results. All of the DICOM radiographs were converted to single-channel grayscale images before model

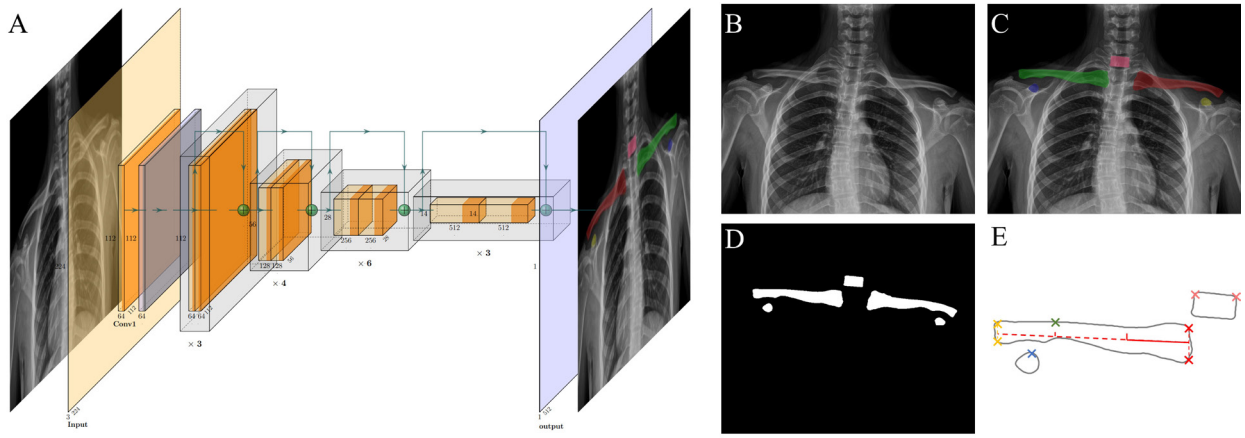


Figure 1. **(A)** Schematic of the segmentation network with a RepVGG-A1 backbone for segmenting the clavicles, coracoids, and T1 vertebra; **(B)** Original X-ray; **(C)** Segment annotations of the clavicles and coracoids; **(D)** Predicted masks of the clavicles and coracoids; **(E)** Key points of shoulder balance including CA, RSH, CHD, CTAD, and T1TA.

training and inference. Image intensities were clipped to reduce the influence of extreme values and then normalized to the range of 0–1. After the initial coarse localization of the T1 vertebra, a local region of interest was cropped from the center of T1. The crop width was set as the original image width, and the crop height was set as half of the crop width. Specifically, 0.2 of the crop height was retained above the center of T1 and 0.8 of the crop height was retained below the center of T1. The cropped local patch was then resized to 512 × 1,024 pixels and used as the input for the final segmentation model.

The model was trained on the 843-patient training set for 20,000 iterations with a batch size of 8. A combined cross-entropy loss and Dice loss was used as the objective function. The cross-entropy loss was used to optimize pixel-wise semantic classification, whereas the Dice loss was used to improve foreground structure overlap and reduce the influence of the class imbalance between the background and relatively small anatomical structures. The total loss was defined as the sum of cross-entropy loss and Dice loss. Common medical image augmentation strategies were applied online during training, including small-angle random rotation within ± 10°, random scaling between 0.9 and 1.1, random translation within ± 5% of the image size, random brightness and contrast adjustment within ± 20%, and mild Gaussian noise. Horizontal flipping was not used because the left and right clavicles and coracoids were defined as separate semantic classes.

The Adam optimizer was used with $\beta1 = 0.9$, $\beta2 = 0.999$, and a cosine annealing schedule was adopted to gradually decrease the learning rate from 1×10^{-3} to 1×10^{-6} over the 20,000 training iterations. Model performance on the internal validation set was monitored during training, and the checkpoint with the highest foreground macro-average Dice coefficient was selected as the final model. All experiments were implemented in Python using PyTorch and performed on an NVIDIA RTX A6000 GPU.

2.3. Shoulder balance parameters

Given the segmentation mask predicted by the network, landmarks were extracted based on morphological image processing, as shown in Figure 1E. The inner and outer upper and lower vertices of the clavicle were detected using a convex point detection algorithm. Moreover, an additional process involved a vertical search upwards from the outer upper vertices on both sides until a notable decrease in grayscale values was observed. This led to the identification of an edge point, signifying the presence of air and corresponding to the acromioclavicular joint. The vertical height difference between these two points was denoted as RSH (Figure 2D), which served as a measure of the vertical soft tissue thickness at the acromioclavicular joints.

The approximate orientation of the clavicle was determined based on the positions of these vertices. Subsequently, the inner one-third median line of the clavicle was calculated. As depicted in Figure 1E, this process entailed determining the midpoint between the inner upper and lower vertices as well as the midpoint between the outer upper and lower vertices. The angle between the inner one-third of the posture line and the horizontal line was then measured, providing the clavicle's angle of inclination. Moreover, the difference in the inclination of the two clavicles was computed and denoted as CTAD (Figure 2B).

The line connecting these two midpoints represented the posture line of the clavicle. In the outer one-third of the clavicle, we searched along the contour for the point farthest from and above the posture line in the vertical direction. This point was identified as the highest point of the clavicle. The angle between the line connecting the highest points on both sides of the clavicle and the

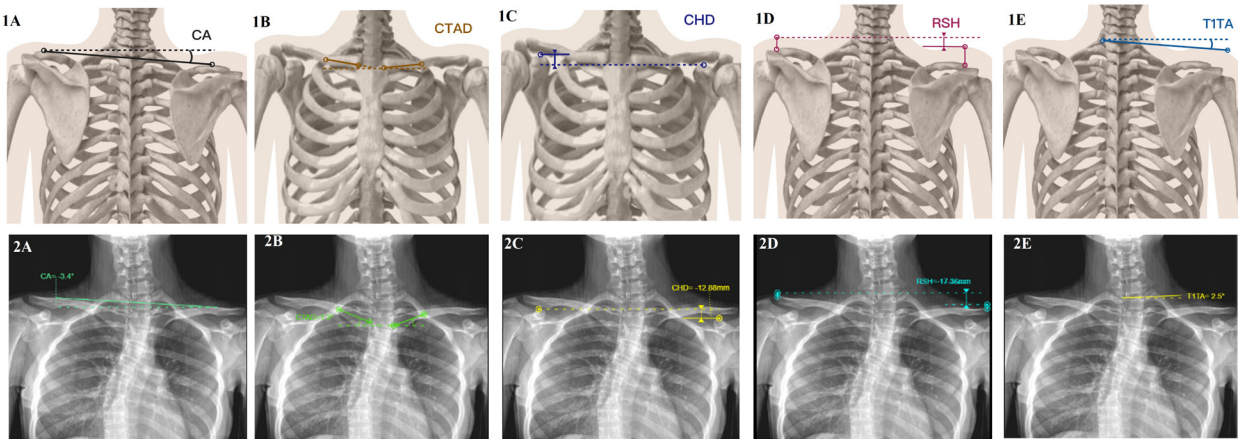


Figure 2. **(1) Schematic of shoulder balance parameters; (2) Self-developed measurement tools for shoulder balance parameters. (A)** Clavicular angle (CA); **(B)** Clavicle tilt angle difference (CTAD); **(C)** Coracoid height difference (CHD); **(D)** Radiological shoulder height (RSH); **(E)** T1 vertebral tilting angle (T1TA).

horizontal line was denoted as CA (Figure 2A).

Similarly, the vertical height difference between the highest points of the two sides of the masks covering the coracoid process segments was referred to as CHD (Figure 2C). In addition, the upper endplates of the T1 vertebra were detected with a convex point detection algorithm. The angle formed by the line connecting these two points and the horizontal line was defined as T1TA (Figure 2E).

For all landmark-based measurements, coordinates detected on resized local patches were first transformed back to the original DICOM image coordinate system using the inverse transformation of cropping and resizing. Linear measurements reported in millimeters, including the MRE, CHD, and RSH, were then calculated using the image-specific DICOM PixelSpacing information. For Euclidean distances such as the MRE, both row and column pixel spacing values were considered. For vertical height differences such as the CHD and RSH, the vertical coordinate difference was multiplied by the row pixel spacing. For angular measurements, landmark coordinates were first transformed into the physical coordinate system before angle calculation, thereby accounting for anisotropic pixel spacing and differences in pixel spacing across radiographic devices and institutions.

2.4. Measurement validation

To validate the accuracy of the automated measurements generated by the deep learning model, we used an external validation dataset consisting of 70 AIS X-ray images collected from Guizhou Provincial People's Hospital. For external segmentation evaluation, the T1 vertebra, both clavicles, and both coracoids in the 70 external radiographs were manually annotated using the same annotation protocol as the model-development cohort. These manual masks served as the reference standard for calculating external segmentation metrics, including recall, precision, IoU, and the Dice coefficient. In the evaluation of our method of automated measurement to detect landmarks, two crucial metrics were used: the successful detection rate (SDR) and the mean radial error (MRE). The SDR represents the percentage of successful detections at different radial error thresholds (2 mm, 3 mm, and 4 mm), while MRE quantifies the mean error in millimeters between landmark positions determined automatically and manually. To minimize the potential biases associated with manual measurements and to streamline the results, we performed a comparative analysis by comparing the positions obtained through the averaging of measurements from three human observers with the positions detected automatically *via* our method.

Three senior spinal surgeons experienced in AIS assessment independently measured the shoulder balance parameters using a custom-developed software platform equipped with specialized measurement tools. The average of the three observers' measurements was used as the observer-averaged reference for agreement analysis. In addition, observer-specific comparisons between the automated measurements and each individual observer's measurements were performed. For these observer-specific comparisons, measurement error was summarized using the mean absolute error (MAE) with standard deviation, root mean square error (RMSE), minimum and maximum absolute errors (MinAE–MaxAE), and median absolute error with interquartile range [median AE (IQR)].

2.5. Statistical analysis

Agreement between automated measurements and manual measurements was evaluated by comparing automated measurements with the average measurements of the three senior spinal surgeons. Intraclass correlation coefficients (ICCs) were calculated using a two-way mixed-effects model for absolute agreement between the automated measurements and observer-averaged reference measurements. Bland–Altman analysis was performed to quantify systematic bias and 95% limits of agreement (LoA). The difference was defined as the automated measurement minus the average measurement of the three observers.

To compare automated measurement error with human observer variability, automated–observer variability was calculated as the mean absolute difference between automated measurements and individual observer measurements. Interobserver variability was calculated as the average pairwise mean absolute difference among the three observers. Interobserver variability was further used as a parameter-specific empirical acceptability threshold. A case within interobserver variability was considered where the absolute difference between the automated measurement and the observer-averaged measurement did not exceed the interobserver variability for the corresponding parameter. Pearson correlation coefficients and p-values from paired *t*-tests were calculated to assess the correlation between automatic measurements and each observer's measurements. In addition, we performed percentage cumulative error analysis. Statistical analysis was performed using the Python package SciPy (version 1.11.3).

3. Results

3.1. Deep learning neural network performance

Table 1 summarizes the segmentation performance of each foreground anatomical structure, including recall, precision, intersection over union (IoU), and the Dice coefficient, on the validation datasets. The left and right clavicles performed well, with recalls of 0.95 and 0.95,

precision scores of 0.92 and 0.94, IoU values of 0.88 and 0.89, and Dice coefficients of 0.94 for both sides. However, segmentation of the left and right coracoids yielded slightly lower scores, with a recall of 0.76 and 0.75, precision scores of 0.82 and 0.85, IoU values of 0.65 and 0.66, and Dice coefficients of 0.79 for both sides. The T1 vertebra performed extremely well, with a recall of 0.89, precision of 0.87, IoU of 0.79, and Dice coefficient of 0.88. The foreground macro-average recall, precision, IoU, and Dice coefficient in the internal validation dataset were 0.86, 0.88, 0.77, and 0.87, respectively. Because the background class was excluded from the average, these values represent the segmentation performance of clinically relevant foreground anatomical structures.

In the external validation dataset, the model displayed slightly lower but comparable segmentation performance. The left and right clavicles resulted in IoUs of 0.84 and 0.85 and Dice coefficients of 0.91 and 0.92, respectively. The left and right coracoids resulted in IoUs of 0.60 and 0.61 and Dice coefficients of 0.75 and 0.76, respectively. The T1 vertebra resulted in an IoU of 0.74 and a Dice coefficient of 0.85. The foreground macro-average recall, precision, IoU, and Dice coefficient in the external validation dataset were 0.83, 0.85, 0.73, and 0.84, respectively. The total processing time was 89.9 ± 18.3 ms per image.

The 2-mm, 3-mm, and 4-mm SDR and the MRE of the average landmarks in the external dataset are shown in Table 2. The superior coracoid points achieved an SDR of 72.62% at a 2-mm threshold, which increased to 79.10% at the 3-mm threshold and 91.65% at the 4-mm threshold. The MRE for these points was 1.83 mm. The superior clavicle points achieved SDR values of 66.70% (2-mm threshold), 83.55% (3-mm threshold), and 87.89% (4-mm threshold). The MRE for these points was 2.75 mm. The superior extraclavicular points on both sides achieved SDR values of 77.33% (2 mm), 83.10% (3 mm), and 92.65% (4 mm). The MRE for these points was 1.59 mm. The soft tissue shoulder points achieved SDR values of 69.68% (2 mm), 78.31% (3 mm), and 92.26% (4 mm), with an MRE of 1.92 mm. Finally, the superior T1 vertebral point achieved SDR values of 65.21% (2 mm), 75.24% (3 mm), and 90.54% (4 mm) along with an MRE of 2.62 mm.

3.2. Agreement and error analysis

The descriptive statistics of shoulder balance parameters in the external dataset are summarized in Table 3. On average, the CA measurement had a mean value of 0.73°, ranging from −3.9 to 12.1°. The mean absolute CA was 2.11°, ranging from 0.2 to 12.1°. The mean CHD was 3.14 mm, ranging from −16.8 to 42.3 mm. The mean

Table 1. Segmentation metrics of each class in the internal and external validation datasets

	Recall	Precision	IoU	Dice
Clavicle (left)	0.95/0.93	0.92/0.90	0.88/0.84	0.94/0.91
Clavicle (right)	0.95/0.93	0.94/0.91	0.89/0.85	0.94/0.92
Coracoid (left)	0.76/0.72	0.82/0.79	0.65/0.60	0.79/0.75
Coracoid (right)	0.75/0.71	0.85/0.81	0.66/0.61	0.79/0.76
T1 vertebrae	0.89/0.86	0.87/0.84	0.79/0.74	0.88/0.85
Average	0.86/0.83	0.88/0.85	0.77/0.73	0.87/0.84

Notes: Values are presented as internal validation/external validation.

Table 2. The 2-mm, 3-mm, 4-mm success detection rate (SDR) and the mean radial error (MRE) of the average landmarks in the external dataset

	2 mm SDR (%)	3 mm SDR (%)	4 mm SDR (%)	MRE (mm)
Superior coracoid points	72.62	79.10	91.65	1.83
Superior clavicle points	66.70	83.55	87.89	2.75
Superior extraclavicular points	77.33	83.10	92.65	1.59
Soft tissue shoulder points	69.68	78.31	92.26	1.92
Superior T1 vertebral points	65.21	75.24	90.54	2.62

Table 3. Summary of the mean and range for shoulder balance parameters in the external dataset

	Mean	Range	Mean (Absolute)	Range (Absolute)
CA (°)	0.73 ± 2.76	−3.9-12.1	2.11 ± 1.90	0.2-12.1
CHD (mm)	3.14 ± 11.30	−16.8-42.3	9.11 ± 7.32	0.0-42.3
CTAD (°)	−1.85 ± 5.90	−24.8-9.2	4.61 ± 4.10	0.1-24.8
RSH (mm)	1.47 ± 12.88	−24.2-42.7	10.12 ± 8.02	0.3-42.7
T1TA (°)	1.11 ± 4.39	−16.9-9.8	3.43 ± 2.93	0.0-16.9

absolute CHD was 9.11 mm, ranging from 0.0 to 42.3 mm. The mean CTAD was −1.85°, ranging from −24.8 to 9.2°. The mean absolute CTAD was 4.61°, ranging from 0.1 to 24.8°. The mean RSH was 1.47 mm, ranging from −24.2 to 42.7 mm. The mean absolute RSH was 10.12 mm, ranging from 0.3 to 42.7 mm. The mean T1TA was 1.11°, ranging from −16.9 to 9.8°. The mean absolute T1TA was 3.43°, ranging from 0.0° to 16.9°.

We assessed the reliability of these measurements by examining interobserver agreement among three different observers. The statistical analysis revealed strong positive relationships between measurements made by different observers, as indicated by high Pearson's correlation coefficients (*r*) ranging from 0.90 to 0.99 for all parameters. Moreover, paired *t*-tests showed no statistically significant systematic differences between observer measurements for most parameters; however, agreement was primarily evaluated using intraclass correlation coefficients and Bland–Altman analyses. The agreement between automatic measurements and each observer is shown in Table 4.

The distribution of absolute measurement errors between the automated method and each individual observer's measurements is summarized in Table 5. Across the three observers, the MAE ranged from 0.67 to 0.97° for CA, the CHD ranged from 1.00 to 2.56 mm, the CTAD ranged from 2.15 to 2.72°, the RSH ranged from 1.25 to 2.56 mm, and the T1TA ranged from 0.58 to 1.10°. The observer-specific RMSE, MinAE–MaxAE, and median AE [IQR] are also reported in Table 5.

Agreement analysis was performed between the automated measurements and the average measurements

Table 4. Pearson correlation coefficients and paired *t*-tests between automated measurements and each observer

	Observer 1		Observer 2		Observer 3	
	Pearson's *r*	Paired *t*-test (*p* value)	Pearson's *r*	Paired *t*-test (*p* value)	Pearson's *r*	Paired *t*-test (*p* value)
CA (°)	0.92	0.35	0.91	0.38	0.96	0.74
CHD (mm)	0.96	0.74	0.98	0.08	0.99	0.37
CTAD (°)	0.90	0.91	0.90	0.40	0.92	0.28
RSH (mm)	0.99	0.40	0.97	0.26	0.97	0.96
T1TA (°)	0.97	0.23	0.99	0.18	0.95	0.12

Table 5. Distribution of measurement errors between automated measurements and each observer

	Observer 1			
	MAE ± SD	RMSE	MinAE–MaxAE	Median AE [IQR]
CA (°)	0.75 ± 0.68	1.01	0.00–4.10	0.60 [0.30–1.00]
CHD (mm)	2.56 ± 2.03	3.26	0.02–9.00	2.09 [1.35–3.37]
CTAD (°)	2.15 ± 1.78	2.78	0.00–8.70	1.65 [1.00–2.67]
RSH (mm)	1.25 ± 1.01	1.60	0.05–3.88	1.09 [0.45–1.71]
T1TA (°)	0.86 ± 0.66	1.08	0.00–2.80	0.70 [0.40–1.20]
	Observer 2			
	MAE ± SD	RMSE	MinAE–MaxAE	Median AE [IQR]
CA (°)	0.97 ± 0.74	1.22	0.00–3.20	0.80 [0.40–1.40]
CHD (mm)	1.47 ± 0.92	1.73	0.04–3.76	1.57 [0.59–2.11]
CTAD (°)	2.72 ± 1.97	3.36	0.10–8.30	2.20 [1.20–3.98]
RSH (mm)	2.34 ± 1.95	3.04	0.01–9.23	2.05 [0.73–3.13]
T1TA (°)	0.58 ± 0.37	0.68	0.00–1.40	0.50 [0.30–0.87]
	Observer 3			
	MAE ± SD	RMSE	MinAE–MaxAE	Median AE [IQR]
CA (°)	0.67 ± 0.48	0.82	0.00–2.40	0.60 [0.30–0.97]
CHD (mm)	1.00 ± 0.76	1.25	0.00–2.90	0.88 [0.43–1.38]
CTAD (°)	2.15 ± 1.45	2.58	0.10–5.60	1.90 [0.93–3.15]
RSH (mm)	2.56 ± 1.72	3.08	0.01–6.73	2.35 [1.27–3.66]
T1TA (°)	1.10 ± 0.86	1.39	0.00–3.60	0.90 [0.50–1.48]

Notes: Absolute errors were calculated between automated measurements and each individual observer. Median AE and the IQR were calculated from the absolute errors. *Abbreviations*: AE, absolute error; MAE, mean absolute error; SD, standard deviation; IQR, interquartile range; RMSE, root mean square error.

Table 6. Agreement between automated measurements and observer-averaged measurements

	ICC	Mean bias	95% LoA	Automated–observer variability	Within interobserver variability rate
CA (°)	0.974	0.09	−1.11-1.30	0.80	63/70 (90.0%)
CHD (mm)	0.994	0.12	−2.34-2.58	1.67	69/70 (98.6%)
CTAD (°)	0.964	0.21	−3.06-3.48	2.34	67/70 (95.7%)
RSH (mm)	0.993	-0.18	−3.29-2.92	2.05	66/70 (94.3%)
T1TA (°)	0.991	-0.07	−1.14-1.00	0.85	68/70 (97.1%)

Notes: Mean bias and 95% LoA were calculated as automated measurements minus the average measurements made by the three observers. Automated–observer variability was expressed as the mean absolute difference between automated measurements and individual observer measurements. Interobserver variability was expressed as the average pairwise mean absolute difference among the three observers. Within interobserver variability was defined as the proportion of cases in which the absolute difference between the automated measurement and the observer-averaged measurement did not exceed the interobserver variability for the corresponding parameter. *Abbreviations*: ICC, intraclass correlation coefficient; LoA, limits of agreement.

of the three observers. The ICCs for absolute agreement ranged from 0.964 to 0.994, indicating high agreement across all shoulder balance parameters, as shown in Table 6. Bland–Altman analysis revealed small mean biases, including 0.09° for the CA, 0.12 mm for the CHD, 0.21° for the CTAD, −0.18 mm for the RSH, and −0.07° for the T1TA. The corresponding 95% LoA was −1.11 to 1.30° for the CA, −2.34 to 2.58 mm for the CHD, −3.06 to 3.48° for the CTAD, −3.29 to 2.92 mm for the RSH, and −1.14 to 1.00° for the T1TA. The corresponding correlation scatter plots, Bland–Altman plots, and cumulative error curves are shown in Figure 3.

The automated–observer variability was smaller than the corresponding interobserver variability for all five parameters. Specifically, the automated–observer variability values were 0.80° for the CA, 1.67 mm for the CHD, 2.34° for the CTAD, 2.05 mm for the RSH, and 0.85° for the T1TA, whereas the corresponding interobserver variability values were 1.11°, 2.44 mm, 3.25°, 2.93 mm, and 1.31°, respectively. When interobserver variability was used as a parameter-specific empirical threshold, 90.0% of CA, 98.6% of CHD, 95.7% of CTAD, 94.3% of RSH, and 97.1% of T1TA measurements were within the range of interobserver variability.

4. Discussion

In this study, we developed an automated deep learning-based method for shoulder balance assessment in AIS radiographs and evaluated its performance at the segmentation, landmark detection, and measurement levels. For anatomical structure segmentation, the model achieved foreground macro-average IoU values of 0.77 and 0.73 and Dice coefficients of 0.87 and 0.84 in the internal and external validation datasets, respectively, indicating stable segmentation performance across the two validation cohorts. Among the segmented structures, the clavicles performed best, with IoUs of 0.88–0.89 and Dice coefficients of 0.94 in the internal validation dataset and IoUs of 0.84–0.85 and Dice coefficients of 0.91–0.92 in the external validation dataset. The T1 vertebra also displayed acceptable segmentation performance, with IoUs of 0.79 and 0.74 and Dice coefficients of 0.88 and 0.85 in the internal and external datasets, respectively. In contrast, the coracoids displayed relatively lower segmentation performance, with IoUs of 0.65–0.66 internally and 0.60–0.61 externally, reflecting the difficulty of segmenting small anatomical structures with overlapping radiographic projections. Previous studies have shown that small structures and class imbalance can disproportionately affect overlap-based metrics such as the IoU and Dice coefficient (*18-20*). Therefore, the relatively lower coracoid IoU and Dice values should be interpreted together with landmark-level and measurement-level performance. Importantly, the CHD depends primarily on the localization of the superior coracoid landmarks rather than complete mask overlap. In this study, the superior coracoid landmarks achieved an MRE of 1.83 mm and an SDR of 91.65% at the 4-mm threshold. The CHD also showed a small mean bias of 0.12 mm, a 95% LoA of −2.34 to 2.58 mm, and 98.6% of cases were within interobserver variability. These findings suggest that although coracoid segmentation remained challenging, its influence on CHD measurement was limited in the present validation cohort. At the measurement level, the automated method demonstrated a high level of agreement with observer-averaged measurements, with ICCs ranging from 0.964 to 0.994 across all five shoulder balance parameters. Bland–Altman analysis revealed small mean biases, including 0.09° for the CA, 0.12 mm for the CHD, 0.21° for the CTAD, −0.18 mm for the RSH, and −0.07° for the T1TA. Moreover, the automated–observer variability was lower than the corresponding interobserver variability for all parameters, and 90.0% to 98.6% of automated measurements were within the range of interobserver variability. These findings suggest that the proposed method achieved not only reliable anatomical segmentation but also measurement accuracy comparable to manual assessment by experienced observers.

The precise segmentation of shoulder structures has the potential for valuable applications in clinical tasks such as fracture diagnosis, joint assessment, and surgical planning. The Japanese Society of Radiological

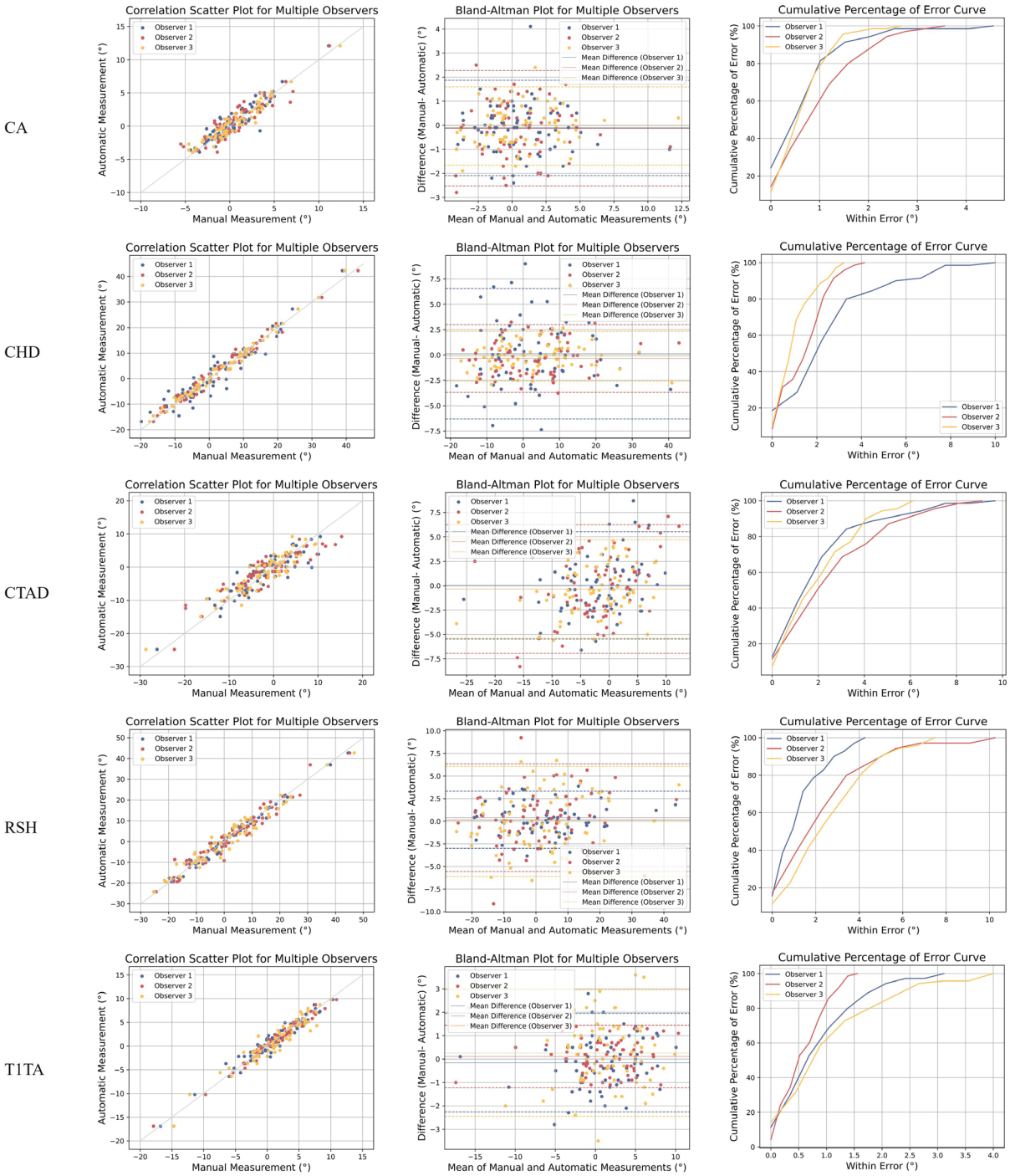


Figure 3. Correlation scatter diagrams (Left), Bland–Altman difference diagrams (Middle) and Cumulative percentage of error curves (Right) for CA, CHD, CTAD, RSH and T1TA measured automatically and by each observer.

Technology (JSRT) dataset (*21*), which includes lung contours, heart contours, and clavicle annotations within the lung fields, has been widely used in studies of clavicle segmentation. van Ginneken *et al*. (*22*) conducted a comparative study that used three supervised segmentation methods on the JSRT dataset: active shape models, active appearance models, and a multi-resolution pixel classification method utilizing Gaussian derivative filters and k-nearest neighbor classification. That method achieved an IoU of 0.736, compared to 0.896 for a human observer. Fully convolutional network (FCN)-based segmentation subsequently improved the reported IoU to 0.868 (*23*). A semi-supervised approach further improved the reported IoU to 0.881 (*24*). Wang *et al*. (*25*) introduced a multi-object segmentation method based on collaborative learning with multiple teacher models. Their model was trained using four heterogeneous partially labeled datasets that included the RCS-CXR dataset, which contains complete clavicle annotations as well as anterior and

posterior rib segmentation labels (*26*). The achieved results, with a Dice coefficient of 0.95 and an IoU of 0.91, outperformed the current state-of-the-art reported in the literature. In comparison, our model achieved Dice coefficients of 0.94 and 0.91–0.92 and IoUs of 0.88–0.89 and 0.84–0.85 for clavicle segmentation in the internal and external validation datasets, respectively.

Moreover, automated measurement of shoulder balance parameters can enable the screening of large cohorts in a reasonable timeframe with good reliability. The T1TA and RSH demonstrated better reader-agreement in previous studies (*9*), while a considerable variation in the RSH and a reduced variation in the T1TA were evident in our automated measurements. Previous studies reported that the CA and CHD displayed a high level of reliability among observers, with MAE values comparable to those observed in our study (*13*). Previous studies have indicated that shoulder balance should be considered a crucial factor in surgical planning and prognosis. The RSH was significantly correlated with the occurrence of the adding-on phenomenon in the shoulder imbalance group at follow-up (*1*). Moreover, there is a significant correlation between the T1TA just after surgery as well as the CA and the recurrence of shoulder imbalance during the 1-year follow-up in AIS patients. However, there is currently a lack of widely accepted diagnostic and assessment standards for shoulder balance parameters (*27*), and more population-based diagnostic studies need to be conducted to elucidate those parameters.

Several biases and limitations may have influenced the interpretation of our findings. Our dataset was obtained from a single institution, which may not represent the broader population of AIS patients. The data may be subject to selection bias based on the patient demographics, geographical location, or referral patterns. The performance of the deep learning algorithm can be influenced by factors such as image quality, patient positioning, and image artifacts. Additionally, when considering shoulder balance assessment, the positioning of patients during X-ray imaging is of paramount importance (*28*). These methods may yield measurements that do not fully reflect the true shoulder balance, as they are based on radiographic measurements rather than actual photographic measurements (*29*). Accurate measurement of X-ray parameters must be complemented by the identification of anatomical landmarks on the patient's body surface and an overall visual assessment. These factors collectively contribute to the comprehensive evaluation of shoulder balance, emphasizing the need for a holistic approach beyond just numerical measurements. In addition, the empirical acceptability thresholds used in this study were derived from interobserver variability among three senior spinal surgeons rather than from universally established clinical thresholds. Although this approach provides an observer-based reference for measurement acceptability, these thresholds should be further validated in larger multicenter cohorts. Future studies should also evaluate how automated measurement errors may affect shoulder balance classification and treatment decision-making.

In conclusion, our deep learning-based automated method provides a reliable and efficient approach for radiographic shoulder balance assessment in AIS patients. By reducing observer-dependent measurement variability and improving measurement reproducibility, this method may aid in clinical assessment, follow-up evaluation, and large-scale radiographic screening. Further multicenter validation is warranted to determine its impact on shoulder balance classification and treatment decision-making.

Funding: This study was funded by the National Key Research and Development Program of China (Grant No. 2025YFE0214102).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Cao K, Watanabe K, Hosogane N, Toyama Y, Yonezawa I, Machida M, Yagi M, Kaneko S, Kawakami N, Tsuji T, Matsumoto M. Association of postoperative shoulder balance with adding-on in Lenke type II adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2014; 39:E705-E712.
2. Chang DG, Kim JH, Kim SS, Lim DJ, Ha KY, Suk SI. How to improve shoulder balance in the surgical correction of double thoracic adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2014; 39:E1359-E1367.
3. Kuklo TR, Lenke LG, Graham EJ, Won DS, Sweet FA, Blanke KM, Bridwell KH. Correlation of radiographic, clinical, and patient assessment of shoulder balance following fusion versus nonfusion of the proximal thoracic curve in adolescent idiopathic scoliosis. Spine (Phila Pa 1976). 2002; 27:2013-2020.
4. Elsebaie HB, Dannawi Z, Altaf F, Zaidan A, Al Mukhtar M, Shaw MJ, Gibson A, Noordeen H. Clinically orientated classification incorporating shoulder balance for the surgical treatment of adolescent idiopathic scoliosis. Eur Spine J. 2016; 25:430-437.
5. Hong JY, Suh SW, Modi HN, Yang JH, Park SY. Analysis of factors that affect shoulder balance after correction surgery in scoliosis: A global analysis of all the curvature types. Eur Spine J. 2013; 22:1273-1285.
6. Uzümcügil O, Atici Y, Ozturkmen Y, Yalcinkaya M, Caniklioglu M. Evaluation of shoulder balance through growing rod intervention for early-onset scoliosis. J Spinal Disord Tech. 2012; 25:391-400.
7. Kurra S, Cahill PJ, Albanese SA, Betz RR, Toole T, Lavelle WF. Evaluation of shoulder balance in early onset scoliosis after definitive fusion and comparison with adolescent idiopathic scoliosis shoulder balance. Spine Deform. 2022; 10:183-188.
8. Gotfryd AO, Silber Caffaro MF, Meves R, Avanzi O. Predictors for postoperative shoulder balance in Lenke 1 adolescent idiopathic scoliosis: A prospective cohort

study. Spine Deform. 2017; 5:66-71.
9. Bagó J, Carrera L, March B, Villanueva C. Four radiological measures to estimate shoulder balance in scoliosis. J Pediatr Orthop B. 1996; 5:31-34.
10. Chiu CK, Chan CYW, Tan PH, Goh SH, Ng SJ, Chian XH, Ng YH, Ler XY, Chandren JR, Chung WH, Kwan MK. Conformity and changes in the radiological neck and shoulder balance parameters throughout 3-year follow-up period: Do they remain the same? Spine (Phila Pa 1976). 2020; 45:E319-E328.
11. Luhmann SJ, Sucato DJ, Johnston CE, Richards BS, Karol LA. Radiographic assessment of shoulder position in 619 idiopathic scoliosis patients: Can T1 tilt be used as an intraoperative proxy to determine postoperative shoulder balance? J Pediatr Orthop. 2016; 36:691-694.
12. Akel I, Pekmezci M, Hayran M, Genc Y, Kocak O, Derman O, Erdogan I, Yazici M. Evaluation of shoulder balance in the normal adolescent population and its correlation with radiological parameters. Eur Spine J. 2008; 17:348-354.
13. Hong JY, Suh SW, Yang JH, Park SY, Han JH. Reliability analysis of shoulder balance measures: Comparison of the 4 available methods. Spine (Phila Pa 1976). 2013; 38:E1684-E1690.
14. Meng N, Cheung JPY, Wong KYK, Dokos S, Li S, Choy RW, To S, Li RJ, Zhang T. An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. EClinicalMedicine. 2022; 43:101252.
15. Wang L, Xie C, Lin Y, *et al*. Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior X-Ray images: The AASCE2019 challenge. Med Image Anal. 2021; 72:102115.
16. Zhang K, Xu N, Guo C, Wu J. MPF-net: An effective framework for automated Cobb angle estimation. Med Image Anal. 2022; 75:102277.
17. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. RepVGG: Making VGG-style ConvNets great again. CVPR 2021. 2021; 13733-13742.
18. Mun C, Lee S, Uh Y, Choe J, Byun H. Small objects matters in weakly-supervised semantic segmentation. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, Waikoloa, HI, USA, 2024; 414-423.
19. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. 2022; 15:210.
20. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015; 15:29.
21. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, Doi K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am J Roentgenol. 2000; 174:71-74.
22. van Ginneken B, Stegmann MB, Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. Med Image Anal. 2006; 10:19-40.
23. Novikov AA, Lenis D, Major D, Hladuvka J, Wimmer M, Buhler K. Fully convolutional architectures for multiclass segmentation in chest radiographs. IEEE Trans Med Imaging. 2018; 37:1865-1876.
24. Bortsova G, Dubost F, Hogeweg L, Katramados I, de Bruijne M. Semi-supervised medical image segmentation *via* learning consistency under transformations. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, eds.). Springer International Publishing, Cham, 2019; pp. 810-818.
25. Wang H, Zhang D, Feng J, Cascone L, Nappi M, Wan S. A multi-objective segmentation method for chest X-rays based on collaborative learning from multiple partially annotated datasets. Information Fusion. 2023; 102:102016.
26. Wang W, Feng H, Bu Q, Cui L, Xie Y, Zhang A, Feng J, Zhu Z, Chen Z. MDU-Net: A convolutional network for clavicle and rib segmentation from a chest radiograph. J Healthc Eng. 2020; 2020:2785464.
27. Clement RC, Anari J, Bartley CE, Bastrom TP, Shah R, Talwar D, Upasani VV. What are normal radiographic spine and shoulder balance parameters among adolescent patients? Spine Deform. 2020; 8:621-627.
28. Marks MC, Stanford CF, Mahar AT, Newton PO. Standing lateral radiographic positioning does not represent customary standing balance. Spine (Phila Pa 1976). 2003; 28:1176-1182.
29. Qiu X, Ma W, Li W, Wang B, Yu Y, Zhu Z, Qian B, Zhu F, Sun X, Ng BKW, Cheng JCY, Qiu Y. Discrepancy between radiographic shoulder balance and cosmetic shoulder balance in adolescent idiopathic scoliosis patients with double thoracic curve. Eur Spine J. 2009; 18:45-51.

Received May 15, 2026; Revised June 11, 2026; Accepted June 15, 2026.

Released online in J-STAGE as advance publication June 20, 2026.

[§]*These authors contributed equally to this work.*
**Address correspondence to*:
Yan Yu, Division of Spine, Department of Orthopaedics, Tongji Hospital, Tongji University School of Medicine, Tongji University, 389 Xincun Road, Putuo District, Shanghai 200065, China.
E-mail: yyu15@tongji.edu.cn

Global Health & Medicine. 2026; 8(3):215-221. PERSPECTIVE
DOI: 10.35772/ghm.2026.01063

From companion technologies to social care infrastructure: A multi-level perspective on loneliness-related support in dementia care in an era of artificial intelligence (AI)

Machiko Uenishi[1], Peipei Song[1,2,*]

[1] Division of Global Health & Medicine, Bureau of Health Emergency and Management, Japan Institute for Health Security, Tokyo, Japan;
[2] National College of Nursing, Japan (NCNJ), Japan Institute for Health Security, Tokyo, Japan.

Abstract: Loneliness and social isolation are important psychosocial concerns in dementia care, but they are difficult to address through pharmacological treatment or episodic social activities alone. Companion technologies, including socially assistive robots, humanoid and conversational robots, avatars, virtual agents, and artificial intelligence (AI)-enabled companions, are increasingly discussed as possible aids for engagement and interaction. However, current evidence does not justify treating these technologies as standalone interventions that directly reduce loneliness in people with dementia. We argue that these technologies may be more appropriately considered not only as devices or interventions, but also as potential components of social care infrastructure: tools that can encourage individual engagement and mediate human relationships but that need to be integrated into care ecosystems. We propose the multi-level care with social technologies (MCST) model, which consists of three connected layers: individual engagement, relational interaction, and the care ecosystem. The model emphasizes that loneliness-related support is produced through the interaction between technology, human facilitation, care workflows, ethical governance, and feedback-based adjustment. This approach is especially relevant as dementia care systems face workforce constraints and an increasing need for home-based care, while psychosocial needs remain and may be overlooked in routine care. AI-enabled and large-language-model-based companions may expand possibilities for personalization and continuous engagement, but dementia-specific evidence remains preliminary and safety concerns are substantial. Future research should validate the MCST model in real-world dementia care and establish evaluation frameworks that address psychosocial outcomes, sustained use, safety, privacy, human oversight, and accountability.

***Keywords*:** dementia, loneliness-related distress, companion technologies, social care infrastructure, artificial intelligence (AI), perspective

1. Introduction

According to the World Health Organization (WHO), 57 million people were living with dementia globally in 2021, with approximately 10 million new cases occurring each year (*1*). As the number of people living with dementia continues to increase, dementia care increasingly requires attention to psychosocial needs that extend beyond cognition, diagnosis, and pharmacological symptom management. Loneliness and social isolation are particularly relevant because they are associated with depressive symptoms, reduced participation, poorer mental and physical health, and cognitive decline (*2-5*). Loneliness refers to the subjective gap between desired and actual relationships, whereas social isolation refers to limited social contact, social participation, interaction, or networks (*3,6*). For people with dementia or mild cognitive impairment, these experiences may be intensified by memory loss, difficulty communicating, reduced confidence, stigma, loss of one's previous role, and dependence on family members or professional caregivers.

Although loneliness is subjective, it is not merely an individual emotional state. It is also shaped by relational and environmental conditions. A person with dementia may feel lonely despite living with others if communication is difficult or interactions become task-focused. Conversely, supportive engagement may be possible even in institutional or home-based settings when care routines, family involvement, and activity opportunities facilitate connection. This means that loneliness-related support is a practical issue for dementia care systems to address: it requires sustained recognition of changing emotional responses and not

merely the provision of occasional social contact.

Companion technologies have attracted attention as possible non-pharmacological approaches to address that issue. These include animal-like socially assistive robots, humanoid robots, conversational robots, artificial intelligence (AI)-enabled avatars, virtual agents, and other digital companions designed to stimulate interaction, provide comfort, or encourage participation in activities (*7,8*). Japan provides an important context for examining companion technologies in dementia care. In 2023, among the 19 OECD countries with available comparable data, Japan had the highest estimated prevalence of dementia among people age 65 years and over at 122 cases per 1,000 population, which is equivalent to approximately 12.2% of this age group (*9*). Japan has also developed dementia policy initiatives and has a long history of care-robot development, including PARO, an animal-like socially assistive robot (*10,11*). These conditions make Japan a relevant setting for considering how companion technologies can be integrated into dementia care systems. And yet the central question is not whether a device can alleviate loneliness by itself. A more relevant question is how such technologies can be embedded within human relationships and care systems to address loneliness-related needs safely, ethically, and sustainably.

This perspective therefore builds on existing discussions of assistive technologies, social care, and relationship-centered dementia care by considering companion technologies not only as devices or interventions but also as potential components of social care infrastructure. To elucidate this perspective, we propose the multi-level care with social technologies (MCST) model. The model clarifies how companion technologies may contribute to individual engagement, relational interaction, and care ecosystem governance, while informing research design, institutional decision-making, and care policy without overstating current evidence.

2. Current evidence and its limits in dementia care

When adopting the aforementioned perspective, we use a pragmatic functional classification informed by existing reviews of socially assistive robots and their applications in dementia care (*7,8,12*). We group companion technologies into four broad categories: animal-like socially assistive robots, humanoid robots, conversational robots, and AI-enabled or large language model (LLM)-based companion technologies. These categories are not mutually exclusive, as some devices combine embodiment, programmed dialogue, and AI-enabled interaction depending on their functions and implementation contexts. Their potential roles, risks, and implementation requirements across the three layers of the MCST model are summarized in Table 1.

Existing evidence suggests that companion technologies may help to achieve loneliness-related psychosocial outcomes, but it does not indicate that they are direct solutions for loneliness in dementia care (*8,12-14*). Among the technologies considered here, animal-like socially assistive robots, and PARO in particular, currently have the strongest evidence base (*11-14*). Their key mechanism is tactile and non-verbal emotional engagement through holding, stroking, sound, and responsive behavior (*11*). Studies and reviews have reported possible benefits in terms of anxiety, agitation, depressive symptoms, sleep, sociability, and the caregiver burden (*11-14*). However, findings on behavioral and psychological symptoms of dementia, cognition, quality of life, and loneliness itself remain inconsistent. Differences in intervention duration, human facilitation, frequency of use, dementia severity, and care setting limit generalizability and prevent the reaching of definite conclusions about a direct reduction in loneliness (*12-14*).

Humanoid robots and communication robots, such as Pepper and RoBoHoN, may encourage participation in activities, engagement in rehabilitation, conversation, and interaction with caregivers or other residents (*15-17*). Their value often lies in embodied social cues, predictable routines, and the ability to become a shared focus of attention (*18*). Conversational robots may also provide games, programmed dialogue, or simple cognitive and behavioral stimulation (*19,20*). Nevertheless, these technologies are highly dependent on usability, speech recognition, staff support, cost, maintenance, and inclusion in daily workflows (*17,21*). If these conditions are not met, the technology may be underused or may add a burden rather than provide support.

AI-enabled companions, including conversational agents, avatars, android robotic media, and potential large-language-model-based systems, create new possibilities for more personalized and continuous engagement. They could adapt the content of a conversation to the user's life history, preferences, emotional state, and response patterns. Early exploratory studies suggest possible effects on loneliness-related distress, anxiety, depressive symptoms, or communication, but the evidence remains preliminary and heterogeneous (*20,22*). Importantly, current dementia-specific evidence for generative AI and large language models is still insufficient. Their potential should therefore be presented as a future direction requiring validation, not as established effectiveness.

Across technology types, the consistent lesson is that benefits are unlikely to emerge from the device alone. Companion technologies may provide opportunities for engagement, but the meaning and safety of that engagement depend on human interpretation, facilitation, and supervision. They should therefore be evaluated beyond device-level outcomes, including their effects on relationships, care routines, staff workload, family involvement, and long-term sustainability. This supports

Table 1. Companion technologies in dementia care: roles, mechanisms, and implementation requirements

Technologies	Examples	Intended users	Care settings	Key mechanisms	Contributions	Implementation requirements
Animal-like socially assistive robots	PARO	People with dementia, including those with limited verbal communication or advanced cognitive impairment	Long-term care facilities, residential care settings, day-care centers	Tactile and non-verbal emotional engagement; sensory comfort through touching, holding, and stroking	Reducing anxiety, agitation, or depressive symptoms in some users; aiding comfort and sociability	Identification of appropriate users; regular but not excessive use; caregiver facilitation; monitoring of individual responses; integration into daily care routines
Humanoid robots	Pepper RoBoHoN	People with mild to moderate dementia; older adults who can respond to visual, verbal, or social cues; caregivers and staff involved in activities	Hospitals, rehabilitation units, long-term care facilities, day-care centers, residential care settings	Embodied social cues; voice interaction; gestures; facilitation of activity	Encouraging engagement in rehabilitation; initiating interactions with users, caregivers, and other residents	Staff training; cost and maintenance planning; workflow integration; adjustment to the user's cognitive status and user acceptance
Conversational robots	Pepper RoBoHoN	People with mild cognitive impairment or mild to moderate dementia; users who can engage in simple verbal interaction	Home care, residential care settings, day-care centers, long-term care facilities	Simple voice interaction; cognitive stimulation; predictable communication	Providing programmed conversations, games, and repeated interaction; offering cognitive and behavioral stimulation	Consistent speech recognition; design of simple and comprehensible interactions; personalization to the user's cognitive level; human monitoring; evaluation of sustained use
AI-enabled conversational companions	AI-enabled conversational robots, avatars, virtual agents, android robotic media	Potentially people with mild cognitive impairment or early-stage dementia, older adults at risk of loneliness, family caregivers, and care staff	Home care, residential care settings, digitally supported community care	AI-mediated dialogue; personalization; conversational continuity; data-informed engagement	Providing individualized and context-sensitive interaction; encouraging reminiscence and emotional expression; potentially alleviating loneliness-related distress and depressive symptoms; enabling data-informed modification of the level of engagement	Human supervision; control of hallucinations; privacy and consent safeguards; audit logs; escalation procedures; accountability framework; real-world validation before clinical or care-system claims

Data source: *Ref.* (*7,8,11-22,26-29*).

a shift from a device-centered view to an infrastructure-centered view.

3. Companion technologies as potential social care infrastructure

The concept of social care infrastructure emphasizes that companion technologies become meaningful when embedded in care relationships and organizational routines, where they aid with human care rather than replacing it. In dementia care, this may include prompts for conversation, opportunities for shared activities, comfort during periods of isolation, support for reminiscence, or signals that help caregivers notice changes in mood, responsiveness, or participation (*11,19,20,22*).

This way of viewing companion technologies is useful because loneliness-related needs are multidimensional (*2,3,23-25*). On the individual level, a person may need emotional reassurance, sensory comfort, or stimulation. On the relational level, the person may need mediated opportunities to interact with family, staff, or other users. On the ecosystem level, the care setting needs workflows, supervision, training, and governance to ensure that technologies are used appropriately. A companion robot or AI avatar cannot deal with these levels automatically. Its value depends on whether it is integrated into a broader care process.

Viewing companion technologies as social care infrastructure also prevents overly broad claims. A positive short-term response to a robot does not prove that loneliness has been reduced. Similarly, increased interaction during a session does not necessarily mean sustained social connection has taken place. The more appropriate claim is that these technologies may create entry points for engagement and relationship-building, which may in turn contribute to alleviating loneliness-related psychological distress in some users when supported by a responsive care environment.

This infrastructure perspective also distinguishes expectations among stakeholders: family members may seek reassurance, care staff may expect support for group activities or management of agitation, and facility managers may focus on workload, maintenance, and risk management. These expectations correspond to different layers of care. If the layer is not made explicit, a correct evaluation may be hampered: a technology may be judged ineffective because it does not reduce loneliness scores even though it improves participation, or it may be judged successful because users enjoy it briefly even though it is not sustainable in daily care.

4. The MCST model

The MCST model organizes companion technologies into three interconnected layers: individual engagement, relational interaction, and the care ecosystem (Figure 1). It draws on social frailty, relationship-centered care, and social ecological approaches (*23-25*) but focuses specifically on the role of companion technologies in loneliness-related support in dementia care. The model is intended to clarify how technology-related benefits may arise, where risks occur, and what should be evaluated before and during implementation.

Layer 1, individual engagement, refers to the direct responses of the person with dementia. Companion technologies may provide tactile, visual, auditory, or conversational stimuli that provide comfort or emotional reassurance or encourage participation in activities, curiosity, reminiscence, or a sense of security (*11-14,19,20,22*). Relevant evaluation domains include distress, anxiety, agitation, engagement, acceptance, confusion, discomfort, and adverse emotional reactions, rather than loneliness scores alone. The central question in this layer is whether the technology provides a safe and meaningful entry point for engagement for a particular person.

Layer 2, relational interaction, refers to the way technology mediates relationships with caregivers, family members, staff, and other users (*18,24*). A robot, avatar, or conversational agent may become a shared object of attention, a prompt for reminiscence, a bridge for communication, or an activity involving more than one person (*15-20*). In this layer, the technology is not a substitute for human contact. Its intended role is to make human interaction easier to initiate, sustain, or personalize. Evaluation should therefore include caregiver-user interaction, family participation, group activities, frequency of communication, the caregiver burden, staff perceptions, and whether the use of the technology enhances or unintentionally reduces human contact.

Layer 3, the care ecosystem, refers to the operational and ethical conditions that make Layers 1 and 2 safe and sustainable. These include staff training, workflow integration, maintenance, cost, documentation, data management, privacy safeguards, consent procedures, human supervision, escalation rules, and accountability (*21*). Without this layer, positive responses in Layer 1 or interactional benefits in Layer 2 may remain temporary, fragmented, or dependent on individual staff enthusiasm. For AI-enabled technologies, Layer 3 is especially important because inaccurate, inconsistent, or emotionally inappropriate responses may confuse or distress users whose memory, judgment, or ability to verify information is impaired (*26-29*).

The MCST model is dynamic rather than static. User responses in Layer 1 should be interpreted through relational interactions in Layer 2 and then translated into operational decisions in Layer 3. Caregivers may adjust the frequency, timing, type of activity, conversation content, supervision, or discontinuation criteria according to observed benefits and risks; Figure 1 represents this feedback loop. In this sense, the model shifts the question

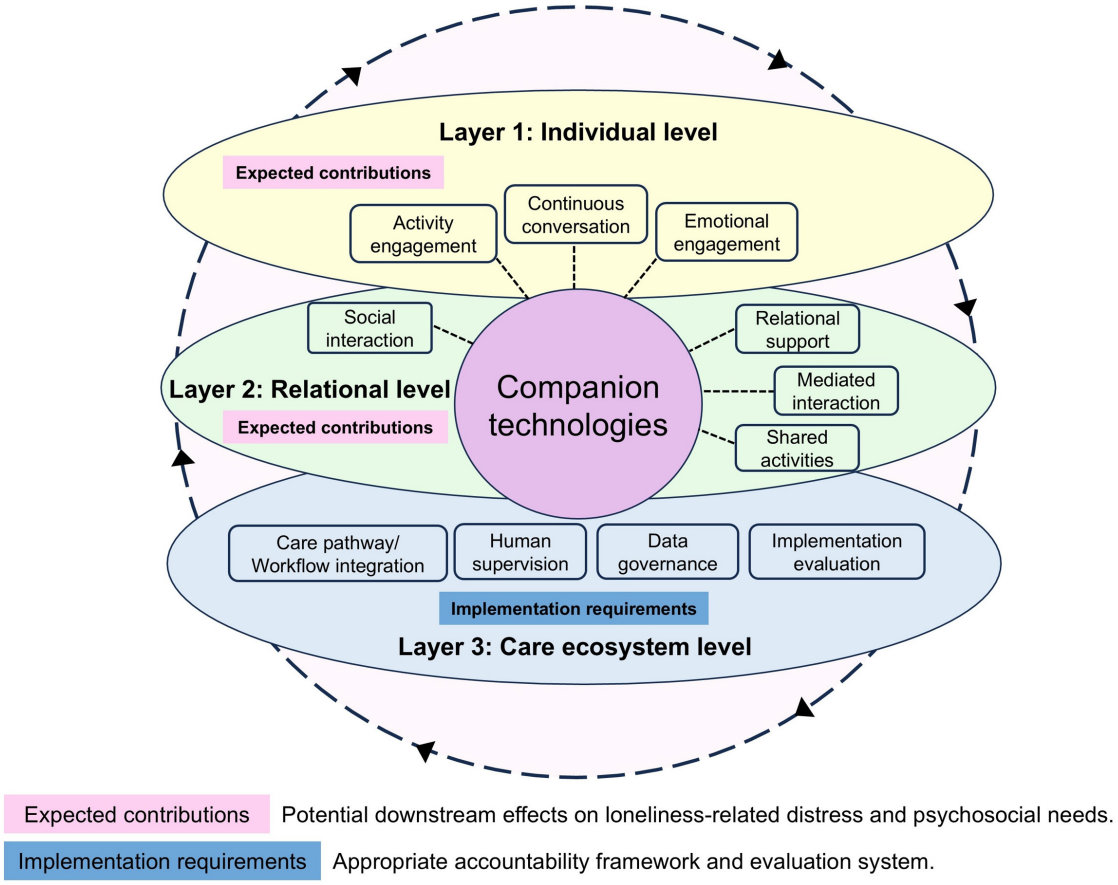


Figure 1. Concept diagram of the MCST model. Safe and sustainable engagement and interaction in Layers 1 and 2 are supported by the stability of Layer 3. The outer circular arrows indicate a feedback loop through which user responses and usage data are reviewed by care providers and used to adjust the level of engagement, supervision, and implementation.

from 'Does the device reduce loneliness?' to 'How can technology, relationships, and care systems work together to address loneliness-related needs?'

The model also helps identify levels of failure: poor personalization in Layer 1, engagement that remains disconnected from human contact in Layer 2, or unsustainable, unsafe, or burdensome implementation in Layer 3. This identification can inform research design, implementation planning, and institutional evaluation.

The same logic can guide outcome selection. Outcome selection should align with each layer: emotional response, distress, engagement, and adverse reactions for Layer 1, interaction patterns, caregiver involvement, shared activities, and human contact for Layer 2, and training, workflow, privacy, cost, maintenance, safety incidents, and accountability for Layer 3. By aligning outcomes with layers, the MCST model can make future studies more interpretable and reduce the risk of drawing broad conclusions from narrow indicators.

5. Future agenda for AI-enabled companion technologies in dementia care

Future research should clarify for whom, when, and for what purpose companion technologies are appropriate, recognizing differences in the stage of disease, communication ability, sensory function, personal history, the living environment, and preferences. They should also distinguish loneliness, social isolation, loneliness-related distress, depression, anxiety, behavioral and psychological symptoms, participation in activities, quality of life, caregiver burden, staff workload, and workflow impact. Longitudinal and real-world designs are needed to separate novel effects from sustained benefits and to examine discontinuation, maintenance, cost, and acceptability.

Second, AI-enabled companions require explicit human oversight. Personalization may improve engagement, but it also increases risks related to privacy, consent, emotional dependence, reduced human contact, and inappropriate reliance on machine-generated responses (*26,27*). Large language models can generate hallucinations or inconsistent statements, which may be particularly harmful to people with dementia (*28,29*). Practical safeguards should include audit logs, procedures for reviewing conversations, criteria for temporary suspension of use, escalation pathways when confusion or anxiety occurs, and clear designation of responsibility among developers, care facilities, clinicians, caregivers, and family members.

Third, implementation and regulatory evaluation should be linked to intended use and claimed outcomes. Technologies used only for daily engagement may be managed as assistive or care technologies. If, however, they are marketed, reimbursed, or implemented with claims of alleviating loneliness, depressive symptoms, or behavioral symptoms, encouraging participation in activities, alleviating the caregiver burden, or achieving other clinical and psychosocial outcomes, stronger evidence may be required under digital health, care-technology, or software-as-a-medical-device frameworks (*30,31*). The MCST model can help institutions specify what should be evaluated at the individual, relational, and ecosystem levels before broad implementation.

Finally, policy and practice should avoid a false choice between technology and human care. The key implementation question is not whether AI companions can replace caregivers but whether they can support caregivers, families, and care teams in recognizing and responding to psychosocial needs that are otherwise missed. This requires not only technical innovation, but also staff education, ethical governance, reimbursement discussion, and institutional accountability.

For this reason, implementation studies should report not only positive user responses but also non-use, withdrawal, technical failures, staff burden, and unintended consequences. Acceptability may vary by context, such as mealtimes, nighttime anxiety, or family visits; real-world evidence should therefore indicate the timing, facilitation, and adaptation of use. Such information would help the field move beyond general claims that companion technologies are beneficial toward a clearer understanding of how, for whom, and under what conditions they may help to provide loneliness-related support.

6. Conclusion

Companion technologies do not cure dementia, replace human relationships, or independently resolve loneliness. Their potential value lies in encouraging emotional engagement, mediating interaction, and improving the operational conditions of dementia care. The MCST model offers a conceptual framework for considering these technologies as potential components of social care infrastructure at the individual, relational, and ecosystem levels. Empirical validation of the MCST model in diverse care contexts and dementia stages, with transparent reporting of implementation conditions, needs to be done to keep claims in line with evidence. Future work should determine whether companion technologies, including AI-enabled systems, can address loneliness-related needs safely, ethically, and sustainably when embedded within human relationships and care systems.

Funding: This work was supported by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (24K14216).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. WHO. Dementia. *https://www.who.int/news-room/fact-sheets/detail/dementia* (accessed April 8, 2026).
2. Hajek A, König HH. Prevalence of loneliness and social isolation among individuals with mild cognitive impairment or dementia: Systematic review and meta-analysis. BJPsych Open. 2025; 11:e44.
3. National Academies of Sciences, Engineering, and Medicine. Social isolation and loneliness in older adults: Opportunities for the health care system. Washington, DC: The National Academies Press; 2020. *https://www.ncbi.nlm.nih.gov/books/NBK557964/* (accessed April 8, 2026).
4. Victor CR, Rippon I, Nelis SM, Martyr A, Litherland R, Pickett J, Hart N, Henley J, Matthews F, Clare L; IDEAL programme team. Prevalence and determinants of loneliness in people living with dementia: Findings from the IDEAL programme. Int J Geriatr Psychiatry. 2020; 35:851-858.
5. Cardona M, Andrés P. Are social isolation and loneliness associated with cognitive decline in ageing? Front Aging Neurosci. 2023; 15:1075563.
6. Berkman LF, Syme SL. Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. Am J Epidemiol. 1979; 109:186-204.
7. Riches S, Azevedo L, Vora A, Kaleva I, Taylor L, Guan P, Jeyarajaguru P, McIntosh H, Petrou C, Pisani S, Hammond N. Therapeutic engagement in robot-assisted psychological interventions: A systematic review. Clin Psychol Psychother. 2022; 29:857-873.
8. Nichol B, McCready J, Erfani G, Comparcini D, Simonetti V, Cicolini G, Mikkonen K, Yamakawa M, Tomietto M. Exploring the impact of socially assistive robots on health and wellbeing across the lifespan: An umbrella review and meta-analysis. Int J Nurs Stud. 2024; 153:104730.
9. OECD. Health at a glance 2025: OECD Indicators. *https://doi.org/10.1787/8f9e3f98-en* (accessed April 8, 2026).
10. Ishihara M, Matsunaga S, Islam R, Shibata O, Chung UI. A policy overview of Japan's progress on dementia care in a super-aged society and future challenges. Glob Health Med. 2024; 6:13-18.
11. Shibata T, Wada K. Robot therapy: A new approach for mental healthcare of the elderly - A mini-review. Gerontology. 2011; 57:378-386.
12. Yu C, Sommerlad A, Sakure L, Livingston G. Socially assistive robots for people with dementia: Systematic review and meta-analysis of feasibility, acceptability and the effect on cognition, neuropsychiatric symptoms and quality of life. Ageing Res Rev. 2022; 78:101633.
13. Hsieh CJ, Li PS, Wang CH, Lin SL, Hsu TC, Tsai CT. Socially assistive robots for people living with dementia in long-term facilities: A systematic review and meta-analysis of randomized controlled trials. Gerontology. 2023; 69:1027-1042.
14. Rashid NLA, Leow Y, Klainin-Yobas P, Itoh S, Wu VX. The effectiveness of a therapeutic robot, 'Paro', on behavioural and psychological symptoms, medication use, total sleep time and sociability in older adults with dementia: A systematic review and meta-analysis. Int J Nurs

Stud. 2023; 145:104530.
15. Sato M, Yasuhara M, Osaka K, Ito H, Dino MJ, Ong IL, Zhao Y, Tanioka T. Rehabilitation care with Pepper humanoid robot: A qualitative case study of older patients with schizophrenia and/or dementia in Japan. Enfermería Clínica. 2020; 30 Suppl 1:32-36.
16. Zuschnegg J, Häussl A, Lodron G, Orgel T, Russegger S, Schneeberger M, Fellner M, Holter M, Prodromou D, Schultz A, Roller-Wirnsberger R, Paletta L, Koini M, Schüssler S. Psychosocial effects of a humanoid robot on informal caregivers of people with dementia: A randomised controlled trial with nested interviews. Int J Nurs Stud. 2025; 162:104967.
17. Cui L, Li Y, Yang X, Liu X, Zhang L, Hou L. Humanoid robot-assisted support for health care in older adults: Systematic scoping review. JMIR Aging. 2026; 9:e83849.
18. Broadbent E. Interactions with robots: The truths we reveal about ourselves. Annu Rev Psychol. 2017; 68:627-652.
19. Sugiyama H, Nakamura K. Temporary improvement of cognitive and behavioral scales for dementia elderly by shiritori word game with a dialogue robot: A pilot study. Front Robot AI. 2022; 9:941056.
20. Nagata Y, Satake Y, Yamazaki R, Nishio S, Suzuki M, Kanemoto H, Yamakawa M, Figueroa D, Maalouly E, Ishiguro H, Ikeda M. A conversational robot for cognitively impaired older people who live alone: An exploratory feasibility study. Psychogeriatrics. 2025; 25:e70076.
21. Papadopoulos I, Koulouglioti C, Lazzarino R, Ali S. Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: A systematic review. BMJ Open. 2020; 10:e033096.
22. Yamazaki R, Kase H, Nishio S, Ishiguro H. Anxiety reduction through close communication with robotic media in dementia patients and healthy older adults. J Robot Mechatron. 2020; 32:32-42.
23. Bunt S, Steverink N, Olthof J, van der Schans CP, Hobbelen JSM. Social frailty in older adults: A scoping review. Eur J Ageing. 2017; 14:323-334.
24. Nolan M, Brown J, Davies S, Nolan J, Keady J. The Senses Framework: Improving care for older people through a relationship-centred approach. Sheffield Hallam University. 2006. *https://shura.shu.ac.uk/280/1/PDF_Senses_Framework_Report.pdf* (accessed April 8, 2026).
25. Golden SD, Earp JA. Social ecological approaches to individuals and their contexts: Twenty years of health education & behavior health promotion interventions. Health Educ Behav. 2012; 39:364-372.
26. Sharkey A, Sharkey N. Granny and the robots: Ethical issues in robot care for the elderly. Ethics Inf Technol. 2012; 14:27-40.
27. Deusdad B. Ethical implications in using robots among older adults living with dementia. Front Psychiatry. 2024; 15:1436273.
28. Granstedt J, Kc P, Deshpande R, Garcia V, Badano A. Hallucinations in medical devices. AI in Medicine and Life Sciences. 2025; 100145.
29. Tanioka T, Yokotani T, Tanioka R, Betriana F, Matsumoto K, Locsin R, Zhao Y, Osaka K, Miyagawa M, Schoenhofer S. Development issues of healthcare robots: Compassionate communication for older adults with dementia. Int J Environ Res Public Health. 2021; 18:4538.
30. Ministry of Economy, Trade, and Industry. AI Guidelines for business Ver. 1.0 Compiled. 2024. *https://www.meti.go.jp/english/press/2024/0419_002.html* (accessed April 8, 2026).
31. Pharmaceuticals and Medical Devices Agency. Report on AI-based software as a medical device (SaMD). 2023. *https://www.pmda.go.jp/files/000266100.pdf* (accessed April 8, 2026).

Received April 28, 2026; Revised May 30, 2026; Accepted June 2, 2026.

Released online in J-STAGE as advance publication June 3, 2026.

**Address correspondence to*:
Peipei Song, Division of Global Health & Medicine, Bureau of Health Emergency and Management, Japan Institute for Health Security, 1-21-1 Toyama, Shinjuku-ku, Tokyo 162-8655, Japan.
E-mail: psong@jihs.go.jp

Global Health & Medicine. 2026; 8(3):222-226.
DOI: 10.35772/ghm.2026.01069

Artificial intelligence (AI)-assisted full-course case management for primary liver cancer: System design, preliminary implementation, and practical considerations

Yanhui Wang[1], Xian Yue[1], Ruishuang Zheng[1], Lu Chen[1], Ying Wang[2], Wanmin Qiang[2,*]

[1] Department of Hepatobiliary Cancer, Tianjin Medical University Cancer Hospital and Institute, National Clinical Research Centre for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Centre for Cancer, Tianjin, China;
[2] Department of Nursing, Tianjin Medical University Cancer Hospital and Institute, National Clinical Research Centre for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Centre for Cancer, Tianjin, China.

Abstract: Primary liver cancer presents substantial management challenges across the surgical trajectory, including high recurrence rates, prolonged rehabilitation, and fragmented post-discharge care. This correspondence presents a multidisciplinary physician–nurse co-led, artificial intelligence (AI)-assisted full-course case management model grounded in just-in-time adaptive intervention (JITAI) theory. The model spans four phases—peri-admission, perioperative, post-discharge home-based care, and long-term follow-up—supported by an intelligent platform enabling automated decision triggering, symptom monitoring, tailored health education, and intervention matching. A pilot study with 25 patients was conducted from March to May 2026 at a tertiary cancer hospital in Tianjin, China. Preliminary results revealed improvements in antiviral medication adherence (80–96%), targeted therapy adherence (96–100%), and satisfaction (99.8%), with reductions in missed follow-ups and symptom reporting delays. Multicenter controlled studies need to be conducted to evaluate this model's effectiveness, cost-effectiveness, and long-term sustainability.

***Keywords*:** artificial intelligence, case management, just-in-time adaptive intervention, primary liver cancer, full-course management

1. Introduction

According to data from the National Cancer Center of China, an estimated 367,700 new cases of primary liver cancer were reported in 2022, representing 42.5% of global cases and ranking fourth among new cancer diagnoses nationally (*1*). Primary liver cancer accounted for 316,500 deaths that year, making it the second leading cause of cancer death in China. The overall 5-year relative survival rate remains at 14.4%, with over half of patients diagnosed in intermediate or advanced stages (*1*). For patients who are eligible for surgery, hepatectomy is a crucial curative treatment modality (*2*), and patients who undergo resection require continuous management spanning preoperative preparation, perioperative care, post-discharge rehabilitation, and long-term surveillance (*3*).

There are major gaps in the full-course management of surgically treated liver cancer patients. Current care pathways often place greater emphasis on in-hospital treatment decisions, whereas preoperative preparation, post-discharge rehabilitation, and recurrence monitoring may receive less systematic attention (*4*). Nurse-led follow-up may also be insufficiently integrated with clinical workflows, which may contribute to delayed symptom reporting, missed appointments, declining medication adherence, and delayed detection of complications (*5*). Full-course management models that integrate multidisciplinary collaboration with health information technology are evolving toward data-driven, personalized, and continuous care delivery (*6-8*).

Artificial intelligence (AI) offers promising tools for long-term cancer care, including remote monitoring, intelligent follow-up, risk stratification, tailored health education, and clinical decision-making support (*9,10*). Just-in-time adaptive intervention (JITAI) theory provides a framework for designing dynamic interventions based on tailoring variables, decision points, decision rules, intervention options, and proximal and distal outcomes (*11*). Integrating JITAI with AI platforms may help operationalize the delivery of the right intervention to the right patient at the right time (*12*).

Guided by JITAI theory, we developed an AI-assisted full-course case management model for patients undergoing surgery for primary liver cancer. In this correspondence, we describe the system design and AI-enabled functions, and we present preliminary feasibility findings from a single-center pilot study.

2. An AI-assisted full-course case management model for primary liver cancer based on JITAI theory

From December 2025 to February 2026, our team developed a multidisciplinary physician–nurse co-led, AI-enabled, multidisciplinary full-course case management program for patients undergoing surgery for primary liver cancer, grounded in JITAI theory. The development team consisted of three hepatobiliary oncologists, one head nurse with 18 years of liver cancer nursing experience, three specialist oncology nurses, and one software engineer; of these members, two oncology nurses served as case managers. The program was informed by our previous work on liver cancer perioperative and follow-up management needs, a comprehensive literature review, group discussion, and two rounds of Delphi expert consultation.

The final program spans four phases—peri-admission, perioperative care, post-discharge home-based care, and long-term follow-up—and is structured around key management nodes, including admission preparations, preoperative assessment, postoperative rehabilitation, discharge planning, medication management, symptom monitoring, appointment scheduling, and long-term surveillance. These are operationalized into 19 decision nodes supporting continuous, stratified, and individualized care.

The program's distinguishing feature is the systematic translation of JITAI's core components—tailoring variables, decision points, decision rules, and intervention options—into a platform-based workflow. Baseline tailoring variables include demographics, disease etiology, nutritional risk scores, Child–Pugh classification, family history, and medical history. Process tailoring variables encompass dynamic data generated across the care trajectory, including high-risk screening results (nutritional, fall, and pressure injury risk), laboratory and imaging findings, patient-reported symptoms, and medication adherence. Decision points correspond to clinically significant time points across all four phases. At each decision point, the platform automatically generates reminders, risk alerts, and intervention recommendations by integrating preset rules, risk tags, and AI-assisted analysis functions, which are then reviewed by specialized nurses and a multidisciplinary team.

3. Platform architecture and AI-enabled functions

The final program was deployed on the Full-Course Case Management Platform operating within our institution's Internet hospital network.

The intelligent information-based management platform and the final program are presented in Figure 1. The platform's functional modules and information infrastructure include AI-driven analytics, knowledge graphs, automated response engines, semantic processing, intelligent interaction interfaces, patient stratification algorithms, and Internet hospital operations. It is also integrated with institutional systems, including the hospital information system (HIS), laboratory information system (LIS), picture archiving and communication system (PACS), electronic medical records, and online payment systems, thereby enabling seamless data flow across the care continuum.

The platform operates around core processes of patient screening and assessment, case enrollment, care plan allocation, plan implementation, and management evaluation. Through coordinated patient and provider portals, it enables stage-specific, precision-oriented case management throughout the disease trajectory.

3.1. Patient portal

The patient portal supports self-directed health management across six functional categories: *i*) individualized records with automated synchronization and dynamic updating of diagnostic, treatment, and care bundle information spanning the perioperative, discharge, and follow-up phases; *ii*) real-time synchronization of laboratory and examination results, with capacity for manual entry of home-based health data shared with the provider portal; *iii*) multimodal patient–provider communication *via* text, voice, and images across all treatment phases, with records retained; *iv*) stage-appropriate health education in text, graphic, and video formats; *v*) intelligent follow-up scheduling with automated reminders, secondary alerts for overdue tasks, and data feedback to providers; and *vi*) recovery stories, experience sharing, and peer support groups to facilitate psychological well-being.

3.2. Healthcare personnel portals

The healthcare personnel portals—serving oncologists, nurses, nutritionists, and other team members—support efficient management and precision intervention. Providers access patient records, risk assessments, and care plans in real time, with group-based management by treatment stage, age, family history, medication regimen, and risk labels to facilitate differentiated care. Communication features include text, image, and voice messaging with quick-reply templates, ensuring both efficiency and traceability.

For follow-up management, the portals enable individualized plan customization, automatic progress tracking, and alerts for incomplete tasks, abnormal

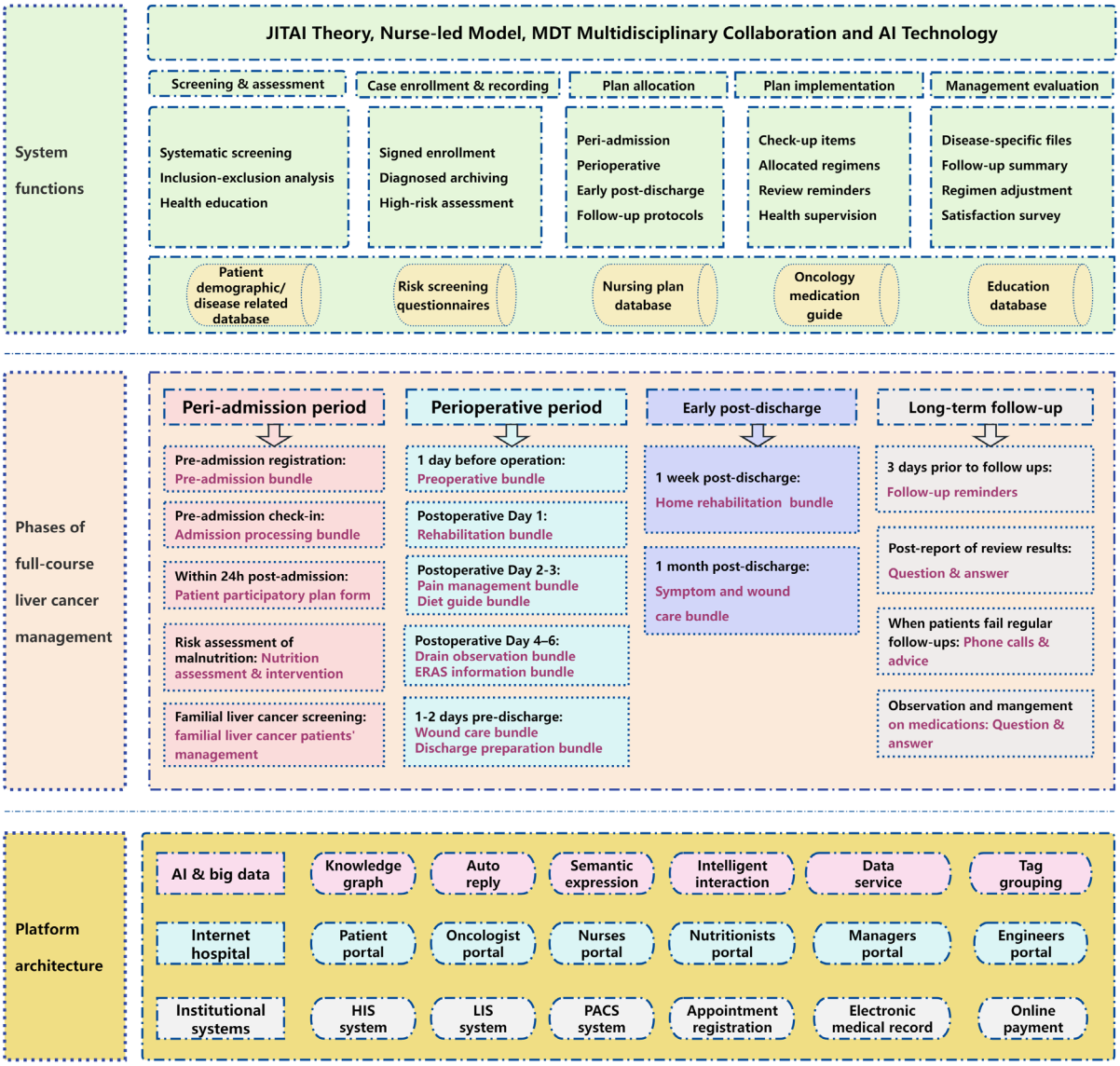


Figure 1. AI-assisted full-course case management model for primary liver cancer patients based on just-in-time adaptive intervention theory. *Abbreviations*: HIS, hospital information system; LIS, laboratory information system; PACS, picture archiving and communication system.

symptom feedback, and high-risk situations. For health education, evidence-based content is delivered on demand with reading status tracking and secondary reminders for non-engaged patients. For risk management, the platform integrates data on pain, nutrition, liver function, medication adherence, appointment completion, and symptom changes, pushing alerts to providers when anomalies are detected to facilitate earlier risk identification.

3.3. AI-assisted decision triggering and risk alerts

AI represents a key differentiating feature of this model. Drawing on Internet hospital data and institutional information system connections and integrating natural language processing, risk prediction algorithms, and label-based grouping, the platform continuously analyzes patient-reported content and clinical data. AI-assisted functions identify risk signals within patient-reported symptoms, match corresponding educational content, trigger medication or appointment reminders, and flag patients requiring priority attention.

Importantly, AI serves as an assistive role rather than replacing clinical judgment. High-risk alerts, intervention adjustments, and treatment-related recommendations require review and confirmation by case managers, specialist nurses, or physicians. This design improves efficiency while preserving humanistic care, professional judgment, and meaningful patient–provider communication.

4. Preliminary practice outcomes

From March to May 2026, we conducted a pilot study with 25 patients undergoing surgery for primary

liver cancer who received individualized full-course management *via* the intelligent platform. Our preliminary observations suggested this model's feasibility: antiviral medication adherence increased from 80 to 96%, oral targeted therapy adherence increased from 96 to 100%, and patient satisfaction reached 99.8%. Missed appointments and delays in symptom reporting—common in conventional management—also decreased.

From the patient's perspective, the platform provided timely reminders for preoperative preparations and examinations, promptly addressed preoperative questions, and helped patients gain a comprehensive understanding of the perioperative process and prevention of complications. It also simplified follow-up workflows, making appointment reminders, medication guidance, rehabilitation education, and symptom reporting easier. For postoperative patients requiring long-term surveillance and home monitoring, the platform offered continuous care support that may enhance the capacity for self-management.

From the perspective of healthcare personnel, this AI-assisted management system reduced certain manual workload components—such as patient registration, data compilation, routine notifications, and repetitive follow-up tasks—thereby allowing providers to focus more effectively on identifying high-risk patients, delivering individualized education, and engaging in meaningful clinical communication. In addition, the standardized workflow, grounded in the JITAI framework, helped reduce inter-provider variability and enhanced the traceability of patient management.

A point worth noting is that these results only represent our preliminary practice findings from 25 cases rather than confirmatory evidence of clinical effectiveness. Our study's limitations include the small sample size, short observation period, and absence of a control group. The impact of this model on liver cancer recurrence, survival, complication rates, provider workload, and cost-effectiveness remains to be determined through rigorous evaluation.

5. Practical considerations and future directions

Although this model remains in a preliminary stage, we believe that AI-assisted full-course case management for primary liver cancer offers practical value in several key respects.

First, at the technological level, AI facilitates the operationalization of JITAI theory. While JITAI emphasizes dynamic adaptation and precision intervention, conventional models are limited by subjective judgment, inefficient manual screening, and insufficient standardization. By drawing on Internet hospital platforms, natural language processing, risk prediction, and automated decision triggering, management can shift from a reactive response to proactive anticipation (*13*).

Second, at the management level, this model has the advantages of multidisciplinary physician–nurse coordination combined with multidisciplinary collaboration. Cancer chronic disease management frequently suffers from a disconnection between oncologists and nurses, disciplinary fragmentation, and gaps between inpatient and outpatient care (*14*). Our model positions case managers and specialist nurses as primary implementers who coordinate with physicians, dietitians, psychologists, and other team members (*15*). Monitoring, follow-up, rehabilitation, and education are integrated through the AI platform, bridging the gap between in-hospital treatment and home-based care.

Third, at the clinical practice level, AI assistance must adhere to a humanistic-first principle. Postoperative patients require not only surveillance reminders and risk alerts but also ongoing support with regard to their disease uncertainty, fear of recurrence, and treatment burden. Intelligent management should augment rather than replace professional judgment and communication, helping providers identify needs earlier, respond to risks more rapidly, and deliver more continuous care.

Fourth, several challenges warrant acknowledgment. Disparities in digital literacy may affect equitable platform access; elderly patients or those with severe symptoms may require caregiver assistance. Provider trust in AI-assisted decision-making needs to be gradually cultivated through sustained use. For symptom recognition, risk prediction, and individualized recommendations specifically, clear mechanisms for manual review, adverse event protocols, and delineation of responsibilities must be established.

Fifth, future optimization should address AI algorithms, stratification rules, health education knowledge bases, and provider workflows, with further integration across HIS, LIS, PACS, electronic medical records, and Internet hospital systems. With regard to research design, multicenter controlled studies with long-term follow-up need to be conducted to evaluate impacts on follow-up adherence, medication adherence, symptom reporting timeliness, recurrence detection, complication identification, patient experience, provider workload, cost-effectiveness, and survival outcomes.

Lastly, this framework may also have broader applicability. Its adaptation to other perioperative cancers—including lung, gastric, and colorectal—could be explored and potentially expanded to high-risk population screening and life-course health management, providing a reference for smart hospital development and refined cancer care delivery.

6. Conclusions

The AI-assisted full-course case management model for patients with primary liver cancer, guided by JITAI theory, offers a potential pathway for improving continuous care across in-hospital and out-of-hospital settings. Through

multidisciplinary physician–nurse coordination and collaboration, platform-based management, and AI-assisted alerts, the model integrates symptom monitoring, medication reminders, appointment management, health education, and risk identification within a unified framework. Our preliminary single-center pilot study involving 25 patients suggests that this model has preliminary feasibility and patient acceptability and may help improve medication reminders, follow-up management, and symptom feedback. However, its clinical effectiveness, safety, cost-effectiveness, and long-term sustainability need to be validated in larger, multicenter, controlled studies. Future efforts should further integrate AI technology with full-course case management while maintaining humanistic care and professional judgment.

Funding: This research was supported by the Tianjin Project for Creation of Key Medical Disciplines (grant no. TJYXZDXK-3-003A).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. National Health Commission of the People's Republic of China. Guideline for diagnosis and treatment of hepatocellular carcinoma (2026 edition). Zhonghua Wai Ke Za Zhi. 2026; 64:549-580. (in Chinese)
2. Sidali S, Trépo E, Sutter O, Nault JC. New concepts in the treatment of hepatocellular carcinoma. United European Gastroenterol J. 2022; 10:765-774.
3. Tu HB, Chen LH, Huang YJ, Feng SY, Lin JL, Zeng YY. Novel model combining contrast-enhanced ultrasound with serology predicts hepatocellular carcinoma recurrence after hepatectomy. World J Clin Cases. 2021; 9:7009-7021.
4. Li J, Su X, Yang MJ, Dai Y, Li F, Yu H, Zhang J. Practice of disease whole-course nursing management model in oncology department. J Nursing Sci. 2021; 36:40-43. (in Chinese)
5. Jiang XL, Chen CL. Establishment and implementation of a full-course care model for single-disease case management in oncology. J Nursing. 2022; 29:32-34. (in Chinese)
6. Zhou Y, Huang LY, Wang BY, Luo Q, Peng DY, Tan JX, Guo YB. Construction and application of post-discharge management for Severe Acute Pancreatitis based on a whole-course intelligent platform. Digital Intelligence in Nursing. 2025; 25:1881-1886.
7. Hepatobiliary Tumor Integrated Nursing Committee of China Anti-Cancer Association, China Anti-Cancer Association Committee of Liver Cancer; Yu JX, Huang ZY, Wu Y. Expert consensus on perioperative nutritional management for patients undergoing hepatec-tomy for primary liver cancer (2025 edition). Chinese Journal of Digestive Surgery, 2025, 24: 1539-1547. (in Chinese)
8. Cui HB, Li XF, Ji YR, Wang SY. Practice of " Internet+ " whole course management with specialized nurses as the main service subject. Chinese Journal of Hospital Administration. 2023; 39:579-583. (in Chinese)
9. Du Y, Yang P, Liu Y, Deng C, Li X. Artificial intelligence in chronic disease self-management: Current applications and future directions. Front Public Health. 2025; 13:1689911.
10. Seif El Dahan K, Reczek A, Daher D, Rich NE, Yang JD, Hsiehchen D, Zhu H, Patel MS, Bayona Molano MDP, Sanford N, Gopal P, Parikh ND, Yopp AC, Singal AG. Multidisciplinary care for patients with HCC: A systematic review and meta-analysis. Hepatol Commun. 2023; 7:e0143.
11. Spruijt-Metz D, Nilsen W. Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Computing. 2014; 13:13-17.
12. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. Ann Behav Med. 2018; 52:446-462.
13. Wang L, Miller L. Assessment and disruption of ruminative episodes to enhance mobile cognitive behavioral therapy just-in-time adaptive interventions in clinical depression: Pilot randomized controlled trial. JMIR Form Res. 2023; 7:e37270.
14. Altom DS, Awad Taha AI, Mahmoud Hussein AAA, Ibrahim Elshiekh MA, Alata Abdelmajed AH, Abdalla Ibrahim FI, Abelgadir Mohammed SM, Elamin Eltain Tifoor MM. Artificial intelligence-based chatbots in chronic disease management: A systematic review of applications and challenges. Cureus. 2025; 17:e81001.
15. Gan BH, Qiang WM, Wang Y, Li MM, Liu CM, Zhu MY, Liu H, Zhang F, Zhen XW, Fan RN. Construction and application of a case manager-led whole-course management model for breast cancer. Tianjin Journal of Nursing. 2026; 34:11-17. (in Chinese)

Received May 1, 2026; Revised June 10, 2026; Accepted June 14, 2026.

Released online in J-STAGE as advance publication June 19, 2026.

**Address correspondence to*:
Wanmin Qiang, Department of Nursing, Tianjin Medical University Cancer Hospital and Institute, Tianjin, West Huan-Hu Road, Ti Yuan Bei, Hexi District, Tianjin 300060, China.
E-mail: nursing1331@sina.com

Global Health & Medicine. 2026; 8(3):227-232.
DOI: 10.35772/ghm.2026.01061

Artificial intelligence (AI)-based pose estimation detects movements linked to unplanned tube removal in ICU patients

Aya Umeda[1,2,*], So Mizuno[3,4], Fumio Ishizaki[5], Tatsuya Okamoto[6]

[1] Department of Adult Health Nursing, National College of Nursing, Japan (NCNJ), Japan Institute for Health Security, Tokyo, Japan;
[2] Department of Nursing, Japan Institute for Health Security, Tokyo, Japan;
[3] Japan Institute of Life Design Counseling, Chiba, Japan;
[4] Division of Educational Sciences, Hiroshima University, Hiroshima, Japan;
[5] Modal Stage Inc, Tokyo, Japan;
[6] Department of Emergency and Critical Care Medicine, Japan Institute for Health Security, Tokyo, Japan.

Abstract: Unplanned removal of life-sustaining tubes in intensive care units (ICUs) poses serious risks, yet existing monitoring methods relying on physical restraints have ethical and clinical drawbacks. Here we applied artificial intelligence (AI)-based pose estimation using MediaPipe to analyze ICU surveillance videos, extracting skeletal coordinates to detect movements associated with tube removal. Using Singular Spectrum Transformation for change-point detection, we identified movement changes corresponding to tube-removal behaviors in three consented cases, achieving average precision values substantially above chance. These preliminary results demonstrate that AI-driven, contactless motion analysis can capture clinically relevant signals from existing ICU infrastructure without additional patient burden. Although limited by sample size and environmental factors, this approach holds promise for real-time, non-invasive monitoring to reduce reliance on physical restraints and enhance patient safety in critical care settings.

***Keywords*:** AI-based monitoring, ICU patient safety, pose estimation technology, unplanned tube removal detection, contactless patient monitoring

1. Introduction

Patients admitted to the intensive care unit (ICU) often depend on life-sustaining devices such as endotracheal tubes, nasogastric tubes, arterial lines, and various drainage catheters. Unplanned removal of these devices is a serious clinical problem, as it can immediately threaten patient survival. For example, approximately 40–60% of patients who remove their endotracheal tube require reintubation, which is associated with prolonged mechanical ventilation, longer ICU and hospital stays, and increased healthcare costs (*1*). Although the incidence of unplanned endotracheal tube removal is relatively low—ranging from 0.05% to 2% among mechanically ventilated ICU patients (*1*)—its consequences can be severe. A pooled prevalence of about 6.7% has been reported across all ICU patients receiving mechanical ventilation (*2*), making reliable prevention a critical clinical priority.

Physical restraints have been widely adopted in ICUs worldwide, with restraint rates reported to range from 8.7% to 59.1% (*3*). However, restraint carries significant ethical and clinical concerns: beyond violating patient autonomy, it is associated with focal skin injury, delirium, post-traumatic stress disorder (PTSD), and other sequelae of post-intensive care syndrome (PICS) (*2,4,5*). Furthermore, restraint does not reliably prevent device removal; even among restrained patients, 61.4% still succeeded in removing their devices (*6*). These findings underscore the urgent need for a fundamentally different approach—specifically, proactive, real-time detection of tube-removal behaviors before they occur. This need highlights the unresolved dilemma faced by clinicians: preventing life-threatening events while also avoiding harms inherent in prolonged restraint.

To our knowledge, no published study has prospectively characterized movement patterns that immediately precede unplanned tube removal in ICU patients, leaving open the question of whether such behaviors are detectable in real time. Since 2020, our group has systematically investigated approaches to quantify these movements with the goal of developing a contactless early warning system. We found that optical motion capture was incompatible with the ICU environment. Inertial motion capture (IMC), while promising for three-dimensional movement data, proved

unreliable in supine patients due to posterior sensor occlusion by the mattress, and also posed challenges related to sensor attachment in critically ill patients. These limitations led us to explore artificial intelligence (AI)-based pose estimation—a contactless approach that reconstructs body landmark positions from standard surveillance footage without need for wearable devices. This builds on our preliminary findings first presented at a national conference (*7*).

2. Real-world ICU video data and ethical challenges

This prospective observational study was conducted between 2021 and 2023 at the ICU of a tertiary hospital in Tokyo, Japan. Security camera footage was reviewed for unplanned tube removal events as they occurred during the study period, and patients who experienced tube removal were approached for consent.

Obtaining informed consent proved exceptionally challenging throughout the study period. Many patients had pre-existing cognitive impairment or dementia, or were in a critical condition from which they did not survive, rendering direct consent impossible. Furthermore, the study overlapped substantially with the COVID-19 pandemic, during which family visitation was severely restricted at the institution, making surrogate consent from family members equally difficult to obtain. As a result, valid consent was obtained from only 13 of the 58 patients who experienced unplanned tube removal during the three-year study period, underscoring the exceptional difficulty of enrolling this patient population in real-world ICU research. Characteristics and imaging conditions of all 13 video cases are summarized in Table 1.

From each consented case, video frames were extracted centered on the time of the tube removal event, using a maximum window of 5 minutes per clip. Videos were recorded at 5 frames per second (fps) with a resolution of 640 × 480 pixels. Facial blurring was applied to video frames used in publication figures to protect patient privacy. The study was approved by ethics committee of National Center for Global Health and Medicine (approval No. 004011), and written informed consent was obtained from all participants or their surrogates in accordance with the Declaration of Helsinki.

3. Pose estimation and change-point analysis of hand-to-face movements

Pose estimation was conducted using MediaPipe (Google LLC), an open-source machine learning framework that applies deep neural networks to extract three-dimensional (3D) coordinates for 33 anatomical landmarks from standard monocular video footage in real time (*8*). This approach enables continuous skeletal tracking—including the face, trunk, limbs, and hands—using existing surveillance cameras, without the need

Table 1. Characteristics and analyzability of ICU videos involving unplanned tube removal

No.	Tube type	Site	Pose est.	Presumed main reason for failure	Camera orientation	Lighting	Body visibility	SST	AP
1	Endotracheal tube	Oral	✓	Low ambient lighting (lights-out)	Frontal	Adequate	Full body	✓	0.65
2	Arterial line	L. radial	✗	—	Frontal	Inadequate	Upper body	✗	—
3	Nasogastric tube	Nasal	✓	—	Frontal	Adequate	Upper body	✓	0.64
4	Nasogastric tube	Nasal	✓	Face obscured by right-rear angle and IV stand	Frontal	Adequate	Upper body	✓	0.32
5	Nasogastric tube	Nasal	✗	Poor facial visibility and limited body information	Right-rear oblique	Adequate	Upper body	✗	—
6	Endotracheal tube	Oral	✗	Low ambient lighting (lights-out)	Frontal	Adequate	Upper body	✗	—
7	Nasogastric tube	Nasal	✗	Low ambient lighting (lights-out)	Frontal	Inadequate	Upper body	✗	—
8	Nasogastric tube	Nasal	✗	Low pose discriminability (tube removal pose in 88% of frames)	Frontal	Inadequate	Full body	✗	—
9	Nasogastric tube	Nasal	✗	Low ambient lighting (lights-out)	Frontal	Adequate	Upper body	✗	—
10	Nasogastric tube	Nasal	✗	Low ambient lighting (lights-out)	Frontal	Inadequate	Upper body	✗	—
11	Arterial line	L. radial	✗	Facial landmarks obscured by surgical mask	Frontal	Inadequate	Full body	✗	—
12	Nasogastric tube	Nasal	✗	Lateral camera angle	Frontal	Adequate	Full body	✗	—
13	Nasogastric tube	Nasal	✗		Right lateral	Adequate	Full body	✗	—

Note: Videos 2, 7, 8, 10, and 11 occurred during nighttime hours after lights-out, accounting for all cases with inadequate lighting. Failure reasons represent the most plausible explanation based on visual inspection; the exact cause could not be determined algorithmically. *Abbreviations:* ICU, intensive care unit; L, left; Pose est, MediaPipe-based pose estimation; SST, Singular Spectrum Transformation; AP, Average Precision; —, not applicable.

for additional sensors or modifications to the clinical environment.

MediaPipe was applied frame-by-frame to each extracted video clip, and the resulting skeletal coordinate data were visually inspected to assess estimation quality. For each video, a trained observer reviewed individual frames and evaluated whether the estimated landmark positions corresponded accurately to the patient's actual body landmarks visible in the footage. Cases in which landmarks were completely undetected or their estimated positions were judged to deviate substantially from the patient's actual posture—due to occlusion, poor lighting, or non-frontal camera angle—were classified as invalid and excluded from further analysis. Valid pose estimation was achieved in 3 of the 13 video clips (23.1%) (Table 1). Of 33 detectable landmarks, nine were selected for analysis: six body landmarks (including both thumbs and mouth corners) and three facial landmarks (nose and mouth corners), as these were considered most relevant to postures associated with tube removal.

For the three analyzable cases—one endotracheal tube removal (Video 1, oral) and two nasogastric tube removals (Videos 3 and 4, nasal)—six Euclidean distances between anatomically relevant landmark pairs were computed as time-series features. These pairs (left thumb–right mouth corner, left thumb–left mouth corner, right thumb–right mouth corner, right thumb–left mouth corner, nose–right mouth corner, and nose–left mouth corner) were chosen for their sensitivity to hand-to-face movements associated with both oral and nasal tube removal, resulting in a six-dimensional time-series dataset for each video.

To examine whether coordinates obtained *via* MediaPipe could capture tube-removal-related movement changes, we performed time-series change-point detection using the Singular Spectrum Transformation (SST). SST is a statistical method that detects shifts in distribution of sequential data by comparing trajectory and test matrices using singular value decomposition (*9*). For each video, SST parameters (window size w and lag L) were manually optimized after reviewing the data to align the peak change score with visually confirmed onset of tube-removal behavior. All computations were performed in Python (Version 3.11) on the AI Bridging Cloud Infrastructure (ABCI) of the National Institute of Advanced Industrial Science and Technology.

Given the severe class imbalance between normal frames and the few frames corresponding to tube-removal-related movements (approximately 10 abnormal frames per video), we used Average Precision (AP) as evaluation metric. AP quantifies how well tube-removal-related frames are correctly identified among the predominantly normal frames. To assess whether the SST-derived scores reflected meaningful signal above chance, AP values were also computed using random score distributions of equal length as a baseline. Absence of negative-control videos (footage from patients who did not attempt tube removal) is acknowledged as a limitation of the current study design.

4. Preliminary detection of tube-removal-related movement signals

Of the 13 consented video clips, valid skeletal coordinate time-series data were obtained from 3 cases (23.1%), comprising one endotracheal tube removal (Video 1) and two nasogastric tube removals (Videos 3 and 4). Failure in the remaining 10 cases was attributable to low ambient lighting after lights-out (Videos 2, 7, 8, 10, 11), non-frontal camera orientation (Videos 5, 13), surgical mask use obscuring facial landmarks (Video 12), poor facial visibility and limited body information (Video 6), and low pose discriminability due to tube removal posture being present in 88% of frames (Video 9) (Table 1). Among the three analyzable cases, all had adequate ambient lighting; one showed the patient's full body (Video 1) and two showed upper body only (Videos 3 and 4), with all patients clearly visible and centrally positioned within the frame.

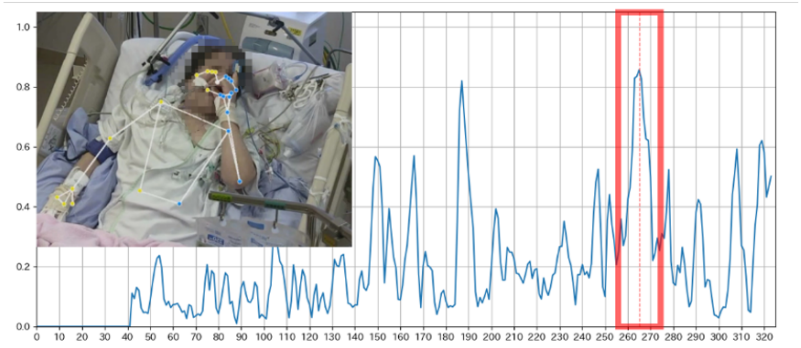
SST applied to the MediaPipe-derived inter-landmark distance data yielded prominent peaks in the change score at or immediately before the visually confirmed onset of tube-removal-related movements in all three analyzable cases (Figure 1). In Video 1 (endotracheal tube), peak change scores of a = 0.82 and a = 0.83 were observed at frames 188 and 265, corresponding to the left hand grasping the tube at the mouth and subsequent withdrawal movement, respectively. In Video 3 (nasogastric tube), consecutive peaks were observed from frame 750 through frame 1,230, corresponding to a sequence of actions including right-hand grasping and withdrawal of the tube, holding the withdrawn tube, reaching movements, and re-contact with the slack tube. In Video 4 (nasogastric tube), a peak of a = 0.65 at frame 1025 corresponded to left-hand tape removal followed by tube withdrawal.

As an exploratory evaluation, AP values for Videos 1 (AP = 0.65) and 3 (AP = 0.64) substantially exceeded the random-baseline AP (AP < 0.10 in all cases), indicating that detected change-score peaks reflected tube-removal-related signals rather than chance variation. In Video 4, the AP was 0.32, which exceeded the random baseline but was lower than the other two cases. This lower value was likely due to large bilateral arm movements occurring after tube removal, which generated higher change scores than the removal onset itself. These results are preliminary and exploratory, given the small sample size and absence of prospective validation.

5. Implications and limitations for future ICU monitoring

The present findings provide preliminary feasibility evidence that AI-based pose estimation using MediaPipe

(A): Video 1 (Endotracheal tube removal, oral)

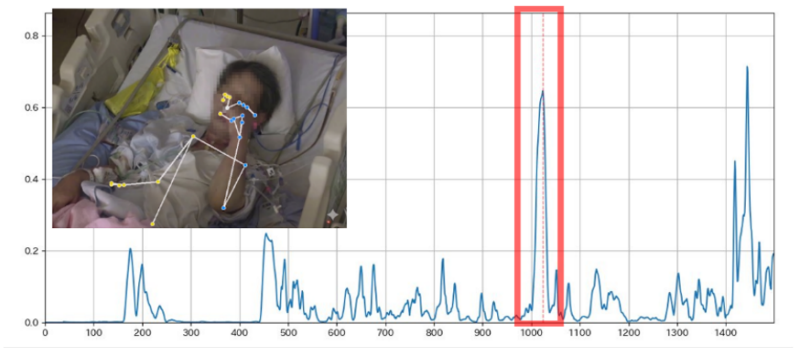


Frame	Degree of change	Motion description
188	0.82	Left-hand grasping of tube at mouth
265	0.83	Left-hand grasping and withdrawal of tube at mouth

(B): Video 3 — Nasogastric tube removal (nasal)

Frame	Degree of change	Motion description
750	0.26	Right-hand grasping and withdrawal of tube at nose
863	0.22	Right-hand holding of withdrawn tube
922	0.23	Right hand at chest, left hand reaching forward-left to grasp object
1230	0.22	Right-hand re-contact with slack of withdrawn tube

(C): Video 4 — Nasogastric tube removal (nasal)



Frame	Degree of change	Motion description
1025	0.65	Left-hand removal of tape around nose followed by tube withdrawal
1445	0.71	Bilateral hand contact with oxygen inhalation tube at nose

Figure 1. SST-based change scores with representative MediaPipe skeletal landmark overlays showing tube-removal-related movement onset in three analyzable ICU videos. (A) Video 1 (endotracheal tube removal, oral); **(B)** Video 3 (nasogastric tube removal, nasal). **(C)** Video 4 (nasogastric tube removal, nasal). In each panel, the blue line represents the SST-based change score at each frame (x-axis: frame number; y-axis: change score). The red box indicates primary SST peak region, which corresponded temporally to tube-removal-related movement observed in the video; the inset photograph shows the corresponding video frame as visual evidence of the detected movement. The dashed red line marks the frame of the primary change-score peak. The table within each panel lists frame number, change score, and corresponding motion description at key detected events. Facial blurring has been applied to all video frames shown for publication to protect patient privacy. *Abbreviation*: ICU, intensive care unit; SST, Singular Spectrum Transformation.

can extract skeletal coordinate data from ICU surveillance footage, and that SST-based change-point detection applied to these data can capture movement signals associated with tube-removal-related behaviors in real clinical ICU videos. To our knowledge, this is among the first demonstrations of such an approach using actual consented clinical footage of real tube removal events.

The potential of image-based motion analysis for unplanned tube removal detection is increasingly recognized internationally. An approach combining YOLOv3-based region-of-interest extraction with optical flow and support vector machine classification provided an early demonstration of video-based detection from RGB surveillance footage (*10*), and the AURA system subsequently used text-to-video diffusion models to generate synthetic ICU footage for training pose estimation-based detection of hand-to-tube proximity and agitation (*11*). The present study complements these approaches by using actual consented clinical footage, providing direct clinical validity that synthetic data cannot replicate. More broadly, Gabriel *et al.* demonstrated large-scale AI-driven patient monitoring across 11 hospitals (*12*), Nahin *et al.* showed deep learning pose estimation from infrared images in bedridden patients (*13*), and Feng *et al.* reported significant correlations between a pose-derived Movement Index (λMI) and GCS and RASS scores in ICU patients (*14*)—collectively affirming the clinical relevance of continuous posture-based monitoring.

Several limitations of the current work must be acknowledged in the interest of transparency. The proportion of videos yielding valid pose data was low (3/13, 23%), attributable to heterogeneous camera placement, low ambient lighting, non-frontal camera angles, and patient-specific factors such as surgical mask use. SST parameters were optimized post-hoc for each video, and absence of negative-control videos from patients who did not attempt tube removal precluded formal assessment of specificity. Notably, in one case (Video 9), the tube removal posture was present in 88% of frames, which likely prevented SST from detecting a meaningful change point. This highlights that change-point detection methods require sufficient contrast between normal and abnormal frames to function effectively. Additional limitations include small sample size, restriction to a single institution, lack of distinction between tube-removal-related and routine movements, and absence of prospective real-time validation. These limitations are particularly relevant given the challenging conditions of real-world ICU care during a pandemic.

Critically, the present results provide preliminary evidence that the core concept is feasible: AI-based pose estimation can capture anatomically meaningful movement signals from existing ICU surveillance infrastructure without modification to the clinical environment or additional burden on patients or staff. Practical barriers identified in this study are primarily technical and environmental, such as camera placement and lighting. Camera standardization protocols have already been demonstrated as feasible in large-scale clinical deployments (*12*), and use of infrared and RGB-D depth cameras has shown promise for reliable operation in low-light environments, including for pose estimation in bedridden patients (*13*). These technologies would directly address the nighttime lighting problem that accounted for most failures in this study, which is particularly relevant given that unplanned tube removal occurs more frequently during night shifts when lighting is reduced (*15,16*).

Looking ahead, a prospective implementation of this approach—with standardized camera placement, infrared imaging, and real-time alert functionality—could enable nursing staff to receive timely notifications before tube removal occurs, allowing targeted intervention without need for continuous physical restraint. As more annotated clinical data become available, supervised learning models could be developed to distinguish pre-removal movements from routine behaviors, advancing from statistical detection to genuine predictive monitoring. At the population level, continuous motion data could support individualized, dynamic risk stratification, replacing the current paradigm of default prolonged restraint with evidence-based, patient-specific care decisions.

The present study represents a first step toward that goal. Despite its limitations, it provides preliminary evidence from real clinical footage that AI-based contactless motion analysis may be a viable direction for ICU patient safety. Ultimately, a validated system of this kind could contribute to reducing unplanned tube removal and the physical and psychological burden of restraint in one of medicine's most vulnerable settings.

Funding: This work was supported by JSPS KAKENHI Grant Number JP.20K10753.

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Trenchat L, Galerneau LM, Ruckly S, *et al*. Self-extubation in critically ill patients: From the French OUTCOMEREA Network. Crit Care. 2025; 29:286.
2. Li P, Sun Z, Xu J. Unplanned extubation among critically ill adults: A systematic review and meta-analysis. Intensive Crit Care Nurs. 2022; 70:103219.
3. Zare-Kaseb A, Sarmadi S, Sanaie N, Emami Zeydi A. Prevalence and variability in use of physical restraints in intensive care units: A systematic review and meta-analysis. Aust Crit Care. 2025; 38:101210.
4. Franks ZM, Alcock JA, Lam T, Haines KJ, Arora N, Ramanan M. Physical restraints and post-traumatic stress disorder in survivors of critical illness. A systematic review and meta-analysis. Ann Am Thorac Soc. 2021; 18:689-697.

5. Acevedo-Nuevo M, Velasco-Sanz T, Del Olmo-Somolinos B, Vía-Clavero G. Ethical and legal considerations and recommendations for action in the physical restraint use in critically ill patients. Enferm Intensiva (Engl Ed). 2025; 36:100497.
6. Galazzi A, Adamini I, Consonni D, Roselli P, Rancati D, Ghilardi G, Greco G, Salinaro G, Laquintana D. Accidental removal of devices in intensive care unit: An eight-year observational study. Intensive Crit Care Nurs. 2019; 54:34-38.
7. Mizuno S, Ishizaki F, Umeda A, Okamoto T. Detection of self-extubation movements in the intensive care unit using a posture estimation model. Proc Annu Conf Jpn Soc Artif Intell. 2024; 38:3E5-GS-10-04.
8. Google AI for Developers. MediaPipe Pose Landmarker. *https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker* (accessed May 24, 2026).
9. Ide T, Sugiyama M. Anomaly Detection and Change Detection. Kodansha, Tokyo, Japan, 2015. (in Japanese)
10. Chen Y, Wang L, Wang G, Yang S, Wang Y, Xiang M, Zhang X, Chen H, Hu D, Cheng H. Spatio-temporal features for fast early warning of unplanned self-extubation in ICU. Eng Appl Artif Intell. 2024; 127:107294.
11. Seo J, Moon H, Jung KH, Oh N, Kim T. AURA: Development and validation of an augmented unplanned removal alert system using synthetic ICU videos. *https://arxiv.org/abs/2511.12241* (accessed May 24, 2026).
12. Gabriel P, Rehani P, Troy T, Wyatt T, Choma M, Singh N. Continuous patient monitoring with AI: Real-time analysis of video in hospital care settings. Front Imaging. 2025; 4:1547166.
13. Nahin SK, Acharjee S, Saha S, Das A, Hossain S, Haque MA. Human sleeping pose estimation from IR images for in-bed patient monitoring using image processing and deep learning techniques. Heliyon. 2024; 10:e36823.
14. Feng R, Richter F, Mari E, Gleason A, Le C, Kellner CP, Shrivastava RK, Fields M, Rapoport BI, Bederson JB, Schadt EE, Glicksberg BS, Richter F, Dangayach NS. Artificial intelligence monitoring of neurological status from patient videos in the neuroscience intensive care unit. Neurosurgery. 2026; doi: 10.1227/neu.0000000000003899.
15. Yeh SH, Lee LN, Ho TH, Chiang MC, Lin LW. Implications of nursing care in the occurrence and consequences of unplanned extubation in adult intensive care units. Int J Nurs Stud. 2004; 41:255-262.
16. Chang LC, Liu PF, Huang YL, Yang SS, Chang WY. Risk factors associated with unplanned endotracheal self-extubation of hospitalized intubated patients: A 3-year retrospective case-control study. Appl Nurs Res. 2011; 24:188-192.

Received May 27, 2026; Revised June 13, 2026; Accepted June 16, 2026.

Released online in J-STAGE as advance publication June 19, 2026.

**Address correspondence to*:
Aya Umeda, National College of Nursing, Japan (NCNJ), Japan Institute for Health Security, 1-2-1 Umezono, Kiyose-shi, Tokyo, 204-0024, Japan.
E-mail: a-umeda@umin.ac.jp

Print ISSN: 2434-9186
Online ISSN: 2434-9194
Issues/Year: 6
Language: English

GHM
Global Health & Medicine

1. Scope of Articles

Global Health & Medicine is (Print ISSN 2434-9186, Online ISSN 2434-9194) is an international, open-access, peer-reviewed journal dedicated to publishing high-quality original research that contributes to advancing global health and medicine, with the goal of creating a global information network for global health, basic science as well as clinical science oriented for clinical application.

We encourage submission of original research findings in the fields of global health, public health, and health care delivery as well as the seminal and latest research on the intersection of biomedical science and clinical practice.

2. Types of Articles

Original Articles should be well-documented, novel, and significant to the field as a whole. They should include an abstract and be structured as follows: Title page, Abstract, Introduction, Materials and Methods, Results, Discussion, Acknowledgments, References, Figures and/or Tables; and Supplementary Data, if appropriate. Original articles should not exceed 5,000 words in length (excluding references) and should be limited to a maximum of 50 references. Articles may contain a maximum of 10 figures and/or tables. Supplementary Data are permitted but should be limited to information that is not essential to the general understanding of the research presented in the main text, such as unaltered blots and source data as well as other file types.

Types of Articles	Words in length (excluding references)	Figures and/or Tables	References
Original Articles	~5,000	~10	~50
Brief Reports	~3,000	~5	~30
Reviews	~8,000	~10	~100
Mini reviews	~4,000	~5	~50
Policy Forum articles	~3,000	~5	~30
Communications	~2,000	~2	~20
Perspectives			
Comments			
Correspondence			
Editorials	~1,000	~1	~10
Letters	~1,000	~1	~10
News	~800	~1	~5

Abstract: ~250 words (Original Articles, Brief Reports, Reviews, Policy Forum); ~150 words (Communications, Editorials, Letters, and News).
Keywords: 3~6 words

Brief Reports definitively documenting either experimental results or informative clinical observations will be considered for publication in this category. Brief Reports are not intended for publication of incomplete or preliminary findings. Brief Reports should not exceed 3,000 words in length (excluding references) and should be limited to a maximum of 5 figures and/or tables and 30 references. Brief Reports should be structured as follows: Title page, Abstract, Introduction, Materials and Methods, Results and Discussion, Acknowledgments, References, Figures and/or Tables; and Supplementary Data, if appropriate.

Reviews should present a full and up-to-date account of recent developments within an area of research. Normally, reviews should not exceed 8,000 words in length (excluding references) and should be limited to a maximum of 100 references and up to 10 figures and/or tables. Mini reviews are also accepted, which should not exceed 4,000 words in length (excluding references), have no more than 50 references, and have up to 5 figures and/or tables.

Policy Forum articles discuss research and policy issues in areas related to global health and medicine, such as public health, medical care, and social science that may address governmental issues at district, national, and international levels of discourse. Policy Forum articles should not exceed 3,000 words in length (excluding references), have no more than 30 references, and have up to 5 figures and/or tables.

Communications are short, timely pieces that spotlight new research findings or policy issues of interest to the field of global health and medical practice that are of immediate importance. Depending on their content, Communications will be published as "Perspectives", "Comments", or "Correspondence". Communications should not exceed 2,000 words in length (excluding references), have no more than 20 references, and have up to 2 figures and/or tables.

Editorials are short, invited opinion pieces that discuss an issue of immediate importance to the fields of global health, medical practice, and basic science oriented for clinical application. Editorials should not exceed 1,000 words in length (excluding references), have no more than 10 references, and have one figure or table.

Letters are articles that provide readers with an opportunity to respond to an article published in *Global Health & Medicine* within the previous two months or to raise issues of general interest to our readers. Letters should provide new information or insights. If appropriate, letters are sent to the authors of the article in question for a response. Letters should not exceed 1,000 words in length (excluding references), have no more than 10 references, and have one figure or table.

News articles should report the latest events in health sciences and medical research from around the world. News should not exceed 800 words in length (excluding references), have no more than 5 references, and have one figure or table.

3. Formatting Guidelines

Manuscripts should be written in clear, grammatically correct English and submitted as a Microsoft Word file in a single-column format. Manuscripts must be paginated and typed in 12-point Times New Roman font with 24-point line spacing. Please do not embed figures in the text. Technical terms should be defined. Abbreviations should be used as little as possible and should be explained at first mention unless the term is a well-known abbreviation (*e.g.* DNA). Single words should not be abbreviated. Please include page numbers in your submitted file. We also encourage use of line numbers.

The submission to *Global Health & Medicine* should include:

1. Cover letter
2. Main manuscript
3. Figures
4. Supplementary Data, if appropriate

The main manuscripts should be assembled in the following order:

1. Title page
2. Abstract
3. Main Text
4. Acknowledgments
5. References
6. Tables
7. Figure Legend
8. List of Supplementary Data, if appropriate

For manuscript samples, please visit *https://www.globalhealthmedicine.com/site/download.html* (Download Center).

Please provide all figures as separate files in an acceptable format (TIFF

or JPEG). Supplementary Data should also be submitted as a single separate file in Microsoft Word format.

An abstract is necessary for all types of articles. An Original Article should be structured as follows: Title page, Abstract, Introduction, Materials and Methods, Results, Discussion, Acknowledgments, References, Figures and/or Tables; and Supplementary Data, if appropriate. A Brief Report contains the same sections as an Original Article, but the Results and Discussion sections should be combined. For manuscripts that are Reviews, Policy Forum articles, Communications, Editorials, Letters, or News, subheadings should be used for increased clarity.

4. Manuscript Preparation

Title page: The title page must include 1) the title of the paper (Please note the title should be short, informative, and contain the major key words); 2) full name(s) and affiliation(s) of the author(s), 3) abbreviated names of the author(s), 4) full name, mailing address, telephone/fax numbers, and e-mail address of the corresponding author; and 5) conflicts of interest (if you have an actual or potential conflict of interest to disclose, it must be included as a footnote on the title page of the manuscript; if no conflict of interest exists for each author, please state "There is no conflict of interest to disclose").

Abstract: The abstract should briefly state the purpose of the study, methods, main findings, and conclusions. For articles that are Original Articles, Brief Reports, Reviews, or Policy Forum articles, a one-paragraph abstract consisting of no more than 250 words must be included in the manuscript. For Communications, Editorials, Letters, and News, a one-paragraph brief summary of the main content in 150 words or less should be included in the manuscript. Abbreviations must be kept to a minimum and non-standard abbreviations should be explained in brackets at first mention. References should be avoided in the abstract. Three to six key words or phrases that do not occur in the title should be included on the Abstract page.

Introduction: The introduction should provide sufficient background information to make the article intelligible to readers in other disciplines and sufficient context clarifying the significance of the experimental findings.

Materials/Patients and Methods: The description should be brief but with sufficient detail to enable others to reproduce the experiments. Procedures that have been published previously should not be described in detail but appropriate references should simply be cited. Only new and significant modifications of previously published procedures require complete description. Names of products and manufacturers with their locations (city and state/country) should be given and sources of animals and cell lines should always be indicated. All clinical investigations must have been conducted in accordance with the Declaration of Helsinki (as revised in 2013, *https://wma.net/what-we-do/medical-ethics/declaration-of-helsinki*). All human and animal studies must have been approved by the appropriate institutional review board(s) and a specific declaration of approval must be made within this section.

Results: The description of the experimental results should be succinct but in sufficient detail to allow the experiments to be analyzed and interpreted by an independent reader. If necessary, subheadings may be used for an orderly presentation. Two levels of subheadings may be used if warranted, please distinguish them clearly. All Figures and Tables should be cited in order, including those in the Supplementary Data.

Discussion: The data should be interpreted concisely without repeating material already presented in the Results section. Speculation is permissible, but it must be well-founded, and discussion of the wider implications of the findings is encouraged. Conclusions derived from the study should be included in this section.

Acknowledgments: All funding sources should be credited in the Acknowledgments section. In addition, people who contributed to the work but who do not meet the criteria for authors should be listed along with their contributions.

References: References should be numbered in the order in which they appear in the text. Two references are cited separated by a comma, with no space, for example (*1,2*). Three or more consecutive references are given as a range with an en rule, for example (*1-3*). Citing of unpublished results, personal communications, conference abstracts, and theses in the reference list is not recommended but these sources may be mentioned in the text. In the reference list, cite the names of all authors when there are fifteen or fewer authors; if there are sixteen or more authors, list the first three followed by *et al*. Names of journals should be abbreviated in the style used in PubMed. Authors are responsible for the accuracy of the references. The EndNote Style of *Global Health & Medicine* could be downloaded at Download Center.

Examples are given below:

Example 1 (Sample journal reference):
Kokudo N, Hara T. "History, Tradition, and Progress": The ceremony of 150th Anniversary of the National Center for Global Health and Medicine held in Tokyo, Japan. BioSci Trends. 2019; 13:105-106.

Example 2 (Sample journal reference with more than 15 authors):
Darby S, Hill D, Auvinen A, *et al*. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ. 2005; 330:223.

Example 3 (Sample book reference):
Shalev AY. Post-traumatic stress disorder: Diagnosis, history and life course. In: Post-traumatic Stress Disorder, Diagnosis, Management and Treatment (Nutt DJ, Davidson JR, Zohar J, eds.). Martin Dunitz, London, UK, 2000; pp. 1-15.

Example 4 (Sample web page reference):
World Health Organization. The World Health Report 2008 – primary health care: Now more than ever. *http://www.who.int/whr/2008/whr08_en.pdf* (accessed March 20, 2022).

Tables: All tables should be prepared in Microsoft Word and should be arranged at the end of the manuscript after the References section. Please note that tables should not be in image format. All tables should have a concise title and should be numbered consecutively with Arabic numerals. Every vertical column should have a heading, consisting of a title with the unit of measure in parentheses. If necessary, additional information should be given below the table.

Figure Legend: The figure legend should be typed on a separate page of the main manuscript and should include a short title and explanation. The legend should be concise but comprehensive and should be understood without referring to the text. Symbols used in figures must be explained. Any individually labeled figure parts or panels (A, B, *etc*.) should be specifically described by part name within the legend.

Figure Preparation: All figures should be clear and cited in numerical order in the text. Figures must fit in a one- or two-column format on the journal page: 8.3 cm (3.3 in.) wide for a single column, 17.3 cm (6.8 in.) wide for a double column; maximum height: 24.0 cm (9.5 in.). Please make sure that the symbols and numbers appearing in the figures are clear. Please make sure that artwork files are in an acceptable format (TIFF or JPEG) at minimum resolution (600 dpi for illustrations, graphs, and annotated artwork, and 300 dpi for micrographs and photographs). Please provide all figures as separate files. Please note that low-resolution images are one of the leading causes of article resubmission and scheduling delays.

Units and Symbols: Units and symbols conforming to the International System of Units (SI) should be used for physicochemical quantities. Solidus notation (*e.g*. mg/kg, mg/mL, mol/mm^2/min) should be used. Please refer to the SI Guide www.bipm.org/en/si/ for standard units.

Supplemental Data: Supplemental data might help to support and enhance your manuscript. *Global Health & Medicine* accepts the submission of these materials, which will be only published online alongside the electronic version of your article. Supplemental files (figures, tables, and other text materials) should be prepared according to the above guidelines, numbered in Arabic numerals (*e.g.*, Figure S1, Figure S2, and Table S1, Table S2), and referred to in the text. All figures and tables should have titles and legends. All figure legends, tables and supplemental text materials should be placed at the end of the paper. Please note all of these supplemental data should be provided at the time of initial submission and note that the editors reserve the right to limit the size and length of Supplemental Data.

5. Cover Letter

The manuscript must be accompanied by a cover letter prepared by the corresponding author on behalf of all authors. The letter should indicate the basic findings of the work and their significance. The letter should also include a statement affirming that all authors concur with the submission and that the material submitted for publication has not been published previously or is not under consideration for publication elsewhere. For example of Cover Letter, please visit *https://www.globalhealthmedicine.com/site/download.html* (Download Center).

6. Submission Checklist

The Submission Checklist will be useful during the final checking of a manuscript prior to sending it to Global Health & Medicine for review. Please visit *https://www.globalhealthmedicine.com/site/download.html* and download the Submission Checklist file.

7. Online Submission

Manuscripts should be submitted to *Global Health & Medicine* online at *https://www.globalhealthmedicine.com/site/login.html*. Receipt of your manuscripts submitted online will be acknowledged by an e-mail from Editorial Office containing a reference number, which should be used in all future communications. If for any reason you are unable to submit a file online, please contact the Editorial Office by e-mail at *office@globalhealthmedicine.com*

8. Editorial Policies

For publishing and ethical standards, *Global Health & Medicine* follows the Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals issued by the International Committee of Medical Journal Editors (ICMJE, *https://icmje.org/recommendations*), and the Principles of Transparency and Best Practice in Scholarly Publishing jointly issued by the Committee on Publication Ethics (COPE, *https://publicationethics.org/resources/guidelines-new/principles-transparency-and-best-practice-scholarly-publishing*), the Directory of Open Access Journals (DOAJ, *https://doaj.org/apply/transparency*), the Open Access Scholarly Publishers Association (OASPA, *https://oaspa.org/principles-of-transparency-and-best-practice-in-scholarly-publishing-4*), and the World Association of Medical Editors (WAME, *https://wame.org/principles-of-transparency-and-best-practice-in-scholarly-publishing*).

Global Health & Medicine will perform an especially prompt review to encourage submissions of innovative work. All original research manuscripts are to be subjected to an expeditious but rigorous standard of peer review, and are to be edited by experienced copy editors to the highest standards.

The publishing is supported by the International Research and Cooperation Association for Bio & Socio-Sciences Advancement (IRCA-BSSA) Group Journals. The editorial office comprises a range of experienced individuals, including managing editor, editorial associates, software specialists, and administrative coordinators to provide a smooth service for authors and reviewers.

Ethical Approval of Studies and Informed Consent: For all manuscripts reporting data from studies involving human participants or animals, formal review and approval, or formal review and waiver, by an appropriate institutional review board or ethics committee is required and should be described in the Methods section. When your manuscript contains any case details, personal information and/or images of patients or other individuals, authors must obtain appropriate written consent, permission, and release in order to comply with all applicable laws and regulations concerning privacy and/or security of personal information. The consent form needs to comply with the relevant legal requirements of your particular jurisdiction, and please do not send the signed consent form to *Global Health & Medicine* in order to respect your patient's and any other individual's privacy. Please instead describe the information clearly in the Methods (patient consent) section of your manuscript while retaining copies of the signed forms in the event they should be needed. Authors should also state that the study conformed to the provisions of the Declaration of Helsinki (as revised in 2013, *https://wma.net/what-we-do/medical-ethics/declaration-of-helsinki*). When reporting experiments on animals, authors should indicate whether the institutional and national guide for the care and use of laboratory animals was followed.

Reporting Clinical Trials: The ICMJE (*https:// icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html*) defines a clinical trial as any research project that prospectively assigns people or a group of people to an intervention, with or without concurrent comparison or control groups, to study the relationship between a health-related intervention and a health outcome. Registration of clinical trials in a public trial registry at or before the time of first patient enrollment is a condition of consideration for publication in *Global Health & Medicine*, and the trial registration number will be published at the end of the Abstract. The registry must be independent of for-profit interest and be publicly accessible. Reports of trials must conform to CONSORT 2010 guidelines (*https://consort-statement.org/consort-2010*). Articles reporting the results of randomized trials must include the CONSORT flow diagram showing the progress of patients throughout the trial.

Conflict of Interest: All authors are required to disclose any actual or potential conflict of interest, including financial interests or relationships with other people or organizations that might raise questions of bias in the work reported. If no conflict of interest exists for each author, please state "There is no conflict of interest to disclose".

Submission Declaration: When a manuscript is considered for submission to *Global Health & Medicine*, the authors should confirm that 1) no part of this manuscript is currently under consideration for publication elsewhere; 2) this manuscript does not contain the same information in whole or in part in manuscripts that have been published, accepted, or are under review elsewhere, except in the form of an abstract, a letter to the editor, or part of a published lecture or academic thesis; 3) authorization for publication has been obtained from the authors' employer or institution; and 4) all contributing authors have agreed to submit this manuscript.

Initial Editorial Check: Immediately after submission, the journal's managing editor will perform an initial check of the manuscript. A suitable academic editor will be notified of the submission and invited to check the manuscript and recommend reviewers. Academic editors will check for plagiarism and duplicate publication at this stage. The journal has a formal recusal process in place to help manage potential conflicts of interest of editors. In the event that an editor has a conflict of interest with a submitted manuscript or with the authors, the manuscript, review, and editorial decisions are managed by another designated editor without a conflict of interest related to the manuscript.

Peer Review: *Global Health & Medicine* operates a single-anonymized review process, which means that reviewers know the names of the authors, but the authors do not know who reviewed their manuscript. All articles are evaluated objectively based on academic

content. External peer review of research articles is performed by at least two reviewers, and sometimes the opinions of more reviewers are sought. Peer reviewers are selected based on their expertise and ability to provide quality, constructive, and fair reviews. For research manuscripts, the editors may, in addition, seek the opinion of a statistical reviewer. Every reviewer is expected to evaluate the manuscript in a timely, transparent, and ethical manner, following the COPE guidelines (*https://publicationethics.org/files/cope-ethical-guidelines-peer-reviewers-v2_0.pdf*). We ask authors for sufficient revisions (with a second round of peer review, when necessary) before a final decision is made. Consideration for publication is based on the article's originality, novelty, and scientific soundness, and the appropriateness of its analysis.

Suggested Reviewers: A list of up to 3 reviewers who are qualified to assess the scientific merit of the study is welcomed. Reviewer information including names, affiliations, addresses, and e-mail addresses should be provided at the same time the manuscript is submitted online. Please do not suggest reviewers with known conflicts of interest, including participants or anyone with a stake in the proposed research; anyone from the same institution; former students, advisors, or research collaborators (within the last three years); or close personal contacts. Please note that the Editor-in-Chief may accept one or more of the proposed reviewers or request a review by other qualified persons.

Submission Turnaround Time:
- From submission to first editorial decision: 1-2 weeks.
- From acceptance to publication ahead of print: 1-4 weeks.
- From acceptance to publication: 2-6 months. Original Articles are listed as priority.

Language Editing: Manuscripts prepared by authors whose native language is not English should have their work proofread by a native English speaker before submission. If not, this might delay the publication of your manuscript in *Global Health & Medicine*.

Copyright and Reuse: Before a manuscript is accepted for publication in *Global Health & Medicine*, authors will be asked to sign a transfer of copyright agreement, which recognizes the common interest that both the journal and author(s) have in the protection of copyright. We accept that some authors (*e.g.*, government employees in some countries) are unable to transfer copyright. A JOURNAL PUBLISHING AGREEMENT (JPA) form will be e-mailed to the authors by the Editorial Office and must be returned by the authors by mail, fax, or as a scan. Only forms with a hand-written signature from the corresponding author are accepted. This copyright will ensure the widest possible dissemination of information. Please note that the manuscript will not proceed to the next step in publication until the JPA Form is received. In addition, if excerpts from other copyrighted works are included, the author(s) must obtain written permission from the copyright owners and credit the source(s) in the article.

9. Accepted Manuscripts

Proofs: Galley proofs in PDF format will be e-mailed to the corresponding author. Corrections must be returned to the editor (*office@globalhealthmedicine.com*) within 3 working days.

Offprints: Authors will be provided with electronic offprints of their article. Paper offprints can be ordered at prices quoted on the order form that accompanies the proofs.

Article Processing Charges: The open-access policy of *Global Health & Medicine* will allow all readers from the medical and scientific community to freely utilize material published in the journal. To support open access, article processing charges will be applied to manuscripts accepted for publication: JPY 165,000 for Original Articles, Brief Reports, Reviews, and Policy Forum articles; and JPY 110,000 for Communications, Editorials, Letters, and News articles. In exceptional circumstances, authors may apply for a waiver of article processing charges by clearly stating the reason in the Cover Letter at the time of initial submission *via* the online submission system. All invited articles are free of charge.

Article processing charges pay for: Immediate, worldwide open access to the full article text; Preparation in various formats for print & online publication; Inclusion in global important platforms, enabling electronic citation in other journals that are available electronically.

Misconduct: *Global Health & Medicine* takes seriously all allegations of potential misconduct and adhere to the ICMJE Guideline (*https://icmje.org/recommendations*) and COPE Guideline (*https://publicationethics.org/files/Code_of_conduct_for_journal_editors.pdf*). In cases of suspected research or publication misconduct, it may be necessary for the Editor or Publisher to contact and share submission details with third parties including authors' institutions and ethics committees. The corrections, retractions, or editorial expressions of concern will be performed in line with above guidelines.

(*As of August 2025*)

Global Health & Medicine
Japan Institute for Health Security,
1-21-1 Toyama Shinjuku-ku, Tokyo 162-8655, Japan
URL: *www.globalhealthmedicine.com*
E-mail: *office@globalhealthmedicine.com*

Print ISSN: 2434-9186 Online ISSN: 2434-9194

GHM

Global Health & Medicine

Volume 1, Number 1
October, 2019

www.globalhealthmedicine.com