Risk Management and Healthcare Policy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Risk Management of Large Language Model-Based Exercise and Health Guidance: A China-Anchored, Comparatively Informed Six-Dimensional Trigger Matrix and Lifecycle Governance Framework for the Wellness-to-SaMD Continuum

Kaijiang Pan[1], Xinyu Lin[2], Shengqi Huang[3], Caihua Huang[4]

[1]School of Marxism, Xiamen Ocean Vocational College, Xiamen, Fujian, People's Republic of China; [2]School of Film and Communication, Xiamen University of Technology, Xiamen, Fujian, People's Republic of China; [3]Huandaolu (Xiamen) Sports and Health Service Co., Ltd, Xiamen, Fujian, People's Republic of China; [4]Research and Communication Center for Exercise and Health, Xiamen University of Technology, Xiamen, Fujian, People's Republic of China

Correspondence: Caihua Huang, Research and Communication Center for Exercise and Health, Xiamen University of Technology, Xiamen, Fujian, People's Republic of China, Email caihua.huang@foxmail.com

Abstract: LLM-based exercise and health guidance creates a distinctive risk-management challenge: seemingly modest changes in product claims, target users, data inputs, personalization, automation, human oversight, or updates may move a tool from general wellness support toward higher-risk medical use. In exercise prescription and rehabilitation, an unsafe recommendation can affect physical load, recognition of warning symptoms, and timely referral. We conducted a structured narrative review and doctrinal/comparative legal analysis, anchored in China's National Medical Products Administration (NMPA) framework and informed by the European Union Medical Device Regulation (MDR), the EU Artificial Intelligence Act, and US Food and Drug Administration (FDA) and International Medical Device Regulators Forum (IMDRF) materials. Peer-reviewed literature was primarily searched for 2019–2026, with foundational regulatory, legal, and technical guidance included where directly relevant. We propose a six-dimensional trigger matrix covering intended use and claims, user context, depth of personalization, data and sensor sources, automation, human oversight and closed-loop control, and upgrade and change pathways. The framework includes anchored Green/Yellow/Red coding rules, non-compensatory aggregation rules, a structured governance checklist, and a lifecycle pathway for evidence generation, risk management, and change control. In a preliminary application exercise, three independent raters applied the coding rules to seven standardized hypothetical scenarios and achieved complete agreement on all dimension-level and overall designations (Fleiss' kappa = 1.00). This small exercise supports initial reproducibility of the rubric but does not establish legal classification accuracy, clinical validity, or real-world effectiveness. The proposed framework is intended to support earlier risk identification, evidence planning, procurement review, and dialogue among developers, healthcare institutions, and regulators; it does not replace product-specific legal analysis or regulatory determination.

Plain Language Summary: Large language models can give advice about exercise, weight management, rehabilitation, and healthy living. A tool may begin as general fitness coaching but become riskier when it starts to use medical records or medical-device data, gives advice for people with a disease, changes plans automatically, or claims to treat or monitor a health condition. In these situations, an unsafe recommendation could ask a person to exercise too hard, fail to respond appropriately to warning symptoms, or delay professional care. This study offers a practical way to identify these changes early. We reviewed Chinese medical-device rules and compared them with European and US approaches. We propose six questions about a tool's purpose, users, tailoring, data, automation, professional review, and updates. Three independent raters used the draft questions on seven example tools and reached the same overall judgments. This early test does not prove that the framework can determine a product's legal status or clinical safety. It can help developers, health services, and regulators ask the right safety questions before a tool is widely used.

https://doi.org/10.2147/RMHP.S624615
Received: 14 May 2026
Accepted: 7 July 2026
Published: 24 July 2026

Keywords: healthcare risk management, artificial intelligence, large language models, software as a medical device, exercise prescription, human oversight

Introduction

LLMs at the Exercise-and-Health Decision Point

Large language models (LLMs) are rapidly entering medicine and public health because they can interpret and generate natural-language content at scale.[1–6] In digital health, they are increasingly used to provide exercise, nutrition, rehabilitation, and health-coaching guidance to the general public and to people using health services.[7–9] These systems may support behavior change and access to information; when linked to user-reported histories, consumer wearable data, or clinical information, they may also generate individualized training, recovery, or follow-up suggestions.[7–14] Their role is therefore moving beyond static health education toward decision points that can influence physical activity, symptom response, and self-management.

This article focuses on LLM-based tools used in exercise prescription, weight management, rehabilitation, chronic-disease support, and workplace health programs.[10–17] The same underlying language model may be deployed as a general wellness coach, a semi-personalized support tool, or a component of a clinical workflow. The resulting risk profile depends not only on the model itself, but also on what the product claims to do, who it is intended for, what data it uses, how strongly it tailors or automates recommendations, whether qualified professionals review safety-critical outputs, and how the system changes after deployment.[1–4,14–18]

Why LLM-Based Exercise Guidance Presents a Distinctive Risk-Management Problem

In many settings, an inaccurate LLM response is principally an information-quality problem. In exercise prescription and rehabilitation, however, an inaccurate response may rapidly translate into physical load. For example, an inappropriate increase in exercise intensity for a person after a coronary intervention, or inadequate escalation when chest discomfort or unusual breathlessness is reported, may expose the user to unsuitable exertion or delay clinical assessment.[16,17] Similarly, generic exercise advice that overlooks contraindications, symptoms, or clinician-imposed restrictions during pregnancy may be inappropriate for the individual concerned.[19] These examples do not imply that all exercise guidance is medical care; rather, they show why the consequences of error depend on the intended use, user context, available data, and degree of automation.

Human oversight is a structural safeguard in this setting, not merely a disclosure item. Professional review, predefined referral thresholds, and mandatory interception of safety-critical outputs can reduce the risk of inappropriate automated action and clarify responsibility within a service pathway.[20–30] At the same time, human oversight does not erase an otherwise explicit medical purpose. A system that claims to diagnose, monitor, treat, or manage a disease may still require a medical-device pathway even when a clinician reviews some outputs. For this reason, the present framework treats human oversight as part of the automation and closed-loop dimension while also recognizing it as a cross-cutting governance requirement.

China-Anchored Regulatory Boundary and Governance Gap

The boundary between general wellness support and medical use is difficult to draw because it can change with product claims and deployment choices, without a major change in the underlying model. A wellness tool may move toward symptom triage, disease management, or clinical decision support when it processes disease-specific indicators, uses electronic health record (EHR) or medical-device data, makes therapeutic or risk-related claims, or changes exercise recommendations without appropriate review.[10,14–18,26–42]

This analysis is anchored in China's NMPA-led medical-device regulatory system, including rules and technical guidance concerning medical-device software, artificial intelligence-enabled medical devices, registration, and data governance.[26,27,31–33,43–45] The European Union MDR and AI Act, together with FDA and IMDRF materials, are used as comparative reference points rather than as a universal classification standard.[28–39] This China-anchored,

comparatively informed approach is intended to identify transferable governance signals while respecting that formal classification, evidence, and post-market requirements remain jurisdiction- and product-specific.

Existing regulations and governance frameworks provide important principles for intended use, data quality, validation, human oversight, documentation, and lifecycle control.[20–30,34–39] However, developers and healthcare institutions still have limited product-level guidance for recognizing when incremental changes in an LLM-based exercise or health tool create a materially different safety and regulatory profile. The gap is particularly relevant before a product is deployed in a hospital, rehabilitation, occupational-health, or consumer-health setting, when design choices can still be revised and evidence plans can still be established.

Aim, Questions, and Bounded Contribution

This article develops a proposed risk-management framework for LLM-based exercise and health guidance. It asks: (1) which product characteristics most strongly move a tool from wellness positioning toward SaMD-like or other higher-risk medical functions; (2) how developers, healthcare institutions, and procurement reviewers can identify these shifts before formal regulatory classification; and (3) what lifecycle governance activities should be considered as the product evolves. The article does not provide a legal classification rule for any product or jurisdiction.

Three linked outputs are presented: a six-dimensional trigger matrix; an anchored Green/Yellow/Red rubric with non-compensatory aggregation rules; and a structured governance checklist and lifecycle pathway for evidence planning, risk control, and change management. The six dimensions are intended use and claims; user context; depth of personalization; data and sensor sources; automation, human oversight and closed-loop control; and upgrade and change pathways. A preliminary independent-rater exercise using seven standardized hypothetical scenarios was added to examine whether the rubric could be applied consistently. This limited exercise supports initial reproducibility only; it does not establish clinical validity, legal classification accuracy, or real-world effectiveness. The framework is therefore intended to support earlier risk identification, structured governance planning, and dialogue among developers, healthcare institutions, and regulators.

Materials and Methods

Structured Narrative Review

This article adopts a structured narrative review approach to synthesize literature and policy materials relevant to LLMs, digital health, and medical-device regulation. Searches were conducted in major bibliographic databases, including CNKI, PubMed, Web of Science, and Google Scholar, together with official websites and repositories of the NMPA, National Health Commission, European institutions, and the FDA.

Representative English-language search terms included combinations of “large language model”, “generative AI”, “ChatGPT”, “health”, “healthcare”, “regulation”, “medical device”, and “SaMD”. For Chinese-language searches, combinations of the Chinese equivalents of “generative AI”, “large language model”, “ChatGPT”, “digital health”, “internet medicine”, “exercise prescription”, “medical device”, and “software as a medical device” were used. The initial search round was completed on 30 June 2024, and the final update was completed on 15 March 2026.

Included materials comprised peer-reviewed articles, official regulations and policy documents, professional consensus statements, and authoritative technical guidance. Peer-reviewed literature was primarily searched for the period 2019–2026, while foundational regulatory, legal, and technical guidance documents published before 2019 were included where directly relevant to SaMD, medical-device software, AI governance, or healthcare risk management. To be included, a source needed to be relevant to the use, governance, regulation, or quality evaluation of AI or LLMs in health or healthcare, with particular attention to digital health, SaMD, exercise prescription, rehabilitation, and health management. Sources focused solely on model architecture or technical performance without meaningful connection to health applications or regulation were excluded.

This review is not a systematic review, scoping review, or meta-analysis. Accordingly, no PRISMA flow diagram was produced. Instead, the purpose was to provide a structured and transparent narrative synthesis that combined database searching, source selection, thematic coding, and comparative regulatory analysis. To improve transparency, this article specifies the search window, source types, inclusion logic, exclusion logic, and analytic dimensions used for source

selection and coding. Sources were not screened for quantitative synthesis; rather, they were selected for conceptual, legal, regulatory, and implementation relevance. After initial retrieval, materials were grouped into four analytic categories: LLMs in healthcare; AI/LLM safety, reporting, and quality evaluation; exercise-, rehabilitation-, and chronic-disease-related digital health applications; and SaMD/AI medical-device regulation and lifecycle governance. These categories informed the six dimensions of the proposed framework.[20–25]

Doctrinal and Comparative Legal Analysis

Building on the review, the article undertakes doctrinal analysis of Chinese legal and regulatory materials together with comparative analysis of EU and US materials.[26–39] On the Chinese side, the analysis focuses on the Regulations on the Supervision and Administration of Medical Devices, the Rules for the Classification of Medical Devices, the Administrative Measures for the Registration and Filing of Medical Devices, AI and software review guidelines, and major data-governance statutes such as the Personal Information Protection Law, Data Security Law, and Cybersecurity Law.[26,27,31–33,43–45] On the comparative side, the analysis draws on the EU MDR, the EU AI Act, and FDA guidance on SaMD, AI-enabled device software functions, GMLP, and lifecycle and change-control expectations.[28–30,34–39]

The purpose of the comparative analysis is not to offer exhaustive commentary on each jurisdiction, but to identify functionally transferable regulatory concerns. These include intended use, product claims, risk stratification, target population, dependence on medical-grade data, degree of automation, human oversight, and governance of updates and post-market monitoring.[26–39]

Framework Synthesis, Source Traceability, and Construct-Retention Logic

Building on the structured narrative review and comparative legal analysis, we undertook a document-level traceability audit to make the derivation of the proposed framework explicit. Rather than beginning with six fixed labels and retrospectively locating supporting citations, we first extracted source-specific governance signals from a core corpus of binding legal instruments, formal regulator guidance, international SaMD and good-machine-learning-practice documents, and health-AI evaluation and reporting guidance. The extraction focused on product characteristics or control arrangements that could materially change the declared medical purpose, the healthcare situation in which outputs are used, the pathway from individual information to a safety-relevant action, or the level of lifecycle governance required. The traceability audit is presented in Supplementary Table 3.[20–45]

A candidate construct was retained as a matrix dimension only when it met all three pre-specified criteria: (1) it was anchored in at least two independent source families, jurisdictions, or international regulatory frameworks; (2) a change in that construct could reasonably alter the product's safety, regulatory, or clinical-governance profile; and (3) it could not be collapsed into another retained construct without losing a distinct decision question. The resulting six dimensions are: intended use and claims; user context; depth of personalization; data and sensor sources; automation, human oversight, and closed-loop control; and upgrades and change pathways. The six dimensions should not be read as statistically orthogonal variables or as a validated additive prediction score. They are a structured set of decision constructs with different roles in early-stage governance.[20–45]

Intended use and claims are retained together as the primary classification anchor because both describe the product's represented purpose. User context, personalization, data and sensor sources, and automation with human oversight are retained as risk-modifying dimensions because they answer separate questions about whom the product is designed for, how far outputs are individualized, what information materially drives those outputs, and whether a person or the system retains control over safety-relevant action. Upgrades and change pathways are retained as a lifecycle-governance dimension because they address whether performance-relevant changes can be introduced and monitored in a controlled manner after deployment. Supplementary Table 4 explains these retention and non-merger decisions.[26–39]

Several related constructs were deliberately not retained as stand-alone dimensions. Human oversight is nested within the automation and closed-loop-control dimension because it modifies who can act on, interrupt, or confirm a safety-relevant output; it does not independently remove an explicit medical intended use. Privacy, cybersecurity, bias, transparency, documentation, and clinical evidence are cross-cutting quality and evidence requirements that apply across multiple dimensions and are therefore addressed in the risk-governance checklist rather than treated as additional

boundary triggers. General model architecture or the use of an LLM is likewise not a stand-alone trigger, because the same underlying model can support wellness or medical-use functions depending on its intended use, context, data inputs, automation, and change controls.[20–45]

The source-to-dimension mapping is not a vote count and does not assert that every source creates the same legal obligation across jurisdictions. Its purpose is to make the conceptual bridge from the source corpus to the proposed framework auditable and challengeable. It should therefore be read alongside the anchored rubric and non-compensatory aggregation rules in Table 1 and Table 2. The framework remains a proposed governance construct and does not replace product-specific legal analysis, clinical review, or a formal determination by competent authorities.[20–45]

Supplementary Material D provides the full traceability audit and construct-retention rationale, including the source-to-dimension mapping and non-merger decisions.

Preliminary Application and Inter-Rater Agreement Exercise

To examine whether the proposed anchored rubric could be applied consistently before broader validation, we conducted a preliminary scenario-based coding exercise in June 2026. Three independent raters with complementary expertise in sports medicine and clinical exercise prescription, digital health and medical-device regulatory affairs, and AI medical informatics and medical-device software quality management participated. For confidentiality, they are referred to as Rater 1–3. Each rater confirmed that they had not developed the matrix, revised the coding rubric, or contributed to the manuscript revision.

Each rater received the same standardized case packet containing seven hypothetical product scenarios (B1–B7), the six-dimensional Green/Yellow/Red anchors, the non-compensatory overall aggregation rules, and the M-flag definition. Raters were instructed to score only the information supplied in the packet; they did not receive the author reference coding, the public-product comparison materials, or other raters' scores. Before analysis, the completed forms were locked and transcribed without modification. We calculated average pairwise agreement, exact three-rater agreement, and Fleiss' kappa[46] for each dimension and the overall profile. M-flag agreement was reported using the same descriptive metrics. This small, deliberately constructed vignette exercise was intended to assess rubric reproducibility, not to establish legal-classification accuracy, clinical validity, or generalizability to real-world products.

The standardized scenario packet, anonymized original ratings, agreement summaries, and rater-characteristic information are provided in Supplementary Material E (Supplementary Tables 5–7).

Regulatory and Risk-Governance Background: China as the Primary Anchor, with EU and US Comparison

China as the Primary Jurisdictional Anchor

This article is anchored in China's medical-device and health-data governance setting. The purpose is not to declare the legal classification of any particular product, but to identify the product characteristics that should prompt earlier risk-management, evidence-planning, and regulatory-consultation activity in China. Formal classification and submission requirements remain matters for the applicable Chinese laws, technical guidance, and regulatory authorities in light of the complete functionality, intended use, risk-control measures, and deployment context of a product.[26,27,31–33]

For software potentially used in exercise prescription, rehabilitation, or disease-related health management, the Chinese regulatory starting point is the medical purpose reflected in the intended use, claims, instructions, and actual product function. Software intended to diagnose, prevent, monitor, treat, or alleviate disease may enter the medical-device pathway; a product's connection to clinical workflow, medical records, or medical-device signals may further intensify the relevant safety and evidentiary questions.[26,27,31–33,40–42] These considerations inform the present matrix's primary intended-use-and-claims axis, as well as the user-context and data-and-sensor-source dimensions.

China's technical review materials also make lifecycle governance material for software and artificial intelligence-enabled medical devices. Algorithm description, data provenance, validation, software configuration, version control, change records, and traceability are therefore treated in this article as prospective governance considerations rather than matters to be addressed only at a final regulatory submission.[26,27] The Personal Information Protection Law, Data

Table 1 Anchored Coding Rubric for the Six-Dimensional Trigger Matrix

Dimension and Decision Role	Green (0): Wellness-Compatible Anchor	Yellow (1): Grey-Zone Anchor	Red (2): High-Concern Anchor	Coding and Governance Notes
1. Intended use and claims Primary classification anchor	Provides general health information, lifestyle education, or fitness/wellness support. Avoids claims to diagnose, prevent, monitor, treat, alleviate, or manage a specific disease, condition, or clinical event.	Claims to improve a specific non-diagnostic physiological or functional metric, or provides bounded risk-awareness prompts, without disease-management, diagnostic, therapeutic, or clinical-monitoring claims.	Claims to diagnose, prevent, monitor, treat, alleviate, predict, risk-stratify, or manage a specific disease, condition, clinical event, or rehabilitation pathway.	A Red score is a primary Red trigger. Review all product-facing language, instructions, marketing, interface wording, and sales materials; function and claims must be interpreted together.
2. User context Risk modifier	Intended for generally healthy adults using the product outside a clinical workflow; no pre-specified high-risk or vulnerable target group.	Intended for people with risk factors, borderline physiological indicators, or potentially vulnerable groups, but without an explicit disease-management or clinical-service purpose.	Explicitly intended for people with established disease, post-procedural or rehabilitation patients, or users in a clinical workflow where outputs may inform care decisions.	A Red user-context score alone does not automatically establish medical-device status. It becomes a Red overall trigger when combined with a relevant Red personalization, data, or automation configuration.
3. Depth of personalization Risk modifier	User selects or receives standard plans based only on preferences, goals, or generic lifestyle information; no individualized risk or clinical adaptation.	Semi-customized recommendations use multi-domain self-report or consumer-wearable information and pre-specified rules, while avoiding disease-specific or dynamically clinical adjustment.	Recommendations dynamically adapt to individual medical history, medication, pathological status, physiological signals, or multi-source risk factors in a way that can change exercise load, nutritional advice, monitoring, or referral.	Code the highest level of personalization actually used in producing a recommendation. Personalization is not equivalent to simple preference matching.
4. Data and sensor sources Risk modifier	Relies on self-reported lifestyle, goals, preferences, and nonclinical activity information. No EHR, diagnostic test, or medical-grade device data are used to direct outputs.	Uses consumer-wearable or home-monitoring data (for example, steps, consumer heart-rate estimates, sleep, or activity patterns) with disclosed limitations and without clinical-data interpretation.	Uses EHR information, diagnostic or laboratory results, medical-grade device signals, or clinically validated physiological data to individualize recommendations, trigger alerts, or support decisions.	The source alone is not decisive; the key question is whether the data materially direct a safety-relevant output. Data protection duties remain relevant at all levels.
5. Automation, human oversight, and closed-loop control Risk modifier	Open loop: provides information or options only. The system does not execute, schedule, or alter an intervention, and no safety-critical output is delivered as an action instruction.	Semi-closed loop: generates recommendations from user data, but implementation requires explicit user confirmation and prespecified guardrails. For elevated-risk outputs, a timely, role-specific human-review or referral pathway is defined.	Closed or near-closed loop: automatically changes an exercise plan, intensity, follow-up, alert, or other safety-relevant action based on individual data; or delivers such outputs in a high-risk/clinical context without qualified human confirmation.	Oversight must be predefined, timely, and performed by a person appropriately qualified for the output. It can mitigate automation risk but cannot neutralize an explicit medical intended use.
6. Upgrades and change pathways Lifecycle-governance dimension	Static model or content-only updates that are versioned and manually checked; no material change to safety-relevant behavior.	Planned batch updates or rule/prompt/content revisions with documented impact assessment, verification/validation, approval, and version control before release.	Online learning, unbounded fine-tuning, or major model/prompt/rule/retrieval changes that can alter high-risk outputs without documented impact assessment, validation, approval, version control, and monitoring.	A Red score here does not establish medical purpose by itself, but it precludes an overall Green profile and triggers a mandatory lifecycle-governance escalation flag.

Notes: Green/Yellow/Red are ordered governance states. They may be recorded as 0/1/2 for traceability but are not summed as a validated risk score.

https://doi.org/10.2147/RMHP.S624615

Table 2 Non-Compensatory Overall Aggregation and Escalation Rules

Decision Step	Rule	Overall Profile/Action
1. Completeness gate	Before assigning an overall profile, confirm that claims, intended users, deployment setting, data inputs, personalization rules, automation/oversight arrangements, and change plans are documented sufficiently to code all six dimensions.	If essential information is missing: "Insufficient information—requires expert/regulatory review". Do not use interval labels such as "Yellow to Red".
2. Primary Red trigger	Dimension 1 (intended use and claims) is Red.	Overall Red: candidate for regulated medical-use pathway and enhanced clinical/governance review.
3. Red interaction trigger A	Dimension 2 (user context) is Red and at least one of Dimensions 3, 4, or 5 is Red.	Overall Red: a high-risk population or clinical setting is combined with highly individualized, clinical-data-driven, or autonomous safety-relevant functionality.
4. Red interaction trigger B	Dimensions 3 and 4 are both Red and the clinical or medical-grade data materially individualize exercise, dietary, monitoring, or referral recommendations.	Overall Red: individual clinical data are used to generate safety-relevant action.
5. Red interaction trigger C	Dimension 5 is Red because the system automatically changes a safety-relevant exercise plan, initiates/suspends exercise, changes referral/escalation timing, or delivers a patient-specific alert/action based on individual health data.	Overall Red: safety-critical closed-loop or near-closed-loop operation requires enhanced review.
6. Yellow profile	No primary Red trigger and no Red interaction trigger is present, but one or more dimensions are Yellow; OR Dimension 6 is Red without another Red trigger.	Overall Yellow: grey-zone profile. Apply proportionate constraints and controls before expansion, including claim review, safety guardrails, human-review/referral arrangements, validation, and/or expert consultation.
7. Green profile	All six dimensions are Green and no essential information is missing.	Overall Green: wellness-compatible profile. Maintain basic privacy, transparency, version logging, and post-deployment monitoring appropriate to the product.
8. Mandatory lifecycle-governance flag	Dimension 6 is Red, whether the overall profile is Yellow or Red.	Add "M flag". The organization must complete impact assessment, configuration/version control, appropriate validation and regression testing, approval, and monitoring before release. Dimension 6 alone does not establish medical purpose.

Notes: The rules guide early-stage governance and evidence planning; they do not replace product-specific legal or regulatory determination.

Security Law, and Cybersecurity Law do not themselves determine whether software is a medical device, but they shape the lawful and secure handling of personal and health-related data, including data used by LLM-based systems.[43–45] National Health Commission rules on internet diagnosis and treatment, internet hospitals, and telemedicine provide additional service-context considerations for products that connect to healthcare delivery.[40–42]

European Union and United States as Comparative Reference Points

The European Union and United States are used in this article as comparative reference points, not as parallel legal anchors or substitutes for Chinese regulatory determinations. Their value lies in making several governance signals more explicit: intended purpose; the clinical significance and context of software-supported decisions; the consequences of reliance on outputs; documentation and evidence; human oversight; and lifecycle management of software changes.[28–39]

In the European Union, the Medical Device Regulation (MDR) is the principal device-regulatory framework. Under Rule 11, software that provides information used to take decisions for diagnostic or therapeutic purposes may be classified as Class IIa or higher depending on the decision and the potential consequences of an erroneous decision.[34] The EU Artificial Intelligence Act adds a horizontal risk-based framework for AI systems within its scope, including

requirements relevant to high-risk AI systems such as risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, and post-market monitoring.[35] These materials inform the matrix but do not provide a direct classification shortcut for China.

In the United States, FDA materials on device software functions and AI-enabled device software functions emphasise intended use, evidence, and total-product-lifecycle management.[28,30,39] The FDA's 2025 final guidance on predetermined change control plans (PCCPs) for AI-enabled device software functions illustrates an approach in which planned modifications, the methodology for developing and validating them, and their anticipated impact are specified in advance.[28] The related FDA lifecycle-management guidance remains draft guidance, and is used here only as a comparative indication of emerging expectations for documentation and risk management.[30] IMDRF materials on SaMD and good machine learning practice similarly inform the comparative discussion of clinical evaluation, traceability, data quality, and lifecycle governance.[29,36–38]

Cross-Jurisdictional Governance Signals and the Role of Human Oversight

The comparative analysis is designed to identify transferable governance signals, not to erase jurisdictional differences. Across the China-anchored analysis and the EU/US comparator materials, six signals recur: intended use and claims; user context; depth of personalization; data and sensor sources; automation, human oversight and closed-loop control; and upgrade and change pathways.[26–39] The six dimensions are functionally distinct governance constructs rather than statistically orthogonal variables, and their derivation and retention are documented in Supplementary Tables 3 and 4.

Human oversight is treated as a structural safeguard within the automation and closed-loop dimension and as a cross-cutting governance requirement. In higher-risk exercise and health settings, an oversight design should specify which outputs require review before delivery, who is competent to conduct that review, when review must occur, what information is available to the reviewer, how the reviewer can override the system, and when a user must be referred or escalated to appropriate healthcare personnel.[20–30,35,39] Nominal human involvement is not sufficient where review is delayed, lacks authority, or cannot meaningfully evaluate the recommendation. Conversely, appropriate professional review can reduce inappropriate automation and over-reliance, but cannot convert an explicit diagnosis-, treatment-, monitoring-, or disease-management claim into a general-wellness function.[26–39]

Accordingly, the matrix is used as a proposed, structured risk-identification tool. It does not replace legal analysis, clinical governance, a quality-management system, or product-specific regulatory interaction. In this article, a Green/Yellow/Red designation indicates an initial governance posture and evidence-planning need, not a binding classification in China, the European Union, or the United States. The anchored coding rubric, non-compensatory aggregation rules, and preliminary independent-rater exercise are described in Six-Dimensional Risk Trigger Matrix: Structure, Anchored Rules, and Aggregation and Application Scenarios and Boundary Cases and in Supplementary Material E.

Empirical Evidence and Quality Requirements for LLM-Based Health Guidance

The technical and governance issues identified above are amplified in LLM-based health guidance. LLM outputs can appear plausible while being inaccurate, unsupported, or poorly calibrated; performance and safety may also vary across user groups, prompt formulations, data contexts, and deployment environments.[2,15,18,47–50] Exercise-related applications add a direct action pathway because recommendations can influence physical load, adherence to restrictions, symptom response, and timing of referral. For this reason, evaluation should address not only text quality but also the appropriateness of intended use, output boundaries, escalation design, and the conditions under which users or professionals may act on a recommendation.

DECIDE-AI, SPIRIT-AI, CONSORT-AI, MI-CLEAR-LLM, IMDRF good machine learning practice, and related materials support a structured approach to intended-use definition, data and model documentation, performance evaluation, subgroup analysis, human oversight, transparency, and post-deployment monitoring.[20–25,29] These materials do not validate the matrix proposed here. They provide the governance and reporting concepts used to design the structured checklist in Structured Risk-Governance Checklist and the source-to-dimension traceability materials in Supplementary Tables 1–4.

Six-Dimensional Risk Trigger Matrix: Structure, Anchored Rules, and Aggregation

The six-dimensional trigger matrix is a proposed governance tool for determining when an LLM-based exercise or health-guidance application remains compatible with a lower-risk wellness position, enters a grey-zone profile requiring additional controls, or becomes a candidate for a regulated medical-use pathway. It is not a substitute for product-specific legal analysis, clinical governance review, or a formal regulatory determination. Its function is to make the relevant triggers visible before a product is deployed, materially changed, procured, or scaled.[26–39]

Framework Architecture and the Role of Each Dimension

The framework distinguishes three decision roles. First, intended use and claims are the primary classification anchor. Explicit claims to diagnose, prevent, monitor, treat, alleviate, or manage a specific disease, condition, or clinical event establish a strong medical-use signal even where the product's technical architecture is relatively simple.[26,27,31–39]

Second, four dimensions modify the risk profile: user context, depth of personalization, data and sensor sources, and automation, human oversight, and closed-loop control. These dimensions capture how the product is actually deployed and whether its outputs can directly influence physical activity intensity, continuation or cessation of exercise, referral timing, or other safety-relevant actions. Their interaction is particularly important in exercise and rehabilitation settings because the same generic recommendation may have materially different consequences for a healthy adult, a post-procedural cardiac patient, or a person whose physiological data are being used to adapt an exercise plan.[10,16,17,20–25]

Third, upgrades and change pathways are treated as a lifecycle-governance dimension. A model, prompt, retrieval source, rule base, or user-interface change can alter the safety profile of an LLM-based system even when the declared intended use is unchanged. This dimension therefore governs the need for documented impact assessment, version control, validation, regression testing, approval, and post-deployment monitoring.[26–30]

Human oversight is incorporated within the automation dimension because its protective effect depends on the actual control arrangement. To lower automation-related concern, oversight must be predefined, timely, role-specific, and carried out by a person appropriately qualified for the risk level of the output. A nominal statement that a professional is "available" is not sufficient where a system can deliver or implement safety-critical recommendations without review. Conversely, human oversight cannot neutralize an explicit disease-directed intended use or convert a product with a clear medical purpose into a general wellness tool.[20–30,36–39]

Anchored Coding Rubric for Individual Dimensions

Each product is coded separately on all six dimensions using the anchored criteria in Table 1. Green denotes a profile that is compatible with general wellness positioning on that dimension. Yellow denotes a boundary condition that requires design constraints, clearer user communication, documented review, or other proportionate controls. Red denotes a high-concern configuration that, depending on the overall profile, may require formal clinical, legal, or regulatory review. The anchoring is based on observable product documentation and behavior, including intended-use statements, marketing language, instructions for use, target-user descriptions, data-flow diagrams, system prompts and rules, interface screens, escalation logic, and change-control records.[26–39]

The rubric should be applied to the product as actually designed and deployed, rather than to isolated wording or a single screen. For example, a general wellness statement does not remain Green if the same product combines clinical data with disease-specific exercise adjustment or safety-critical automated action. Conversely, the incidental use of a general wellness product by an individual who happens to have a chronic condition does not, by itself, establish a Red user-context score unless the product is intentionally targeted to that population or embedded in a clinical workflow.

Non-Compensatory Aggregation and Escalation Rules

The overall profile is not determined by adding the six codes or by treating all dimensions as equally weighted. Instead, the framework uses a non-compensatory logic. First, a Red intended-use-and-claims score is sufficient to designate the product as a Red, regulated-pathway candidate. Second, where intended use is not Red, combinations of Red risk

modifiers can still produce a Red overall profile when they create a clinically meaningful pathway from individual health information to a safety-relevant action. The interaction configurations are specified in Table 2.

A product is designated Yellow when no primary Red trigger or Red interaction configuration is present, but at least one dimension is Yellow, or when the lifecycle-governance dimension is Red. This identifies a grey-zone profile that should not proceed as unrestricted wellness deployment. It requires proportionate constraints, such as claim revision, tighter inclusion criteria, defined human-review and referral thresholds, validation of key outputs, or expert/regulatory consultation before expansion. A product is designated Green only when all six dimensions are Green and no essential information is missing.[26–39]

A Red upgrade-and-change-pathways score does not, by itself, establish a medical purpose. It nevertheless precludes a Green designation and triggers a mandatory lifecycle-governance escalation flag. This flag requires the organization to address the uncontrolled update pathway before the relevant change is released, including a documented impact assessment, version and configuration control, appropriate validation and regression testing, approval, and monitoring.[26–30]

The terms Green, Yellow, and Red describe proposed risk-management states, not formal device classes. A Red overall profile means "candidate for a regulated medical-use pathway and enhanced clinical/governance review"; it does not state that the product has been legally classified as SaMD in China, the European Union, or the United States. Products with missing information are not assigned a range such as "Yellow to Red". They are classified as "insufficient information—requires expert/regulatory review", and the missing documentation becomes a required action item.

Applying the Matrix in Practice

Application proceeds in four steps. First, assemble the evidence needed to code each dimension: product claims and user instructions; intended user population and deployment setting; inputs and data flows; personalization and safety rules; automation, human-review, and referral arrangements; and planned changes after deployment. Second, assign a Green, Yellow, or Red code to each dimension using the anchored criteria in Table 1. Third, apply the non-compensatory aggregation rules in Table 2, including the completeness gate and any mandatory lifecycle-governance escalation flag. Fourth, translate the resulting profile into actions proportionate to the identified concern, such as maintaining wellness constraints, strengthening governance controls, seeking clinical or legal review, or preparing for a regulated pathway.[26–39]

This logic also clarifies common escalation pathways. A tool may move from Green to Yellow when it begins to claim improvement in a specific physiological metric, incorporates consumer-wearable inputs, or adds semi-customized recommendations that require stronger safeguards. It may move from Yellow to Red when it adds disease-specific claims, integrates medical-grade or EHR data to individualize exercise or dietary advice, or automates safety-critical exercise adjustment, referral, or alerting without appropriate qualified review. These transitions should be assessed before deployment and after every material change to claims, data, automation, or model configuration.[10,16,17,26–39]

Figure 1 presents the six dimensions and their traffic-light anchors. The figure should be read together with Table 1 and Table 2: the visual matrix supports rapid orientation, whereas the tables provide the coding and aggregation logic needed for a traceable assessment.

Structured Risk-Governance Checklist

The structured risk-governance checklist translates the six-dimensional trigger matrix into reviewable questions for developers, healthcare institutions, procurement teams, and governance bodies. It is intended to support earlier documentation and evidence planning before pilot deployment, procurement, or regulatory consultation. It is not a legal classification instrument, a complete quality-management system, or a substitute for product-specific regulatory assessment.[26–39]

The checklist has three linked functions. First, it documents the product profile that underlies the matrix assessment: intended use and claims; intended users and deployment context; degree of personalization; data and sensor inputs; automation and professional oversight; and update and change pathways. Second, it links each domain to evidence artifacts that can be assembled and reviewed, such as a claims inventory, user-population definition, data-flow map, oversight protocol, validation plan, configuration baseline, change-impact assessment, and monitoring plan. Third, it helps determine whether unresolved gaps should lead to a design constraint, additional evidence generation, a hold on deployment, professional review, or regulatory consultation.

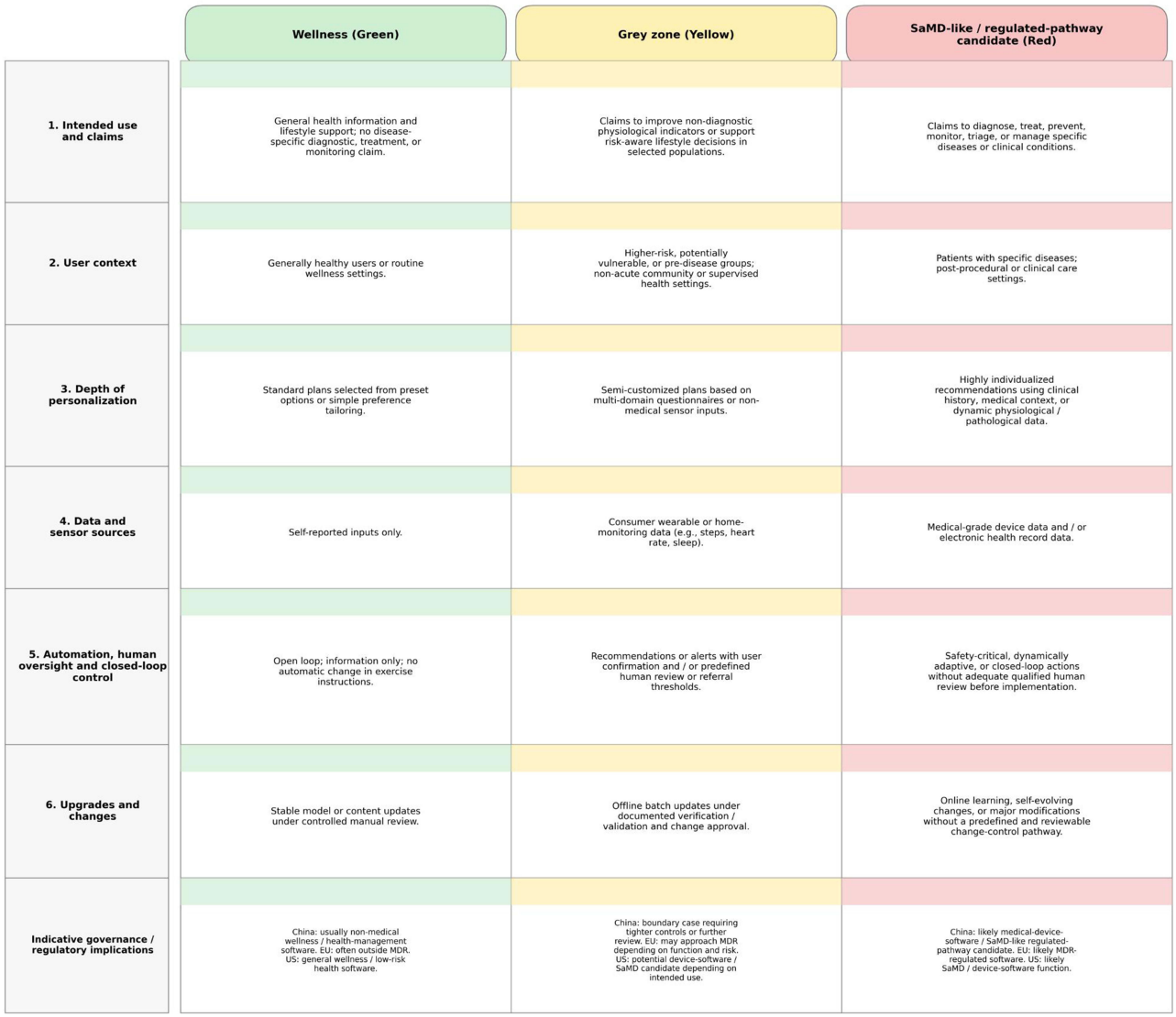

Figure 1 Six-dimensional trigger matrix for early risk identification of LLM-based exercise and health guidance tools. The matrix is anchored primarily in the Chinese NMPA regulatory and governance context and is informed comparatively by EU and US materials. It is intended to support preliminary governance screening, product-design review, and evidence-planning discussions rather than to provide a formal legal classification outcome. The three column headers identify the Green, Yellow, and Red governance profiles. Green indicates a lower-risk wellness profile; Yellow indicates a grey-zone profile that may require tighter design constraints, clearer claims, stronger human oversight, or further review; and Red indicates features more consistent with SaMD-like or regulated-pathway candidate functions, especially when several high-risk features occur together. Dimension 1 addresses intended use and claims; Dimension 2, user context; Dimension 3, depth of personalization; Dimension 4, data and sensor sources; Dimension 5, automation, human oversight and closed-loop control; and Dimension 6, upgrades and changes.

Abbreviations: LLM, large language model; SaMD, software as a medical device; NMPA, National Medical Products Administration; MDR, Medical Device Regulation; EHR, electronic health record.

The depth of evidence should be proportionate to risk. A low-risk wellness tool may require clear claims, user-facing limitations, basic privacy documentation, and controlled content/version records. A Yellow-profile tool may require stronger documentation of personalization logic, consumer-wearable inputs, professional review pathways, output boundaries, and pre-deployment testing. A Red-profile or SaMD-like candidate may require formal risk management, quality-system alignment, clinical-evaluation planning, robust traceability, documented human-oversight and referral controls, and lifecycle monitoring.[26–39] The absence of evidence for a safety-critical function should not be interpreted as evidence of low risk.

For LLM-based systems, the controlled configuration is broader than the base model version. The governance baseline should identify, where relevant, the model and model version; system prompts and prompt templates; retrieval sources; tool or application-programming-interface connections; rule logic and thresholds; output templates; safety filters; user-interface constraints; and the sources and routing of high-risk data. A material change to any of these

https://doi.org/10.2147/RMHP.S624615

elements may alter output behavior or the product's risk profile and should therefore be assessed, documented, and verified before release in a manner proportionate to the intended use and risk.[20,26–30,47–50]

Table 3 presents the checklist domains, their linkage to the six-dimensional matrix, core governance questions, illustrative evidence artifacts, and examples of decision evidence or review indicators. The source/regulation-to-dimension mapping and construct-retention rationale are provided in Supplementary Tables 3 and 4. Supplementary Table 1 maps selected checklist domains to illustrative NMPA review focuses, while Supplementary Table 2 provides illustrative, product-specific KPI considerations. All measures and examples should be adapted to the product, risk level, and regulatory materials in force at the time of implementation. Supplementary Material C presents the illustrative mapping of checklist domains to NMPA review focuses and the associated KPI considerations.

Lifecycle Pathway for Evidence Generation, Risk Management, and Change Control

The checklist should not be read as requiring identical evidence for every product. A low-risk wellness product may require only light documentation of intended use, user instructions, privacy assessment, and version logging. A grey-zone system may require stronger validation, human-oversight protocols, and risk communication. A red-zone or SaMD-like system should require formal risk management, clinical evaluation evidence, QMS alignment, robust change control, post-market monitoring, and documentation of model update governance. This staged interpretation is consistent with risk-proportionate governance and avoids treating all LLM-enabled wellness tools as if they were high-risk SaMD.[19,34–36,43–45,47–51]

Figure 2 presents the proposed lifecycle pathway, showing how evidence generation, risk management, human oversight, and change control may become progressively more formal as an LLM-based exercise and health-guidance product moves from wellness-oriented prototyping to grey-zone exploration, possible regulated-pathway preparation, and post-deployment monitoring.

Phase 1: Concept Validation and Prototyping

In Phase 1, the product remains clearly positioned as a wellness or lifestyle-support tool.[10,51] The focus is on user value, technical feasibility, and basic usability rather than on medical claims. Product language should avoid diagnosis, treatment, triage, or disease-specific optimization.[26,27,31–39]

Evidence generation at this stage may include user-testing reports, technical feasibility assessments, and preliminary privacy-impact review of basic data flows.[20–25,43–45] Change control can remain relatively lightweight, but major changes should still be logged so that the evolution of functionality can later be reconstructed.[26–30]

Phase 2: Pre-Market Validation and Grey-Zone Exploration

In Phase 2, some features may begin to move into a grey zone, for example through consumer wearable integration, more detailed personalization, or use in higher-risk subgroups.[10–13,16,17,19] The product may still be positioned as a wellness or health-guidance tool, but the likelihood of movement toward medical-use interpretation increases.[26,27,31–39]

At this stage, more formal risk management and technical documentation should begin.[20–27,29,30] This may include initial risk files, validation protocols, usability records, and small prospective studies. Where exercise prescription or rehabilitation is involved, subgroup-specific safety evaluation becomes especially important.[10,16,17] Early dialogue with regulators or participation in pilot or sandbox settings may also be valuable where available.[26,27,31–33,40–42]

Phase 3: Regulatory Submission and Controlled Deployment

Phase 3 begins when the product's profile is sufficiently close to medical use that formal regulatory planning becomes necessary.[26–39] At this point, intended use, target population, and product claims should be tightly aligned with the proposed submission pathway.[26,27,31–39]

Required evidence is likely to include formal technical documentation, risk analysis, cybersecurity and privacy materials, software lifecycle documentation, and clinical evaluation or real-world evidence appropriate to the product's function and risk level.[20–39] Organizational quality management and change-control mechanisms also become central.[26–33]

Table 3 Structured Risk-Governance Checklist for LLM-Based Exercise and Health Guidance

Checklist Domain and Matrix Linkage	Core Governance Question	Illustrative Controls and Evidence Artifacts	Illustrative Decision Evidence/ Review Indicator	China-Anchored Traceability Link
1. Intended use, claims, and user context D1-D2	Do the stated purpose, interface language, marketing claims, instructions, and deployment setting remain within general wellness support, or do they indicate diagnosis, treatment, monitoring, disease management, risk assessment, or clinical decision support? Who is expected to use the product and in which care context?	Current and proposed claims inventory; intended-use statement; instructions for use; screenshots and marketing-review record; user-population definition; deployment/ workflow map; contraindications and exclusions.	Documented alignment between claims, functionality, intended users, and available evidence. Any disease-specific or medical-purpose claim triggers a formal review of pathway, evidence, and oversight needs before deployment.	Primary: 25–26,30-32,39–41 Comparative: 33,35–38 See Supplementary Tables 1,3 and 4
2. Personalization and data/sensor inputs D3-D4	What information changes the recommendation? Does tailoring remain preference based, or does it use multi-domain health history, dynamic physiological data, EHR information, or medical-device signals?	Data-flow map; data dictionary; source and quality description; consent/permission basis; personalization/rule specification; device and EHR interface documentation; data-minimisation assessment.	A documented explanation of which inputs affect which outputs, and why. Medical-grade data, EHR integration, or high-consequence data-driven adaptation require enhanced validation and governance review.	Primary: 25–26,30-32,42–44 Comparative: 28,33–38 See Supplementary Tables 1,3 and 4
3. Automation, human oversight, referral, and closed-loop control D5; cross-cutting	Which outputs may reach a user without professional review? Which require user confirmation, qualified review, or referral? Are reviewer competence, timeliness, authority to override, and escalation thresholds defined?	Risk-stratified human-oversight protocol; reviewer role/competency record; mandatory pre-delivery review list; escalation/referral algorithm; emergency stop or safe-default design; override and audit logs; staff training materials.	Pre-specified coverage of safety-critical outputs and referral triggers. No autonomous adjustment or action should be released where the required review, escalation, or safe-default controls are absent. Human review reduces automation risk but does not negate an explicit medical intended use.	Primary: 25–26,30-32,39–41 Comparative: 27–29,33-38 See Supplementary Tables 1,3 and 4
4. Prompt, model, and configuration traceability D3-D6; cross-cutting	Can a safety-relevant output be reconstructed? Is the controlled configuration defined beyond the base model, including prompts, retrieval sources, tool connections, rule thresholds, filters, and output templates?	Model card; model/version register; system-prompt and prompt-template register; retrieval and tool-connection inventory; safety-policy/rule specification; configuration baseline; traceable logs for high-risk outputs; challenge-test/red-team records.	A reviewer can identify the configuration that generated a high-risk output and reproduce the applicable rules, prompts, data route, and version. Material configuration changes undergo documented impact assessment and proportionate verification before release.	Primary: 25–26 Comparative: 27–29,35–37 See Supplementary Tables 1,3 and 4

(*Continued*)

https://doi.org/10.2147/RMHP.S624615

Table 3 (Continued).

Checklist Domain and Matrix Linkage	Core Governance Question	Illustrative Controls and Evidence Artifacts	Illustrative Decision Evidence/ Review Indicator	China-Anchored Traceability Link
5. Data governance, privacy, and cybersecurity D4; cross-cutting	Are collection, consent, access, transfer, storage, retention, deletion, and logging proportionate to the data sensitivity and intended use? Are security controls appropriate for connected devices and health data?	Privacy notice; data-protection impact or risk assessment where appropriate; consent/ authorisation record; access-control matrix; retention/deletion schedule; security-testing record; incident-response procedure; vendor/ data-processing documentation.	Documented lawful basis or authorisation and minimisation of identifiable health data. Security incidents and material vulnerabilities trigger documented response, corrective action, and reassessment of affected functions.	Primary: 30,42–44 Comparative: 28,34 See Supplementary Tables 1, 3 and 4
6. Evidence, performance, subgroup safety, and clinical plausibility D1-D5; cross-cutting	What evidence supports the intended use and the safety of outputs? Has performance been examined across relevant user subgroups and foreseeable high-risk situations?	Evaluation plan; test-set/data documentation; clinically plausible scenario set; subgroup analysis plan; unsafe-output and hallucination test set; expert review record; usability/ human-factors evidence; limitations statement.	Pre-specified outcome measures and acceptable error/risk criteria that fit the intended use. Evidence gaps, subgroup concerns, or unresolved unsafe-output patterns trigger design constraints, additional testing, or a hold on deployment.	Primary: 25–26 Comparative: 28,35-37 Reporting: 19–24,45-48 See Supplementary Tables 1–4
7. Transparency, user understanding, and deployment governance D1-D2 -D5; cross-cutting	Do users, professionals, and purchasing institutions understand the validated scope, limitations, warning signs, roles, and conditions for use? Is the product deployed only in an environment that can support the stated controls?	User-facing limitation and warning text; clinician/institutional instructions; procurement due-diligence record; role and responsibility matrix; comprehension/usability test; marketing-claims review; service-level escalation pathway.	Evidence that limitations and escalation instructions are understandable to intended users and that the deployment site can provide the promised human oversight, referral, documentation, and incident response.	Primary: 25–26,30-32,39–41 Comparative: 33–34,38 See Supplementary Tables 1,3 and 4
8. Lifecycle monitoring and controlled change D6; M flag where applicable	Are feedback, adverse events, near misses, drift, content changes, prompts, model updates, data-source changes, and feature changes monitored and governed across the lifecycle?	Monitoring plan; complaint/incident/near-miss log; post-deployment challenge-test set; performance-drift review; change register; change-impact assessment; regression-test evidence; release approval; CAPA record; PCCP-like planning where appropriate.	Every safety-relevant change is logged, assessed, and approved before release at a level proportionate to risk. A Red D6 result activates the mandatory lifecycle-governance escalation flag (M flag), requiring formal impact assessment, verification/validation, approval, and enhanced monitoring.	Primary: 25–26,30-32 Comparative: 27–29,34-38 See Supplementary Tables 1–4

Notes: The checklist is a proposed structured governance and evidence-planning aid. It is not a binding classification rule, formal legal opinion, complete quality-management system, or guarantee of safety. "Primary" identifies China-anchored references; "Comparative" identifies EU/US/IMDRF reference points. M flag denotes mandatory lifecycle-governance escalation.

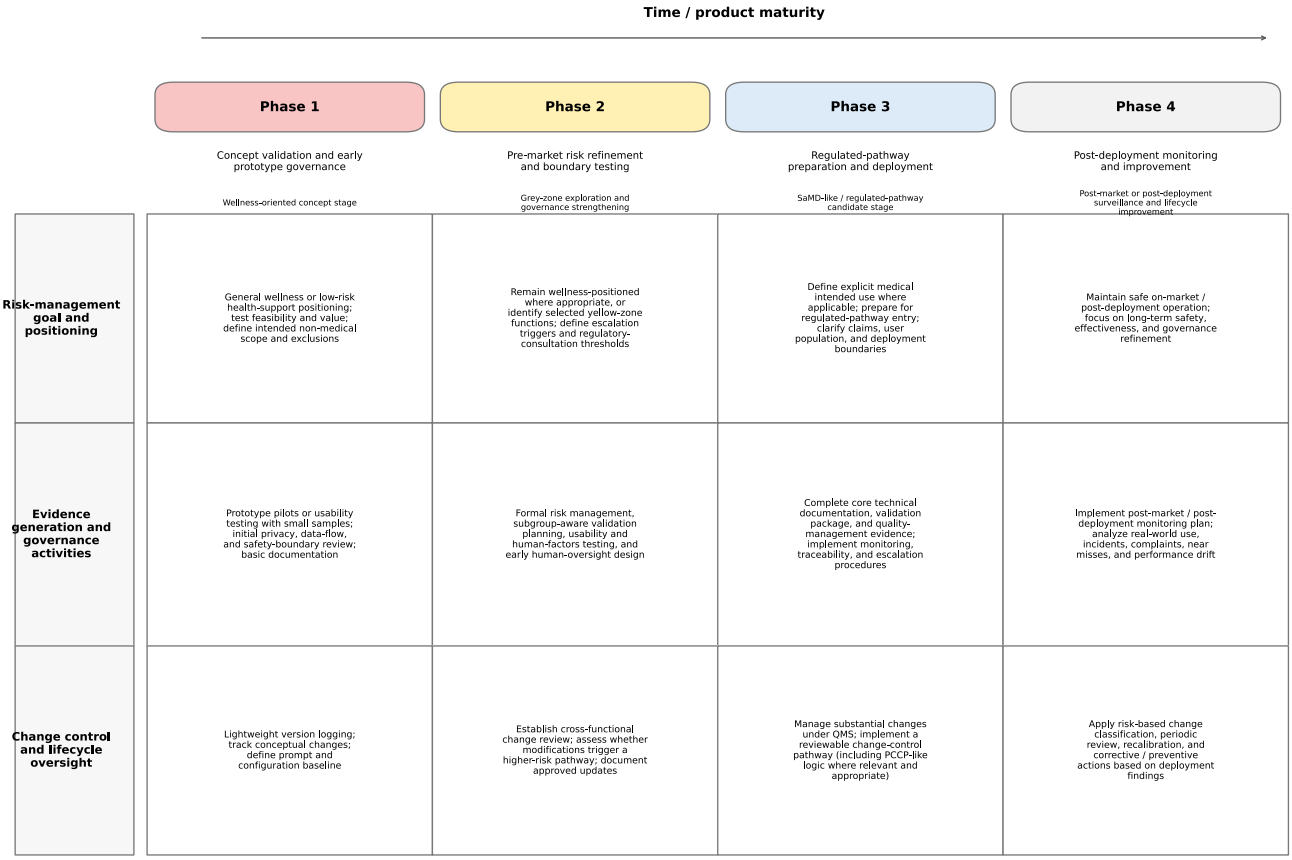

Figure 2 Phased lifecycle governance pathway for LLM-based exercise and health guidance tools. The pathway is anchored primarily in the Chinese regulatory and governance context and is informed comparatively by EU and US concepts relevant to AI-enabled medical software. Phase 1 represents concept validation and early prototype governance; Phase 2, pre-market risk refinement and boundary testing; Phase 3, regulated-pathway preparation and deployment; and Phase 4, post-deployment monitoring and improvement. The figure illustrates how governance expectations may intensify as products move across these phases. The colored phase headers are visual aids only and do not denote a separate risk category or regulatory classification. It is intended to support planning for evidence generation, risk management, human oversight, and change control; it does not itself determine legal classification or guarantee that a product will enter or remain outside any specific regulatory pathway.
Abbreviations: QMS, quality management system; PCCP, predetermined change control plan; LLM, large language model; SaMD, software as a medical device.

Phase 4: Post-Market Surveillance and Continuous Improvement

In Phase 4, the priority shifts to maintaining safety and effectiveness in real-world use.[20–25,29,30] For LLM-enabled exercise and health-guidance systems, post-deployment change may arise from shifts in users, settings, prompt design, versioning, or fine-tuning, even where the core system appears stable.[26–30]

This creates a need for systematic logging, monitoring, incident analysis, performance review, and evidence-based change management.[20–25,29,30] Where a PCCP-like logic is used, post-market evidence should feed back into periodic review of update boundaries and monitoring indicators.[28–30]

Application Scenarios and Boundary Cases

The framework becomes more concrete when applied to plausible borderline cases. The examples below are hypothetical composite scenarios constructed for analytical purposes. They are illustrative only and are not intended to provide definitive legal classification or formal compliance advice for any specific real-world product.[26–39]

Illustrative composite applications of the anchored rubric are provided in Supplementary Material A. Representative mapping examples are provided in Supplementary Material B and Supplementary Figure 1 (Panels A–D. The cases are hypothetical composite scenarios and are intended for illustrative governance screening only.

Case 1: Weight-Management App

A weight-management mini-program may begin as a relatively low-risk service providing calorie calculators, diet logging, and generic exercise videos.[10,51] At that stage, it is more likely to be viewed as a wellness or health-management tool.[10,40–42,51]

The profile may change if the product later integrates chest-strap heart-rate data and introduces an LLM coach that automatically adjusts exercise intensity or duration to keep users within a defined "fat-burning" or cardiovascular-response zone.[10–13] Here, intended use becomes more intervention-oriented, data sources become more physiologically specific, and automation strengthens.[10–13,16,17]

Taken together, such changes may move the product toward the regulated medical-device pathway under the Chinese regulatory system, although final classification would still depend on precise intended use, claims, technical implementation, risk controls, and regulator assessment.[26,27,31–33] Appropriate next steps would likely include stronger risk management, clearer user-facing warnings, more systematic validation of exercise-adjustment logic, and early regulatory consultation.[20–27,31–33]

Case 2: Pregnancy Exercise and Weight-Management App

A pregnancy app may initially offer gestational-age-specific education, low-risk activity suggestions, and general weight-management guidance.[19] Such a profile is more readily understood as health-promotion support than as regulated medical functionality.[19,26,27,31–33]

The risk profile may change if an LLM-based question-and-answer function begins generating concrete and individualized exercise prescriptions, including higher-intensity recommendations, without adequate screening for pregnancy-specific contraindications or complications.[19] In that context, the user group is clinically more vulnerable, recommendation error may carry higher consequences, and general-wellness framing may become insufficient.[19,26,27,31–33]

Under such circumstances, the product could be interpreted as moving toward the regulated medical-device pathway, particularly if the recommendations are individualized and operational rather than educational.[19,26,27,31–39] A more cautious governance approach would include encoded contraindication logic, mandatory escalation or referral thresholds, stronger human oversight, and validation in relevant obstetric subgroups.[19–25]

Case 3: Cardiovascular Rehabilitation Guidance Platform

A cardiovascular rehabilitation platform may begin by offering generic rehabilitation education, exercise videos, and basic recovery advice for post-procedural patients.[16,17] Such a tool may still be framed as educational support if it does not rely on medical-grade signals or make individualized clinical claims.[16,17,26,27,31–33]

The profile may become substantially more regulatory-sensitive if the product later integrates blood-pressure devices, heart-rate sensors, glucose data, or EHR-derived information and uses them to generate or adjust individualized home-based exercise prescriptions.[10,16,17] If marketing then suggests that the system can reduce recurrence risk or optimize rehabilitation for cardiac patients, intended use, user context, data source, and automation may all shift toward a much higher-risk configuration.[16,17,26,27,31–33]

In such a scenario, the product may move into the regulated medical-device pathway and could require higher-risk classification, although final classification would still depend on the exact intended use, claims, technical implementation, risk controls, and regulator assessment.[26–39] A proportionate response would likely require formal risk management, structured clinical or real-world evidence generation, emergency escalation logic, robust post-market monitoring, and strong change-control governance.[16,17,20–25,28–30]

Extended Scenario: Symptom Self-Assessment in Employer Wellness Settings

Similar issues may arise in employer wellness or mental-health settings.[40–42] A platform that initially provides health articles and basic self-management resources may begin to look more medically significant if an LLM advisor starts generating lists of possible conditions and advising employees when to seek care based on symptom descriptions.[26,27,31–33,39–42]

Whether such a tool remains outside the medical-device pathway or moves toward a regulated status would depend on the exact form of its outputs, the specificity of its claims, the data it relies upon, and the extent to which the outputs influence care-seeking decisions.[26,27,31–39] Here again, the matrix is useful not because it offers automatic classification, but because it helps surface why a product may be moving toward a different regulatory category.

Preliminary Application Findings and Inter-Rater Agreement

All three raters completed all seven hypothetical scenarios without missing data or use of the "Insufficient information" category. The raters produced identical scores for every dimension, overall profile, and M flag. Accordingly, for each of the six dimensions and the overall Green/Yellow/Red profile, the average pairwise agreement and exact three-rater agreement were both 100.0%, and Fleiss' kappa was 1.00 (Table 4). M-flag ratings were also identical across all cases (100.0% agreement; Fleiss' kappa = 1.00).

The common final pattern was Green for B1; Yellow for B2, B3, B4, and B6; and Red for B5 and B7. An M flag was assigned only to B5 and B7, where the scenarios combined high-risk medical-use functions with material lifecycle-governance concerns. Notably, B6 was consistently rated Yellow rather than Red: although it concerned post-operative users, it remained a static, clinician-configured education portal without individualized adaptation, medical-grade data integration, autonomous control, or uncontrolled model changes.

These findings support the internal clarity of the anchored rubric and aggregation rules for the supplied vignettes. They should nevertheless be interpreted cautiously. The sample was small, the scenarios were intentionally standardized, and the exercise did not test real-world product files, regulatory determinations, clinical outcomes, or prospective deployment. The framework therefore remains a proposed decision-support and governance scaffold requiring broader external assessment.

Known-Group Contextual Comparison Using Publicly Documented Regulatory Products

To provide a transparent contextual comparison beyond hypothetical scenarios, we examined publicly documented FDA records for SMART (SMART-D, SMART-DX) and BT-001, two US-authorized products with rehabilitation-oriented or disease-management intended uses.[52,53] The comparison did not treat public regulatory status as a gold standard for the entire matrix. Rather, it examined whether the primary intended-use trigger generated an overall Red profile when the public record itself identified an interactive rehabilitation exercise device or a prescription digital therapeutic for type 2 diabetes.

The comparison should be interpreted cautiously. Neither product is an LLM system, and the public records do not disclose all dimensions of the matrix. A US authorization does not determine a product's status in China or the European Union. The exercise is therefore presented as a known-group contextual illustration rather than as a test of regulatory-

Table 4 Preliminary Inter-Rater Agreement for the Six-Dimensional Trigger Matrix and Overall Profile (N = 7 Scenarios)

Outcome	Complete Cases (N)	Average Pairwise Agreement	Exact 3-Rater Agreement	Fleiss' Kappa	I Ratings
D1 Intended use and claims	7	100.0%	100.0%	1.00	0
D2 User context	7	100.0%	100.0%	1.00	0
D3 Depth of personalization	7	100.0%	100.0%	1.00	0
D4 Data and sensor sources	7	100.0%	100.0%	1.00	0
D5 Automation, oversight and closed-loop control	7	100.0%	100.0%	1.00	0
D6 Upgrades and change pathways	7	100.0%	100.0%	1.00	0
Overall profile	7	100.0%	100.0%	1.00	0

Notes: Fleiss' kappa was calculated from original locked categorical ratings. I denotes Insufficient information. M-flag agreement was 100.0% (Fleiss' kappa = 1.00), with M flags assigned only to B5 and B7. This descriptive scenario-based exercise is not a validation of regulatory classification or clinical performance.

https://doi.org/10.2147/RMHP.S624615

classification accuracy, clinical validity, or external validation. A transparent extraction table is provided in Supplementary Table 8. The underlying comparator extraction is provided in Supplementary Material F (Supplementary Table 8).

Discussion

The proposed framework is intended to help these organizations identify when a seemingly low-risk wellness tool begins to acquire the features of medical device software. This is particularly important for procurement because institutions may buy or deploy digital health tools before their regulatory status has been fully examined.

Implications for Healthcare Risk Management

For developers, the framework can function as an early design-control tool. Before making disease-related claims, integrating medical-grade data, adding automatic adjustment loops, or allowing frequent model updates, developers can use the six dimensions to ask whether they are still building a wellness tool or are approaching SaMD territory.

For healthcare institutions, the framework offers a structured way to review procurement claims, vendor documentation, data protection arrangements, staff responsibilities, and escalation pathways. It may also support internal governance committees in distinguishing low-risk health promotion tools from software that requires clinical oversight, procurement controls, or legal and regulatory consultation.

Implications for Healthcare Institutions and Procurement Governance

The six-dimensional trigger matrix and the structured checklist may also support earlier and more structured dialogue with regulators. Instead of asking whether an AI tool is simply "regulated" or "unregulated", developers can present the intended use, data sources, degree of personalization, automation design, update plan, and human-oversight model in a more transparent form.

Regulators and notified bodies may not adopt the exact categories proposed here. However, the framework may still help align pre-submission communication, sandbox evaluation, procurement review, and institutional governance around the same set of risk-relevant product characteristics.

Implications for Preventive Healthcare, Rehabilitation, and Chronic-Disease Services

Exercise prescription and rehabilitation raise distinctive governance concerns because recommendation errors may directly affect physiological load and bodily risk.[10,16,17] In contrast to some informational LLM use cases, errors in exercise intensity, progression, contraindication screening, or response to warning symptoms may create immediate safety consequences.[10,16,17,19]

This means that user context, personalization, data source, and automation are especially important in exercise-related applications.[10–13,16,17,26,27,31–33] In high-risk populations, such as cardiac patients, pregnant individuals, frail older adults, or people with significant metabolic disease, even well-intentioned wellness-oriented design may become inadequate if the tool effectively functions as individualized exercise intervention.[10,16,17,19,26,27,31–33]

From a policy perspective, this suggests that exercise-related LLM tools may benefit from more conservative thresholds in higher-risk populations, stronger referral and escalation logic, and closer linkage to formal medical or sports-medicine governance structures where intervention risk is non-trivial.[16,17,19–27,31–33]

These issues are particularly important in preventive healthcare and chronic-disease services, where LLM tools may be deployed at scale before they are perceived as clinical technologies.[10–13,16,17] A risk-management approach can help avoid both extremes: treating all wellness tools as medical devices, or allowing disease-specific and highly automated functions to remain governed only as general lifestyle advice.

Policy Implications for Risk-Based Oversight

At a broader level, the framework aligns with the increasingly visible policy logic of tiered governance, a pilot-first regulatory approach, and lifecycle supervision in digital health and AI regulation.[26,27,31–33,40–45] It suggests that future guidance could use trigger-matrix-like tools to make regulatory and risk-management expectations more legible for

innovators and institutions.[26–39] It also suggests that domains such as exercise prescription and rehabilitation may be suitable for controlled pilots or sandbox-like testing where demand is high but risk is heterogeneous.[16,17,40–42]

For policymakers, the framework can support a more proportionate approach to oversight. Rather than relying only on final product classification, regulators and healthcare administrators could encourage earlier documentation of claims, data sources, user groups, automation logic, human oversight, and change-control plans.[26–39] Such documentation may make it easier to distinguish low-risk wellness innovation from products that require stronger evidence, professional supervision, or formal regulatory submission.

The framework is China-anchored and comparatively informed: it identifies transferable governance signals that may be mapped onto local regulatory pathways, but it does not provide a jurisdiction-neutral classification rule. It aligns conceptually with IMDRF SaMD materials by distinguishing intended use, user context, data sources, automation and human oversight, evidence planning, and lifecycle change management. In particular, IMDRF N12 discusses the significance of information provided by software and the healthcare situation, while IMDRF N41 addresses clinical association, analytical validation, and clinical validation.[36–38] Formal thresholds for general wellness, clinical decision support, and medical-device software remain jurisdiction- and product-specific.[26,27,31–39]

Limitations and Future Research

Several limitations should be acknowledged. First, the framework is time-bounded by literature and policy materials reviewed through March 2026, and both LLM capabilities and regulatory practice are changing quickly.[1–4,7–9,28–30,47,49,50] Second, the preliminary application exercise involved only three raters and seven intentionally standardized hypothetical scenarios. Although agreement was complete, it supports only initial reproducibility of the supplied rubric and does not establish legal-classification accuracy, clinical validity, or real-world effectiveness. Third, the known-group contextual comparison used public US regulatory records for non-LLM products and did not permit complete coding of all dimensions. Fourth, although the analysis is anchored in China and informed by EU and US comparison, cross-border operational coordination requires dedicated future analysis.[26–39]

Future work should include broader expert-consensus exercises, larger samples of real product documentation, prospective deployment studies, and jurisdiction-specific legal and regulatory analyses. Additional sub-frameworks may be needed for high-risk exercise domains, including cardiovascular rehabilitation, metabolic disease, musculoskeletal disorders, and pregnancy-related exercise.[10,16,17,19]

Conclusion

LLM-enabled exercise and health guidance can move from general wellness support toward higher-risk medical use when product claims, target users, data inputs, personalization, automation, human oversight, or update pathways change. In exercise prescription and rehabilitation, these changes matter because an unsafe recommendation can directly affect physical load, symptom response, and timely referral.

This article presents a proposed six-dimensional trigger matrix, anchored Green/Yellow/Red rubric, non-compensatory aggregation rules, structured risk-governance checklist, and lifecycle pathway. The framework is China-anchored and comparatively informed by EU, US, and IMDRF materials. It is intended to support earlier risk identification, evidence planning, procurement review, and structured dialogue; it does not replace product-specific legal analysis, regulatory determination, clinical evaluation, or a quality-management system.

The preliminary independent-rater exercise supports initial reproducibility of the rubric for standardized hypothetical scenarios, but broader external assessment remains necessary. In particular, future research should test the framework against real-world product documentation, prospective deployment experience, and jurisdiction-specific regulatory decisions.

Generative AI Use Declaration

Google Gemini 3.1 Pro (web-based Gemini application; accessed April 2026) was used during early conceptual visual drafting of the schematic figures. OpenAI ChatGPT (web-based application; accessed June 2026) was used in a limited manner for language, structural, and visual editing assistance during manuscript preparation and figure revision. Generative AI was not used to generate or manipulate research data, clinical images, statistical analyses, scoring

matrices, or the independent-rater agreement results. The authors checked, revised, and approved all final text, figures, tables, and references, and take full responsibility for the accuracy and integrity of the submitted work.

Abbreviations

AI, artificial intelligence; CGM, continuous glucose monitoring; EHR, electronic health record; EU, European Union; FDA, Food and Drug Administration; GMLP, Good Machine Learning Practice; KPI, key performance indicator; LLM, large language model; MDR, Medical Device Regulation; NMPA, National Medical Products Administration; PCCP, Predetermined Change Control Plan; QMS, quality management system; SaMD, software as a medical device.

Data Sharing Statement

The anonymized scenario-level coding data are reported in Supplementary Tables 5 and 6. Original locked rating forms are retained by the authors and can be made available to the Editor for verification of the coding exercise, subject to confidentiality. Other sources discussed are available from the cited literature, regulations, and policy documents.

Ethics Approval and Informed Consent

This study primarily comprised a structured narrative review and doctrinal/comparative legal analysis. The preliminary coding exercise used standardized hypothetical product scenarios and collected no patient data, personal health data, or clinical information. Raters participated voluntarily and were anonymized in reporting. Formal ethics review was not required for this non-interventional methodological exercise.

Acknowledgments

The authors thank the three independent raters for their voluntary contribution to the preliminary scenario-coding exercise.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by the National Key Research and Development Plan Project "Active Health and Technological Response to Population Aging" (Grant No. 2025YFC3610200) and the Xiamen Municipal Medical and Health Key Project (Grant No. 3502Z20234009). The funders had no role in the study design, analysis, interpretation, manuscript preparation, or the authors' decision to submit the article for publication.

Disclosure

All authors report supports from Fujian Provincial Health and Wellness Commission (Grant No. 2021ZQNZD00) and National Intelligent Social Governance Experimental Base Project (Grant No. 350201), outside of this work. The authors declare that they have no other competing interests in this work.

References

1. Raiaan MAK, Mukta MSH, Fatema K, et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access*. 2024;12:26839–26874. doi:10.1109/ACCESS.2024.3365742
2. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177(2):210–220. doi:10.7326/M23-2772
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8

4. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3(1):141. doi:10.1038/s43856-023-00370-1
5. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33. doi:10.1007/s10916-023-01925-4
6. Singhal K, Tu T, Gottweis J, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2
7. Abbasian M, Azimi I, Rahmani AM, Jain R. Conversational health agents: a personalized large language model-powered agent framework. *JAMIA Open*. 2025;8(4):ooaf067. doi:10.1093/jamiaopen/ooaf067
8. Qin H, Tong Y. Opportunities and challenges for large language models in primary health care. *J Prim Care Community Health*. 2025;16:21501319241312571. doi:10.1177/21501319241312571
9. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med*. 2025;5(1):26. doi:10.1038/s43856-024-00717-2
10. Wang Y, Min J, Khuri J, et al. Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. *JMIR Mhealth Uhealth*. 2020;8(4):e15400. doi:10.2196/15400
11. Sun L, Liu D, Wang M, et al. Taming unleashed large language models with blockchain for massive personalized reliable healthcare. *IEEE J Biomed Health Inform*. 2025;29(6):4498–4511. doi:10.1109/JBHI.2025.3528526
12. Wang B, Zheng Y, Han X, et al. A systematic literature review on integrating AI-powered smart glasses into digital health management for proactive healthcare solutions. *NPJ Digit Med*. 2025;8(1):410. doi:10.1038/s41746-025-01715-x
13. Tavabi N, Singh M, Pruneski J, et al. Systematic evaluation of common natural language processing techniques to codify clinical notes. *PLoS One*. 2024;19(3):e0298892. doi:10.1371/journal.pone.0298892
14. Belkhouribchia J, Pen JJ. Large language models in clinical nutrition: an overview of its applications, capabilities, limitations, and potential future prospects. *Front Nutr*. 2025;12:1635682. doi:10.3389/fnut.2025.1635682
15. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
16. Thomas RJ, Beatty AL, Beckie TM, et al. Home-based cardiac rehabilitation: a scientific statement from the American Association of Cardiovascular and Pulmonary Rehabilitation, the American Heart Association, and the American College of Cardiology. *J Am Coll Cardiol*. 2019;74(1):133–153. doi:10.1016/j.jacc.2019.03.008
17. Cotie LM, Vanzella L, Pakosh MT, et al. A systematic review of clinical practice guidelines and consensus statements for cardiac rehabilitation delivery: consensus, divergence, and important knowledge gaps. *Can J Cardiol*. 2024;40(3):330–346. doi:10.1016/j.cjca.2023.10.016
18. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digit Med*. 2024;7(1):183. doi:10.1038/s41746-024-01157-x
19. American College of Obstetricians and Gynecologists. Physical activity and exercise during pregnancy and the postpartum period: ACOG Committee Opinion, Number 804. *Obstet Gynecol*. 2020;135(4):e178–e188. doi:10.1097/AOG.0000000000003772
20. Park SH, Suh CH, Lee JH, et al. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865–868. doi:10.3348/kjr.2024.0843
21. 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. *BMJ*. 2022;377:e070904. doi:10.1136/bmj-2022-070904
22. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–1363. doi:10.1038/s41591-020-1037-7
23. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health*. 2020;2(10):e537–e548. doi:10.1016/S2589-7500(20)30218-1
24. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. 2019;25(10):1467–1468. doi:10.1038/s41591-019-0603-3
25. Solebo AL, Braithwaite T. Implications of the artificial intelligence extensions to the guidelines for consolidated standards of reporting trials and for standard protocol item recommendations for interventional trials (the CONSORT-AI and SPIRIT-AI extensions). *EClinicalMedicine*. 2020;26:100536. doi:10.1016/j.eclinm.2020.100536
26. Center for Medical Device Evaluation, National Medical Products Administration. Guideline for the review of artificial intelligence medical devices. Beijing: CMDE; 2022.
27. Center for Medical Device Evaluation, National Medical Products Administration. Guideline for the review of medical device software registration (2022 revision). Beijing: CMDE; 2022.
28. US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and food and drug administration staff. Silver Spring (MD): FDA; 2025.
29. International Medical Device Regulators Forum. Good machine learning practice for medical device development: guiding principles. IMDRF/AIML WG/N88FINAL:2025. 2025.
30. US Food and Drug Administration. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations: draft guidance for industry and food and drug administration staff. Silver Spring (MD): FDA; 2025.
31. National Medical Products Administration. Regulations on supervision and administration of medical devices. Beijing: NMPA; 2021.
32. National Medical Products Administration. Rules for classification of medical devices. Beijing: NMPA; 2015.
33. National Medical Products Administration. Provisions for medical device registration and filing. Beijing: NMPA; 2021.
34. European Parliament, Council of the European Union. Regulation (EU) 2017/745 of 5 April 2017 on medical devices. *Off J Eur Union*. 2017;L117:1–175.
35. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union*. 2024;2024:1689.
36. International Medical Device Regulators Forum. Software as a Medical Device (SaMD): key definitions. IMDRF/SaMD WG/N10FINAL:2013. 2013.
37. International Medical Device Regulators Forum. Software as a Medical Device (SaMD): possible framework for risk categorization and corresponding considerations. IMDRF/SaMD WG/N12FINAL:2014. 2014.

38. International Medical Device Regulators Forum. Software as a Medical Device (SaMD): clinical evaluation. IMDRF/SaMD WG/N41FINAL:2017. 2017.
39. US Food and Drug Administration. Policy for device software functions and mobile medical applications: guidance for industry and food and drug administration staff. Silver Spring (MD): FDA; 2022.
40. National Health Commission. National administration of traditional Chinese medicine. measures for the administration of internet diagnosis and treatment (Trial). Beijing: NHC; 2018.
41. National Health Commission. National administration of traditional Chinese medicine. measures for the administration of internet hospitals (Trial). Beijing: NHC; 2018.
42. National Health Commission. National administration of traditional Chinese medicine. telemedicine service management specification (Trial). Beijing: NHC; 2018.
43. Standing Committee of the National People's Congress. Personal information protection law of the People's Republic of China. Beijing: National People's Congress; 2021.
44. Standing Committee of the National People's Congress. Data security law of the People's Republic of China. Beijing: National People's Congress; 2021.
45. Standing Committee of the National People's Congress. Cybersecurity law of the People's Republic of China. Beijing: National People's Congress; 2016.
46. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378–382. doi:10.1037/h0031619
47. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319–328. doi:10.1001/jama.2024.21700
48. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195. doi:10.1038/s41746-023-00939-z
49. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat Med*. 2025;31(2):618–626. doi:10.1038/s41591-024-03445-1
50. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. *Medicina*. 2024;60(3):445. doi:10.3390/medicina60030445
51. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *Br J Sports Med*. 2020;54(24):1451–1462. doi:10.1136/bjsports-2020-102955
52. U.S. Food and Drug Administration. 510(k) Premarket Notification: SMART (SMART-D, SMART-DX), K131660. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K131660. Accessed June 20, 2026.
53. U.S. Food and Drug Administration. De Novo classification request for BT-001, DEN220058. Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN220058.pdf. Accessed June 20, 2026.