

Clinical Tasks Explain AI Readiness Across Medical Specialties

Bertalan Mesko

berci@medicalfuturist.com

The Medical Futurist Institute

Gellért Katonai

The Medical Futurist Institute

Shangchu Chen

Tsinghua University

Bin Sheng

Shanghai Jiao Tong University

Tien Yin Wong

Tsinghua University

Eric Topol

Scripps Research Translational Institute, Scripps Research

Article

Keywords: foresight, artificial intelligence, medical specialties, healthcare, medicine

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11164578/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) is increasingly transforming clinical work, yet its potential impact is often considered at the level of medical specialties rather than their constituent clinical tasks.

We developed a clinical task-based framework, medical specialty AI readiness (MSAIR), for evaluating AI readiness across 22 medical specialties, scoring 330 representative clinical tasks across six dimensions.

We found AI-readiness scores varied more widely across clinical task categories than across specialties. Clinical documentation and diagnostic interpretation had the highest task-category scores, whereas emergency decision-making and medical procedures had the lowest. Clinical task explained 71.6% of the variation in AI readiness, compared with only 6.4% explained by medical specialty.

These findings suggest that AI readiness is more strongly associated with the nature of clinical work than with medical specialty. The MSAIR framework is useful for planning and anticipating how AI may reshape medical practice.

Introduction

Artificial intelligence (AI) has rapidly progressed from proof-of-concept systems to clinical deployment across healthcare¹. Over the past decade, AI technologies have expanded from image interpretation algorithms and clinical decision support (CDS) tools to ambient clinical documentation, workflow automation, predictive analytics, and increasingly sophisticated foundation models capable of supporting a broad range of clinical, operational, and administrative tasks, with health systems beginning to integrate AI into routine clinical practice². At the same time, regulatory pathways have matured, and the number of approved AI-enabled medical technologies has increased substantially³. There has also been an expansion of reporting guidelines and recommendations for the responsible use of AI⁴. The current discussion has shifted from evaluating isolated, individual AI models to understanding how AI may reshape and transform clinical work across the medical profession and the broader scope of different medical specialties, such as radiology, pathology, primary care and internal medicine⁵.

These trends have raised two increasingly important questions: which medical specialties are most ready for AI-enabled transformation, and how will these changes reshape medical training and the workforce by augmenting or replacing specific clinical tasks?

There are suggestions that image-based specialties like radiology and pathology, as well as language-based specialties such as primary care and psychiatry, are more susceptible to AI-driven automation or even replacement⁶. However, radiology provides a useful illustration of this paradox. Because AI has achieved remarkable performance in image interpretation, radiology has frequently been cited as one of the specialties most susceptible to automation^{7,8}. However, despite substantial investment in imaging AI

and repeated predictions of workforce disruption, radiology continues to experience strong demand, expanding residency programs in many regions, and sustained clinical relevance⁹. This apparent paradox reflects the fact that the clinical task of image interpretation (which may be replaced by AI) represents only one task of the profession of radiology.

Importantly, these debates overlook a fundamental limitation of current understanding of AI: AI does not transform the medical professions directly but likely transforms the specific clinical tasks that constitute the professional practice of that specialty. A medical specialty is a historical and traditional classification of the medical profession that includes many similar yet heterogeneous collections of clinical tasks, from diagnostic reasoning, procedures, communication, and documentation, to multidisciplinary collaboration, and longitudinal patient management. Some clinical tasks (e.g., image interpretation and data analysis) may become more prone to the use and impact of AI, whereas others (e.g., longitudinal management and physical examination) may still remain predominantly human-led for the foreseeable future¹⁰. Meaningful comparisons between medical specialties should therefore start by examining the individual clinical tasks that constitute them, rather than treating each specialty as a homogeneous entity¹¹.

To our knowledge, there has been no previous study that has proposed a systematic analysis and a clinical task-based framework for comparing AI readiness across major medical specialties. We suggest that the more informative question is how the collection of clinical tasks defining each medical specialty differs in its readiness for AI-enabled transformation.

In this study, we aimed to develop and demonstrate a framework for estimating medical specialty AI readiness (MSAIR) by systematically identifying the core clinical tasks within each specialty and examining how their readiness for AI-enabled transformation differs both across types of clinical work and across medical specialties. Importantly, MSAIR represents the aggregated AI readiness of the clinical tasks within a specialty rather than an intrinsic property of the specialty itself. We envision this framework as a tool to support healthcare leaders, policymakers, medical educators, and professional societies in anticipating how AI may reshape clinical work and preparing the healthcare workforce accordingly.

Results

Specialty-level and task-category AI readiness

MSAIR values across medical specialties ranged from 2.010 to 2.543 (Fig. 1), summarizing heterogeneous task portfolios. Pathology (2.543) and radiology (2.543) were at the upper end, and anesthesiology (2.010) and orthopaedic surgery (2.019) at the lower end. However, the wide within-specialty task ranges and differences in task-category composition illustrated that these specialty means aggregated heterogeneous clinical work rather than representing uniform specialty-level readiness (Supplementary Table S1).

Mean task-level AI-readiness scores across task categories ranged from 1.639 for emergency decision making to 2.874 for documentation (Fig. 2). Documentation and diagnostic interpretation (2.700) had the highest means, whereas emergency decision making and procedures (1.720) had the lowest. The spread between task-category means (1.235) was more than twice the spread between specialty means (0.533; Supplementary Table S2).

Task category explains substantially more variation than specialty

Clinical task category explained substantially more variation in the primary task-level AI-readiness score than medical specialty (adjusted R^2 0.716 vs 0.064; Fig. 3). Adding specialty to task category increased adjusted R^2 only modestly to 0.734, an additional 1.7 percentage points (Supplementary Table S3).

The same pattern was observed after excluding specialty centrality. Adjusted R^2 was 0.778 for task category, 0.098 for specialty, and 0.799 for the combined model. These results show that the principal finding did not depend on the inclusion or double weighting of specialty centrality (Supplementary Table S3).

Principal component analysis and sensitivity analyses

In the task-level PCA, PC1 explained 52.8% of the total variance and was primarily associated with data-structure dependence (loading 0.910), transferability of decision authority (0.836), clinical governability (0.774), codifiability (0.765), and physical dependence (0.630) (Supplementary Tables S4 and S5).

The PCA supported an empirical distinction between task transformability and specialty centrality. The five transformability dimensions loaded predominantly on PC1 (loadings 0.630–0.910) but only weakly on PC2 (– 0.186 to 0.091), whereas specialty centrality loaded weakly on PC1 (0.269) but strongly on PC2 (0.951) (Supplementary Table S4). PC2 and PC3 explained 16.1% and 14.1% of the variance, respectively, and the first three components together explained 83.0% of total variance (Supplementary Table S5). PC1 showed strong alignment with the primary task-level AI-readiness score (Pearson r = 0.957).

Across alternative scoring and aggregation methods, 13 of 22 specialties remained in the same tertile across all methods. Stability was greatest among specialties at the upper and lower ends of the score distribution, whereas changes were concentrated among middle-ranked specialties and those close to grouping boundaries. Specialty-level comparisons between the primary score and PC1-derived profiles are provided in Supplementary Table S6. Detailed rank, tertile, and quartile stability results are reported in Supplementary Table S7.

Together, these analyses support interpreting MSAIR as the aggregated readiness of representative clinical tasks within a specialty rather than as an intrinsic property of the specialty itself.

Discussion

The growing use of AI across medicine has intensified debate about which clinical tasks or which medical specialties are most ready for AI augmentation. Such predictions have often been broad: public figures and investors have argued that AI could eventually perform 80% of the work currently done by physicians, while others have claimed that AI already outperforms most of them^{12,13}. Similar narratives have repeatedly singled out data-intensive specialties such as radiology as candidates for substantial automation or even replacement.

In this study, AI-readiness varied more widely across clinical task categories than across medical specialties. Clinical documentation and diagnostic interpretation had the highest mean task-level AI-readiness scores, whereas emergency decision making and medical procedures had the lowest. Among medical specialties, pathology and radiology had the highest MSAIR values, whereas anesthesiology and orthopaedic surgery had the lowest. These findings are not unexpected^{14,15}. More importantly, we demonstrate that the category a clinical task belongs to explained substantially more variation in task-level AI readiness than the specialty itself. In fact, clinical task category explained more than ten times as much adjusted R^2 in AI readiness as medical specialty (71.6% vs 6.4%).

Thus, our work shows AI affects individual clinical tasks within each specialty differently, making specialty-level predictions potentially misleading. It means that the readiness of medical AI applications depends more on the nature of specific clinical tasks than on medical specialty, suggesting that AI readiness is more strongly associated with the characteristics of clinical work than with the specialty in which that work is performed.

We argue that discussions about AI's impact on medicine should shift from asking which specialties will be most affected or even replaced towards identifying which clinical tasks are most likely to be augmented, transformed, or automated by AI. As we indicated above, radiology provides a clear illustration of why AI readiness should be examined at the task rather than the specialty level. Radiology was among the highest-scoring specialties in our analysis, consistent with its concentration of data-intensive tasks such as image interpretation. However, our task-category analysis showed substantially greater variation across types of clinical work (mean scores 1.639–2.874) than across specialties (2.010–2.543). Diagnostic interpretation was among the highest-scoring task categories, whereas procedures and emergency decision-making were among the lowest. Radiology combines these highly AI-ready diagnostic tasks with activities involving physical intervention, multidisciplinary collaboration, and clinical decision-making, illustrating why high readiness of individual tasks should not be interpreted as uniform readiness across the specialty or as evidence of professional replacement^{16,17}.

This pattern extended beyond radiology. Across all 330 tasks, task category explained 71.6% of the adjusted R^2 in AI-readiness scores, compared with only 6.4% explained by specialty, and adding specialty to task category increased the explained variation only modestly to 73.4%. Thus, whether a physician was performing documentation, diagnostic interpretation, a procedure, or an emergency decision was

considerably more informative about AI readiness than the specialty in which that task was performed. This finding helps explain why specialties such as dermatology or psychiatry can contain individual tasks highly amenable to AI while the overall specialty remains substantially more heterogeneous^{18,19}.

The task-category analysis provides a more generalizable perspective across medicine. Documentation, a workflow shared broadly across specialties, had the highest mean AI-readiness score (2.874) and relatively limited dispersion. This finding suggests that some of the earliest and most widespread effects of AI may occur across, rather than within, medical specialties²⁰. AI may therefore transform medicine horizontally before it transforms it vertically, reshaping common workflows across many specialties rather than uniformly transforming individual medical professions. At the same time, peer-to-peer collaboration, medical procedures, and emergency decisions showed the lowest affinity towards automation, highlighting the continuing importance of physicality: specialties containing a greater proportion of tasks requiring physical presence, manual skills, or direct interaction with patients and colleagues may be less amenable to AI-enabled transformation.

Diagnostic interpretation provides a complementary finding. Although it had the second-highest mean AI-readiness score (2.700), scores were more dispersed than for documentation, indicating that diagnostic tasks should not themselves be regarded as uniformly AI-ready. The characteristics of the information being interpreted, the clinical context, and the consequences and governability of the resulting decision remain important even within a task category generally considered highly amenable to AI²¹.

The PCA provides a further perspective on the structure of the framework. The first principal component was predominantly associated with the five dimensions describing task transformability, whereas specialty centrality loaded strongly on the second component. This pattern suggests that two related, but distinct questions underlie MSAR: how amenable a clinical task is to AI-enabled transformation, and how important transformation of that task would be to the specialty in which it is performed. The distinction may be important when interpreting specialty-level estimates, because transformation of a highly AI-ready but peripheral task is unlikely to have the same implications for a specialty as transformation of one of its defining clinical activities.

Viewing AI through the lens of clinical tasks may provide a more useful foundation for workforce planning than specialty-level predictions. Rather than asking which specialties face the greatest disruption, policymakers may identify which components of clinical work are most likely to change and where education, regulation, and workforce development should be prioritized. Our findings suggest that AI is therefore more likely to augment, transform, or replace specific components of clinical work than medical specialties as a whole.

Medical education may similarly benefit from a task-oriented perspective. As AI becomes integrated into clinical workflows, training programs may need to prepare physicians to supervise, validate, and collaborate with AI in those tasks where AI support is most likely to emerge, while continuing to

strengthen uniquely human clinical capabilities. Medical AI education may therefore be better organized around broad principles of AI use across clinical tasks such as diagnosis, documentation, monitoring, and decision support rather than around its application within existing specialty boundaries.

Professional societies and medical associations may use such frameworks to identify areas where specialty-specific guidance or competency standards should evolve. Likewise, developers and healthcare organizations may better anticipate where AI applications are most likely to achieve meaningful clinical adoption by considering the characteristics of individual tasks rather than broad professional categories. More broadly, AI models may be more usefully developed around clinical functions that extend across specialties, such as diagnostic interpretation or clinical documentation, rather than around individual specialties themselves.

A similar principle may apply to governance and regulation, where claims about AI systems may be more meaningfully evaluated according to the clinical tasks they perform and the associated risks than according to the specialties in which they are deployed.

Previous discussions of AI and the medical workforce have largely relied on expert opinion, technology-specific case studies, or broad economic projections. Influential reports from academic, governmental, and industry organizations have highlighted the transformative potential of AI for healthcare, while thought leaders have emphasized augmentation rather than replacement of physicians^{22,23}. Our framework is consistent with these perspectives but offers a contribution by providing a structured method for comparing specialties through their constituent clinical tasks. Rather than predicting future workforce outcomes, it provides a structured framework for discussing them and anticipating how AI may reshape clinical work across medical specialties.

Our findings suggest that understanding the future impact of AI on medicine requires looking beyond medical specialties to the clinical tasks that constitute medical practice.

Limitations

Several limitations should be acknowledged. First, the selection and scoring of representative clinical tasks necessarily involved expert judgement. Although tasks were selected to capture broad and characteristic forms of clinical work, the database was not developed through a formal Delphi process, and alternative investigators might reasonably select, define, or score some tasks differently. Consequently, specialty-level estimates should be interpreted as summaries of the selected task portfolios rather than definitive measurements of entire specialties.

Second, the six dimensions and their weighting represent methodological choices, and alternative frameworks could produce different estimates. We sought to limit unwarranted precision through predefined criteria and a three-point ordinal scale, while sensitivity analyses assessed the stability of the results under alternative weighting and aggregation approaches.

Finally, AI capabilities and medical practice are evolving rapidly. AI readiness should therefore be regarded as dynamic rather than fixed, and the framework will require reassessment as technologies, clinical tasks, and specialty practice evolve.

Future work should evaluate the framework using broader international expert consensus, larger specialty-specific task inventories, and periodic updates reflecting advances in AI capabilities. The same task-based approach could also be adapted to other healthcare professions, health-system functions, or specific categories of AI-enabled technologies.

Conclusion

We introduce the Medical Specialty AI Readiness (MSAIR) framework, a clinical task-based approach for estimating the aggregated AI readiness of representative clinical tasks within medical specialties. Our study findings suggest that the effects of AI on medicine will be shaped primarily by the composition of clinical work within each specialty, rather than by specialty labels alone. By shifting the focus from professions to the clinical tasks that define them, the framework provides a foundation for workforce planning, medical education, and future AI implementation research.

Rather than providing a definitive ranking of specialties, we view the framework as an evolving methodology that can be refined as AI capabilities, clinical practice, and healthcare systems change.

Methods

Study design and specialty selection

We developed a cross-sectional, task-based scoring framework to estimate the potential readiness of medical specialties for AI. The unit of analysis was the individual clinical task rather than the specialty as a whole. Specialty-level estimates were subsequently calculated by aggregating scores across the tasks selected to represent each specialty.

We used the official list of specialty certificates published by the American Board of Medical Specialties as the basis for specialty selection. From this list, we included 22 major medical specialties representing a broad range of diagnostic, procedural, cognitive, and longitudinal forms of clinical practice. Specialties were included when they represented a distinct field of physician practice and had a sufficiently developed competency framework to support the systematic identification of specialty-defining tasks. Subspecialties were not assessed separately because their inclusion would have introduced substantial overlap in clinical activities and resulted in unequal levels of granularity across fields.

Development of the clinical task database

For each specialty, we identified 15 clinical tasks intended to represent the breadth of work performed by physicians in that field. Task identification was based primarily on specialty-specific competency

frameworks and supplemental guides published by the Accreditation Council for Graduate Medical Education.

The task-selection process was designed to capture the principal forms of clinical work that define each specialty, rather than to produce an exhaustive inventory of all activities performed by its physicians. Tasks were included when they were sufficiently specific to permit consistent scoring, commonly performed or central to the specialty, primarily undertaken by physicians, distinguishable from other selected tasks, and collectively representative of practice across different clinical settings.

Tasks were excluded when they were too broad to score reliably, represented general professional competencies shared by almost all physicians, substantially duplicated another selected task, were performed predominantly by non-physician professionals, or reflected a narrow subspecialty activity that was not representative of the wider specialty.

To improve consistency across specialties, tasks were mapped to a common set of functional categories: patient assessment; diagnostic interpretation; clinical reasoning; treatment planning; communication; documentation; monitoring; procedures; emergency decision making; multidisciplinary collaboration; prevention; and quality improvement and systems-based practice.

These categories were used to support balanced task selection and cross-specialty comparison. They were not treated as mutually exclusive classifications because some clinical activities incorporated more than one functional component.

The resulting database contained 330 clinical tasks across 22 specialties, with 15 tasks assigned to each specialty.

The six dimensions were selected to capture complementary characteristics relevant to the potential transformation of clinical tasks by AI. Codifiability and data-structure dependence address whether a task can be represented in forms amenable to AI; transferability of decision authority and clinical governability address whether AI involvement can be delegated and safely supervised in clinical practice; physical dependence captures constraints arising from embodied or procedural work; and specialty centrality reflects the significance of the task to the specialty itself. Together, these dimensions were intended to capture technical, clinical, physical, and professional aspects of AI-enabled task transformation rather than to provide an exhaustive account of all factors influencing real-world AI adoption (Supplementary Table S8).

Development of the AI-readiness dimensions

We assessed each task across six dimensions intended to capture different characteristics relevant to its potential transformation by AI. Five dimensions described properties affecting the technical and clinical applicability of AI, whereas specialty centrality described the importance of the task to the identity and value of the specialty.

Each dimension was scored on a three-point ordinal scale. The restricted scale was chosen to support interpretable comparisons and to avoid implying a level of numerical precision that the assessment process could not provide.

1) Codifiability

Codifiability represented the extent to which a task could be described through explicit rules, templates, examples, or labelled data. Tasks that depended heavily on tacit knowledge, contextual judgment, or poorly specified processes received lower scores. Tasks that could be described consistently and represented through repeatable patterns received higher scores.

Scores were assigned as follows:

1: low codifiability;

2: moderate codifiability;

3: high codifiability.

2) Data-structure dependence

Data-structure dependence represented the extent to which task performance relied on information available in structured or digitally accessible forms. Tasks based predominantly on interpersonal interaction, direct observation, or information not readily captured digitally received lower scores. Tasks using a mixture of digital data and contextual or interpersonal information received intermediate scores. Tasks performed predominantly through structured digital data received higher scores.

Scores were assigned as follows:

1: predominantly interpersonal or non-digital;

2: mixed data and interpersonal dependence;

3: predominantly data-based.

3) Transferability of decision authority

Transferability of decision authority represented the difficulty of transferring a meaningful component of decision-making from the physician to an AI system. The dimension incorporated the degree of professional discretion involved, the consequences of error, and the extent to which clinicians, patients, and institutions might reasonably accept AI participation in the decision.

Scores were assigned as follows:

1: high difficulty transferring decision authority;

2: moderate difficulty;

3: low difficulty.

4) Physical dependence

Physical dependence represented the extent to which successful task completion required physical co-presence, manual examination, procedural skill, or direct interaction with the patient or clinical environment. Tasks with substantial physical or procedural requirements received lower scores, whereas tasks that could be conducted primarily through digital information or remote interaction received higher scores.

Scores were assigned as follows:

1: high physical dependence;

2: moderate physical dependence;

3: low physical dependence.

5) Clinical governability

Clinical governability represented the extent to which AI-generated outputs could be safely supervised, verified, corrected, and governed in clinical practice. Tasks received lower scores when errors could be difficult to detect, correction opportunities were limited, or the outputs had immediate and potentially serious consequences. Tasks received higher scores when outputs could be independently reviewed, compared with established standards, corrected before implementation, or introduced within a clearly defined governance process.

Scores were assigned as follows:

1: difficult to supervise, verify, or correct;

2: moderately governable;

3: readily governable.

6) Specialty centrality

Specialty centrality represented the importance of a task to the specialty's professional identity, distinctive expertise, and contribution to patient care. Peripheral tasks received lower scores, tasks representing an important but non-defining component of practice received intermediate scores, and tasks central to the specialty's identity and value creation received higher scores.

Scores were assigned as follows:

1: peripheral;

2: important;

3: core.

Specialty centrality was conceptually distinct from the other dimensions. It reflected how consequential the transformation of that task would be for the specialty. AI transformation of a peripheral administrative activity and transformation of a specialty-defining clinical activity should not contribute equally to the interpretation of specialty-level readiness, even when their technical characteristics are otherwise similar.

For this reason, specialty centrality was assigned twice the weight of each of the other dimensions in the composite task score. This weighting was specified before the principal analyses and was examined in sensitivity analyses.

Task scoring and calculation of the composite score

Each clinical task was assessed across the six dimensions. The primary task-level AI-readiness score for each task was calculated as:

$$AI_{readiness_i} = \frac{C_i + D_i + T_i + P_i + G_i + 2S_i}{7}$$

where C represented codifiability, D data-structure dependence, T transferability of decision authority, P physical dependence, G clinical governability, and S specialty centrality.

This calculation produced a task-level score ranging from 1 to 3, with higher scores indicating greater estimated readiness for AI-enabled transformation. Primary specialty-level estimates, representing MSAIR, were calculated as the arithmetic mean of the 15 primary task-level AI-readiness scores assigned to each specialty.

These scores were intended to support relative comparison rather than to represent absolute probabilities of AI adoption, automation, or workforce displacement.

Cross-checking, consensus, and uncertainty

The initial task lists and dimension scores were reviewed by members of the study team to assess whether the selected tasks adequately represented each specialty and whether the scoring criteria had been applied consistently. Disagreements were discussed and resolved through consensus.

Statistical analysis

The analysis was designed as a task-level evaluation of medical specialty AI-readiness, with specialty-level estimates treated as summaries of the selected task portfolios. The clinical task was treated as the primary unit of analysis and the specialty-level estimates were interpreted as aggregations of task-level scores.

Primary score construction and specialty-level aggregation

For each task, three task-level scores were calculated. The primary score was calculated as $(C + D + T + P + G + 2S) / 7$, with specialty centrality double-weighted. The equal-weight score was calculated as $(C + D + T + P + G + S) / 6$, and the no-centrality score as $(C + D + T + P + G) / 5$.

The primary specialty-level score was the arithmetic mean of primary task-level scores within each specialty. Additional specialty-level sensitivity estimates included the equal-weight score, the no-centrality score, the centrality-weighted primary score, and the centrality-weighted no-centrality score. For the two centrality-weighted scores, task scores were averaged using the corresponding specialty-centrality ratings as weights. Task-category scores were calculated as the arithmetic mean of the primary task scores within each category.

Specialties were ordered from highest to lowest according to the primary specialty-level score. Tied specialties shared the same competition rank. The observed tertiles were labelled as low, moderate, or

high AI-Readiness.

Comparison of task category and specialty

To provide a descriptive comparison of AI-readiness between task category and specialty, we fitted task-level ordinary least-squares models containing specialty alone, primary task category alone, or both factors. Specialty and task category were represented as categorical variables. Model performance was summarized using R^2 and adjusted R^2 . The analysis was repeated using the transformability score excluding specialty centrality to assess whether the findings depended on the inclusion and double weighting of centrality. (Fig. 3.)

Task-level principal component analysis

To explore the empirical structure of the six scoring dimensions, we performed task-level principal component analysis. PCA was conducted on the six-dimensional task-level matrix after z-score standardization. Six principal components were estimated using full singular value decomposition. PCA component interpretation was based on the PCA loading matrix, which was calculated by multiplying the transposed PCA component coefficient matrix by the square root of the corresponding eigenvalues. Variables with loadings of at least 0.40 or at most -0.40 were treated as dominant positive or negative contributors, respectively. Components without dominant variables were described using the variables with the largest absolute loadings.

Alignment between the PCA-derived structure and the prespecified scoring framework was assessed descriptively by calculating Pearson correlations between each task-level principal component and the primary task-level score. Task-level PCA scores were then averaged within specialties to generate specialty-level PCA profiles. Spearman correlation was calculated between specialty-level PC1 scores and the primary specialty-level score. PC1-based specialty ranks and tertiles were compared with those based on the primary specialty-level score.

Sensitivity and stability analyses

Sensitivity and stability analyses compared specialty ranks, primary tertiles, and supplemental quartiles across alternative scoring and aggregation schemes. The methods compared were the primary specialty-level score, equal-weight specialty score, no-centrality specialty score, centrality-weighted score, and centrality-weighted no-centrality score. For each specialty, we calculated the maximum rank change across methods and assessed whether the primary tertile and supplemental quartile assignments remained unchanged across methods.

Declarations

Funding: Not applicable.

Author contributions. B.M. conceived the study, developed the methodological framework, and drafted the manuscript. S.C. and B.S. performed the statistical analyses and calculations. G.K., S.C., B.S., T.Y.W., and E.T. contributed to the interpretation of the findings and critically revised the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript.

Competing interests: There are no competing interests.

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* vol. 25 44–56 (2019).
2. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
3. Benjamens, S., Dhunoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digit. Med.* **3**, 118 (2020).
4. Kolbinger, F. R., Veldhuizen, G. P., Zhu, J., Truhn, D. & Kather, J. N. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun. Med.* **4**, 71 (2024).
5. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
6. Popa, S. L. et al. Artificial Intelligence and medical specialties: support or substitution? *Med. Pharm. reports* **97**, 409–418 (2024).
7. Obuchowicz, R. et al. Artificial Intelligence-Empowered Radiology—Current Status and Critical Review. *Diagnostics* **15**, (2025).
8. Katal, S., York, B. & Gholamrezanezhad, A. AI in radiology: From promise to practice – A guide to effective integration. *Eur. J. Radiol.* **181**, 111798 (2024).
9. Rozenshtein, A., Findeiss, L. K., Wood, M. J., Shih, G. & Parikh, J. R. The U.S. Radiologist Workforce: AJR Expert Panel Narrative Review. *AJR. Am. J. Roentgenol.* **224**, e2432085 (2025).
10. Obuchowicz, R. et al. Will AI Replace Physicians in the Near Future? AI Adoption Barriers in Medicine. *Diagnostics (Basel, Switzerland)* **16**, (2026).
11. Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
12. Bruce, G. ‘AI is already better than most doctors,’ says Elon Musk. *Becker’s Healthcare* (2025).
13. On with Kara Swisher - Vinod Khosla’s View of the Future, From AI to China. *Podscripts* <https://podscripts.co/podcasts/on-with-kara-swisher/vinod-khoslas-view-of-the-future-from-ai-to-china> (2023).
14. Tierney, A. A. et al. Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. *NEJM Catal.* **6**, CAT.25.0040 (2025).

15. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
16. Pesapane, F., Codari, M. & Sardanelli, F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* **2**, 35 (2018).
17. Yu, F. *et al.* Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
18. Udrea, A. *et al.* Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *J. Eur. Acad. Dermatology Venereol.* **34**, 648–655 (2020).
19. Volkmer, S., Meyer-Lindenberg, A. & Schwarz, E. Large language models in psychiatry: Opportunities and challenges. *Psychiatry Res.* **339**, 116026 (2024).
20. Rotenstein, L. S. *et al.* Changes in Clinician Time Expenditure and Visit Quantity With Adoption of Artificial Intelligence-Powered Scribes: A Multisite Study. *JAMA* **335**, 1408–1417 (2026).
21. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
22. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
23. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).

Figures

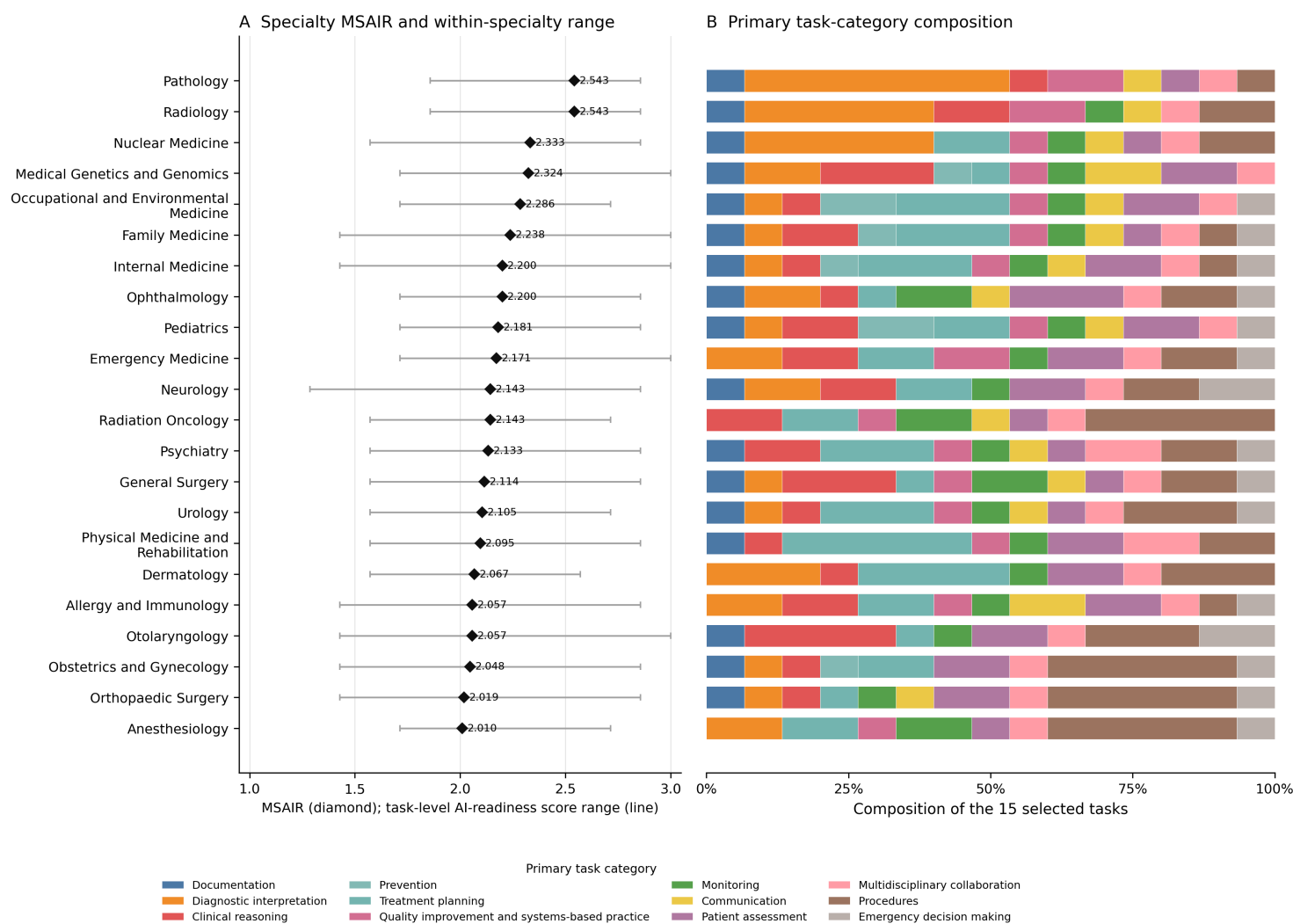


Figure 1

MSAIR and Composition of Selected Clinical Tasks.

Panel A, MSAIR (diamond) and observed minimum-to-maximum task-level AI-readiness score range among the 15 selected tasks. Panel B, Distribution of the 15 selected tasks across primary task categories within each specialty.

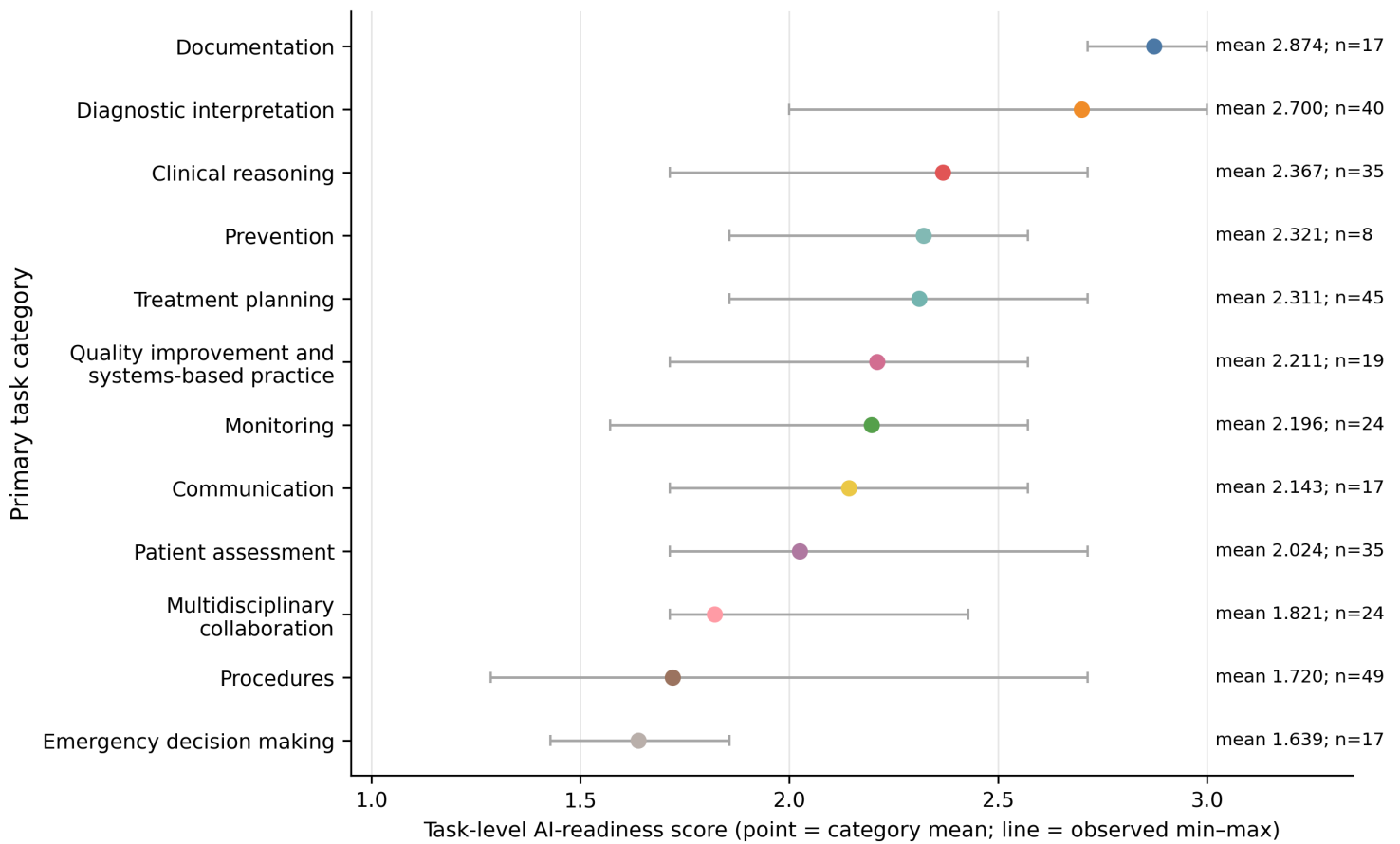


Figure 2

Distribution of Primary AI-Readiness Scores Across Clinical Task Categories. Points indicate mean task-level AI-readiness scores, and horizontal lines indicate the observed minimum-to-maximum range. The number of tasks assigned to each primary task category is shown on the right.

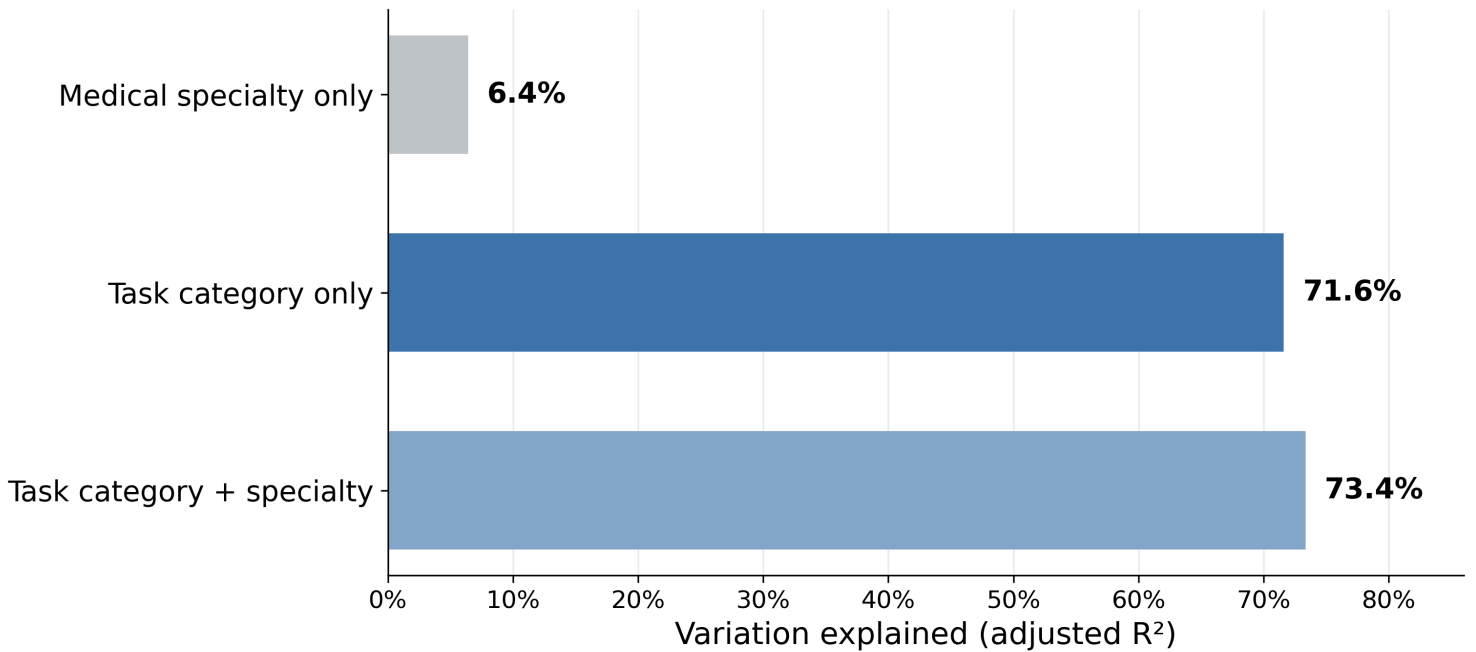


Figure 3

Explanatory Value of Task Category and Specialty for Task-Level AI-Readiness Scores.

Bars show adjusted R^2 values from ordinary least-squares models containing specialty alone, primary task category alone, or both factors, using the primary task-level AI-readiness score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationTablesS1S7.pdf](#)
- [SupplementaryDataS8.pdf](#)