Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnostic specificity of a retrieval augmented language model on complex clinical cases: a retrospective benchmark study

Evans Kwabena Abor
University of Cincinnati College of Medicine
Philip M. Diller
University of Cincinnati College of Medicine
Pius Kwasi Sam
University of Cape Coast
Loretta Yaa Ema Afful
University of Cape Coast
Leslie Afutu
University of Cape Coast
George Ghartey-Kwansah
University of Cape Coast
Francis Ackah Armah
francis.armah@ucc.edu.gh
University of Cape Coast

Research Article

Keywords: Artificial intelligence, Large language models, Diagnostic accuracy, Diagnostic specificity, Retrieval augmented generation, Clinical decision support

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11793294/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. E.K.A. developed MediDx AI, the platform evaluated in this study. P.M.D., L.A., G.G.-K. and F.A.A. contributed to its development. P.K.S. and L.Y.E.A. reported no platform-development role. No other competing interests were reported.

Abstract

Background

Diagnostic benchmarks can reward recognition of a disease family without requiring the specificity of the final diagnosis. We evaluated the separation between broad diagnostic concordance and strict diagnostic correctness in a retrieval augmented language model.

Methods

This retrospective benchmark evaluated MediDx AI, which couples Claude Sonnet 4.6 with retrieval from seven biomedical sources without fine-tuning, on presentation-only text from 100 consecutive published Massachusetts General Hospital Case Records, issued from 26 October 2023 to 11 June 2026. Published expert discussions, final diagnoses, source tables and images were withheld. Five ranked suggestions were adjudicated under broad concordance and strict correct, partial or incorrect scoring. Case-level proportions and Wilson 95% confidence intervals were calculated. A separate evaluation used 1,273 medical multiple-choice questions, counting unanswered questions as incorrect.

Results

Broad top-ranked concordance was 71% (95% confidence interval 61.5–79.0%), whereas 53% (43.3–62.5%) of top-ranked diagnoses were strictly correct; 19% were partial and 28% incorrect. Within five suggestions, broad concordance was 88% and strict correctness was 63%. The 18 broadly concordant but partial top-ranked diagnoses accounted for the difference between the two top-ranked outcomes. Failures included missed causative organisms, rare neoplasms and structural conditions; partial responses often omitted an underlying mechanism or a component of a multicomponent diagnosis. Medical question-answering accuracy was 86.5% across all submitted questions, compared with 92.0% among parseable responses; response coverage was 94.0%.

Conclusions

Broad diagnostic concordance and strict diagnostic correctness describe different aspects of performance. Reporting both, alongside ranked inclusion and response completion, exposes information concealed by aggregate accuracy. These results do not establish the incremental benefit of retrieval or clinical effectiveness, which require controlled evaluation.

Background

Published evaluations have established that language models can generate plausible differential diagnoses for complex clinical narratives [1, 2]. Their estimates nevertheless depend on the case

population, information supplied, permitted differential-list length and definition of correctness. Barile and colleagues distinguished correct diagnoses from partial and incorrect responses in pediatric cases [3], underscoring the need to state what diagnostic agreement represents. These studies provide a foundation for evaluation rather than interchangeable performance standards.

A diagnostic suggestion can identify the right disease family and still omit the organism, subtype or mechanism needed to explain a patient's illness. This distinction matters for language-model evaluation: a system can appear accurate under a permissive definition while remaining incomplete under a more specific one. Diagnostic error is an established concern in patient safety [4, 5], but an agreement score on published cases is not itself a measure of avoided error in clinical practice.

Retrieval-augmented generation supplies external information at inference and has been investigated in biomedical reasoning [6, 7]. Its diagnostic contribution must be distinguished from that of the underlying model, prompting and reasoning configuration. We therefore evaluated a defined combined system rather than attributing its performance to an individual component. MediDx AI couples Claude Sonnet 4.6 with retrieval from seven biomedical sources and produces five ranked diagnostic suggestions. Using 100 consecutive published Case Records, we quantified broad concordance, stricter diagnostic correctness and the errors concealed by an aggregate agreement score. Medical question answering and published discussant diagnoses were assessed separately to preserve the distinction between tasks and information conditions. MediDx AI is intended to assist licensed clinicians by suggesting a ranked differential diagnosis for further evaluation; it is not intended for patient use or autonomous diagnosis.

Methods

Study design and source material

This retrospective, single-arm benchmark evaluated 100 consecutive Case Records of the Massachusetts General Hospital published in the New England Journal of Medicine from Case 33-2023 through Case 16-2026, corresponding to issue publication dates of 26 October 2023 through 11 June 2026 (volumes 389–394). Cases were included without exclusion based on specialty or diagnostic complexity. Full texts were accessed through journal subscriptions. Three independent reviewers confirmed the reference diagnoses extracted from each article's Pathological Diagnosis or Final Diagnosis section. The study was reported with reference to TRIPOD-LLM guidance [8, 9]. It used published educational narratives without direct patient recruitment or acquisition of new clinical records. All case texts were in English and originated from a single US academic medical center. A completed TRIPOD-LLM checklist is provided as an additional file. No study protocol was published, the study was not registered, and patients and the public were not involved in its design, conduct or reporting.

The case-level workbook contained 100 unique combinations of publication year and case identifier: eight cases from 2023 (Cases 33–40), 40 from 2024 (Cases 1–40), 36 from 2025 (Cases 1–36) and 16 from 2026 (Cases 1–16). Case numbering was uninterrupted within each included annual segment. The

earliest included record was Case 33-2023, “An 86-Year-Old Man with Shortness of Breath” (doi:10.1056/NEJMcpc2300908), and the latest was Case 16-2026, “A 14-Year-Old Girl with Hypertension” (doi:10.1056/NEJMcpc2517866). Dates refer to the journal issue, rather than the preceding online publication date. Publication year and case identifier were used together because case numbering restarts annually.

The workbook did not contain a specialty variable. For descriptive case-mix reporting, a post hoc classification assigned each confirmed final diagnosis to one primary clinical specialty domain. Infection-defined diagnoses were generally assigned to infectious diseases, neoplastic and hematologic disorders to hematology and oncology, and systemic autoimmune diseases to rheumatology. Organ-specific endocrine, neurologic and other disorders were assigned to the corresponding domain; multisystem diagnoses required judgment about the dominant diagnostic entity. This classification was separate from diagnostic outcome scoring, was not an eligibility criterion and was not used for specialty-specific performance comparisons. It does not imply that each case belongs exclusively to one specialty in clinical practice.

System and inference settings

MediDx AI is a web-based platform with a Python FastAPI backend and React 18 frontend. The case evaluation used Claude Sonnet 4.6 through the Anthropic Messages API and Python software development kit, version 0.97.0. The API model identifier was claude-sonnet-4-6. Anthropic reports a training data cutoff of January 2026 for this model; 84 of the 100 cases were published from 2023 to 2025, before that cutoff, and the 16 cases from 2026 were published in or after the cutoff month. The investigators did not train, fine-tune or change the weights of the foundation model and did not have access to the developer's pretraining procedures. Each case was submitted as an independent request without conversational history from other cases. The 100 cases were run between August and October 2026 on the final platform configuration, with one complete run analyzed per case. All inference used the Anthropic API without local computation; the case evaluation comprised 100 inference requests and the MedQA evaluation 1,273.

Extended thinking was enabled for the case evaluation, with an 8,000-token thinking budget and a maximum response length of 16,000 tokens. The reported default temperature was 1.0; no alternative temperature was specified. The model returned separate thinking and final-response content [10]. Only the final structured diagnostic response was parsed and scored. Internal thinking content was not used in outcome classification.

Retrieval and structured prompting

Before model inference, seven sources were queried asynchronously using the chief complaint as the search term: PubMed/MEDLINE through NCBI Entrez, with up to three retrieved citations per query [11];

SNOMED CT US Edition through the UMLS REST API [12]; the UMLS Metathesaurus [13]; RxNorm through the National Library of Medicine REST API [14]; ClinVar through NCBI Entrez [15]; Online Mendelian Inheritance in Man through its API [16]; and the Human Phenotype Ontology through the Jackson Laboratory API [17]. Returned material was processed by a knowledge-injection module and incorporated into the model prompt (Fig. 1).

The system prompt required five ranked differential diagnoses, ICD-11 codes, model-generated confidence estimates, clinical reasoning, proposed actions, urgency classification and guideline citations with publication years, returned in a predefined JSON schema. The user prompt supplied age, sex, chief complaint and vital signs, followed by the retained clinical narrative, and instructed the model to consider the presentation as a whole and seek unifying patterns. A structured five-step reasoning instruction preceded the differential. No urgency label, target diagnosis or case-specific diagnostic hint was supplied. The standard outputs carried a disclaimer that the platform supports clinical judgment and does not provide an autonomous diagnosis. No MGH Case Record from the study period was used during prompt development.

Clinical input protocol

The original presentation prose was copied verbatim into the corresponding clinical input fields without paraphrasing or summarization. The reason for presentation or admission was entered as the chief complaint. Available history, medications, allergies, examination findings and initial investigations were entered into the relevant fields. Expert diagnostic discussions, their differential reasoning, confirmed final diagnoses, source tables and images were excluded. Information contained in the retained prose was preserved. Presentation-only therefore denotes a text-only narrative protocol, not an assumption that the submitted text represented every item available at the first clinical encounter.

Diagnostic scoring

Each of the five ranked diagnoses was adjudicated against the published reference diagnosis under two systems. The primary system classified a suggestion as broadly concordant if it named the final diagnosis, an equivalent term, the same disease entity at broader or narrower specificity, or a major confirmed component of a multicomponent diagnosis. Other suggestions were discordant. Top-k concordance was the proportion of cases containing at least one concordant suggestion within the first k positions.

The secondary system followed the three-category framework of Barile and colleagues [3]. A correct suggestion matched the reference diagnosis exactly or was clinically equivalent with comparable specificity. A partial suggestion identified a relevant organ system or disease category but lacked the required specificity, such as lymphoma instead of diffuse large B-cell lymphoma. An incorrect suggestion did not correspond to the reference diagnosis. An uninterpretable differential entry was

scored as discordant under broad scoring and incorrect under strict scoring. Strict top-k correctness required at least one strictly correct suggestion within the first k positions.

Three independent reviewers (one physician and two biomedical scientists) each classified all 100 cases using the stated criteria, and disagreements were resolved by consensus. Reviewers were not blinded to the identity of the system, and all three had contributed to its development. We assigned failure categories to cases with discordant top-ranked diagnoses and classified the specificity gaps among partial responses. The categories are reported descriptively. Model-generated confidence, urgency, clinical reasoning, red flags, recommended actions, specialist referral and clinical summaries were not evaluated.

Published discussant analysis

In an exploratory secondary analysis, we extracted the primary diagnosis stated in each published discussant's conclusion after review of the complete case narrative, including imaging and laboratory results. Both scoring systems were applied, with reviewer disagreements resolved by consensus. The analysis did not blind discussants to information or reproduce a clinical encounter. Discussants were chosen for diagnostic expertise and had substantially more information than the model. Results were summarized descriptively without hypothesis testing.

Medical question answering

MediDx AI was separately evaluated on the four-option MedQA USMLE test set of 1,273 questions on 15 August 2026. Each question was submitted with a structured prompt requesting one answer choice in JSON. This evaluation used Claude Sonnet 4.6 with five retrieval sources (PubMed/MEDLINE, UMLS, SNOMED CT, RxNorm and ClinVar) and without extended thinking. No investigator-led model training or fine-tuning was performed for either evaluation.

Questions without a parseable answer, including API errors, were classified as unanswered. Full-sample accuracy used all 1,273 questions as the denominator and counted unanswered items as incorrect. Answered-only accuracy used the 1,197 parseable responses as the denominator. Response coverage was the proportion of submitted questions yielding a parseable answer. The available aggregate counts do not separate the causes of non-response.

Statistical analysis

Performance was summarized as case or question counts and proportions. Case-level 95% confidence intervals were calculated using the Wilson score method [18]. The diagnostic analysis unit was the case; ranked suggestions from a case were not treated as independent observations. Case records were joined across the Clinician Comparison, Barile Scoring and Head-to-Head Barile worksheets using publication year and case identifier. The three worksheets contained the same 100 unique records.

Broad concordance, strict top-k correctness, rank distributions, published-discussant classifications and paired scoring cross-tabulations were recalculated from those records. Percentage-point differences and gains from expanding the ranked list are descriptive contrasts, not paired significance tests. No formal comparisons with published benchmarks or discussants were performed. No inter-rater agreement statistic or repeated-inference analysis was reported.

Results

Case mix

A post hoc classification of the confirmed diagnoses into mutually exclusive primary clinical specialty domains yielded infectious diseases (32 cases), hematology and oncology (21), rheumatology (10), endocrinology and metabolism (9), neurology (5), gastroenterology and hepatology (5), cardiology (4), nephrology (3), pulmonology (3), psychiatry (2), clinical immunology (2), toxicology (2), dermatology (1) and ophthalmology (1). Each of the 100 cases was counted once. These diagnosis-derived categories describe case mix rather than the clinical services involved in the original patients' care.

Concordance and diagnostic specificity

Under the primary broad criterion, 71 of 100 top-ranked diagnoses were concordant with the published final diagnosis (71.0%; 95% confidence interval, 61.5–79.0%). Concordance increased to 83.0% within three positions (74.5–89.1%) and 88.0% within five (80.2–93.0%; Table 1). Broad concordance accepted an equivalent diagnosis, broader or narrower specificity, or a major confirmed component of a multicomponent diagnosis.

The stricter framework required comparable diagnostic specificity [3]. It classified 53 top-ranked diagnoses as correct (53.0%; 43.3–62.5%), 19 as partial (19.0%; 12.5–27.8%) and 28 as incorrect (28.0%; 20.1–37.5%). Strictly correct diagnoses occurred within three positions in 61 cases and within five in 63. Thus, the aggregate difference between broad concordance and strict correctness was 18 percentage points at rank one and 25 points within five suggestions (Fig. 2). The case-level cross-tabulation showed that all 53 strictly correct top-ranked diagnoses were also broadly concordant. Of the 71 broadly concordant top-ranked diagnoses, 18 were partial under strict scoring; the other 53 were strictly correct. Among the 29 broadly discordant top-ranked diagnoses, 28 were strictly incorrect and one was partial. Thus, partial and broadly concordant are not interchangeable categories.

The exceptional partial but broadly discordant response occurred in Case 37-2023: the confirmed diagnosis was Campylobacter fetus bacteremia and osteomyelitis, whereas the top-ranked suggestion was a prosthetic joint infection. The workbook classified this suggestion as partial under strict scoring but did not credit broad concordance at any of the five positions. At the five-suggestion level, 63 cases met both broad concordance and strict correctness, 25 met broad concordance without a strictly correct suggestion, and 12 met neither criterion.

The first broadly concordant suggestion appeared at rank one in 71 cases, rank two in eight, rank three in four, rank four in four and rank five in one. Twelve cases had no broadly concordant suggestion among the five generated diagnoses. Expanding the list from one to five therefore added 17 cases under broad scoring and ten cases under strict scoring. List inclusion does not establish that a clinician would select or act on the appropriate suggestion.

Broad top-ranked discordance occurred in 29 cases (29%; 95% CI, 21.0–38.5%), and 12 cases had no broad concordance within five suggestions (12%; 95% CI, 7.0–19.8%). Under strict scoring, 47 top-ranked responses were partial or incorrect (47%; 95% CI, 37.5–56.7%), distinct from the 28 responses classified as strictly incorrect.

Table 1
Diagnostic outcomes under broad and strict scoring

Outcome	Cases, n/N (%)	95% CI (%)
Broad top-1 concordance	71/100 (71.0)	61.5–79.0
Broad top-3 concordance	83/100 (83.0)	74.5–89.1
Broad top-5 concordance	88/100 (88.0)	80.2–93.0
Broad top-1 discordant	29/100 (29.0)	21.0–38.5
No broad concordance within top 5	12/100 (12.0)	7.0–19.8
Strict top-1 correct	53/100 (53.0)	43.3–62.5
Strict top-1 partial	19/100 (19.0)	12.5–27.8
Strict top-1 incorrect	28/100 (28.0)	20.1–37.5
Strict top-1 partial or incorrect	47/100 (47.0)	37.5–56.7
Strict top-3 correct	61/100 (61.0)	51.2–70.0
Strict top-5 correct	63/100 (63.0)	53.2–71.8

Table footnotes. N = 100 cases; n = cases meeting the specified outcome; CI = confidence interval. Top-k denotes inclusion within the first k ranked suggestions. Broad concordance credits diagnostically related answers under the definition in Methods; strict correctness requires an exact or clinically equivalent diagnosis with comparable specificity. All 95% CIs use the Wilson score method. The scoring systems are not assumed to be nested: 18 broadly concordant top-ranked diagnoses were partial under strict scoring, and one broadly discordant diagnosis was partial. Within five suggestions, 25 cases met broad concordance without a strictly correct suggestion.

Broad top-1 discordance and absence of broad concordance within five suggestions are complements of the corresponding concordance outcomes. The combined strict partial-or-incorrect outcome includes

all top-ranked responses that were not strictly correct; it must not be confused with the strictly incorrect category alone.

Where diagnostic specificity was lost

The 12 cases without broad concordance within five suggestions fell into five descriptive categories (Fig. 3a). Four lacked the specific infectious organism, three lacked a rare neoplasm, two lacked a mechanical, foreign-body or structural explanation, two substituted a related mimic, and one lacked the primary immunodeficiency. Examples included failure to generate disseminated strongyloidiasis, reninoma, wooden-toothpick duodenal perforation and common variable immunodeficiency. Related mimics included atrio-esophageal fistula instead of esophageal–pericardial fistula and NK/T-cell lymphoma instead of diffuse large B-cell lymphoma. Anatomical or histological proximity did not make these diagnoses equivalent.

Among the remaining 17 cases with a discordant first suggestion but broad concordance at a lower rank, the top-ranked diagnosis was a related mimic in seven, a complication or concurrent condition in four, a neoplastic versus infectious or inflammatory alternative in three, and an incorrect causative organism in three. The case-level classifications are reported in Supplementary Table 2.

The 19 partial first suggestions under strict scoring exposed a different form of incompleteness (Fig. 3b). Six identified only one component of a multicomponent final diagnosis, six incompletely specified the underlying cause or mechanism, four omitted or misidentified the causative organism, and three specified an incorrect neoplasm subtype or histology. These categories describe the adjudicated errors in this sample and were not tested as predictors of future failure.

Published discussants as context

Published discussant diagnoses were broadly concordant in 97 cases and strictly correct in 71; 28 were partial and one was incorrect under strict scoring (Table 2). At the first model rank, both the model and discussant were broadly concordant in 69 cases, the discussant alone in 28, the model alone in two and neither in one. When all five model suggestions were considered, both identified a broadly concordant diagnosis in 85 cases, the discussant alone in 12 and the model alone in three.

Under strict scoring, the model and discussant received the same category in 62 cases; the discussant received a higher category in 32 and the model in six. These are descriptive comparisons with published diagnostic conclusions, not a blinded or controlled clinician benchmark. Discussants were selected for expertise and evaluated substantially more information, including imaging and laboratory findings. Their conclusions were produced for educational publication rather than under the model's input protocol.

Case-level examples illustrate the distinction between broad concordance and diagnostic specificity. In Case 8-2026, the reference diagnosis was fungal endocarditis with an aortic-root abscess and contained rupture. The model suggested bacterial pericarditis with tamponade, likely secondary to infective

endocarditis; the discussant identified hemopericardium in the setting of aortic-root or ascending aortic rupture. Only the model response was credited under broad scoring, but both were partial under strict scoring. Similarly, in Case 11-2025, the model suggested non-Hodgkin lymphoma rather than the reference diagnosis of extranodal marginal-zone lymphoma, whereas the discussant suggested a cancer-related pleural effusion; both were strictly partial. In Case 16-2025, granulomatosis with polyangiitis appeared fourth in the model differential, while the top-ranked suggestion was nasopharyngeal carcinoma and the discussant conclusion was autoimmune disease; both top-ranked conclusions were strictly incorrect. These examples reflect the recorded scoring decisions under unequal information conditions.

Table 2
Published discussant context under unequal information access

Outcome	MediDx AI	Published discussant
Broad top-1 concordance	71/100	97/100
Strict top-1 correct	53/100	71/100
Strict top-1 partial	19/100	28/100
Strict top-1 incorrect	28/100	1/100
Input information	Presentation prose only	Complete published case information
Evaluation setting	One model run per case	Educational diagnostic discussion

Table footnotes. Both columns summarize the same 100 cases. Top-1 denotes the first model suggestion or the primary published discussant conclusion. Broad and strict outcomes follow the definitions in Methods. Discussant diagnoses were extracted retrospectively from published educational discussions that used more information than was supplied to MediDx AI. These results are descriptive and do not represent a blinded, controlled or information-matched assessment of physician versus model performance. No statistical test of superiority was performed.

Knowledge accuracy and completion

The separate MedQA evaluation yielded 1,101 correct answers among 1,273 questions, for full-sample accuracy of 86.5%. Of 1,197 parseable responses, 96 were incorrect; 76 questions produced no parseable answer. Accuracy conditional on a parseable response was 92.0% (1,101/1,197), with response coverage of 94.0% (1,197/1,273; Fig. 4). The full-sample estimate incorporates failures to return an answer, whereas the conditional estimate excludes them.

MedQA used five retrieval sources without extended thinking, whereas the case benchmark used seven sources with extended thinking. The two evaluations therefore assess different tasks and configurations. Their results cannot identify whether retrieval breadth or extended thinking improved performance.

Discussion

The principal finding is the separation between identifying a diagnostically related answer and reproducing the specificity of the final diagnosis. On the same 100 case presentations, broad top-ranked concordance was 71%, whereas strict correctness was 53%. At five suggestions, the separation widened to 88% versus 63%. Case-level reconciliation showed that the top-ranked difference comprised 18 broadly concordant diagnoses classified as partial under strict scoring. One further diagnosis was partial but broadly discordant, demonstrating why the scoring systems should be cross-tabulated rather than treated as equivalent binary classifications. Reporting the larger percentage alone would obscure a substantial difference in what was credited as a successful response. Conversely, reporting strict correctness alone would discard information about diagnostically related suggestions that might still be useful for further investigation. The two outcomes answer different questions and should be named accordingly.

The error analysis makes that distinction concrete. A diagnosis can capture a syndrome while leaving its cause unspecified, describe one component while missing another, or name a related neoplasm without the correct histology. These are not merely stylistic differences in terminology. They locate the boundary between recognition and diagnostic resolution. Whether a partial suggestion helps or distracts a clinician cannot be determined from the suggestion alone; it requires evaluation of subsequent decisions, testing and harms. The present categories can guide such studies, but their small counts do not justify claims about systematic disease-specific weaknesses.

A second distinction concerns retrieval. External biomedical information is an established component of language-model systems [6, 7], and the present benchmark evaluates a particular implementation of that approach. It does not demonstrate that seven sources outperform fewer sources, that retrieval improves on the underlying model, or that extended thinking increases correctness. A direct comparison should hold the model, cases, output schema and scoring constant while varying retrieval and reasoning. Retrieved material should also be audited for relevance, provenance and inadvertent disclosure of case-specific answers. Without those controls, system performance cannot be converted into a mechanistic claim about why the system performed as it did.

Presentation-only input addresses one route to direct answer disclosure: the submitted text omitted the published diagnostic discussion and final diagnosis. It does not establish that the case or its answer was absent from foundation-model pretraining or inference-time retrieval. Subscription access to the full article likewise does not establish that identifying details were unavailable elsewhere. Validation on unpublished cases, with retrieval sources and returned content recorded, would better separate diagnostic inference from recognition of previously available material. Replicated runs would quantify variability that a single response per case cannot reveal.

Published model studies and the discussant analysis serve different contextual roles. Prior diagnostic benchmarks differ in case populations, information conditions and adjudication [1–3]. Shared terminology such as top-1 accuracy does not make their estimates exchangeable. Similarly, the

discussant analysis characterizes agreement with an educational diagnostic conclusion under unequal information access. It does not measure physician performance under the model's conditions. Neither source of context supports a ranking of MediDx AI against clinicians or contemporary models. The estimates are therefore interpreted within the evaluated cohort and protocol.

Kanjee and colleagues assessed a longer differential list, McDuff and colleagues tested clinician assistance within a controlled study, and Barile and colleagues distinguished incomplete from correct pediatric diagnoses. These studies motivate explicit scoring and information conditions rather than cross-study claims of superiority (Table 3).

Table 3
Contextual comparison of diagnostic evaluations

Study or evaluation	Cases and inputs	Reported outcomes	Interpretive differences
MediDx AI (present study)	100 consecutive MGH cases; presentation-only text; five ranked suggestions.	Broad top-1 71%, top-5 88%; strict top-1 53%, top-5 63%.	Two scoring systems applied to identical outputs; no matched model-only arm.
Kanjee et al. [1]	70 NEJM cases from 2021–2022; case text before the discussant response; GPT-4.	Top diagnosis correct in 39%; final diagnosis included in 64% of differentials; mean list length 9.0.	Different cases, model, adjudication and list length; differential inclusion is not a fixed top-5 endpoint.
McDuff et al. [2]	302 NEJM cases; AMIE alone and 20 clinicians before and after assistance.	Top-1/top-10: AMIE 29.2%/59.1%; unassisted clinicians 15.9%/33.6%; search-assisted 24.3%/44.5%; AMIE-assisted 25.2%/51.8%.	Within-study assistance comparisons used a controlled design; they cannot serve as a matched comparator for MediDx.
Barile et al. [3]	100 pediatric cases: 60 JAMA Pediatrics and 40 NEJM; case text entered into ChatGPT-3.5.	Final diagnosis correct 17%, incomplete 11%, incorrect 72%; reference diagnosis in the differential in 36%.	Pediatric-only population; final diagnosis and differential inclusion are separate endpoints; scoring framework informed the present strict assessment.
Published MGH discussants (present study)	Same 100 case records; diagnoses extracted from published discussions with more information than supplied to MediDx.	Broad top-1 concordance 97%; strict correct 71%, partial 28%, incorrect 1%.	Educational expert conclusions under unequal information access; not a controlled physician-performance comparison.

Table footnotes. AI = artificial intelligence; AMIE = Articulate Medical Intelligence Explorer; GPT = Generative Pre-trained Transformer; MGH = Massachusetts General Hospital; NEJM = New England Journal of Medicine. Top-k denotes inclusion within the first k suggestions. Values retain each study's

own scoring criteria, denominators and list lengths. The 64% differential-inclusion result from Kanjee et al. and the 36% result from Barile et al. are not fixed top-5 estimates. No cross-study significance tests or pooled estimates were calculated, and the rows do not establish a performance ranking.

The MedQA analysis adds response completion as a separate dimension. Its 92.0% answered-only accuracy applies to the subset with a parseable response; the 86.5% full-sample estimate is the relevant summary when every submitted question must be accounted for. The difference does not show that the unanswered questions would have been answered incorrectly, but it prevents technical non-response from disappearing from the evaluation. Medical question answering also differs from generating a differential diagnosis, and the configuration change between tasks prevents a direct test of the relationship between the two capabilities.

Several limits constrain translation. The cases were selected for educational publication and do not represent the prevalence or workflow of routine care. Adjudication involved clinical judgment, with disagreements resolved by consensus; no inter-rater agreement estimate was reported. Reviewers were not blinded to the system, and all three had contributed to its development, which may have favored generous classification. The analysis used one run per case, without controlled model comparators, calibration assessment, retrieval-quality measurement, or evaluation of resource use and clinical harms. The study did not test whether clinicians use the suggestions appropriately or whether patients benefit. These limits define the next evidential steps rather than negate the descriptive benchmark: independent case validation, component ablation, repeated inference and prospective clinician-assisted evaluation.

Conclusions

The practical contribution is an evaluation profile that keeps diagnostic direction, specificity, ranked inclusion and response completion visible. For this combined system, broadly related answers were more frequent than strictly correct diagnoses, and adding suggestions recovered more cases without resolving all specificity gaps. These findings support transparent outcome definitions as a prerequisite for interpreting diagnostic language-model benchmarks. Establishing clinical value will require evidence that the suggestions improve decisions under the conditions in which the system is intended to be used.

Abbreviations

AI
artificial intelligence
API
application programming interface
CI
confidence interval
ICD
International Classification of Diseases

JSON
JavaScript Object Notation
LLM
large language model
MEDLINE
Medical Literature Analysis and Retrieval System Online
MGH
Massachusetts General Hospital
SNOMED CT
Systematized Nomenclature of Medicine Clinical Terms
UMLS
Unified Medical Language System
USMLE
United States Medical Licensing Examination
TRIPOD-LLM
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis for large language models.

Declarations

Ethics approval and consent to participate

Not applicable. The study used only previously published educational case reports, involved no patient recruitment, interaction or collection of new clinical records, and therefore did not constitute human subjects research.

Consent for publication

Not applicable. No new patient narratives or identifiable patient images are reproduced. The analysis used previously published educational cases.

Availability of data and materials

The accompanying case-level scoring workbook contains 100 records identified by publication year and MGH case number, with published final diagnoses, top-ranked model and discussant diagnoses, the rank of first broad concordance, and strict classifications at each of five differential positions. Its worksheets are Clinician Comparison, Error Cases, Head-to-Head Barile and Barile Scoring. Failure categories are listed in Supplementary Table 2. Model output files are provided in Supplementary Appendix 1. Original clinical narratives, published diagnostic discussions, tables and images are not reproduced. Full source case records are available through the New England Journal of Medicine subscription platform.

Code availability

The MediDx AI source code and deployment infrastructure are not publicly released. Requests for technical documentation and controlled access for research validation may be directed to the corresponding author for consideration through appropriate collaborations or approved studies.

Competing interests

E.K.A. developed MediDx AI, the platform evaluated in this study. P.M.D., L.A., G.G.-K. and F.A.A. contributed to its development. P.K.S. and L.Y.E.A. reported no platform-development role. No other competing interests were reported.

Funding

This study received no external funding. The authors were responsible for study design, data collection, analysis and interpretation, manuscript preparation and the decision to submit.

Authors' contributions

E.K.A. conceived the study, developed MediDx AI, took part in data curation and analysis, and drafted and revised the manuscript. P.M.D., L.A., G.G.-K. and F.A.A. contributed to platform development, validation and manuscript revision. L.A., G.G.-K. and F.A.A. performed outcome classification. P.M.D., G.G.-K. and F.A.A. provided supervision. P.K.S. and L.Y.E.A. contributed to validation and manuscript revision.

Acknowledgements

Not applicable.

Supplementary information

The accompanying materials comprise the case-level diagnostic scoring workbook, including the error-case classifications, and the model-output appendix identified in the Data availability statement. A completed TRIPOD-LLM checklist is also provided.

References

1. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330:78–80. https://doi.org/10.1001/JAMA.2023.8288.

2. McDuff D, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642:451–7. https://doi.org/10.1038/S41586-025-08869-4.
3. Barile J, et al. Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies. JAMA Pediatr. 2024;178:313–5. https://doi.org/10.1001/JAMAPEDIATRICS.2023.5750.
4. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf. 2014;23:727–31. https://doi.org/10.1136/BMJQS-2013-002627.
5. Bates DW, Singh H. Two Decades Since To Err Is Human: An Assessment Of Progress And Emerging Priorities In Patient Safety. Health Aff (Millwood). 2018;37:1736–43. https://doi.org/10.1377/HLTHAFF.2018.0738.
6. Zakka C, et al. Almanac — Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI. 2024;1. https://doi.org/10.1056/aioa2300068.
7. Lewis P, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv Neural Inf Process Syst. 2020;33:9459–74. https://arxiv.org/abs/2005.11401v4.
8. Collins GS, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385. https://doi.org/10.1136/BMJ-2023-078378.
9. Gallifant J, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31:60–9. https://doi.org/10.1038/S41591-024-03425-5.
10. Anthropic. Extended thinking. Claude Platform Docs. (2026). https://platform.claude.com/docs/en/build-with-claude/extended-thinking
11. Sayers EW, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res 53. 2025;D20–9. https://doi.org/10.1093/nar/gkae979.
12. Chang E, Mostafa J. The use of SNOMED CT, 2013–2020: A literature review. J Am Med Inform Assoc. 2021;28. https://doi.org/10.1093/jamia/ocab084.
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–70. https://doi.org/10.1093/nar/gkh061.
14. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011;18. https://doi.org/10.1136/amiajnl-2011-000116.
15. Landrum MJ, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46. https://doi.org/10.1093/nar/gkx1153.
16. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47:D1038–43. https://doi.org/10.1093/NAR/GKY1151.
17. Köhler S, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49:D1207–17. https://doi.org/10.1093/NAR/GKAA1043.
18. Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. J Am Stat Assoc. 1927;22:209–12. https://doi.org/10.1080/01621459.1927.10502953.

Figures

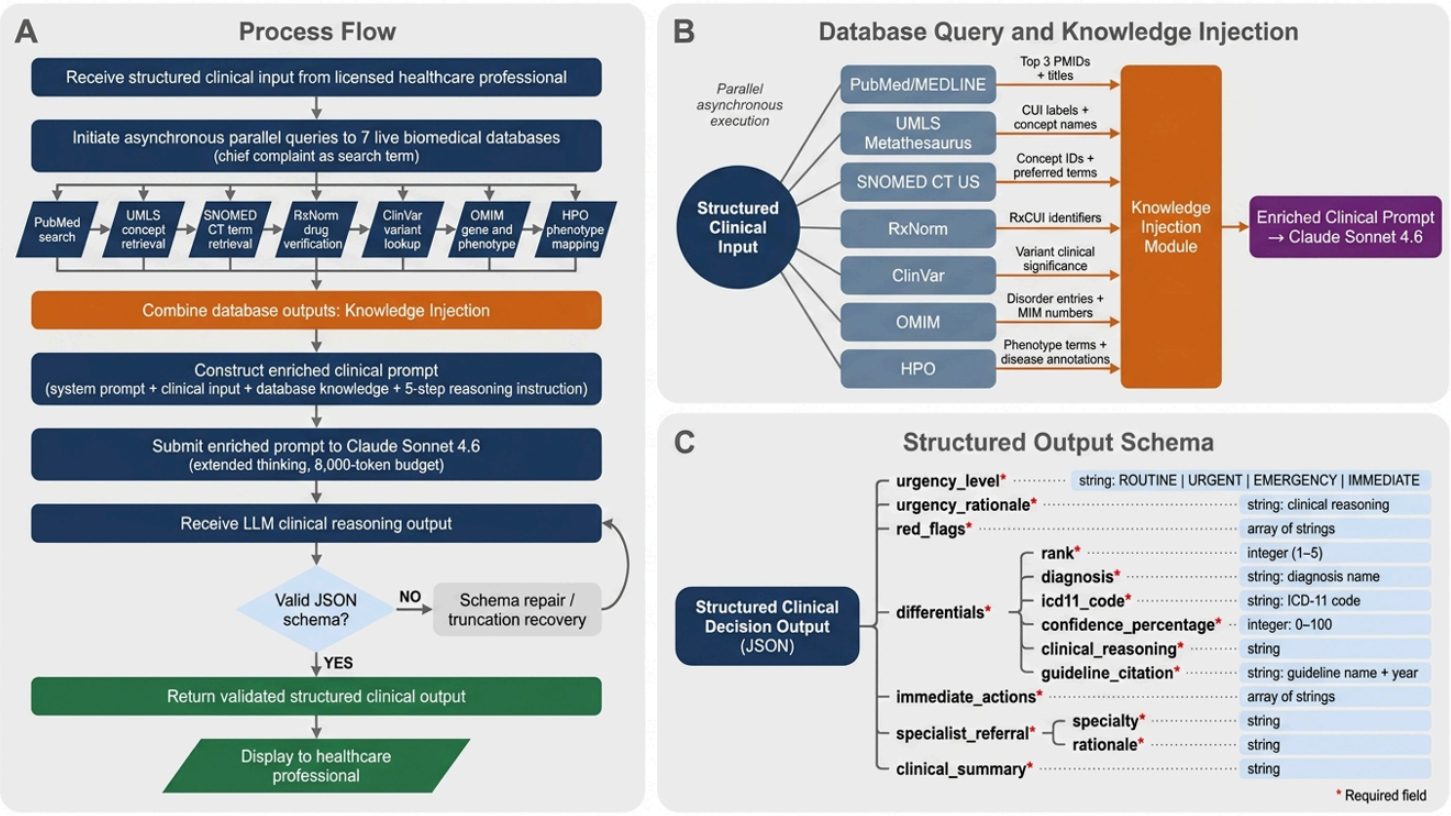


Figure 1

System configuration and inference workflow. a, Structured clinical input is followed by parallel retrieval, prompt construction, Claude Sonnet 4.6 inference with extended thinking and structured-response validation. b, Seven biomedical sources supply information to the knowledge-injection module before inference. c, The standard output schema includes ranked diagnoses and additional clinical decision-support fields. Only the ranked differential diagnoses were evaluated in the case benchmark. Schema validation is a technical check and does not establish diagnostic validity.

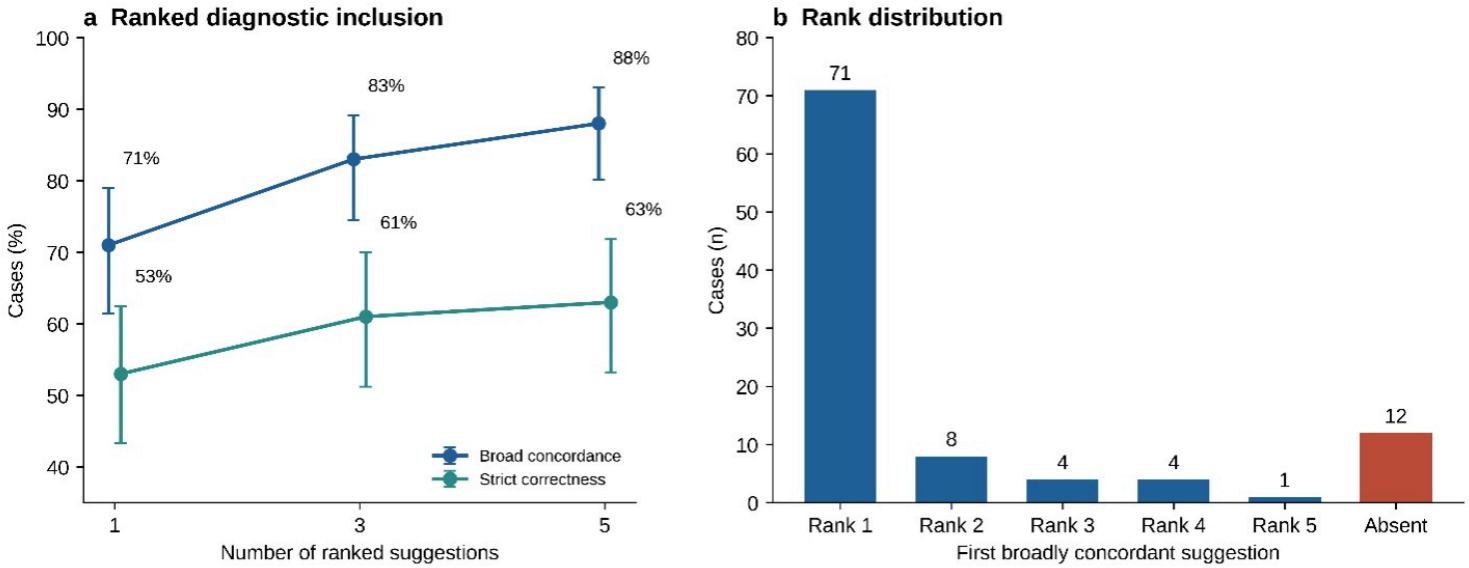


Figure 2

Ranked inclusion depends on diagnostic specificity. a, Broad concordance and strict correctness at ranks one, three and five. Points show proportions of 100 cases, with 95% Wilson confidence intervals. The lines guide the eye and do not imply continuous change or independent repeated observations. b, Rank of the first broadly concordant suggestion, including the 12 cases without concordance within five suggestions. All values derive from the aggregate case counts.

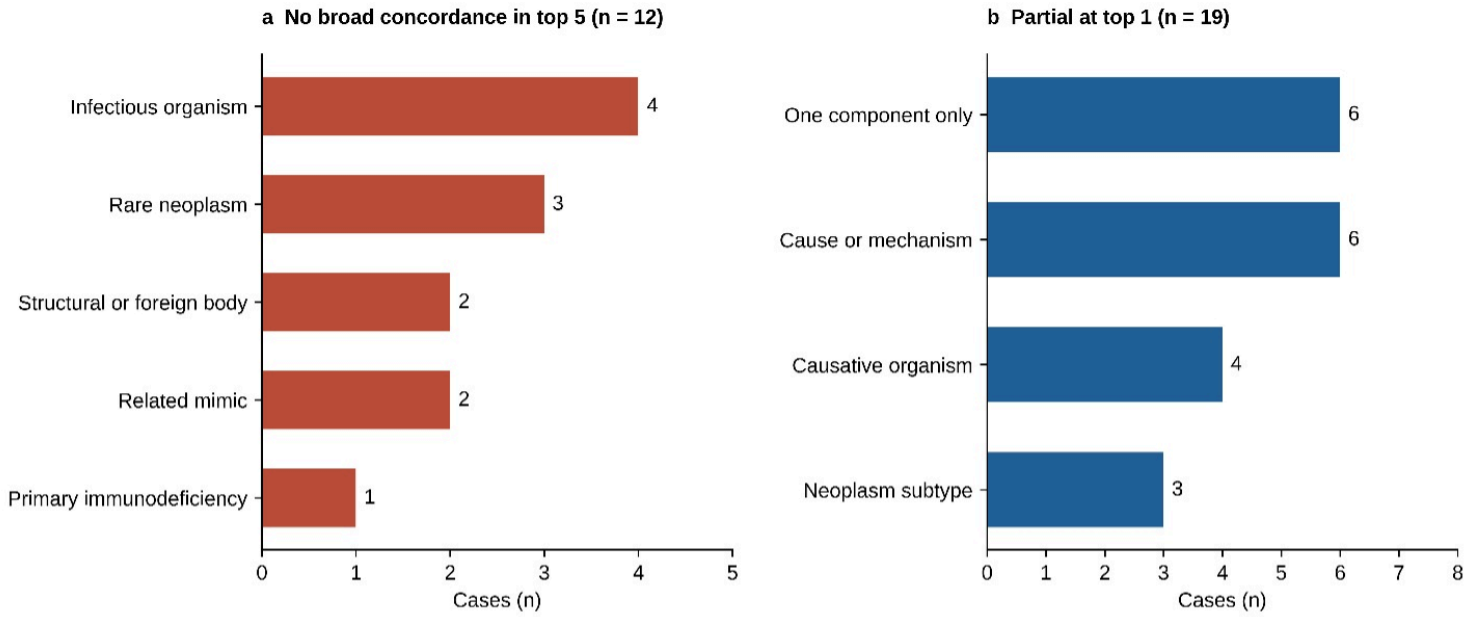


Figure 3

Diagnostic failures and specificity gaps. a, Descriptive categories for 12 cases without a broadly concordant suggestion among the first five. b, Categories for 19 top-ranked diagnoses classified as partial under strict scoring. Bars show case counts. These panels use different denominators and

describe different outcomes; they are not a single pooled error distribution. Categories were taken from the case-level error summary and were not used for inferential testing.

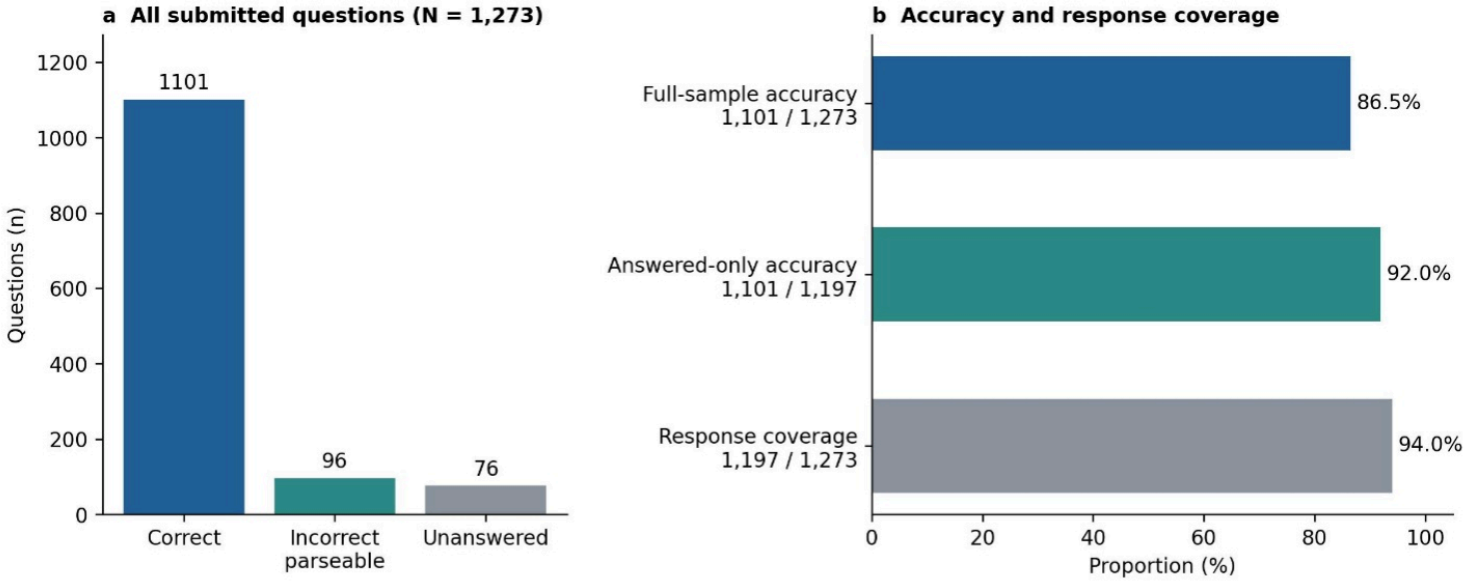


Figure 4

Medical question answering and response completion. a, Outcomes across all 1,273 MedQA questions: 1,101 correct, 96 incorrect parseable responses and 76 unanswered or unparseable responses. b, Full-sample accuracy, accuracy conditional on a parseable response and response coverage. Denominators are printed below each estimate. Answered-only accuracy excludes the 76 unanswered items; the full-sample estimate counts them as incorrect. This task used five retrieval sources without extended thinking and is separate from the diagnostic case configuration.

Supplementary Files

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

- MediDxAIMGHFinalScoringv6.xlsx
- TripodLLMChecklist.pdf
- eAppendix1.docx