AI-Assisted Clinical Decision Support for Rare Disease Diagnosis: Evidence from China's First Disease-Specific Large Language Model and Real-World Physician Adoption Data

Adele Li, MBA, David Wang, MBA, Yixuan Zhou, MSc.

MSR26

Background

Rare disease diagnosis in China faces a structural crisis. Average diagnostic delay exceeds four years; primary care physicians lack disease-specific knowledge for conditions they may encounter once in a career; and specialist capacity is concentrated in tier-one cities, so confirmation routinely requires the patient to travel.

The resulting diagnostic odyssey is not merely slow — it is a period during which treatable disease progresses untreated. Where an approved orphan therapy exists, diagnostic delay operates as a de facto access barrier.

China has built the policy scaffolding around rare disease: a national rare disease catalogue, designated diagnosis and treatment networks, and a widening set of orphan therapies reaching NRDL negotiation. The binding constraint has shifted upstream — patients cannot access what they have not been diagnosed with.

Objective

To evaluate the clinical value and health outcomes impact of AI-assisted diagnostic support tools — specifically disease-specific large language models — for rare disease physicians in China, drawing on real-world adoption data and a documented clinical deployment case.

Methods

Two data sources were synthesised. First, the 2025 Digital Life Physician (DLP) cross-sectional survey (n=4,500 Chinese physicians) provided real-world adoption metrics for AI tools in rare disease clinical practice, including usage rates, functional preferences and perceived clinical value.

Second, a documented deployment case of a rare disease-specific LLM built on a national rare disease knowledge base (Xiehe·Taichu model, Peking Union Medical College Hospital) was analysed for clinical architecture, hallucination mitigation mechanisms, and reported outcome metrics including diagnostic timeline compression, differential diagnosis accuracy and hospital integration pathways across 200+ partner institutions.

Adoption data were analysed descriptively by specialty and city tier. Deployment evidence was assessed against four dimensions — physician demand, model architecture and safety controls, reported diagnostic outcomes, and integration pathway — then interpreted within an HEOR value framework.

Figure 1. Evidence base — two synthesised data sources

Population-level adoption data triangulated against a documented single-model deployment

SOURCE 1 · POPULATION 2025 DLP cross-sectional survey	SOURCE 2 · DEPLOYMENT Disease-specific rare disease LLM
BASE n=4,500 Chinese physicians; AI users n=4,246	CASE Xiehe·Taichu — built on China's national rare disease knowledge base and population-specific genetic data
MEASURES AI usage rates by specialty and city tier, functional preferences, perceived clinical value	MEASURES Clinical architecture, hallucination mitigation, diagnostic timeline compression, differential accuracy, hospital integration
ROLE IN ANALYSIS Establishes whether physician demand for diagnostic AI exists at scale, and where it concentrates	ROLE IN ANALYSIS Tests whether a verticalised model converts that demand into measurable diagnostic outcomes

Synthesis Adoption evidence answers *who will use it*; the deployment case answers *what it changes*. Read together they define an HEOR value category — AI-enabled diagnostic equity — that neither source establishes alone.

Analytical framework — four evaluation dimensions

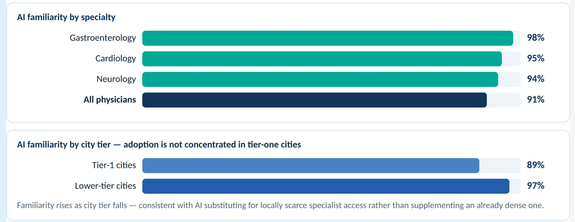
DIMENSION 01 Physician demand	DIMENSION 02 Architecture & safety	DIMENSION 03 Diagnostic outcomes	DIMENSION 04 Integration & scale
Is adoption real, and where does it concentrate?	What design features make the output clinically usable?	Does time to diagnosis and accuracy actually change?	Can it reach the settings where the gap is widest?

Results

Among AI-using physicians (n=4,246), rare disease and cognition-intensive specialties demonstrated the highest AI adoption — gastroenterology 98%, cardiology 95%, neurology 94%. Familiarity was higher in lower-tier cities (up to 97%) than in tier-one cities (89%), indicating that demand is greatest where specialist access is scarcest.

Figure 2. AI adoption concentrates in cognition-intensive specialties

2025 DLP survey; AI-using physicians n=4,246 of n=4,500 surveyed



The inverse relationship between city tier and AI familiarity is the finding that matters for equity. It suggests physicians are not adopting AI as a convenience layer on top of adequate specialist support, but as a substitute for specialist access they do not have locally.

Figure 3. Three architectural breakthroughs and the hallucination controls that make them usable

Design features distinguishing a disease-specific model from general-purpose AI

BREAKTHROUGH 01 Explainable reasoning chain	BREAKTHROUGH 02 Extreme few-shot cold start	BREAKTHROUGH 03 Traceable knowledge repository
Symptom → examination → differential diagnosis, with key decision points made visible. Reasoning is inspectable at each step rather than delivered as an opaque output, so clinicians can audit the path, not just the conclusion.	Small-sample reference standards bootstrap a skeleton knowledge base — the constraint that ordinarily blocks rare disease AI, where case volume per condition is inherently tiny. Federated learning across hospitals shares parameters without moving patient data.	Multi-dimensional sourcing from central authority guidance, guidelines, genomic data and real-time updates, built on China's national rare disease knowledge base and population-specific genetic testing data.

Hallucination governance — rate maintained below 5%

Chain-of-thought review	Confidence thresholding	Model cross-verification
Transparent reasoning paths exposed for medical validation, including AI-assisted checking	Outputs below threshold automatically trigger secondary validation	Suspicious output re-checked against authoritative literature by model ensemble

Disease-specific LLMs incorporating explainable symptom-to-differential-diagnosis reasoning chains, multi-dimensional traceable verification and extreme few-shot training on curated rare disease case repositories addressed the constraint that ordinarily defeats rare disease AI: too few cases per condition to train on.

Figure 4. Compression of the rare disease diagnostic pathway

Reported outcome in documented deployment

CONVENTIONAL PATHWAY	AI-ASSISTED PATHWAY
> 4 years average time to confirmed diagnosis	**< 4 weeks** ~93% accuracy across top-10 suggestions in simulated testing
1 Symptom onset; presentation to primary care	1 Multi-turn triage dialogue captures clinical status in <3 minutes
2 Rare disease not in the differential — misattribution	2 Pattern recognition raises rare disease into the differential early
3 Serial referrals across specialties and hospital tiers	3 Top-N differential with visible reasoning chain at point of care
4 Travel to a tier-one specialist centre	4 Targeted referral and confirmatory testing, no serial detour
5 Genetic testing; confirmation — often years after onset	The gain is not faster confirmation — it is **earlier suspicion**. Time is recovered at the point where the conventional pathway fails, not at the end of it.

Average diagnostic timelines fell from over four years to under four weeks in documented deployments, with approximately 93% accuracy across top-10 diagnostic suggestions in simulated testing.

The model is deployed through both a patient-facing mobile assistant and a physician-facing decision support layer inside hospital systems, so a rare disease can enter the differential without a clinician having first suspected it.

Figure 5. Patient-facing entry point to the disease-specific model

Deployed Chinese-language interface, currently in invitation-based clinical testing

Dual-audience design
The same model serves a patient-facing assistant and a physician-facing decision support layer. Patients reach it through a consumer mobile entry; clinicians reach it inside the hospital information system.

Conversational triage as the front door
Multi-turn dialogue collects initial clinical status in under three minutes and generates preliminary suggestions — positioned ahead of the referral chain rather than inside it.

Onward care navigation
Beyond consultation, the assistant plans the subsequent care pathway and supports booking into the rare disease multidisciplinary clinic, closing the gap between suspicion and specialist contact.

Controlled release
Access is invitation-gated during the testing phase, with capped daily enrolment — a deliberate constraint on scale while diagnostic reliability is validated.

Patient-initiated entry matters for equity: it does not require a physician to have already suspected a rare disease, which is precisely the step at which the conventional pathway fails.

Platform integration via SaaS or on-premises deployment across 200+ hospitals enabled patient triage, clinical decision support and rare disease knowledge retrieval at scale, with genetic interpretation and automated record generation in development.

Figure 6. Integration pathway and clinical application scenarios

SaaS or on-premises deployment into existing hospital information systems

200+	<3 min	<5%	~93%
partner hospitals via SaaS or on-premises deployment	multi-turn triage dialogue to initial suggestion	hallucination rate under governance controls	top-10 differential accuracy, simulated testing

Patient triage & classification	Physician decision support	Genetic interpretation	Automated record generation
Front-line real-time triage identifies high-risk patients and reduces unnecessary referrals. DEPLOYED	Integrates into outpatient and inpatient workflow; locates key symptoms and returns Top-N diagnostic suggestions. DEPLOYED	Converts WES/WGS output into HPO phenotype terms to support genetic screening. IN DEVELOPMENT	Structures patient Q&A into EMR entries, improving documentation completeness. PLANNED

Hallucination rates were maintained below 5% through cross-referencing with authoritative medical literature, chain-of-thought review and confidence scoring — controls that make an explainable output auditable rather than merely visible.

Figure 7. Disease-specific model versus general-purpose LLM

Why verticalisation, not model scale, is the determining variable in rare disease

Dimension	General-purpose LLM	Disease-specific rare disease LLM
Knowledge depth	Broad general coverage; medical content shallow and not rare-disease weighted	Trained on a national rare disease knowledge base with population-specific genetic data
Reasoning	Answer returned without an inspectable clinical reasoning path	Symptom → examination → differential chain with visible decision points
Hallucination control	Variable; requires manual review to detect plausible but incorrect content	<5% via traceable sourcing, confidence thresholds and cross-verification
Data sparsity	Performance degrades where per-condition case volume is small	Extreme few-shot cold start designed for exactly this constraint
Governance	Source provenance uncontrolled; weak fit to clinical compliance requirements	Traceable to authoritative guidance; federated learning preserves data privacy

The comparison is not a matter of model scale. In a domain defined by data sparsity, dispersed authoritative sources and low tolerance for unverifiable output, verticalisation is the determining variable — which is why general-purpose AI adoption, however high, does not by itself close the diagnostic gap.

Read together, the adoption data and the deployment case define a value category that neither establishes alone: physician demand exists at scale, and a verticalised model converts it into measurable diagnostic outcomes in the settings where the gap is widest.

Figure 8. AI-enabled diagnostic equity as an HEOR value category

Where value accrues, and why conventional evaluation models do not capture it

VALUE DRIVER	MECHANISM	EVALUATION IMPLICATION
DRIVER 01 **Geographic reach**	Diagnostic capability delivered to lower-tier and rural settings where specialist density is lowest — the settings where survey familiarity is in fact highest (97%).	**Access equity** Value accrues outside the tier-one centres where trials are run
DRIVER 02 **Time-to-diagnosis**	Compression from over four years to under four weeks shifts treatment initiation earlier in the disease course, before irreversible progression accrues.	**Avoided burden** Diagnostic odyssey cost is rarely modelled as a comparator
DRIVER 03 **Reasoning standardisation**	A common differential pathway applied across hospital tiers reduces variance in whether a rare disease is considered at all.	**Quality floor** Raises the worst case, not the average
DRIVER 04 **Real-world data yield**	Anonymised interaction data aggregates into disease classification and phenotype profiling — an evidence asset for precision medicine and orphan drug development.	**Evidence generation** Downstream value not attributable to any single diagnosis

>4 yrs → <4 wks TIME TO DIAGNOSIS — Where a diagnosis is the gate to an orphan therapy, diagnostic delay functions as an access barrier. Payers and policymakers should treat AI-assisted diagnostic efficiency as an input to rare disease access and reimbursement evaluation, not as an operational nicety.

None of these four drivers is captured by evaluation models that begin at the point of confirmed diagnosis.

Conclusions

Disease-specific LLMs represent a structurally superior approach to rare disease physician support compared with general-purpose AI, demonstrating clinically meaningful reductions in diagnostic delay with compliance-ready traceability.

For HEOR frameworks these tools introduce a new evidence category — AI-enabled diagnostic equity — where value accrues from geographic reach, time-to-diagnosis compression and standardisation of clinical reasoning across tiered hospital systems.

Payers and policymakers should incorporate AI-assisted diagnostic efficiency into rare disease access and reimbursement evaluation models, recognising that for a rare disease the diagnosis itself is the gate to therapy.

Reference

1. Li A, Wang D, Zhou Y. Evolution of AI Value in Chinese Clinical Practice: 2025 Survey and Case Series. 2025.
2. 2025 Digital Life Physician (DLP) Survey. Base: n=4,500 Chinese physicians; AI users n=4,246 (question G2.2).
3. Xiehe·Taichu Rare Disease Large Model, Peking Union Medical College Hospital: model documentation and reported deployment outcomes, 2025.
4. 2025 DLP AI Case Study Report, Case 10: Rare Disease Diagnosis. December 2025.

Acknowledgement

The authors thank the Digital Life Physician research team and collaborating fieldwork partners for survey design, recruitment and data collection across China. This study was conducted in collaboration with JKT (健康通). The authors report no affiliation with the developers of the deployment case described.