Medical and Health Care Applications, Vol. 211, 2026, 31–41
https://doi.org/10.54941/ahfe1007466

When One in a Million Matters: Developing Metrics for Human-AI Collaboration in Rare Disease Diagnosis

Anna Biłyk

SGH Warsaw School of Economics, Warsaw, Poland

ABSTRACT

Rare diseases affect approximately 300 million people worldwide, yet physicians rarely encounter individual conditions, creating significant risk of delayed or missed diagnosis. AI diagnostic tools offer potential to reduce this uncertainty. However, metrics for assessing Human-AI Collaboration (HAIC) quality in this context remain underexplored — existing frameworks lack empirical operationalisation for the Human-Centric collaboration mode, where the physician retains full decision-making authority. This study aims to operationalise quality metrics for Human-Centric HAIC within rare neuromuscular disease diagnosis by exploring neurologists' experiences with a conversational AI diagnostic assistant. An exploratory qualitative design employing thematic analysis is planned. Semi-structured interviews following a critical incident technique protocol will be conducted with 10–12 neurologists. Questions address ten collaboration quality dimensions: clarity of communication, ease of use, user satisfaction, feedback frequency, teaching efficiency, error reduction rate, task completion time, confidence, trust score, and safety incidents. Thematic analysis will identify context-specific subdimensions of each metric. The study will lay the groundwork for a domain-specific HAIC assessment instrument, provide design recommendations for clinical AI systems, and establish a basis for future psychometric validation and research on AI adoption in healthcare.

Keywords: Human-AI collaboration assessment, Human-centric mode, Rare disease diagnosis, Qualitative methodology, Collaboration quality metrics

INTRODUCTION

AI systems are increasingly used to support physicians in complex diagnostic tasks. Rare disease diagnosis is one such context. With over 6,000 distinct conditions, it places substantial cognitive demands on clinicians.

Research on Human-AI Collaboration (HAIC) in medicine has focused mainly on technical performance—accuracy, sensitivity, specificity. Less attention has been paid to the quality of the collaboration process itself: how physicians work with AI in practice, and how that collaboration can be meaningfully assessed.

This paper examines one specific collaboration mode—Human-Centric, where the physician retains full decision-making authority and AI acts as a supportive tool. Drawing on the framework by Fragiadakis et al. (2024), we assess whether existing HAIC quality metrics can be measured within a

Received February 21, 2026; Revised April 2, 2026; Accepted April 19, 2026; Available online July 20, 2026

conversational AI diagnostic assistant for rare neuromuscular diseases, and design a qualitative study to explore their context-specific subdimensions.

This work represents an early step toward developing context-sensitive tools for assessing HAIC quality in clinical settings.

THEORETICAL BACKGROUND

The Rare Disease Diagnosis Challenge

Rare diseases, comprising approximately more than 6,000 distinct disorders, affect 30 million people in Europe and approximately 300 million individuals worldwide (Nguengang Wakap et al., 2020; Schieppati et al., 2008).

The diagnostic process for rare diseases is both lengthy and multi-staged. Recent evidence from a large-scale European survey shows that the average time from symptom onset to diagnosis is 4.7 years, with 56% of patients waiting more than 6 months after their first medical contact to receive a diagnosis (Faye et al., 2024).

Several factors contribute to this situation (Schieppati et al., 2008). A primary factor is symptom overlap with common conditions, leading clinicians to initially attribute symptoms to frequently occurring diseases—a phenomenon known as particularizing. Another contributing factor is the necessity for patients to consult multiple specialists, with patients who consult more than 10 general practitioners experiencing diagnostic delays of over 42 months (Benito-Lozano et al., 2022). Moreover, physicians have limited experience with rare diseases as they see them infrequently in their practice, or may never encounter them at all.

Finally, diagnostic uncertainty creates a tendency toward "low-risk" prioritization, where clinicians make conservative, reversible clinical decisions. This approach enables symptomatic treatment while the underlying disease remains unidentified. However, this uncertainty can sometimes lead to misdiagnosis, inappropriate treatment, or lack of proper treatment, resulting in adverse side effects (Faye et al., 2024). These diagnostic challenges result in misdiagnosis or delayed diagnosis, delayed treatment, and reduced quality of life for patients. AI diagnostic systems offer potential to shorten this period of diagnostic uncertainty.

AI as a Potential Solution

AI systems can transcend human cognitive limitations by solving complex tasks through their capacity to analyze vast amounts of data and recognize patterns (Korteling et al., 2021). Machine learning enables the execution of high-level cognitive tasks previously reserved exclusively for humans (Mnih et al., 2015). Medicine represents one of the domains most significantly affected by AI advancements. Currently, numerous AI solutions support physicians by enhancing diagnostic efficiency in medical imaging, radiological interpretation, genetic analysis, clinical decision-making, and early disease detection.

In the context of neurological rare diseases, AI promises faster, more accurate, and more precise diagnosis, along with the development of personalized treatment plans. Research by Mao et al. (2025) demonstrated

that a phenotype-based AI pipeline outperformed human experts in differentiating rare diseases, achieving higher accuracy particularly in cases involving diseases with overlapping symptoms.

Theoretical Gap: Technical Success Does Not Equal Collaboration Success

Most research focuses on the technical and quantitative performance of Human-AI Collaboration in medicine, examining metrics such as accuracy and sensitivity (Cai et al., 2019; Tschandl et al., 2020; Goh et al., 2024). Yet technical effectiveness does not automatically translate into effective collaboration between AI and human clinicians. How clinicians actually work with AI systems in diagnostic practice remains insufficiently understood. This reveals that Human-AI Collaboration is a multidimensional phenomenon extending beyond technical metrics. While Fragiadakis et al. (2024) provide a domain-agnostic framework of collaboration quality metrics for the Human-Centric mode, the authors themselves call for empirical verification of these metrics in specific contexts. Rare disease diagnosis represents a particularly demanding test case: diagnostic complexity, ambiguity, and high cognitive stakes may shape how collaboration quality is experienced in ways that universal metrics cannot anticipate. This gap motivates the present study.

Human-AI Collaboration: Framework

Researchers in Human-AI Collaboration recognize collaboration as a multifactorial (Vössing et al., 2022; Li et al., 2022; Xiong et al., 2022) and multi-domain phenomenon (healthcare, finance, etc.). This complicates assessing collaboration quality and comparing results across studies. To address this challenge, Fragiadakis et al. (2024) proposed a domain-agnostic theoretical framework that defines universal metrics for collaboration quality. Such a framework enables cross-study comparison, standardizes assessment practices, and supports the development of generalizations.

The ability to comprehensively assess collaboration quality is crucial, particularly as integrating AI into work processes represents one of the main challenges identified in HAIC literature (Łabędzki et al., 2025).

The Fragiadakis et al. (2024) framework builds on the existing literature by distinguishing three collaboration modes. The first mode—AI-centric—operates largely autonomously, assuming primary responsibility for decision-making with minimal human intervention. The second mode represents symbiotic collaboration, where human and AI contribute comparable effort and decision-making authority is equally distributed. This mode is characterized by dynamic task allocation, where the system learns to recognize its limitations and actively defers to the human when appropriate. The third mode—Human-centric collaboration—is characterized by decision-making authority and control residing with the human. AI offloads computationally demanding, repetitive, or cognitively burdensome tasks, enabling humans to leverage their unique capabilities.

This study focuses on the Human-centric collaboration mode. According to this mode, the authors identified metrics in two categories: (1) Communication: clarity of communication, ease of use, user satisfaction, and frequency of feedback, and (2) Efficiency: teaching efficiency, error reduction rate, confidence, task completion time, trust score, and safety incidents (Fragiadakis et al., 2024, p. 14). However, they emphasize the need for empirical verification that operationalizes these metrics within specific contexts. This study addresses this gap by examining Human-Centric mode metrics in rare neurological disease diagnosis.

AI Tools for Rare Disease Diagnosis: Landscape and Case Selection

Scope and Rationale

A conversational AI diagnostic assistant for rare neuromuscular diseases was selected as the empirical case. The selection was guided by three criteria: the system operates in Human-Centric mode; it is actively used in clinical practice; and it targets a condition category characterized by high diagnostic complexity and prolonged diagnostic timelines—a theoretically rich context for examining collaboration quality.

The AI Diagnostic System

The study examines a conversational AI diagnostic assistant developed for rare neuromuscular diseases. The mobile application supports early-stage diagnostic workup, particularly when differential diagnosis remains broad—helping general practitioners identify appropriate subspecialty referrals or assisting neurologists with complex initial presentations.

Physicians manually enter symptom information gathered during patient encounters. The system conducts a dialogue, asking adaptive questions based on previous responses. It analyzes symptoms against a structured knowledge base encompassing over 300 clinical features (ocular and bulbar symptoms, limb weakness patterns, temporal characteristics, family history, biomarkers) mapped to more than 25 disease entities, developed in collaboration with neuromuscular disorder specialists.

The application generates a ranked differential diagnosis list with explanatory reasoning linking reported symptoms to characteristic disease features. It provides diagnostic recommendations (laboratory tests, imaging studies, specialist referrals) to confirm or exclude specific conditions. The system operates as a standalone application, not integrated with electronic health records.

Human-Centric Mode Classification

The system aligns with the Human-Centric mode defined by Fragiadakis et al. (2024), where humans retain primary decision-making authority while AI serves as a supportive tool. This alignment manifests across four dimensions. All outputs are framed as suggestions, with final decisions remaining entirely

with the treating physician. The system offloads cognitively burdensome tasks, extending physician reach without replacing clinical reasoning. It cannot initiate autonomous actions. Each diagnostic suggestion includes transparent reasoning, enabling critical evaluation rather than "black box" acceptance.

METHODOLOGY

Research Objective and Questions

This study aims to operationalize metrics for assessing the quality of Human-AI collaboration in the Human-Centric mode within the context of rare neuromuscular disease diagnosis. The study addresses two research questions:

RQ1: How do physicians experience Human-Centric AI collaboration in rare neuromuscular disease diagnosis, and what subdimensions of collaboration quality does this experience surface?

RQ2: How can these subdimensions be operationalized into measurable indicators suitable for future quantitative assessment?

Preliminary Metric Assessment

The examined system is a conversational AI diagnostic assistant, in which the physician retains full clinical authority over all diagnostic decisions. To determine the scope of the qualitative study, we assessed whether Human-Centric collaboration quality metrics from the Fragiadakis et al. (2024) framework can be directly measured within this system. Each metric was evaluated against the application's functionalities and available system-generated data.

Several metrics permit quantitative measurement—through system-generated data (e.g., session duration, iteration count) or standardised instruments (e.g., SUS, Confidence Delta). However, two remain partially inaccessible: feedback frequency cannot be reliably captured as the system provides no error-reporting mechanism, and correction rate misses physician interventions made outside the system. Importantly, even technically accessible metrics require clinical context to be interpretable—qualitative exploration is therefore warranted for all ten metrics.

Qualitative exploration is therefore necessary: to interpret quantitative indicators that are clinically uninterpretable in isolation, to capture phenomena invisible to system logs, and to generate subdimensions informing future questionnaire development.

Table 1 presents all Human-Centric mode metrics examined in this study, the available quantitative measures, and their limitations in this clinical context. These limitations justify the qualitative approach and will shape the semi-structured interview guide.

Table 1: Human-centric collaboration quality metrics and limitations of quantitative measurement.

Category	Metric	Importance	Definition	Quantitative Measurement	Quantitative Limitations / Qualitative Need
Communication	**Clarity of Communication**	**High**	"Assesses how understandable the communication is (...) evaluated using qualitative methods." [Fragiadakis, 2024] "Maintenance of a common frame of reference (...) without needing a long and tedious elaboration." [Hoc, 2000]	Applicable Clarification Count	The number of clarification requests shows how often communication failed, but not why. The same AI message may be clear to a neurologist yet unclear to a general practitioner. A count of "3 clarification requests" cannot distinguish a terminology problem from an overly complex response structure or missing reasoning — yet each requires a different system improvement. **Goal:** to identify the types of communication breakdown underlying clarification requests.
	Ease of Use	**High**	"Gauges the user's subjective experience (...) Often measured through Likert scale questionnaires." [Fragiadakis, 2024]	Applicable SUS (System Usability Scale)	SUS measures general software usability (e.g., "the system is easy to use", "I would use it regularly", "the complexity level is appropriate"). A high usability score does not rule out that pre-defined response options led the physician to oversimplify the patient description — when the actual clinical picture did not fit any of the offered categories. **Goal:** to uncover the specific sources of low ease-of-use ratings in a clinical AI context.
	User Satisfaction	**High**	"Measures the user's subjective experience (...) satisfaction with the interaction and system's performance." [Fragiadakis, 2024]	Applicable NPS (Net Promoter Score)	A single score does not explain the sources of satisfaction or frustration. A physician may give a high rating despite serious concerns about specific moments in the session. In rare disease diagnostics, satisfaction depends on the quality of the most challenging interactions — and a single number cannot capture those. **Goal:** to identify critical satisfaction and frustration moments that should become anchors in future domain-specific satisfaction scales.
	Frequency of Feedback	**Medium**	"Monitors how often users provide feedback (...) Number of feedback instances." [Fragiadakis, 2024] "Improving the performance of the system by giving continuous feedback (...) dynamic dashboards to visualize metrics." [Dehghani et al., 2024]	Applicable Interaction Count	The feedback action count does not capture situations where the physician wanted to correct something but had no way to do so — because the tool offers no error-reporting mechanism. Such "silent frustrations" are invisible to quantitative measurement, yet may be the most clinically important. **Goal:** to identify feedback needs and inform future measurement design.
Efficiency	**Teaching Efficiency** (In-Context Learning)	**Low**	"Gain in AI performance per unit of teaching effort." [Fragiadakis, 2024] "Achieve better sample efficiency (...) measured by learning curves showing how accuracy changes as examples are sequentially provided (...) and session duration." [Cakmak & Thomaz, 2014]	Applicable Iteration Count (Prompt-to-Diagnosis Ratio)	Model is frozen (pre-trained LLM) and does not "learn" new weights. The number of conversational exchanges before reaching a diagnosis does not account for case complexity — some conditions objectively require more differential questions than others. A high iteration count may indicate either AI inefficiency or necessary diagnostic thoroughness. Only the physician can judge which questions were unnecessary. **Goal:** to identify measures better reflecting clinical work context than raw iteration count.
	Error Reduction Rate	**High**	"Decrease in errors made by AI following human intervention." [Fragiadakis, 2024] "Risk associated with a decision outcome can be defined as the cost of an error multiplied by the probability of that error." [Parasuraman et al., 2000]	Applicable Correction Rate (Error Rejection %)	Correction Rate only captures corrections visible in the system, while some physician interventions remain beyond its reach — the physician rejects a diagnosis in their own judgement, but the system has no access to that decision. Furthermore, the metric does not distinguish trivial errors from clinically dangerous ones — and the consequences of a mistake may differ dramatically depending on the condition. **Goal:** to capture physicians' definitions of AI error and error recognition patterns.
	Task Completion Time	**High**	"Compares task completion times with and without AI support." [Fragiadakis, 2024] "Measure employee productivity in a more objective manner (...) [using] data such as time spent on tasks and the completion rate." [Murugesan et al., 2023]	Applicable Session Duration	Shorter session time does not equal better diagnostic quality. Session duration conflates efficiency with thoroughness — only the physician can judge whether the time spent was diagnostically valuable. **Goal:** to capture physicians' subjective experience of session pace and identify what extended or wasted diagnostic time.
	Confidence	**High**	"Level of user trust in AI recommendations." [Fragiadakis, 2024] "Self-confidence (...) trust in one's own abilities [vs automation] (...) Rated on a scale: 'How high was your self-confidence?'" [Lee & Moray, 1994]	Applicable Confidence Delta (Pre/Post Self-Assessment)	Confidence Delta captures the change in certainty between the start and end of a session, ignoring the baseline and the mechanism of that change. The same score may reflect hypothesis confirmation in a confident specialist or merely uncertainty reduction in a physician lacking experience in the field — and even a negative value may indicate a diagnostically valuable effect when AI challenges an initial diagnosis and prompts the physician to revise it (Lee & Moray, 1994). **Goal:** to uncover mechanisms behind confidence change and develop separate items for self-confidence and trust in automation in a rare disease clinical context.
	Trust Score	**High**	"Assesses user trust (...) through surveys and trust scales." [Fragiadakis, 2024] "Subjective ratings of trust (...) asking 'How much did you trust the automatic controller?' rated on a scale [0-100]." [Lee & Moray, 1994]	Applicable Jian Trust Scale	The trust measure by Lee and Moray (1994), designed for industrial automation systems, may not capture the specifics of trust in AI for medical diagnostics, where trust may be shaped by different factors. **Goal:** identification of trust dimensions relevant to physicians in rare disease diagnostics.
	Safety Incidents	**High**	"Tracks number of safety-related issues encountered." [Fragiadakis, 2024] "Costs of decision/action consequences (...) preventing the operator from intervening successfully or in a timely manner." [Parasuraman et al., 2000]	Applicable Critical Error Count	The number of AI errors is uninterpretable without considering their clinical weight. Confusing two similar conditions may in one case require an immediate change in clinical management, and in another — have no significant consequences (Parasuraman et al., 2000: risk = error cost × probability). **Goal:** to establish a clinician-derived taxonomy of AI error severity, enabling weighted interpretation of error frequency by clinical consequence.

RESEARCH DESIGN

Overall Approach

This study employs an exploratory qualitative design to uncover context-specific subdimensions within the quality metrics identified above. This approach is appropriate given the early stage of HAIC research in clinical contexts and the need for rich, contextualized understanding of neurologists' experiences with AI-assisted rare disease diagnosis.

We adopt a hybrid deductive–inductive analytical framework. Deductively, we begin with the quality metrics operationalized for the Human-Centric collaboration mode—including Communication, Ease of Use, User Satisfaction, Feedback Frequency, Teaching Efficiency, Error Reduction, Task Completion Time, Confidence, Trust, and Safety—as the organizing structure for data collection. Inductively, within each metric, we remain open to identifying context-specific subdimensions reflecting the realities of rare disease diagnosis not captured by existing measurement approaches.

Sampling Strategy and Participant Recruitment

Purposeful sampling will be used to recruit participants best positioned to illuminate the phenomenon (Brewerton and Millward, 2001). Eligible participants are neurologists employed at clinical hospitals, with at least two years of post-residency experience, documented involvement in diagnosing rare neurological conditions (prevalence <5 per 10,000), and at least six months of experience using the application under study or a comparable AI-assisted diagnostic tool in a Human-Centric collaboration mode.

We plan to recruit 10–12 neurologists, consistent with recommendations for thematic saturation in homogeneous professional samples (Guest et al., 2006), with data collection proceeding iteratively until theoretical saturation is achieved. Participants will be recruited through neurology departments at clinical hospitals affiliated with academic medical centres.

Data Collection Procedures

Semi-structured interviews (approximately 60 minutes each) will be conducted following a critical incident technique protocol. Participants will be asked to recall specific diagnostic episodes involving rare or suspected rare conditions where they collaborated with a conversational AI diagnostic tool.

For each recalled episode, the interview will systematically explore the quality dimensions corresponding to the metrics under investigation through open-ended questions tailored to the clinical context of rare disease diagnosis. Example questions include: "Walk me through a recent case where AI helped you consider a rare diagnosis… What made you trust—or distrust—the AI's suggestion?" Relevant contextual factors (case complexity, diagnostic urgency, information availability) will be documented for each episode to enable pattern analysis.

All participants will provide written informed consent. Interviews will be audio-recorded with permission and transcribed verbatim under confidentiality agreements. Data will be stored on encrypted devices and destroyed following transcription verification.

Qualitative Data Analysis

Thematic analysis (Braun and Clarke, 2006) will be employed to identify context-specific subdimensions within each quality metric. The analytical process moves through four stages: (1) initial familiarisation and open coding close to participants' language; (2) focused coding organised deductively within the Human-Centric mode quality metrics, with inductive identification of recurring patterns; (3) subdimension formulation and definition with supporting quotes; and (4) cross-case pattern analysis examining whether subdimensions vary by contextual factors documented during interviews. To enhance analytical trustworthiness, preliminary findings will be returned to selected participants for member checking (Lincoln and Guba, 1985).

Ethical Considerations

The study will be conducted in accordance with ethical principles governing research with human participants. Informed consent will be obtained from all participants and from the institution providing access to the AI system. Data will be anonymized with all personal and institutional identifiers removed. Audio recordings require explicit permission from participants. The study complies with GDPR regulations. Participation is voluntary, and participants retain the right to withdraw at any stage without consequence.

IMPLICATIONS FOR RESEARCH AND PRACTICE

Methodological Implications

Applying a domain-agnostic HAIC framework in a specific clinical context is not straightforward and carries methodological consequences. Systems operating in Human-Centric mode may differ fundamentally in properties such as model adaptability (frozen, session-bound, or continuously learning), workflow integration, or interaction structure (conversational vs. one-shot). These differences determine not only how metrics are measured, but whether they remain conceptually valid—as illustrated by Teaching Efficiency, which requires reconceptualization when the AI model operates with session-bound context, learning within a single conversation but resetting completely across sessions. Critically, this ontological assessment cannot be conducted by the researcher alone—it requires active involvement of a domain expert from the study design stage, not only during data collection. In clinical contexts, this repositions the clinician from a research participant to a co-designer of the measurement instrument. We therefore recommend that future researchers conduct a preliminary technical and ontological audit of the system, in collaboration with domain experts, prior to applying any HAIC assessment framework.

Practical Implications

The findings also carry implications for AI adoption strategy in healthcare institutions. Research on computational phenotyping in rare disease diagnosis shows that trust is a prerequisite for clinician acceptance of

machine learning tools, and is conditioned by perceived reliability, validity, and accuracy (Hallowell et al., 2022). We argue that trust formation may not be independent of the underlying AI mechanism. The system described in this study operates on a frozen model, which limits error variability and makes it a relatively stable tool for early clinical adoption. By contrast, continuously learning systems—such as those examined by Hallowell et al. (2022), where insufficient demographic diversity in training data reduces diagnostic accuracy across racial groups—carry a qualitatively different risk profile.

We therefore suggest that healthcare institutions consider prioritising frozen or knowledge-based AI systems in early adoption phases, where the primary goal is building clinician familiarity while limiting the risk of trust erosion and overinvestment. Where more adaptive systems are subsequently introduced, clinician training should explicitly address algorithm aversion and trust calibration. Critically, regardless of the underlying mechanism, structured onboarding should communicate not only system functionality but also its limitations and the boundaries of safe use.

Directions for Future Research

Beyond adoption strategy, the quality of clinician trust is also shaped by moment-to-moment interaction experience—including factors such as response time. Research demonstrates that response time affects trust in AI asymmetrically—unlike human experts, algorithms are trusted more when they respond quickly (Efendic et al., 2020). However, in rare disease diagnosis, where diagnostic complexity is high, it remains unknown whether this pattern holds: a rapid differential diagnosis may signal insufficient reasoning rather than superior capability. Empirical investigation of optimal response time in this clinical context is therefore warranted. This applies not only across disease types, but also across collaboration modes (Human-Centric, Symbiotic, and AI-Centric) and underlying AI mechanisms—since response time expectations may vary depending on how the system operates. Without empirical calibration of response time, it is difficult to determine why a clinician distrusts or abandons the system. Reluctance may reflect genuine concerns—for instance, that the differential diagnosis failed to account for a key symptom—or it may be partly driven by response time expectations alone. Disentangling these factors is a prerequisite for meaningful interpretation of adoption patterns.

CONCLUSION

This paper proposes a qualitative study examining how physicians experience Human-Centric AI collaboration in rare neuromuscular disease diagnosis. Building on the Fragiadakis et al. (2024) framework, we identify ten collaboration quality metrics applicable to the examined system and design a semi-structured interview study to uncover their context-specific subdimensions.

The preliminary metric assessment reveals that quantitative measurement of Human-Centric collaboration quality is limited in this context—by system architecture, by the clinical meaninglessness of raw indicators without interpretive context, or by both. This finding has implications beyond the present study: even at this early stage, it is clear that a domain-agnostic framework requires contextual adaptation before its metrics can be applied in clinical settings. The methodological approach—combining preliminary quantitative assessment with qualitative exploration to develop measurable indicators—may be useful in other HAIC contexts in healthcare.

The planned qualitative study aims to provide that grounding, laying the foundation for future instrument development and psychometric validation.

REFERENCES

Benito-Lozano, J., López-Villalba, B., Arias-Merino, G., et al. (2022). Diagnostic Process in Rare Diseases: Determinants Associated with Diagnostic Delay. International Journal of Environmental Research and Public Health, Volume 19, No. 11, 6456.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, Volume 3, No. 2, pp. 77–87.

Brewerton, P., & Millward, L. (2001). Organizational Research Methods. Thousand Oaks, CA: SAGE Publications.

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C., Terry, M. (2019). Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. CHI '19, pp. 1–14.

Cakmak, M., and Thomaz, A. L. (2014). Eliciting good teaching from humans for machine learners. Artificial Intelligence, Volume 217, pp. 198–215.

Dehghani, F., Dibaji, M., Anzum, F., et al. (2024). Trustworthy and responsible AI for human-centric autonomous decision-making systems. Available at: arXiv:2408.15550.

Efendić, E., Van de Calseyde, P. P. F. M. and Evans, A. M. (2020), 'Slow response times undermine trust in algorithmic (but not human) predictions', Organizational Behavior and Human Decision Processes 157, 103–114.

Faye, F., Crocione, C., Anido de Peña, R., et al. (2024). Time to diagnosis and determinants of diagnostic delays of people living with a rare disease. European Journal of Human Genetics, Volume 32, pp. 1116–1126.

Fragiadakis, G., Diou, C., Kousiouris, G., and Nikolaidou, M. (2024). Evaluating human-AI collaboration: A review and methodological framework. Available at: arXiv:2407.19098.

Goh, E., Gallo, R., Hom, J., et al. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Network Open, Volume 7, No. 10, e2440969.

Guest, G., Bunce, A. and Johnson, L. (2006). How many interviews are enough? Field Methods, Volume 18, No. 1, pp. 59–82.

Hallowell, N., Badger, S., Sauerbrei, A., Nellåker, C. and Kerasidou, A. (2022), 'I don't think people are ready to trust these algorithms at face value: trust and the use of machine learning algorithms in the diagnosis of rare disease', BMC Medical Ethics 23(1), 112.

Hoc, J.-M. (2000). From human–machine interaction to human–machine cooperation. Ergonomics, Volume 43, No. 7, pp. 833–843.

Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus artificial intelligence. Frontiers in Artificial Intelligence, Volume 4, 622364.

Łabędzki, R., Mikołajczyk, K., Biłyk, A., & Trojanowska, M. (2025). "Understanding human-AI collaboration: A systematic review", in: HCI International 2025 Posters. Communications in Computer and Information Science, Vol. 2529, pp. 264–275. Springer.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, Volume 40, No. 1, pp. 153–184.

Li, J., Huang, J., Liu, J., & Zheng, T. (2022). Human-AI cooperation: Modes and their effects on attitudes. Telematics and Informatics, Volume 73, 101862.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic Inquiry. Thousand Oaks, CA: SAGE Publications.

Mao, X., Huang, Y., Jin, Y., et al. (2025). A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. npj Digital Medicine.

Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. Nature, Volume 518, pp. 529–533.

Murugesan, U., Subramanian, P., Srivastava, S., & Dwivedi, A. (2023). A study of artificial intelligence impacts on human resource digitalization in Industry 4.0. Decision Analytics Journal, Volume 7, 100249.

Nguengang Wakap, S., Lambert, D. M., Olry, A., et al. (2020). Estimating cumulative point prevalence of rare diseases. European Journal of Human Genetics, Volume 28, pp. 165–173.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A, Volume 30, No. 3, pp. 286–297.

Schieppati, A., Henter, J.-I., Daina, E., & Aperia, A. (2008). Why rare diseases are an important medical and social issue. The Lancet, Volume 371, No. 9629, pp. 2039–2041.

Tschandl, P., Rinner, C., Apalla, Z., et al. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine, Volume 26, pp. 1229–1234.

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. Information Systems Frontiers, Volume 24, No. 3, pp. 877–895.

Xiong, W., Fan, H., Ma, L., & Wang, C. (2022). Challenges of human–machine collaboration in risky decision-making. Frontiers of Engineering Management, Volume 9, No. 1, pp. 89–103.