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.

Reducing Diagnostic Odyssey in Rare Diseases: Validation of an Artificial Intelligence-Based Diagnostic Support Tool (Pilot Study)

Lara Mroueh
DIGITAL MEDICAL HUB
Josue Srougbo
DIGITAL MEDICAL HUB
Valentin Hess
Nancy University hospital
Anne Mirguet
Nancy University hospital
Lionel Donier
Medical office
Loïc Etienne
Medvir
Etienne Van Der Elst
Sanofi (France)
Eliane Le Peillet Feuillet
Sanofi (France)
Hubert De Boysson
Caen University hospital
Joris Galland
joris.galland@digitalmedicalhub.com

DIGITAL MEDICAL HUB

Research Article

Keywords: Artificial intelligence, rare diseases

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11319104/v1

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

Additional Declarations: Competing interest reported. This study was funded by Sanofi. EVDE and ELPF are affiliated with Sanofi (Department of Innovation). LE developed the medical algorithm underlying AccelRare and is affiliated with Medvir. LM, JS and JG are affiliated with Digital Medical Hub, which received funding from Sanofi to conduct this study. VH, AM, LD and HdB declare that they have no competing interests

Title : Reducing Diagnostic Odyssey in Rare Diseases: Validation of an Artificial Intelligence-Based Diagnostic Support Tool (Pilot Study)

Authors : Lara Mroueh [1], Josue Srougbo[1], Valentin Hess [2], Anne Mirguet [2], Lionel Donier [3], Loïc Etienne [4], Etienne Van Der Elst [5], Eliane Le Peillet Feuillet[5], Hubert de Boysson[6], Joris Galland [1,7]*

1. Digital Medical Hub, Paris, France
2. Department of pediatry, Nancy University hospital , Vandoeuvre-Lès-Nancy, France
3. General practitioner, Villefranche-sur-Saône, France
4. Medvir, medical intelligence service, Paris, France
5. Department of Innovation, Sanofi, Gentilly, France
6. Department of internal medicine, Caen University hospital, Caen, France
7. Department of internal medicine, Hospital center of Bourg-en-Bresse , Viriat, France

*corresponding author :

Galland Joris M.D.,

Digital Medical Hub, Paris, France

Hôtel Dieu, Biolabs, 1 parvis Notre-Dame - place Jean-Paul-II, 75004 Paris

joris.galland@digitalmedicalhub.com

Tel : +33 6 21 19 11 90

Manuscript word count : 4944

Key words : Artificial intelligence, rare diseases

Funding : Sanofi

Abstract

Background: Rare diseases affect over 3 million people in France and remain associated with substantial diagnostic delay, averaging 1.5 years, partly because of diagnostic complexity and limited expertise in primary care. We evaluated AccelRare, an artificial intelligence-based symptom checker using expert-system logic, to support diagnostic orientation and referral decisions for rare diseases.

Methods: We conducted a cluster-randomized, simulation-based diagnostic performance study involving general practitioners and pediatricians. Thirty adult and 30 pediatric clinical vignettes were developed, each set comprising 24 rare and 6 common disease cases. Junior and senior physicians were randomized to "Physician alone" or "Physician with AccelRare." A post hoc "Physician then AccelRare" analysis assessed diagnostic revision after tool use. The primary outcome was concordance between the top-ranked diagnostic hypothesis and the reference diagnosis. Secondary outcomes included referral accuracy, repeatability, reproducibility, usability, and decision time. McNemar's test was used for comparisons of paired categorical data

Results: Across all vignettes, top-diagnosis concordance was 70.8% for "Physician then AccelRare," compared with 47.5% for "Physician alone" ($p<0.001$) and 64.2% for "Physician with AccelRare" ($p=0.003$). In rare disease cases, concordance was 63.8%, 35.1%, and 58.5%, respectively, while AccelRare alone achieved 36.2%. Referral accuracy was 92.5% with sequential physician–AccelRare use versus 76.7% for physicians alone ($p<0.001$), and 90.4% versus 70.2% in rare disease cases. After AccelRare use, 23.3% of initially incorrect diagnoses became concordant, while no initially correct diagnoses were lost. Improvement was greater among general practitioners than pediatricians (38.3% vs. 8.3%). Reproducibility of the top hypothesis was 35.0% overall and 40.4% for rare disease cases; repeatability was 33.3%. The mean System Usability Scale score was 67.2.

Conclusions: AccelRare improved diagnostic concordance and referral orientation for rare diseases when integrated into physician decision-making. Its greatest value appears to be as a complementary decision-support tool rather than a replacement for clinical expertise.

Background

Rare diseases represent a major public health challenge in Europe, with significant medical, social, and economic repercussions. Although each disease affects a limited number of individuals, their cumulative prevalence is substantial. Among the 7,000 rare diseases identified to date, more than 3 million individuals are affected in France—accounting for 4.5% of the population—and approximately 25 million people across Europe. These conditions primarily affect young children, with 50% of cases occurring in patients under the age of five and causing 10% of deaths between ages one and five. Around 80% of rare diseases have a genetic origin and are characterized by severity, chronicity, and progressive evolution, profoundly impacting patients' quality of life. In 50% of cases, they lead to motor, sensory, or intellectual impairments, and total loss of autonomy is observed in 9% of cases (1).

The diagnosis and management of rare diseases pose numerous challenges due to clinical complexity, lack of specialized expertise, and difficulties in accessing appropriate care. Diagnostic delay remains a major obstacle: in France, 25% of patients wait more than five years to receive a diagnosis, with an average diagnostic delay estimated at 1.5 years (2).

To address these challenges, France launched the first national plan for rare diseases in 2005, structured around 387 Rare Disease Reference Centers (RDRC) and 1,800 Rare Disease Competence Centers (RDCC), ensuring specialized care and local follow-up. In 2015, a second plan reinforced this framework by establishing 23 rare disease health networks, involving a broad network of stakeholders, including healthcare professionals, laboratories, research units, patient associations, and medico-social structures (3). In 2024, the fourth national plan for rare diseases emphasizes the use of artificial intelligence (AI) and health data repositories to accelerate rare disease diagnosis and propose innovative treatments (4).

However, despite these advancements, diagnostic delay remains a persistent issue, generating significant financial costs due to inappropriate treatments, repeated medical consultations, and multiple examinations before referral to a specialized center. Improved diagnostic support, particularly for general practitioners (GP) and pediatricians, as well as optimized referral pathways to expert centers, are key levers to enhance patient care. The use of AI for diagnostic support and patient referral warrants further evaluation in the field of rare diseases. Historically developed and assessed on emergency medicine cohorts, AI-powered *symptom checkers (SC)* have shown variable diagnostic performance depending on the solution, with some models

occasionally surpassing human physicians (5,6). In this context, AccelRare is an innovative SC solution utilizing Medvir® technology, designed to provide a diagnostic and referral support tool specifically for primary care physicians. The goal is to reduce diagnostic delays and improve the management of rare diseases. The AccelRare algorithm has been trained on 272 rare diseases.

Through this simulation study using clinical case vignettes, we evaluate the performance of AccelRare in assisting diagnosis and referral of patients with rare diseases among a population of GPs and primary care pediatricians.

Methods

Study design and justification

This is a comparative, diagnostic performance study, designed as a cluster-randomized, multicenter trial using simulated patients.

The study design is a two-arm trial with cluster randomization of investigator physicians, stratified by experience level.

Four GP (two junior <3 years post-licensure and two senior >5 years post-licensure) who did not write the clinical vignettes were assessed the 30 created clinical vignettes, and four private-practice pediatricians (two junior <3 years post-licensure and two senior >5 years post-licensure) who did not contribute to the writing of the clinical vignettes were also assessed the 30 created clinical vignettes.

Each matched pair (one junior physician and one senior physician in each speciality) was randomized into one of two groups: the "physician alone" cluster or the "physician + AccelRare" cluster (Figure 1).

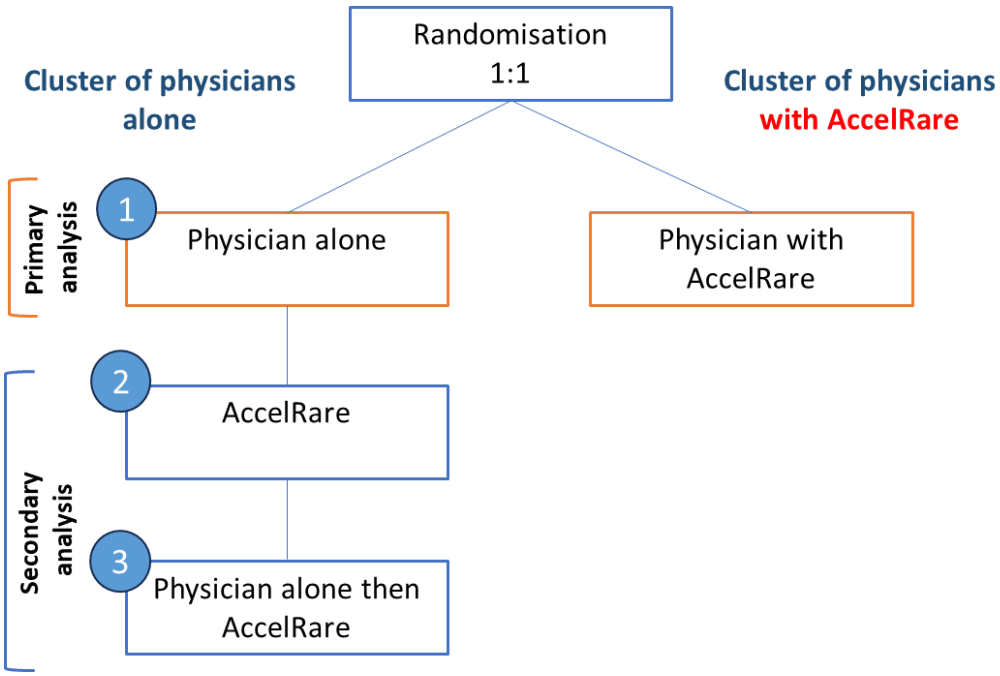

***Figure 1**: Study design*

Investigators recruited for randomization were required to be board-certified physicians, GP or pediatricians working in primary care, with no involvement in vignette creation and no conflict of interest with the industry partner. Clinical vignettes were authored by experienced hospital-based pediatricians or internal medicine specialists, who were also free of any conflicts of interest.

We chose to conduct a simulation-based study, as simulation offers a valuable alternative to real-patient evaluation, particularly in complex clinical situations such as rare diseases. Simulation enables clinical research in settings where real-life implementation would be logistically or ethically challenging (7–10). In the context of rare diseases, it is nearly impossible to gather in one site at least 60 adult or pediatric patients with orphan conditions.

Furthermore, simulation methods have become increasingly relevant for the evaluation of digital decision support tools such as symptom checkers (SC). Several studies highlight their advantages in terms of cost-effectiveness, time efficiency, and safety. Simulation allows controlled comparisons without exposing real patients to risk. These approaches include both clinical vignettes and simulated tasks (7–10).

Primary and secondary objectives

The main objectives of the study were to determine and compare the diagnostic and triage performance for rare diseases in adult and pediatric cases between physicians using a symptom

checker (AccelRare) and those working alone without AccelRare. Diagnostic accuracy and triage performance were evaluated based on a gold standard defined by a panel of clinical vignettes written by expert physicians specialized in rare diseases. The primary outcome was defined as the concordance rate between diagnoses and care orientation with and without the support of the SC.

Secondary objectives included the evaluation of correct primary diagnoses (based on the top three diagnostic hypotheses), evaluation of the change in diagnostic concordance when physicians alone reassess their diagnosis using AccelRare, assessment of the repeatability and reproducibility of the device, assessment of usability and perceived utility of the SC via a standardized System Usability Scale (SUS), evaluation of digital burden and timing with and without SC support (timing from history-taking to diagnosis and SC usage).

Study procedure

In the "physician alone" arm, 3 steps were conducted :

- For the primary analysis “Evaluation by physicians alone” : Physicians classified each clinical vignette as a rare or common disease, proposed up to 10 diagnostic hypotheses, and suggested a triage level (primary care, specialist, or Rare Disease Reference Center). They will record their responses in an electronic case report form (eCRF), with timing monitored by a clinical research associate (CRA), before entering the data into the AccelRare pre-diagnostic tool.

  These assessments and proposed diagnoses were analyzed to evaluate the primary objective and compared with those from the group of physicians using AccelRare.

- For the secondary analysis (Figure 1) :
  - “Evaluation by AccelRare” Diagnoses proposed only by AccelRare.

    Physicians were trained in the use of AccelRare through an online course provided by the principal investigator. They entered clinical vignette data into AccelRare, which generated responses for disease type, diagnostic hypotheses, and patient referral. The physicians then recorded these responses in an eCRF.

  - In the second phase of vignette evaluation, physicians assessed the credibility and relevance of the responses provided by the AccelRare tool. They could

either maintain their initial responses or propose new ones based on their evaluation of the tool's suggestions.

In the "physician + AccelRare" arm, participants used the SC concurrently with vignette review, entering data directly into AccelRare as part of their diagnostic process and documenting their decision-making accordingly.

The time spent on evaluating the clinical vignette was recorded in each study arm. The tools used during the evaluation process, leading up to the final diagnosis, were also documented in each study arm.

Description of the device and comparator (gold standard)

AccelRare is an AI-based symptom checker that leverages expert system technology. For each disease, a set of symptoms is associated with weighted values according to their reported frequency. The tool is freely accessible online for all healthcare professionals, currently available in France.

Healthcare providers input the patient's symptoms, and the algorithm iteratively asks follow-up questions to refine diagnostic hypotheses. In this study, AccelRare was evaluated using clinical vignettes written by specialists in rare diseases. Each vignette included a principal diagnosis, along with one to five possible differential diagnoses. We compared the diagnostic hypotheses generated by the AI tool with those provided by expert specialists and primary care physicians.

Study population

As patients were simulated, no inclusion criteria were applicable. The study population consists of simulated patients represented by clinical vignettes. These vignettes will be designed to include:

- Rare but "common" diseases: Conditions with a prevalence between 1 in 10,000 and 1 in 100,000 individuals, which can be managed in a specialty service without requiring a dedicated rare disease reference center.
- Rare and "complex" diseases: Conditions with a prevalence of less than 1 in 100,000 individuals, requiring management in a specialized rare disease reference center.

- Common diseases: Conditions with a prevalence of more than 1 in 2,000 individuals, frequently encountered in general practice but potentially challenging in terms of diagnostic orientation based on the patient's symptoms.

The definitions of prevalence-based disease were arbitrarily determined by the authors of the clinical vignettes.

Description and creation of vignettes

The clinical vignettes were consist of simulated patient cases containing these information: Age, sex, presence of overweight or obesity, family history, personal medical history, list of treatments, complaints, signs, and symptoms reported by the patient, known pathologies and comorbidities, known clinical data (collected by the GP or a specialist),

Other information optionally: known medical imaging data, known biological data, negative signs and symptoms (differential diagnosis).

The vignettes were written by clinicians and directly inspired by real cases from the clinical practice of the three authors. Signs, symptoms, biological data, and medical imaging findings must be as "stereotypical" as possible and should not be limited to atypical presentations. Each clinical vignette, along with its proposed principal and differential diagnoses, was reviewed and validated by a second expert in rare diseases.

Number of vignettes

In this pilot study, no statistical power calculation is required for this pilot study. Thirty adult and thirty pediatric vignettes were developed, including 24 rare disease cases (6 complex, 18 common rare diseases) and 6 common disease cases for each group.

Thirty adult vignettes were assessed by two GP (one junior and one senior) using AccelRare, while the same vignettes were evaluated by two other GP (one junior and one senior) in the non-AccelRare arm. Similarly, thirty pediatric vignettes were evaluated by two pediatricians (one junior and one senior) with AccelRare, and by two different pediatricians (one junior and one senior) without AccelRare. Each physician evaluated 30 clinical vignettes. Each vignette was assessed twice - once by a junior physician and once by a senior physician - to mitigate potential bias related to years of clinical experience.

Evaluation of reproducibility and repeatability

We assessed the reproducibility of the solution by measuring the percentage of concordance in the diagnostic hypotheses across all investigators who used AccelRare during the study, for each clinical vignette.

Repeatability was also evaluated: prior to the study, a clinical research associate (non-physician) was trained to enter six clinical vignettes into the system. These vignettes were entered in random order, one after the other, and the same process was repeated a second time. We then measured the percentage of concordance in the hypothesis between the two sessions.

Statistical analyses

Statistical analyses included descriptive statistics to assess diagnostic and triage concordance, defined as the proportion of correct diagnoses across different modalities (physician alone, AccelRare, physician followed by AccelRare, and physician with AccelRare). Concordance was calculated for primary diagnosis, patient orientation (reference center, specialist center, or primary care), and the top three differential diagnoses for all clinical vignettes, with a focus on rare vignettes and specialty-specific analysis. McNemar's test was used for comparisons of paired categorical data.

We reported this cluster-randomized vignette study in accordance with CONSORT 2025, using the extensions for cluster-randomized trials and healthcare simulation research where applicable.

Results

Primary endpoints

Correct first diagnostic hypothesis with and without AccelRare

The main endpoint is the concordance between the diagnoses and referrals established with and without the SC, with the gold standard diagnosis being the one specified in the simulated patient vignette.

All Clinical Vignettes

For the diagnostic performance, there is a trend toward a higher concordance rate for the first hypothesis in the "Physician then AccelRare" group, with 70.8% of first hypotheses being concordant, compared to 29.2% for "AccelRare," **47.5**% for "Physician alone," and **64.2**% for "Physician + AccelRare" (Table 1).

McNemar's test reveals a significant difference between the "Physician alone" and "Physician then AccelRare" groups ($p < 0.001$). Similarly, a significant difference is observed between the "Physician + AccelRare" and "Physician alone" groups ($p = 0.003$).

		Physician alone	Physician then AccelRare	Physician with AccelRare	AccelRare
Concordance of 1st hypothesis	N	120	120	120	120
	Yes	57 (47.5%)	85 (70.8%)	77 (64.2%)	35 (29.2%)
	No	63 (52.5%)	35 (29.2%)	43 (35.8%)	85 (70.8%)

Table 1 *: Proportion of concordant diagnoses (all clinical vignettes)*

Rare vignettes

Focusing only on "Rare" diseases: there is a tendency for better concordance of the first hypothesis in the "Physician then AccelRare" group with 63.8% of concordant first hypotheses, compared to **35.1%** for "Physician alone," 36.2% for "AccelRare," and **58.5%** for "Physician + AccelRare" (Annex 1; Table 1.a). The concordance rate is thus decreased across all groups when only rare diseases are considered.

By speciality

When focusing on GP, there is a tendency for better concordance of the first hypothesis in the "Physician then AccelRare" group, with 73.3% of concordant first hypotheses compared to **35%** for "Physician alone", 40% for "AccelRare", and **65%** for "Physician + AccelRare". For pediatricians, the "Physician then AccelRare" group shows better concordance of the first hypothesis with 68.3% concordance, compared to 18.3% for "AccelRare", **60%** for "Physician alone", and **63.3%** for "Physician + AccelRare". In the "Physician + AccelRare" group, concordance rates are fairly similar (65% for GP vs. 63.3% for pediatricians) (Annex 1; Table 1.b).

Exact orientation with and without AccelRare

All Clinical Vignettes

The group "Physician then AccelRare" showed a tendency for better concordance in orientation, with 92.5% concordance compared to 76.7% for "Physician alone" and 91.7% for "Physician +

AccelRare". McNemar's test indicated significant differences between "Physician alone" and "Physician then AccelRare" ($p < 0.001$) and between "Physician + AccelRare" and "Physician alone" ($p = 0.001$) (Table 2).

		Physician alone	Physician then AccelRare	Physician with AccelRare
Concordance of orientation	N	120	120	120
	Yes	92 (76.7%)	111 (92.5%)	110 (91.7%)
	No	28 (23.3%)	9 (7.5%)	10 (8.3%)

Table 2 *: Proportion of concordant orientation (all clinical vignettes)*

Rare vignettes

When focusing on "Rare" diseases, there was a slight decrease in concordance across all groups, with "Physician then AccelRare" showing 90.4% concordance, compared to 70.2% for "Physician alone" and 89.4% for "Physician + AccelRare" (Annex 2; Table 2.a).

By speciality

In the GP group, the best concordance was observed in "Physician + AccelRare" and "Physician then AccelRare," with 90% concordance compared to 61.7% for "Physician alone". In the pediatrician group, "Physician + AccelRare" had the highest concordance with 93.3%, compared to 68.3% for "Physician then AccelRare" and 91.7% for "Physician alone." The concordance rates between "Physician alone" and "Physician + AccelRare" were not significantly different.
Concordance rates were relatively similar in the "Physician + AccelRare" group (90% for GP vs. 93.3% for pediatricians) (Annex 2; Table 2.b).

Secondary endpoints

Correct primary diagnoses (based on the top three diagnostic hypotheses)

There is a tendency for better concordance among the first three hypotheses in the "Physician + AccelRare" group, with 95% of at least one matching hypothesis compared to 54.2% for "AccelRare," 63.3% for "Physician only," and 85.8% for "Physician then AccelRare".

The McNemar test indicates that "Physician only" versus "Physician + AccelRare" produces significantly different results ($p < 0.001$). The same applies to "Physician followed by AccelRare" versus "Physician only" ($p < 0.001$) (Table 3).

Table 3 *: Proportion of concordance diagnoses among the first three hypotheses*

		Physician alone	**Physician then AccelRare**	**Physician with AccelRare**	**AccelRare**
At least one hypothesis concordant among the first three hypothesis	N	120	120	120	120
	Oui	76 (63.3%)	103 (85.8%)	114 (95%)	65 (54.2%)
	Non	44 (36.7%)	17 (14.2%)	6 (5%)	55 (45.8%)

When focusing exclusively on "Rare" diseases, the "Physician + AccelRare" group shows a better concordance for the first three hypotheses, with 93.6% of at least one matching hypothesis, compared to 53.2% for "Physician only," 67% for "AccelRare," and 81.9% for "Physician followed by AccelRare." Concordance rates are slightly lower in all groups when considering only rare diseases, except for the "AccelRare" group (Annex 3; Table 3.a).

When considering only GP groups, there is a trend toward better concordance in the "Physician + AccelRare" group, with 98.3% of at least one matching hypothesis, compared to 48.3% for "Physician only," 66.7% for "AccelRare," and 86.7% for "Physician followed by AccelRare." For Pediatrician groups, the "Physician + AccelRare" group shows a trend toward better concordance with 91.7% of at least one matching hypothesis, compared to 41.7% for "AccelRare," 78.3% for "Physician only," and 85% for "Physician followed by AccelRare."

When comparing GP and Pediatricians, the results indicate a tendency for better concordance among Pediatricians in the "Physician only" group (78.3% for Pediatricians vs. 48.3% for GP), and a tendency for better concordance among GP in the "Physician + AccelRare" group (98.3% for GP vs. 91.7% for Pediatricians) (Annex 3; Table 3.b).

Evolution of concordance in the group of "physicians alone" after using AccelRare

There was a tendency for a higher percentage of initial hypotheses to become concordant after using AccelRare, with 47.5% becoming concordant, compared to 23.3% of initially non-

concordant hypotheses that turned concordant, and 29.2% that remained non-concordant (Table 4).

		General Practitioners	Pediatricians	Total
Evolution of concordance	N	60	60	120
	Hypothesis non-concordant before, then non-concordant after AccelRare	16 (26.7%)	19 (31.7%)	35 (29.2%)
	Hypothesis concordant before, then concordant after AccelRare	21 (35%)	36 (60%)	57 (47.5%)
	Hypothesis non-concordant before, then concordant after AccelRare	23 (38.3%)	5 (8.3%)	28 (23.3%)
	Hypothesis concordant before, then non-concordant after AccelRare	0 (0%)	0 (0%)	0 (0%)

Table 4 *: Evolution of concordance in the group "Physician alone"*

When focusing only on "Rare" diseases, there was a tendency for more initial non-concordant hypotheses to remain non-concordant after using AccelRare, with 36.2%, compared to 28.7% of non-concordant hypotheses becoming concordant and 35.1% of concordant hypotheses remaining concordance.

For GP, 38.3% of hypotheses that were initially non-concordant became concordant after using AccelRare, while 26.7% remained non-concordant, and 35% of hypotheses that were initially concordant remained concordant.

For pediatricians, 60% of hypotheses that were initially concordant remained concordant after using AccelRare, whereas 8.3% of concordant hypotheses became non-concordant, and 31.7% of non-concordant hypotheses remained non-concordant.

Reproducibility and repeatability

The first hypothesis is reproducible in 35% of cases (40% for GP, 30% for pediatricians, and 40.4% for rare vignette cases).

At least one hypothesis is reproducible among the first three in 76.7% of cases overall, with 80.9% for rare vignette cases.

At least two hypotheses are reproducible among the first three in 40% of cases (43.3% for GP, 36.7% for pediatricians, and 40.4% for rare vignette cases).

The first three hypotheses are reproducible in 8.3% of cases (10% for GP, 6.7% for pediatricians, and 8.5% for rare vignette cases) (Figure 2).

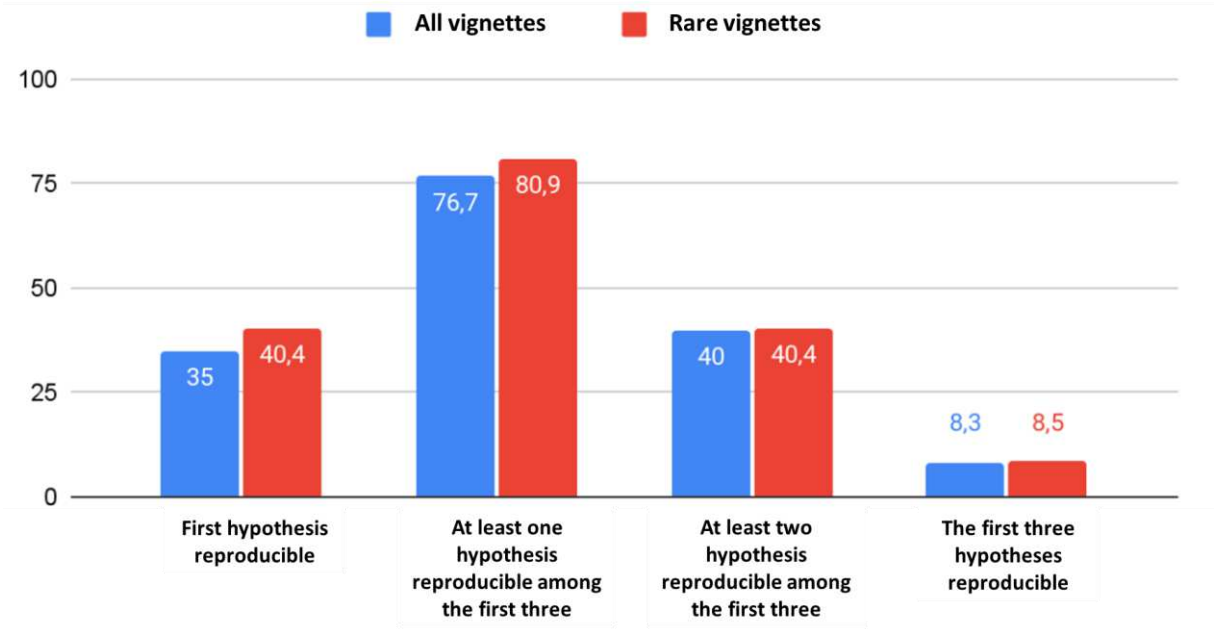

Figure 2 : *Percentage of reproducibility*

The first hypothesis is repeatable in 28.6% of cases (14.3% for GP, 42.9% for pediatricians, and 33.3% for rare vignette cases).

At least one hypothesis is repeatable among the first three in 78.6% of cases (57.1% for GP, 100% for pediatricians, and 75% for rare vignette cases).

At least two hypotheses are repeatable among the first three in 28.6% of cases (14.3% for GP, 42.9% for pediatricians, and 33.3% for rare vignette cases).

The first three hypotheses are repeatable in 7.1% of cases (0% for GP, 14.3% for pediatricians, and 8.3% for rare vignette cases) (Figure 3).

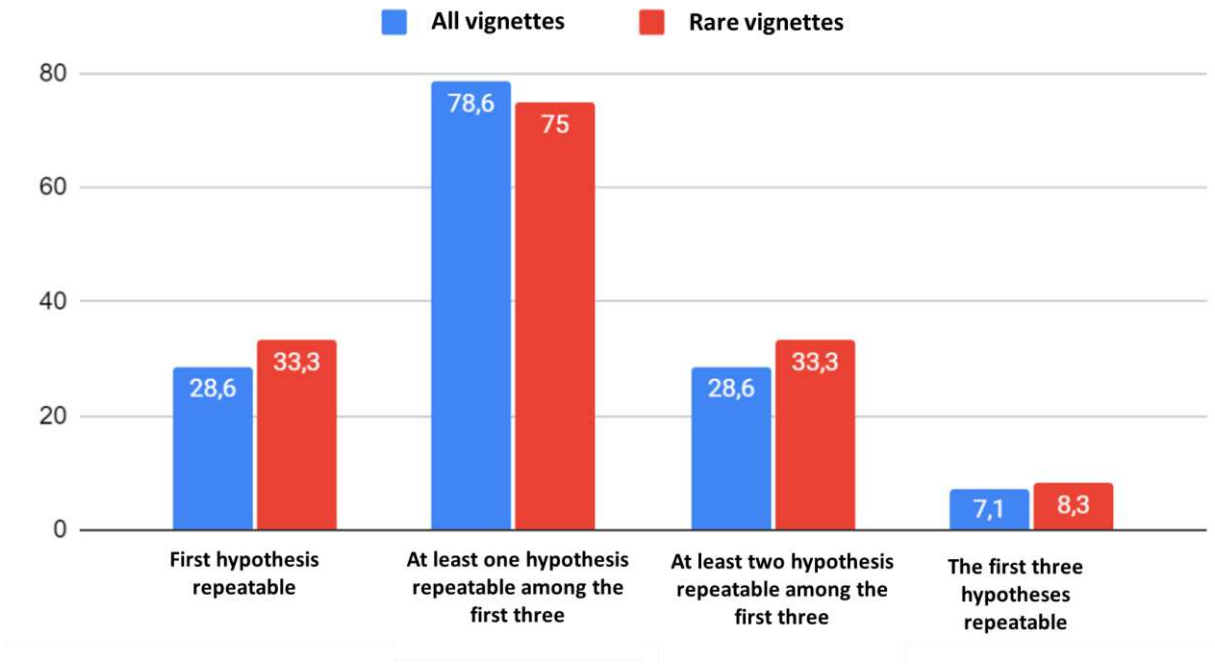

Figure 3 : *Percentage of repeatability*

Semiological Time and Total Decision-Making Time

The results show that both the semiological time and total decision-making time are significantly higher in the "Physician + AccelRare" arm compared to the "Physician alone" arm (6.2 ± 3.8 min vs 3.3 ± 2.4 min ; $p < 0.0001$). As for the total decision-making time, the average is 5.3 min in the "Physician alone" arm and 8.4 in the "Physician + AccelRare" arm. When focusing solely on rare diseases, these differences are even more pronounced, with averages of 6.9 vs. 3.9 for semiological time and 9.6 vs. 6.0 for total decision-making time in the "Physician + AccelRare" and "Physician alone" groups, respectively.

Satisfaction and Usability Ergonomics

User satisfaction with the tool was measured using the SUS score, which revealed an average score of 67.2, indicating overall good satisfaction. The scores were slightly higher among pediatricians (67.5) compared to GP (66.9), but the difference is minimal.

Regarding the use of interfaces, the results show that the number of interfaces used is significantly higher in the "Physician alone" arm (average of 0.9) compared to the "Physician + AccelRare" arm (average of 0.6). Among the most used interfaces, Google, Orphanet, and pediatric steps are the top choices, with differences in usage between the two arms.

Discussion

This clinical vignette simulation study evaluated the diagnostic performance of the AI tool AccelRare in the context of rare diseases. In the "Physician alone" group, diagnostic concordance for the primary hypothesis significantly improved, reaching 63.8% correct responses after AI utilization ($p<0.001$) (see Table 1). AccelRare facilitated pertinent diagnostic revisions in 31.7% of cases among GP and 26.7% among pediatricians, where clinicians modified their initial incorrect hypotheses following AI consultation (see Table 4). Importantly, AccelRare never directed users toward an erroneous diagnosis. The positive impact of AI on medical decision-making is further supported by similar diagnostic concordance results between the "Physician + AccelRare" group using the algorithm upfront and the "Physician alone" group, with no statistically significant differences observed, underscoring AccelRare's value as a second medical opinion. A study conducted in France across six emergency departments demonstrated that implementing regular cross-checks between two physicians reduced medical error rates by 40%, thereby enhancing diagnostic accuracy (11). While the benefits of second medical opinions and collaborative consultations among healthcare professionals in enhancing diagnostic accuracy are well-documented, there is a scarcity of data regarding the impact of physician and artificial intelligence (AI) collaborations on diagnostic efficiency.

In the realm of rare diseases, AccelRare could help reduce diagnostic delays through its consultative role. It is important to note, however, that AccelRare is a medical device intended to support physicians and is not designed for direct use by patients. Clarifying the intended use of this AI-based tool is essential, as its deployment among the general population could lead to an increased number of referrals to rare disease specialists, potentially overburdening specialized care pathways. Beyond rare diseases, AI assistance in medical diagnostics has been documented in various fields, incorporating decision support tools in medical imaging, dermatology, nephrology, and pediatrics. Initially focused on image analysis (radiology, pathology, dermatology), diagnostic models have expanded to clinical decision-making through algorithm training on clinical data. In several specialties, AI is already routinely used as a complement to physicians' clinical judgment in order to enhance diagnostic accuracy during consultations (12–15). For example, in oncology, Krakowski et al. demonstrated in a meta-analysis that the use of an image recognition-based AI improved the sensitivity of skin cancer diagnoses from 74.8% to 81.1%, and the specificity from 81.5% to 86.1% (16). In pulmonology, collaboration between pulmonologists and an AI system for interpreting pulmonary function tests led to a significant 10.4% increase in diagnostic accuracy. In these cases, AI appears to

refine clinical semiology, acting as a form of cognitive support that enhances clinical reasoning in the diagnostic process. In certain conditions, AI can even surpass clinicians, as evidenced by algorithms demonstrating higher sensitivity and specificity in diagnosing neonatal sepsis (17). This example illustrates how AI-based predictive models, developed from routinely collected clinical and EMR data, can outperform clinicians for the diagnosis of complex conditions highlighting the potential of AI to enhance clinical decision-making in critical care settings. In our study, the diagnostic concordance for the primary hypothesis—and across the three primary hypotheses—between physicians alone without AccelRare and the hypotheses generated exclusively by AccelRare alone demonstrated comparable performance (see Table 1). These advancements highlight the complementarity between AI and clinicians: while physicians remain essential for clinical data collection and initial hypothesis formulation, AI can refine diagnostic precision through hypothesis cross-referencing. However, the presence of biases in training data underscores the necessity for balanced data collection to ensure reliable results. From a technological standpoint, AccelRare is based on an expert system architecture, meaning its reasoning process relies on a weighted analysis of clinical features according to their prevalence within a given disease. The more frequently a clinical sign is associated with a specific condition—for instance, headache in giant cell arteritis—the higher its diagnostic weight. AccelRare also incorporates a boosting mechanism, whereby the presence of multiple co-occurring signs that are pathognomonic of a specific disease leads to an amplified cumulative weight. This expert system foundation provides a robust initial framework for AccelRare. However, future training of the algorithm on rare disease health data repositories will further enhance its performance, enabling more refined diagnostic support for rare diseases within the physician + AccelRare collaborative model (18).

The results concerning patient referrals to appropriate rare disease centers are comparable to diagnostic concordance in both groups. This similarity can be attributed to the fact that AccelRare lacks a specific patient referral module based on clinical features. Consequently, the investigators referred patients based on their diagnostic hypotheses. Therefore, physicians without AccelRare, exhibiting lower diagnostic performance, referred patients less appropriately.

A subgroup analysis based on the specialties of the investigator physicians revealed that pediatricians achieved the highest diagnostic and referral accuracy, regardless of the group or the use of AccelRare. This finding can be attributed to the longer duration of pediatric residency

training in France compared to GP, as well as greater exposure to rare diseases during their education.

This study evaluated the impact of AccelRare in assisting the diagnosis of rare diseases. However, the inclusion of clinical vignettes featuring common diseases (e.g., pyelonephritis, cholangitis) revealed that AccelRare does not effectively distinguish between rare and common diseases. Analysis of the AI-generated responses showed a diagnostic concordance rate of 36.2% for rare disease vignettes, which decreased to 29.2% when considering all vignettes. Notably, AccelRare exhibited a 100% error rate for common disease vignettes. The Medvir technology underlying AccelRare was originally developed as a symptom checker designed to assist with diagnostic support and patient triage in emergency care settings, primarily for common conditions. However, the emergency care data used to develop Medvir were not transferred to AccelRare. Although we intended to evaluate AccelRare using common disease vignettes, the system does not currently demonstrate adequate diagnostic performance for such conditions. Moreover, when tested with common diseases, AccelRare does not generate a warning indicating the absence of a rare disease ("no rare diseases detected." warning); instead, it may incorrectly suggest a rare disease diagnosis from its database of 272 conditions. In these situations, clinical reasoning by the physician remains essential. The collaboration between physicians and AI enabled the distinction between rare and common cases, with a negative predictive value (NPV) of 76% and a positive predictive value (PPV) of 99%. Moreover, the significant improvement in diagnostic concordance rates after using AccelRare across all vignettes (70.8% for the Physician alone group and 64.2% for the Physician + AccelRare group) underscores the imperative that AccelRare's use be reserved for healthcare professionals rather than patients.

Another area for improvement concerns the usability of the AccelRare interface. Currently offered as a web application, symptom entry was reported by investigators as only moderately ergonomic, with potential for errors in symptom selection. Enhancing the interface’s usability, particularly through improved classification of symptoms and syndromes, could strengthen the repeatability and reproducibility of the tool (repeatability for the top diagnostic hypothesis in rare disease vignettes was 33.3%, and reproducibility for the top diagnostic hypothesis across investigators was 40,4%.).

Several limitations of this pilot study must be acknowledged. First, the study was based on a simulation model using clinical vignettes, which may not fully capture the complexity and

variability of real-life clinical encounters. Additionally, the influence of other factors - such as physicians' level of experience or their familiarity with AI-based tools - warrants further investigation in future studies. In real-world settings, the time required to use AccelRare may be significantly longer, as the clinical case would not be pre-summarized, requiring physicians to perform an additional cognitive task of data synthesis and potentially interact with the patient during tool use. A real-world evaluation of AccelRare would therefore be of particular interest.

Despite these limitations, the results are encouraging, as the physician + AccelRare pairing improved both diagnostic accuracy and patient triage in suspected rare disease cases among GP and primary care pediatricians. Ultimately, AccelRare could serve as a complementary support tool for rare disease specialists, and future evaluation of its performance in internal medicine or rheumatology departments could be particularly valuable - especially in helping to differentiate between functional disorders and rare diseases in patients experiencing diagnostic uncertainty.

Conclusions

These findings support the relevance of a collaborative approach between artificial intelligence and medical expertise. While AI alone does not outperform clinicians in diagnostic performance, its integration as a decision-support tool significantly enhances diagnostic accuracy and patient triage. This synergy highlights the need to develop tailored implementation strategies aimed at optimizing the acceptance and effective use of AI tools by healthcare professionals.

Appendices

Appendix 1: Correct diagnoses

		Physician alone	Physician then AccelRare	Physician with AccelRare	AccelRare
Concordance of 1st hypothesis	N	94	94	94	94
	Oui	33 (35.1%)	60 (63.8%)	55 (58.5%)	34 (36.2%)
	Non	61 (64.9%)	34 (36.2%)	39 (41.5%)	60 (63.8%)

Table 1.a *: Proportion of concordance diagnoses (rare diseases)*

		Physician alone		Physician then AccelRare		Physician with AccelRare		AccelRare	
		GP	Ped	GP	Ped	GP	Ped	GP	Ped
Concordance of 1st hypothesis	N	60	60	60	60	60	60	60	60
	Yes	21 (35%)	36 (60%)	44 (73.3%)	41 (68.3%)	39 (65%)	38 (63.3%)	24 (40%)	11 (18.3%)
	No	39 (65%)	24 (40%)	16 (26.7%)	19 (31.7%)	21 (35%)	22 (36.7%)	36 (60%)	49 (81.7%)

Table 1.b *: Proportion of concordance diagnoses (according to speciality)*

Appendix 2: Exact orientation

		Physician alone	Physician then AccelRare	Physician with AccelRare
Concordance of orientation	N	94	94	94
	Oui	66 (70.2%)	85 (90.4%)	84 (89.4%)
	Non	28 (29.8%)	9 (9.6%)	10 (10.6%)

Table 2.a *: Proportion of concordance orientation (rare diseases)*

		Physician alone		Physician then AccelRare		Physician with AccelRare	
		GP	**Ped**	**GP**	**Ped**	**GP**	**Ped**
Concordance of orientation	N	60	60	60	60	60	60
	Yes	37 (61.7%)	55 (91.7%)	54 (90%)	57 (95%)	54 (90%)	56 (93.3%)
	No	23 (38.3%)	5 (8.3%)	6 (10%)	3 (5%)	6 (10%)	4 (6.7%)

Table 2.b *: Proportion of concordance orientation (according to speciality)*

Appendix 3: First three hypothesis

Table 3.a *: Proportion of concordance diagnoses among the first three hypotheses (rare vignettes)*

		Physician alone	Physician then AccelRare	Physician with AccelRare	AccelRare
At least one hypothesis concordant among the first three hypothesis - Rare vignettes	N	94	94	94	94
	Oui	50 (53.2%)	77 (81.9%)	88 (93.6%)	63 (67%)
	Non	44 (46.8%)	17 (18.1%)	6 (6.4%)	31 (33%)

		Physician alone		Physician then AccelRare		Physician with AccelRare		AccelRare	
		GP	**Ped**	**GP**	**Ped**	**GP**	**Ped**	**GP**	**Ped**
At least one hypothesis concordant among the first three hypothesis	N	60	60	60	60	60	60	60	60
	Yes	29 (48.3%)	47 (78.3%)	52 (86.7%)	51 (85%)	59 (98.3%)	55 (91.7%)	40 (66.7%)	25 (41.7%)
	No	31 (51.7%)	13 (21.7%)	8 (13.3%)	9 (15%)	1 (1.7%)	5 (8.3%)	20 (33.3%)	35 (58.3%)

Table 3.b *: Proportion of concordance diagnoses among the first three hypotheses (by speciality)*

Declarations

Ethics approval and consent to participate

As this research was conducted within the context of programme assessment, it was exempted from review by our institutional review board. As the questionnaire did not collect any personal data, the European Union General Data Protection Regulation (GDPR) did not apply. All participants gave their consent to participate in this work.

Human Ethics and Consent to Participate declarations:

This study did not involve patients. All clinical cases were fictional vignettes written by the investigators; although informed by clinical experience, they were composite cases containing no identifiable patient information, and no patient data were collected. The only human participants were the eight physicians (four general practitioners and four primary care paediatricians) whose diagnostic performance was assessed.

Under French law, this study does not qualify as research involving the human person (recherche impliquant la personne humaine) as defined in Articles L.1121-1 and R.1121-1 of the French Public Health Code (Code de la santé publique): it was not conducted on patients or healthy volunteers with the aim of developing biological or medical knowledge, but aimed to evaluate the practice of healthcare professionals (Article R.1121-1, II, 2°). Accordingly, review by a Committee for the Protection of Persons (Comité de Protection des Personnes) was not required.

All participating physicians were informed of the objectives and procedures of the study, and written informed consent to participate was obtained from each of them prior to inclusion. Participation was voluntary. Participant responses were recorded in a pseudonymised electronic case report form and processed in accordance with the EU General Data Protection Regulation (Regulation (EU) 2016/679).

The study was conducted in accordance with the principles of the Declaration of Helsinki and with relevant national guidelines and regulations.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analysed during the current study (pseudonymised investigator responses), as well as the clinical vignettes and their reference diagnoses, are available from the corresponding author on reasonable request. The AccelRare algorithm is proprietary and is not publicly available; the tool is freely accessible online to healthcare professionals in France.

Competing interests :

This study was funded by Sanofi. EVDE and ELPF are affiliated with Sanofi (Department of Innovation). LE developed the medical algorithm underlying AccelRare and is affiliated with Medvir. LM, JS and JG are affiliated with Digital Medical Hub, which received funding from Sanofi to conduct this study. VH, AM, LD and HdB declare that they have no competing interests.

Funding :

This study was funded by Sanofi. Two co-authors affiliated with Sanofi (EVDE, ELPF) contributed to improving the performance of the AccelRare tool and reviewed the manuscript. The funder had no role in the writing of the clinical vignettes, the recruitment of investigators, data collection, statistical analysis or interpretation of the results.

Authors' contributions :

JG was responsible for writing the introduction, discussion, and conclusion sections, as well as designing the research protocol. LM co-authored the methods and results sections. VH, AM, LD, and HDB drafted and reviewed the rare disease vignettes. EVDE and ELPF are members of Sanofi's scientific committee, which independently supported this research and contributed to improving the performance of AccelRare. JS reviewed and edited the manuscript for submission. LE was the physician who developed the medical algorithm behind AccelRare. All authors read and approved the final manuscript

Acknowledgements :

The authors wish to thank the general practitioners and pediatricians who participated in the clinical investigation with the seriousness and rigor required. We also acknowledge the contributions of the clinical research associates from the Digital Medical Hub: Lucas Thiery, Mathieu Grajoszex, Jean-Edouard Monnerville, Emmanuel Stranadica, and Julie Mataranga.

600

References

1. Filière de santé maladies rares. Rapport d'activité - Filière maladies Rares 2021 [Internet]. 2022 [cité 5 oct 2023]. Disponible sur: sante.gouv.fr/IMG/pdf/rapport_d_activite_des_filieres_de_sante_maladies_rares_2021.pdf

2. Erradiag-l-errance-diagnostic-dans-les-maladies-rares1.pdf [Internet]. [cité 6 oct 2023]. Disponible sur: https://www.alliance-maladies-rares.org/wp-content/uploads/2020/05/Erradiag-l-errance-diagnostic-dans-les-maladies-rares1.pdf

3. DGOS. Ministère du Travail, de la Santé, des Solidarités et des Familles. [cité 31 mars 2025]. Les maladies rares. Disponible sur: https://sante.gouv.fr/soins-et-maladies/prises-en-charge-specialisees/maladies-rares/article/les-maladies-rares

4. pnmr4.pdf [Internet]. [cité 31 mars 2025]. Disponible sur: https://sante.gouv.fr/IMG/pdf/pnmr4.pdf

5. Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, et al. A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. Front Artif Intell. 2020;3:543405.

6. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med. 17 août 2022;5(1):118.

7. Cross S, Mourad A, Zuccon G, Koopman B. Search Engines vs. Symptom Checkers: A Comparison of their Effectiveness for Online Health Advice. In: Proceedings of the Web Conference 2021 [Internet]. New York, NY, USA: Association for Computing Machinery; 2021 [cité 20 déc 2024]. p. 206-16. (WWW '21). Disponible sur: https://doi.org/10.1145/3442381.3450140

8. Fritsche G, Peabody J. Methods to improve quality performance at scale in lower- and middle-income countries. J Glob Health. 8(2):021002.

9. Benarama A, Prigent E, Martin JC. Designing Interactions with an Alzheimer Virtual Patient for Caregiver Training [Internet]. LISN | Laboratoire Interdisciplinaire des Sciences du Numérique. 2023 [cité 20 déc 2024]. Disponible sur: https://www.lisn.upsaclay.fr/publications/designing-interactions-with-an-alzheimer-virtual-patient-for-caregiver-training/?lang=en

10. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Sharabiani MTA, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. 1 avr 2022 [cité 20 déc 2024]; Disponible sur: https://bmjopen.bmj.com/content/12/4/e053566.long

11. Freund Y, Goulet H, Leblanc J, Bokobza J, Ray P, Maignan M, et al. Effect of Systematic Physician Cross-checking on Reducing Adverse Events in the Emergency Department: The CHARMED Cluster Randomized Trial. JAMA Intern Med. 1 juin 2018;178(6):812-9.

12. Schena FP, Anelli VW, Abbrescia DI, Di Noia T. Prediction of chronic kidney disease and its progression by artificial intelligence algorithms. J Nephrol. nov

2022;35(8):1953-71.

13. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MSB, Maqsood A, et al. Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry-A Systematic Review. BioMed Res Int. 2021;2021:9751564.

14. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial Intelligence in Cardiology: Present and Future. Mayo Clin Proc. mai 2020;95(5):1015-39.

15. Cesaretti M, Izzo A, Mavrothalassitis O, Pellegrino RA. Artificial intelligence for liver diseases: The urgency of collaboration. Dig Liver Dis. 1 juin 2024;56(6):1110-1.

16. Krakowski I, Kim J, Cai ZR, Daneshjou R, Lapins J, Eriksson H, et al. Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. Npj Digit Med. 9 avr 2024;7(1):1-10.

17. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc. 1 mars 2014;21(2):326-36.

18. Fong S, Fortino G, Ghista D, Piccialli F. Special issue on deep learning and big data analytics for medical e-diagnosis/AI-based e-diagnosis. Neural Comput Appl. 1 août 2023;35(22):15961-5.