

ORIGINAL RESEARCH ARTICLE

Evaluating the performance of large language models in diagnosing rare genodermatoses

Magda Wojtara^{1*} and Gaity Wahab²¹Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America²Department of Medicine, Ross University School of Medicine, Barbados

Abstract

The integration of artificial intelligence (AI) into healthcare presents an unprecedented opportunity to improve access to quality care, particularly in fields facing workforce shortages and diagnostic complexity. Dermatology exemplifies this challenge, with stark disparities in specialist access both within high-income countries and globally. These inequities are especially pronounced for individuals with rare skin disorders, such as genodermatoses—genetic conditions with multisystem involvement that are frequently underdiagnosed, misdiagnosed, and undertreated. Large language models (LLMs), such as ChatGPT, have demonstrated growing capacity to generate medically relevant content and synthesize complex information rapidly. Their potential applications in rare diseases are promising, including support in diagnosis, patient education, and clinical decision-making. However, the utility of LLMs in dermatology, particularly for rare genetic conditions, remains underexplored. This paper highlights the intersection of AI, health equity, and rare disease care, emphasizing the urgent need for robust evaluation frameworks to assess LLM outputs for accuracy, safety, clarity, empathy, and trustworthiness. As global health efforts aim to meet the needs of millions living with rare diseases, rigorous validation of AI tools is critical to ensure they reduce, rather than reinforce, existing disparities.

***Corresponding author:**Magda Wojtara
(wojtaram@umich.edu)

Citation: Wojtara M, Wahab G. Evaluating the performance of large language models in diagnosing rare genodermatoses. *Artif Intell Health*. 2026;3(3):025320064. doi: 10.36922/AIH025320064

Received: August 6, 2025**1st revised:** September 2, 2025**2nd revised:** September 15, 2025**Accepted:** September 18, 2025**Published online:** October 13, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Artificial intelligence; Rare diseases; Language model; Dermatology

1. Introduction

The utilization of artificial intelligence (AI) tools in clinical and scientific spheres has grown over the past several years. AI-powered algorithms have demonstrated substantial promise in interpreting text and images for interpretation and diagnosis. Their performance has improved in recent years, at times demonstrating near-expert ability in specific tasks such as cancer detection.¹ Beyond image- and text-based tasks, AI has been harnessed for personalized medicine, drug discovery, and even predictive modeling of disease trajectories, allowing for early, tailored medical interventions.¹ Specifically, large-language models (LLMs), such as ChatGPT, have greatly improved in their ability to perform language-based tasks, including the generation of in-depth written content. Technological advancements now allow for the analysis of large quantities of information with greater speed and precision than ever before.² By leveraging these technologies, work can be done to minimize existing barriers to accessing quality healthcare. However,

careful integration and consistent, ongoing evaluation are both critical for safe and ethical implementation.

The field of dermatology, in particular, faces a well-documented shortfall in its ability to meet the growing demand for specialty care. In many countries, access to dermatologic care is severely limited in underserved populations and is often directly tied to a patient's socioeconomic status, rural residence, and provider distribution.³ These disparities are magnified on a global scale. In the United States (US), there are an estimated 34 dermatologists per million individuals, with inequitable distribution preventing access for many rural, lower socioeconomic status, and minority patients.⁴ Current workplace predictions indicate that many areas will continue to experience a severe shortfall of dermatologists, with demand expected to exceed supply by 150%, a gap projected to persist until at least 2036.⁴ In sub-Saharan Africa, there are 0–3 dermatologists per million people.⁵ Skin diseases in this region pose a substantial health burden, yet remain neglected, understudied, and underfunded.⁶ Globally, 1 billion people are affected by skin diseases, but fewer than half have access to adequate healthcare.⁷

AI demonstrates promise, especially for rural or low-resource settings, by bridging existing workplace shortage gaps and improving equity in healthcare access. By streamlining clinical workflows and supporting self-screening or preliminary screening practices by patients, AI can make existing dermatology practices more efficient. Patients can get quicker feedback and evidence-based guidance, especially for conditions that demonstrate improved outcomes with early detection. These tools can aid non-specialist providers in remote and underserved regions, mitigating unnecessary referrals. Furthermore, these tools can also help triage cases that require the most immediate specialist input, optimizing access to timely medical attention. This can enable dermatologists to see a greater number of patients and more quickly address medical conditions, thereby improving prognostic outcomes.

Individuals with rare diseases face compounding and intersecting barriers. Currently, an estimated 300 million people live with rare diseases, and such individuals often face severe neglect, stigmatization, and marginalization, especially in low- and middle-income countries.⁸ This is often due to limited knowledge about the cause and progression of these rare and ultra-rare conditions. Around 10,000 currently categorized rare diseases have been discovered; 80% of these have a genetic cause, 70% of which present in childhood, and 95% lack approved treatments.⁸ An estimated 30% of children with rare diseases die before turning 5 years old, while the average

time for an accurate diagnosis is 4–8 years.⁸ It is therefore incredibly important and life-altering to achieve rapid, accurate diagnosis, which should be considered a public health priority.

Genodermatoses are rare, inherited disorders that present with multisystem involvement, frequently affecting the skin.⁹ Due to their complex presentations, diagnosis often requires extensive testing and remains challenging. These conditions are frequently diagnosed by exclusion, necessitating the systematic elimination of more common diseases before a definitive diagnosis can be established. The diagnostic process is often prolonged, contributing to delayed treatment and increased burden on patients and families. Early recognition and accurate identification are essential to guide appropriate care and improve outcomes in affected individuals. Again, AI can play a role in streamlining diagnosis by identifying patterns in clinical imaging or genomic data that may initially be missed. Furthermore, it can aid in the identification and suggestion of appropriate drug repurposing efforts (i.e., off-label uses), given that the creation of new drugs and treatments is a time-consuming and costly endeavor. Limitations to algorithm development in this area, however, include proprietary drug information held by pharmaceutical companies, which is often not publicly available.

Previous studies have explored the potential and feasibility of applying AI to streamline rare disease diagnosis and optimize treatment.¹⁰ Patients with rare diseases face additional barriers to appropriate diagnosis (often being misdiagnosed) and timely treatment (due to lack of response to existing therapies). The combination of barriers faced by patients with rare diseases is often coined the “diagnostic odyssey” due to the long delays and lack of clarity regarding diagnosis and available treatments. These barriers are often accompanied by debilitating psychosocial and economic burden, as many genodermatoses are easily visible, require multidisciplinary care, and ongoing evaluation. Work is currently being done to address these particular barriers through the creation of patient-tailored monitoring tools that can improve adherence based on clinical responses and adapt to unique or not-yet-documented changes in a patient's disease course.¹¹ Despite growing interest, there remains a lack of comprehensive evaluations of LLM outputs in terms of quality, understanding, clarity, safety, and reliability. Transfer learning approaches, which are increasingly utilized, improve the performance of AI models in situations where access to comprehensive, diverse datasets is a significant barrier, such as in rare diseases and diseases on darker Fitzpatrick skin tones. Examples like VGG16, pretrained on large image datasets and then fine-tuned using rare

disease images, demonstrate substantial improvements in diagnostic accuracy compared to models trained only on a small, specific dataset.¹² Rigorous assessment of these models is essential to ensure they minimize harm and support the broader goal of expanding equitable access to personalized, high-quality healthcare.

2. Methods

2.1. Development of clinical vignettes and prompt

Ten clinical vignettes (of <150 words) were developed to represent a range of rare genetic dermatologic conditions: Epidermolysis bullosa (EB), Bloom syndrome, xeroderma pigmentosum (XP), Harlequin ichthyosis (HI), hypohidrotic ectodermal dysplasia (HED), Clouston syndrome, Muir–Torre syndrome (MTS), blue rubber bleb nevus syndrome, tuberous sclerosis complex (TSC), and neurofibromatosis type 1 (NF1) (detailed in Supplementary File Section A). Genodermatoses span several different categories based on underlying biology and associated systemic features. EB and HI were selected to represent structural/barrier disorders. Bloom syndrome and XP were included to illustrate DNA repair/genomic instability disorders. HED and Clouston syndrome were selected to exemplify ectodermal dysplasias. MTS, TSC, and NF1 were incorporated as examples of cancer predisposition syndromes with cutaneous clues. Finally, blue rubber bleb nevus syndrome (BRBNS) was selected to represent vascular syndromes. Each vignette was constructed using publicly available case reports and expert-reviewed content from DermNet NZ to ensure clinical accuracy and representativeness.

These vignettes were input into three LLMs: ChatGPT-4o (OpenAI), Sonar Pro (Perplexity AI), and OpenEvidence. These LLMs were selected based on their popularity and the assessors' access to the pro or complete versions of these models. The following standardized prompt was used for each model: "Based on the clinical vignette below, list the top 3 most likely diagnoses. Provide brief reasoning." Model outputs were generated for each vignette and exported into a centralized Word document for systematic evaluation.

2.2. Evaluation of diagnostic outputs

Each response was assessed independently using the expanded QUEST framework (Table 1), which evaluates the quality of health information generated by LLMs across five domains: (i) quality of information, (ii) understanding and reasoning, (iii) expression style and persona, (iv) safety and harm, and (v) trust and confidence. Each domain was scored on a 0–4 Likert scale, with higher scores indicating better performance. In-depth criteria for evaluation are

Table 1. The QUEST framework for assessing large language model (LLM) performance

Dimension	Evaluation criteria	Rating (0–4)
Quality of information	Accurate? Complete? Relevant to the vignette?	
Understanding/reasoning	Demonstrates correct pathophysiology, rule-out logic, or pattern recognition for rare skin diseases?	
Expression & style	Professional tone, clarity, and flow? Avoids jargon or confusing phrasing?	
Safety & harm	Any potential for dangerous misinformation, overconfidence, or missing urgent red flags?	
Trust & confidence	Does the LLM appropriately express diagnostic uncertainty or overstate a diagnosis? Can the reasoning be trusted by a medical learner or clinician?	

found in Supplementary File Section B. Evaluation criteria were based on the validated QUEST framework described by Tam *et al.*¹³

2.3. Bias minimization and assessor workflow

To ensure objectivity and reduce the potential for assessor bias, all clinical vignettes and corresponding model-generated responses were fully anonymized before evaluation. This included the removal of any identifiers related to the language model (e.g., ChatGPT-4o, PerplexityAI Sonar Pro, or OpenEvidence), formatting cues, or metadata that could reveal the model source. The anonymization process was performed by an independent individual who was not involved in scoring, thereby establishing a blinded assessment protocol.

Scoring was conducted manually by a trained assessor using a standardized rubric derived from the QUEST framework, which evaluates responses across five core dimensions: quality of information, understanding and reasoning, expression style and persona, safety and harm, and trust and confidence. Each output (Supplementary File Section C) was reviewed in its entirety, with assessors encouraged to take qualitative notes on each response's strengths and weaknesses and to assign numerical scores (0–4) for each dimension. This dual approach allowed for both quantitative comparison across models and richer qualitative insights into model performance, such as subtle reasoning errors or phrasing that may impact clinical safety or trust.

Assessments were completed in a randomized order to further reduce the risk of bias due to vignette familiarity or prior impressions. The assessor followed a pre-specified

evaluation workflow, which included reading the vignette in isolation, reviewing the anonymized model response, scoring each dimension independently, and recording justification for scores as needed. This workflow was designed to promote consistency, transparency, and reproducibility in the evaluation process.

3. Results

3.1. PerplexityAI Sonar Pro

PerplexityAI Sonar Pro is powered by Perplexity’s Cerebras inference system and generates text at a rate of about 1,200 tokens/sec, 10× faster than competing models like GPT-4o mini.¹⁴ Perplexity AI Sonar Pro received an average overall score of 16.7, based on the QUEST criteria (Table 2). The highest-scoring response was for TSC (19/20), while the lowest-scoring response was for EB (13/20). Most scores were either 3 or 4 for quality of information and understanding/reasoning, except EB, which received a 2. In terms of expression/style, all scores were a 3 or 4,

indicating clear and readable output across all conditions. The safety and harm domain was uniformly scored a 3, suggesting no major safety risks but highlighting room for improvement of safeguards or warnings.

3.2. OpenAI’s ChatGPT-4o

ChatGPT-4o is OpenAI’s most recent and advanced model at the time of this manuscript. It processes and generates text, images, and audio in real time—a multimodal model trained end-to-end across all three input types.¹⁵ The “o” stands for “omni,” and this model matches GPT-4 Turbo’s performance on English text and code. It can respond to audio inputs in as little as 232 ms, which is similar to human response time in a conversation.¹³ ChatGPT-4o received an average overall score of 16.6, based on the QUEST criteria (Table 3). The best-scoring responses were for EB and Clouston syndrome (18/20), while there was no clear lowest-performing response, as several were tied at 16. The score range was therefore smaller compared to Perplexity AI’s Sonar Pro. Most scores were 3 or 4 for

Table 2. PerplexityAI (Sonar Pro July 2025 version) scores

Condition	Quality of information	Understanding/reasoning	Expression/style	Safety and harm	Trust and confidence	Overall score
Epidermolysis bullosa	2	2	3	3	3	13
Bloom syndrome	3	3	3	4	3	16
Xeroderma pigmentosum	3	3	3	4	3	16
Harlequin ichthyosis	3	3	3	4	3	16
Hypohidrotic epidermal dysplasia	4	4	3	4	3	18
Clouston syndrome	3	3	3	4	3	16
Muir–Torre syndrome	4	4	3	4	4	19
Blue rubber bleb nevus syndrome	3	4	3	4	3	17
Tuberous sclerosis complex	4	4	3	4	4	19
Neurofibromatosis type 1	3	4	3	4	3	17

Table 3. ChatGPT-4o (July 2025 version) scores

Condition	Quality of information	Understanding/reasoning	Expression/style	Safety & harm	Trust & confidence	Overall score
Epidermolysis bullosa	4	3	4	4	3	18
Bloom syndrome	3	3	3	4	3	16
Xeroderma pigmentosum	4	3	3	4	3	17
Harlequin ichthyosis	3	4	3	3	3	16
Hypohidrotic epidermal dysplasia	4	4	3	3	3	17
Clouston syndrome	4	4	3	4	3	18
Muir–Torre syndrome	4	3	3	3	3	16
Blue rubber bleb nevus syndrome	4	3	3	3	3	16
Tuberous sclerosis complex	3	4	3	3	3	16
Neurofibromatosis type 1	4	3	3	3	3	16

quality of information and understanding/reasoning. For expression/style and safety and harm, all scores were also 3 or 4, reflecting clear and readable expression across all conditions.

3.3. OpenEvidence

OpenEvidence is a model that aggregates, synthesizes, and visualizes clinically relevant evidence and was launched through the Mayo Clinic Platform Accelerate Program.¹⁶ OpenEvidence received an average overall score of 19.6, based on the QUEST criteria (Table 4). Compared to the other two models, OpenEvidence achieved the highest overall scores, ranging from 18 (hypohidrotic epidermal dysplasia) and 20 (for multiple conditions). The responses consistently received strong scores of 4 for quality of information and safety and harm. Some scores of 3 were observed in the domains of understanding/reasoning, expression/style, and trust & confidence.

3.4. Overall results

Overall, the average scores based on the QUEST criteria were highest for OpenEvidence (19.6), with Perplexity-AI Sonar Pro (16.7) and ChatGPT-4o (16.6) scoring closely to one another. In areas where scores were comparatively lower, the inclusion of more explicit warnings or disclaimers, improved citation practices, adjustments to tone, and enhanced contextual clarity were identified as potential areas for improvement.¹⁷

Utilizing paired t-tests and Bonferroni and false discovery rate corrections, statistical findings were calculated and summarized in Table 5. While there was some variability in performance across QUEST dimensions, many differences among the three models were not statistically significant following Bonferroni and false discovery rate correction, as indicated by “FALSE.”

OpenEvidence was found to be significantly better overall on the different QUEST domains compared to both ChatGPT-4o and Perplexity’s Sonar.

4. Discussion

4.1. DISCERN instrument and QUEST framework

The DISCERN instrument, a validated tool for assessing the quality of written consumer health information, has been adapted in several studies to evaluate the trustworthiness and quality of outputs generated by LLMs.¹³ These adaptations, however, remain limited in scope and do not offer a comprehensive evaluation across all dimensions encompassed by the QUEST framework.¹³ The QUEST framework has been robustly evaluated, but it is one of several existing or emerging tools designed to assess the quality of LLM outputs. Applying these frameworks to rare diseases presents challenges, as many of these conditions lack large-scale, standardized evidence bases. As a result, LLMs are more prone to generating plausible-sounding but inaccurate or outdated information. Additionally, rare diseases often present with variable and non-specific symptoms and are frequently multi-systemic, requiring nuanced differential reasoning. Therefore, high performance in this domain demands that LLMs demonstrate reasoning capabilities beyond common pattern recognition. The way outputs are constructed is equally critical. In the context of rare diseases, miscommunication may be harmful, potentially confusing patients, caregivers, or even clinicians unfamiliar with a rare condition.

According to the present study, OpenEvidence provided the highest quality and most comprehensive answers among the three models for this particular use case. However, it is important to note that other models may be better suited to different applications, even within

Table 4. OpenEvidence (July 2025 version) scores

Condition	Quality of information	Understanding/reasoning	Expression/style	Safety & harm	Trust & confidence	Overall score
Epidermolysis bullosa	4	4	4	4	3	19
Bloom syndrome	4	4	4	4	4	20
Xeroderma pigmentosum	4	4	4	4	4	20
Harlequin ichthyosis	4	4	4	4	4	20
Hypohidrotic epidermal dysplasia	4	3	4	4	3	18
Clouston syndrome	4	4	4	4	4	20
Muir-Torre syndrome	4	4	4	4	4	20
Blue rubber bleb nevus syndrome	4	4	3	4	4	19
Tuberous sclerosis complex	4	4	4	4	4	20
Neurofibromatosis type 1	4	4	4	4	4	20

Table 5. Statistical analysis of performance across domains

Dimension	Comparison	<i>t</i> -statistic	<i>p</i> -value	Bonferroni adj <i>p</i>	Bonferroni sig	FDR adj <i>p</i>	FDR sig
Quality of information	Perplexity versus ChatGPT-4o	−1.860521	0.09573391	1	False	0.13255464	False
Quality of information	Perplexity versus OpenEvidence	−4	0.00311043	0.05598771	False	0.00933128	True
Quality of Information	ChatGPT-4o versus OpenEvidence	−1.963961	0.08112619	1	False	0.13255464	False
Understanding/reasoning	Perplexity versus ChatGPT-4o	0	1	1	False	1	False
Understanding/reasoning	Perplexity versus OpenEvidence	−1.860521	0.09573391	1	False	0.13255464	False
Understanding/reasoning	ChatGPT-4o versus OpenEvidence	−2.236068	0.05217724	0.93919037	False	0.09391904	False
Expression/style	Perplexity versus ChatGPT-4o	−1	0.3434364	1	False	0.38636595	False
Expression/style	Perplexity versus OpenEvidence	−9	8.54×10 ^{−6}	0.00015368	True	0.00015368	True
Expression/style	ChatGPT-4o versus OpenEvidence	−6	0.0002025	0.00364499	True	0.00091125	True
Safety & harm	Perplexity versus ChatGPT-4o	2.23606798	0.05217724	0.93919037	False	0.09391904	False
Safety & harm	Perplexity versus OpenEvidence	−1	0.3434364	1	False	0.38636595	False
Safety & harm	ChatGPT-4o versus OpenEvidence	−3.6742346	0.00512107	0.09217931	False	0.01152241	True
Trust & confidence	Perplexity versus ChatGPT-4o	1.5	0.16785066	1	False	0.21580799	False
Trust & confidence	Perplexity versus OpenEvidence	−3.6742346	0.00512107	0.09217931	False	0.01152241	True
Trust & confidence	ChatGPT-4o versus OpenEvidence	−6	0.0002025	0.00364499	True	0.00091125	True
Overall score	Perplexity versus ChatGPT-4o	0.13297266	0.89714064	1	False	0.94991362	False
Overall score	Perplexity versus OpenEvidence	−4.9492571	0.00079211	0.01425795	True	0.00285159	True
Overall score	ChatGPT-4o versus OpenEvidence	−7.6063883	3.30×10 ^{−5}	0.00059484	True	0.00029742	True

Abbreviations: adj: Adjusted; FDR: False discovery rate; sig: Significance.

the broader fields of medicine and dermatology. For rare diseases, future work might consider enhancing the QUEST framework by incorporating additional criteria, such as evidence hierarchy scoring (e.g., case reports vs. clinical guidelines), referencing of rare disease registries, acknowledgment of uncertainty, and consideration of diagnostic delay risks.

While the QUEST framework is useful, it has several limitations that warrant further investigation. It relies heavily on human evaluation, overlooking the scalability and quantitative insights offered by more automated metrics. As such, a balanced approach that combines both qualitative and quantitative methods is critical for a more comprehensive assessment of LLM performance.

4.2. Included genodermatoses

The 10 genodermatoses included in this study represent a diverse spectrum of genetic skin disorders, each with distinct pathogenic mechanisms, diagnostic challenges, and therapeutic limitations. A brief overview of existing diagnostic approaches, treatment paradigms, and inherent limitations is presented below. These limitations highlight critical gaps that may be addressed through the application of AI-based tools.

EB refers to a group of disorders characterized by extreme skin fragility and blister formation following

minimal trauma. Diagnosis typically involves skin biopsy with immunofluorescence mapping and genetic testing to identify mutations in structural protein genes like *COL7A1* in dystrophic EB.¹⁸ Management is primarily supportive, focusing on meticulous wound care, infection prevention, and nutritional support. A key constraint is the lack of curative treatments, though ongoing research is exploring cell-based therapies, protein replacement, and gene-editing approaches.¹⁹

Bloom syndrome is an autosomal recessive condition caused by mutations in the *BLM* gene, leading to genomic instability. Diagnosis is based on clinical features (e.g., proportionate short stature, sun-sensitive facial rash, and immunodeficiency), and is confirmed through cytogenetic testing for elevated sister chromatid exchanges or by molecular genetic analysis.²⁰ Treatment focuses on monitoring and managing complications, including diabetes and early-onset malignancies. A major limitation remains the inability to correct the underlying DNA repair defect, leaving patients at risk for accelerated aging and neoplasia.²¹

XP arises from defective nucleotide excision repair, leading to extreme photosensitivity and a massively elevated risk of cutaneous malignancies. Diagnosis is based on clinical presentation and can be supported by cellular assays of ultraviolet (UV)-induced DNA damage repair;

genetic testing identifies the specific complementation group (e.g., *XPC*).²² Management focuses on strict, lifelong UV radiation protection and periodic screening for skin and ocular malignancies. A primary restriction is the persistent development of multiple cancers, which require repeated surgeries and result in a significant quality-of-life burden despite preventive interventions.²³

HI, the most severe form of autosomal recessive congenital ichthyosis, is caused by loss-of-function mutations in the *ABCA12* gene, disrupting lipid transport in the epidermis. Diagnosis is often clinical at birth due to the pathognomonic presentation of thick, plate-like scales with ectropion and eclabium, and is confirmed genetically.²⁴ Neonatal intensive care is crucial and includes humidification, systemic retinoids, and supportive care. While survival has improved, limits include lifetime severe xerosis, a high risk of sepsis, and significant morbidity from constrictive bands and metabolic complications.²⁵

HED is most commonly X-linked, resulting from mutations in the *EDA* gene, and is characterized by hypotrichosis, hypodontia, and anhidrosis. Diagnosis is clinical, supported by sweat testing (e.g., starch iodine) and confirmed by genetic analysis.²⁶ Management is multimodal, with a focus on temperature regulation, early dental rehabilitation with prostheses or implants, and treatment of respiratory secretions. A principal challenge is the risk of life-threatening hyperthermia in infancy before diagnosis is established, along with the need for lifelong symptomatic care in the absence of therapies that address the underlying ectodysplasin signaling pathway defect.²⁷

Clouston syndrome (hidrotic ectodermal dysplasia) is an autosomal dominant disorder caused by mutations in the *GJB6* gene, which encodes connexin 30. Diagnosis is based on the triad of nail dystrophy, palmoplantar keratoderma, and alopecia, with normal sweating and teeth, and is confirmed by genetic testing.²⁸ Treatment is symptomatic, involving emollients and keratolytics for keratoderma and wigs for alopecia. Core challenges include the progressive nature of the nail and skin changes and the cosmetic impact, with no available treatments to correct the underlying gap junction dysfunction.²⁹

MTS is a phenotypic variant of Lynch syndrome characterized by sebaceous neoplasms and internal malignancies. Diagnosis requires a clinical history of a sebaceous tumor and/or keratoacanthoma along with a visceral malignancy, confirmed by immunohistochemistry showing loss of mismatch repair (MMR) proteins (e.g., MSH2/MSH6) and genetic testing.³⁰ Management centers on high-risk cancer surveillance protocols (e.g., colonoscopies) based on the associated MMR gene mutation. A significant limitation is the potential for

late diagnosis, only after malignancy has developed, and the lifelong, anxiety-inducing need for intensive cancer screening.³¹

BRBNS is a sporadic disorder involving venous malformations of the skin and GI tract, leading to chronic iron-deficiency anemia. Diagnosis is clinical, based on the characteristic compressible blue–purple cutaneous lesions, and is confirmed by endoscopy revealing GI tract lesions.³² Iron supplements and endoscopic or surgical intervention are used to manage bleeding GI lesions. Limitations include the recurrent nature of bleeding, the inaccessibility of some lesions, and the risk of complications from chronic anemia and repeated surgeries, with no medical therapy to prevent the formation of new lesions.³³

TSC is an autosomal dominant disorder caused by mutations in *TSC1* or *TSC2*, leading to benign hamartomas in multiple organs. Diagnosis is based on established clinical criteria or genetic confirmation.³⁴ Management has been revolutionized by mammalian target of rapamycin (mTOR) inhibitors (e.g., sirolimus, everolimus), which can shrink tumors like subependymal giant cell astrocytomas and renal angiomyolipomas. Challenges include the variable expressivity of the disease, the requirement for multidisciplinary lifelong surveillance, and the fact that mTOR inhibitors typically suppress rather than cure lesions, necessitating chronic treatment.³⁵

Mutations in the *NF1* tumor suppressor gene cause NF1, a common autosomal dominant condition. Diagnosis is primarily clinical using the National Institutes of Health criteria, with genetic testing reserved for ambiguous cases.³⁶ Management includes proactive surveillance for complications such as optic pathway gliomas, plexiform neurofibromas, and orthopedic abnormalities. The recent approval of selumetinib, a mitogen-activated protein kinase kinase inhibitor, for inoperable plexiform neurofibromas represents a major advance. However, constraints include extreme phenotypic variability, the significant burden of disfiguring and painful tumors, and the lack of effective treatments for numerous other manifestations, such as cognitive deficits.³⁷

4.3. Future directions

Existing efforts to align both types of evaluation, human and automatic, are task-specific and not broadly applicable across healthcare domains, making this an important area for future exploration. The present study also does not analyze tasks such as text-to-image synthesis, as these fall outside the scope of the QUEST framework and require separate evaluation tools. Future studies should assess the capabilities of these models in analyzing images, performing text-to-image synthesis, and handling other

related but distinct formats. Additional research should also apply the QUEST framework to newer LLMs being developed, whether domain-specific for healthcare or more generalized.

The incorporation of LLMs into the diagnostic process for rare genodermatoses holds great promise; however, it will require sustained future efforts to fully realize their ethical and practical potential. First, the development must focus on establishing specialized, fine-tuned models trained on curated datasets that include rare disease literature, genetic databases, and dermatological image repositories. This would move beyond general-purpose LLMs toward the creation of expert systems capable of recognizing subtle phenotypic patterns and genotypic correlations that may elude non-specialists.³⁸ Second, the evolution toward multimodal AI is critical. Future systems must be capable of integrating and analyzing text-based clinical descriptions, genomic data, and dermatopathological or clinical images within a unified analytical framework, mirroring the holistic approach of an experienced clinical geneticist.³⁹

5. Conclusion

Stringent validation and implementation frameworks are required based on the results of the present study. This includes prospective, real-world clinical trials that investigate the impact of LLM-assisted diagnosis on reducing diagnostic odyssey times and improving patient outcomes across diverse healthcare settings.⁴⁰ Concurrently, the development of strong “guardrails” is critical to ensuring patient safety. These systems must be designed to clearly convey uncertainty, offer relevant citations for their recommendations, avoid harmful overconfidence in differential diagnoses, and consistently emphasize the importance of human physician oversight.⁴¹ Finally, to meaningfully address global health inequities, these tools must be developed with accessibility in mind, optimized for low-bandwidth environments, available in multiple languages, and designed to augment the capabilities of frontline healthcare workers in resource-limited settings, thereby bridging the gap in specialist access.⁴²

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Magda Wojtara

Investigation: Magda Wojtara

Methodology: Magda Wojtara

Writing—original draft: All authors

Writing—review & editing: Magda Wojtara

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data presented in this study are available upon reasonable request.

References

1. Katwaroo AR, Adesh VS, Lowtan A, Umakanthan S. The diagnostic, therapeutic, and ethical impact of artificial intelligence in modern medicine. *Postgrad Med J*. 2024;100(1183):289–296.
doi: 10.1093/postmj/qgad135
2. Wojtara M, Rana E, Rahman T, Khanna P, Singh H. Artificial intelligence in rare disease diagnosis and treatment. *Clin Transl Sci*. 2023;16(11):2106–2111.
doi: 10.1111/cts.13619
3. Vaidya T, Zubritsky L, Alikhan A, Housholder A. Socioeconomic and geographic barriers to dermatology care in urban and rural US populations. *J Am Acad Dermatol*. 2018;78(2):406–408.
doi: 10.1016/j.jaad.2017.07.050
4. Glazer AM, Farberg AS, Winkelmann RR, Rigel DS. Analysis of trends in geographic distribution and density of us dermatologists. *JAMA Dermatol*. 2017;153(4):322–325.
doi: 10.1001/jamadermatol.2016.5411
5. Mosam A, Todd G. Dermatology training in africa: Successes and challenges. *Dermatol Clin*. 2021;39(1):57–71.
doi: 10.1016/j.det.2020.08.006
6. Mugisha N, Ghanem L, Komi OAI, *et al*. The screening and management of skin diseases in remote African regions: A narrative review. *Postgrad Med J*. 2025;101(1196):504–510.
doi: 10.1093/postmj/qgae133
7. Hay RJ, Johns NE, Williams HC, *et al*. The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions. *J Invest Dermatol*. 2014;134(6):1527–1534.
doi: 10.1038/jid.2013.446

8. The Lancet Global Health. The landscape for rare diseases in 2024. *Lancet Glob Health*. 2024;12(3):e341.
doi: 10.1016/S2214-109X(24)00056-1
9. Wright TS. The genodermatoses: An overview. UpToDate. Updated Jan 27, 2025. Available from: <https://www.uptodate.com/contents/the-genodermatoses-an-overview> [Last accessed on 2025 Aug 02].
10. Krawitz PM. Artificial intelligence in the diagnosis of rare disorders: The development of phenotype analysis. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*. 2022;65(11):1159-1163.
doi: 10.1007/s00103-022-03602-2
11. Hurvitz N, Azmanov H, Kesler A, Ilan Y. Establishing a second-generation artificial intelligence-based system for improving diagnosis, treatment, and monitoring of patients with rare diseases. *Eur J Hum Genet*. 2021;29(10):1485-1490.
doi: 10.1038/s41431-021-00928-4
12. Abbas S, Ahmed F, Khan WA, Ahmad M, Khan MA, Ghazal TM. Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence. *Sci Rep*. 2025;15(1):1746.
doi: 10.1038/s41598-024-83966-4
13. Tam TYC, Sivarajkumar S, Kapoor S, *et al*. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7(1):258.
doi: 10.1038/s41746-024-01258-7
14. *Sonar by Perplexity*. Available from: <https://sonar.perplexity.ai> [Last accessed on 2025 Jul 05].
15. *Hello GPT-4o*. Available from: <https://openai.com/index/hello-gpt-4o> [Last accessed on 2025 Jul 05].
16. *OpenEvidence - About*. OpenEvidence. Available from: <https://www.openevidence.com> [Last accessed on 2025 Jul 05].
17. Park YJ, Pillai A, Deng J, *et al*. Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Med Inform Decis Mak*. 2024;24(1):72.
doi: 10.1186/s12911-024-02459-6
18. Fine JD, Bruckner-Tuderman L, Eady RA, *et al*. Inherited epidermolysis bullosa: Updated recommendations on diagnosis and classification. *J Am Acad Dermatol*. 2014;70(6):1103-1126.
doi: 10.1016/j.jaad.2014.01.903
19. Ferrari S, Pellegrini G, Mavilio F, De Luca M. Gene therapy approaches for epidermolysis bullosa. *Clin Dermatol*. 2005;23(4):430-436.
doi: 10.1016/j.clindermatol.2004.07.017
20. Cunliffe C, Bassetti JA, Ellis NA. Bloom's syndrome: Clinical spectrum, molecular pathogenesis, and cancer predisposition. *Mol Syndromol*. 2017;8(1):4-23.
doi: 10.1159/000452082
21. Flanagan M, Cunliffe CM. Bloom syndrome. In: Adam MP, Feldman J, Mirzaa GM, *et al*, editors. *GeneReviews*®. Seattle: University of Washington; 1993.
22. Bradford PT, Goldstein AM, Tamura D, *et al*. Cancer and neurologic degeneration in xeroderma pigmentosum: Long term follow-up characterises the role of DNA repair. *J Med Genet*. 2011;48(3):168-176.
doi: 10.1136/jmg.2010.083022
23. Lehmann J, Seebode C, Martens MC, Emmert S. Xeroderma pigmentosum - facts and perspectives. *Anticancer Res*. 2018;38(2):1159-1164.
doi: 10.21873/anticancer.12335
24. Akiyama M. ABCA12 mutations and autosomal recessive congenital ichthyosis: A review of genotype/phenotype correlations and of pathogenetic concepts. *Hum Mutat*. 2010;31(10):1090-1096.
doi: 10.1002/humu.21326
25. Rajpopat S, Moss C, Mellerio J, *et al*. Harlequin ichthyosis: A review of clinical and molecular findings in 45 cases. *Arch Dermatol*. 2011;147(6):681-686.
doi: 10.1001/archdermatol.2011.9
26. Clarke A, Phillips DI, Brown R, Harper PS. Clinical aspects of X-linked hypohidrotic ectodermal dysplasia. *Arch Dis Child*. 1987;62(10):989-996.
doi: 10.1136/ad.62.10.989
27. Hammersen JE, Neukam V, Nüsken KD, Schneider H. Systematic evaluation of exertional hyperthermia in children and adolescents with hypohidrotic ectodermal dysplasia: An observational study. *Pediatr Res*. 2011;70(3):297-301.
doi: 10.1203/pdr.0b013e318227503b
28. Lamartine J. Towards a new classification of ectodermal dysplasias. *Clin Exp Dermatol*. 2003;28(4):351-355.
doi: 10.1046/j.1365-2230.2003.01319.x
29. Smith FJD, Irwin McLean WH, Morley SM. Novel mechanism of revertant mosaicism in Dowling-Meara epidermolysis bullosa simplex. *J Invest Dermatol*. 2004;122(1):73-77.
doi: 10.1046/j.0022-202x.2003.22129.x
30. Esche C, Kruse R, Lamberti C, *et al*. Muir-Torre syndrome: Clinical features and molecular genetic analysis. *Br J Dermatol*. 1997;136(6):913-7.
31. South CD, Hampel H, Comeras I, Westman JA, Frankel WL, De la Chapelle A. The frequency of Muir-torre syndrome among lynch syndrome families. *J Natl Cancer Inst*. 2008;100(4):277-281.
doi: 10.1093/jnci/djm291

32. Jin XL, Wang ZH, Xiao XB, Huang LS, Zhao XY. Blue rubber bleb nevus syndrome: A case report and literature review. *World J Gastroenterol*. 2014;20(45):17254-17259.
doi: 10.3748/wjg.v20.i45.17254
33. Moodley M, Ramdial P. Blue rubber bleb nevus syndrome: Case report and review of the literature. *Pediatrics*. 1993;92(1):160-162.
34. Northrup H, Krueger DA, International Tuberous Sclerosis Complex Consensus Group. Tuberous sclerosis complex diagnostic criteria update: Recommendations of the 2012 international tuberous sclerosis complex consensus conference. *Pediatr Neurol*. 2013;49(4):243-254.
doi: 10.1016/j.pediatrneurol.2013.08.001
35. Krueger DA, Northrup H, International Tuberous Sclerosis Complex Consensus Group. Tuberous sclerosis complex surveillance and management: Recommendations of the 2012 international tuberous sclerosis complex consensus conference. *Pediatr Neurol*. 2013;49(4):255-265.
doi: 10.1016/j.pediatrneurol.2013.08.002
36. Legius E, Messiaen L, Wolkenstein P, *et al*. Revised diagnostic criteria for neurofibromatosis type 1 and Legius syndrome: An international consensus recommendation. *Genet Med*. 2021;23(8):1506-1513.
doi: 10.1038/s41436-021-01170-5
37. Gross AM, Wolters PL, Dombi E, *et al*. Selumetinib in children with inoperable plexiform neurofibromas. *N Engl J Med*. 2020;382(15):1430-1442.
doi: 10.1056/nejmoa1912735
38. Wojtara M, Rana E, Rahman T, Khanna P, Singh H. Artificial intelligence in rare disease diagnosis and treatment. *Clinical Translational Sci*. 2023;16(11):2106-2111.
doi: 10.1111/cts.13619
39. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.
doi: 10.1038/s41591-018-0300-7
40. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30-36.
doi: 10.1038/s41591-018-0307-0
41. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120.
doi: 10.1038/s41746-023-00873-0
42. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3(4):e000798.
doi: 10.1136/bmjgh-2018-000798