Journal of Hand Surgery Global Online 6 (2024) 847–854

Contents lists available at ScienceDirect

Journal of Hand Surgery Global Online

journal homepage: www.JHSGO.org

Original Research

Large Language Models in the Diagnosis of Hand and Peripheral Nerve Injuries: An Evaluation of ChatGPT and the Isabel Differential Diagnosis Generator

Abdullah AlShenaiber, BHSc, [*] Shaishav Datta, HBSc, MD, [*,†] Adam J. Mosa, MD, MSc, [*,†] Paul A. Binhammer, MD, MSc, [*,†] Edsel B. Ing, MD, MPH, PhD [‡,§]

[*] Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[†] Division of Plastic, Reconstructive & Aesthetic Surgery, Department of Surgery, University of Toronto, Toronto, ON, Canada
[‡] Department of Ophthalmology & Vision Sciences, University of Toronto, Toronto, ON, Canada
[§] Department of Ophthalmology & Visual Sciences, University of Alberta, Edmonton, AB, Canada

ARTICLE INFO

Article history:
Available online September 3, 2024

Key words:
Artificial intelligence
ChatGPT
Diagnosis
Hand surgery
Peripheral nerve injury

Purpose: Tools using artificial intelligence may help reduce missed or delayed diagnoses and improve patient care in hand surgery. This study aimed to compare and evaluate the performance of two natural language processing programs, Isabel and ChatGPT-4, in diagnosing hand and peripheral nerve injuries from a set of clinical vignettes.

Methods: Cases from a virtual library of hand surgery case reports with no history of trauma or previous surgery were included in this study. The clinical details (age, sex, symptoms, signs, and medical history) of 16 hand cases were entered into Isabel and ChatGPT-4 to generate top 10 differential diagnosis lists. Isabel and ChatGPT-4's inclusion and median rank of the correct diagnosis within each list were compared. Two hand surgeons were then provided each list and asked to independently evaluate the performance of the two systems.

Results: Isabel correctly identified 7/16 (44%) cases with a median rank of two (interquartile range = 3). ChatGPT-4 correctly identified 14/16 (88%) of cases with a median rank of one (interquartile range = 1). Physicians one and two, respectively, preferred the lists generated by ChatGPT-4 in 12/16 (75%) and 13/16 (81%) of cases and had no preference in 2/16 (13%) cases.

Conclusions: ChatGPT-4 had significantly greater diagnostic accuracy within our sample ($P < .05$) and generated higher quality differential diagnoses than Isabel. Isabel produced several inappropriate and imprecise differential diagnoses.

Clinical relevance: Despite large language models' potential utility in generating medical diagnoses, physicians must continue to exercise high caution and use their clinical judgment when making diagnostic decisions.

Hand and peripheral nerve injuries are common and can cause profound impacts on quality of life and daily living.[1,2] Early diagnosis and treatment are imperative to preserve function and prevent complications.[3] However, the diagnostic process can be challenging, costly, and time-consuming, requiring a detailed history, physical examination, and imaging.[3,4] Additionally, access to hand specialists is limited, and patients often rely on primary care physicians, who play an important role in their initial evaluation and referral.[5–8] Given these challenges, patients may face missed or delayed diagnoses, negatively impacting their medical care. Hence, it is of value to find diagnostic aids for physicians to improve patient care.

One such solution is the use of artificial intelligence (AI)-based electronic differential diagnosis support (EDS) systems, such as Isabel.[9–11] Isabel is a web-based EDS that is trained with data of over 6,000 medical conditions; it takes user input of patients' age,

Corresponding author: Abdullah AlShenaiber, BHSc, Temerty Faculty of Medicine, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.
E-mail address: abdullah.alshenaiber@mail.utoronto.ca (A. AlShenaiber).

https://doi.org/10.1016/j.jhsg.2024.07.011

sex, travel history, signs, and symptoms to generate a ranked list of possible differential diagnoses.[12] Its algorithm uses statistical natural language processing techniques that analyze the frequency of text and match it against a supervised and manually curated database specific to its domain (in this case, medical diagnoses).[13] Isabel has been extensively researched and was previously found to have a higher rate of accurate diagnoses compared with other EDS systems.[14–16] To our knowledge, beyond its use in research settings, it is currently used by a limited number of institutions. For example, UTHealth Houston uses Isabel as a diagnostic aid and educational tool in internal medicine and primary care.[17] However, EDS tools like Isabel are limited, as they may not improve existing diagnostic practice, use older natural language processing techniques that are limited in input and contextual understanding, and are prone to generating inaccurate differential diagnoses.[14,15,18,19]

A promising solution to overcome these limitations is the use of large language models (LLMs). LLMs are a new subset of natural language processing, which are distinct, as they are unsupervised and trained on vastly larger quantities of text to process user input, integrate context in conversations, and produce human-like responses to dialog.[20] The most prevalent LLM is the generative pretrained transformer chatbot, ChatGPT. Although not specifically intended for medical use, ChatGPT has shown promise in improving diagnostic practice. In one study, ChatGPT version 3 (ChatGPT-3) was found to include the correct diagnosis within a top 10 list of differential diagnoses with 93.3% accuracy.[19] Additionally, several preprint studies investigating ChatGPT's diagnostic ability have found results ranging from 75.6% to 90.0% accuracy in generating the correct diagnosis from a diverse set of clinical vignettes.[21–24]

ChatGPT-4, its latest version, is trained with a larger set of data and has greater problem-solving ability; however, research investigating its ability to generate differential diagnoses is limited.[25–29] Furthermore, there are currently no studies that have assessed the diagnostic capabilities of AI tools such as Isabel or ChatGPT in hand surgery (Supplementary Fig. S1, available online on the Journal's website at https://www.jhsgo.org). Given the challenges of evaluating hand and peripheral nerve injuries, assessing the diagnostic utility of AI tools is valuable to help reduce missed or delayed diagnoses and improve patient care. Thus, this study aims to evaluate the performance of two natural language processing programs, ChatGPT-4 and Isabel, in diagnosing hand and peripheral nerve injuries from a set of clinical vignettes and compare their performance. We hypothesize that ChatGPT-4 will have higher diagnostic performance and greater accuracy than Isabel. Exploring the ability and limitations of both tools will help appraise their potential utility in clinical practice and in improving patient care.

Materials and Methods

Clinical vignettes

The study was approved by the Michael Garron Research Ethics Board (NR-347) and compliant with the Declaration of Helsinki. Cases from the senior coauthor's (P.B.) virtual hand surgery case report repository (www.ihand.ca) were selected for our study. A total of 212 cases were screened for the following exclusion criteria: (1) history of trauma (fracture, dislocation, and open wound—which contain images that are not accepted as input), (2) previous surgery at the site of interest, (3) diagnosis provided in the case stem, and (4) repeated diagnoses. These cases were excluded from the study as they did not sufficiently test the differential diagnosis (DDx) capabilities of Isabel or ChatGPT.

Diagnostic utility assessment

To generate DDx lists on Isabel, the patients' demographic details (age, sex, pregnancy status, and travel history) were selected from a menu of options provided in its DDx Companion tool (https://www.isabelhealthcare.com/products/isabel-ddx-companion). Clinical details (symptoms, signs, and medical history) were then entered as free text in the form of an itemized list, per the tool's instructions. Pertinent negatives were not included as they are not supported by Isabel. The outputs from the "Top 10" mode were then recorded. An example of a case entered into Isabel is shown in Figure 1. Given ChatGPT-4 takes conversational input, the case stem was directly entered as free text with the question: "What is your diagnosis?" followed by a separate prompt asking, "What are your top 10 differential diagnoses for this patient?" Pertinent negatives were included to test ChatGPT's full diagnostic capabilities. The resulting list of differential diagnoses was then recorded, which was listed in order of likelihood. Given ChatGPT can produce variable responses with the same input, only results from the first attempt were included. An example of a case entered into ChatGPT is shown in Figure 2. Full inputs of the text entered into Isabel and ChatGPT for each case can be found in Supplementary Table S1, available online on the Journal's website at https://www.jhsgo.org.

Physician evaluation

As an exploratory analysis, to assess the quality of differentials generated beyond the correct diagnosis, two independent hand surgeons ("physician one" and "physician two") were asked to evaluate and compare the lists generated by each tool. The physicians were provided the case stem with two anonymized lists of the top 10 differential diagnoses, list "A" and list "B," generated by Isabel or ChatGPT. The physicians were blinded to the tools and their allocation to lists "A" or "B." Allocation was randomized for each case to minimize performance bias. The physicians were then asked to respond to the following question: "Based on the case description, which of the following provides a more accurate list of differential diagnoses?" [options: list A, list B, or indifferent]. The responses were reassigned post-hoc to Isabel or ChatGPT for data reporting.

Statistical analysis

Results were considered "correct" if the diagnosis was contained in the "Top 10" diagnoses output ("Top 10" accuracy). "Partially correct" answers were recorded as those that correctly identified the broader category of the diagnosis but failed to specify the exact diagnosis. To compare the tools' ability to rank diagnoses in order of likelihood, the median rank of the correct diagnosis was reported with an interquartile range (IQR). Additionally, each tool's ability to capture the correct diagnosis within its top differential, top three differentials, and top ten differentials were reported as its "Top 1," "Top 3," and "Top 10" accuracy, respectively. Data for diagnostic accuracy were reported as frequencies in the form of a fraction or percentage, rounded to the nearest integer. A two-sample Z-test was used to detect a statistically significant difference between the "Top 1," "Top 3," and "Top 10" accuracies of ChatGPT and Isabel. Statistical significance was defined as a P value $< .05$.

Results

From a total of 212 clinical cases, 196 were excluded because of meeting the exclusion criteria (Fig. 3). A total of 16 cases were included in the final sample for analysis.

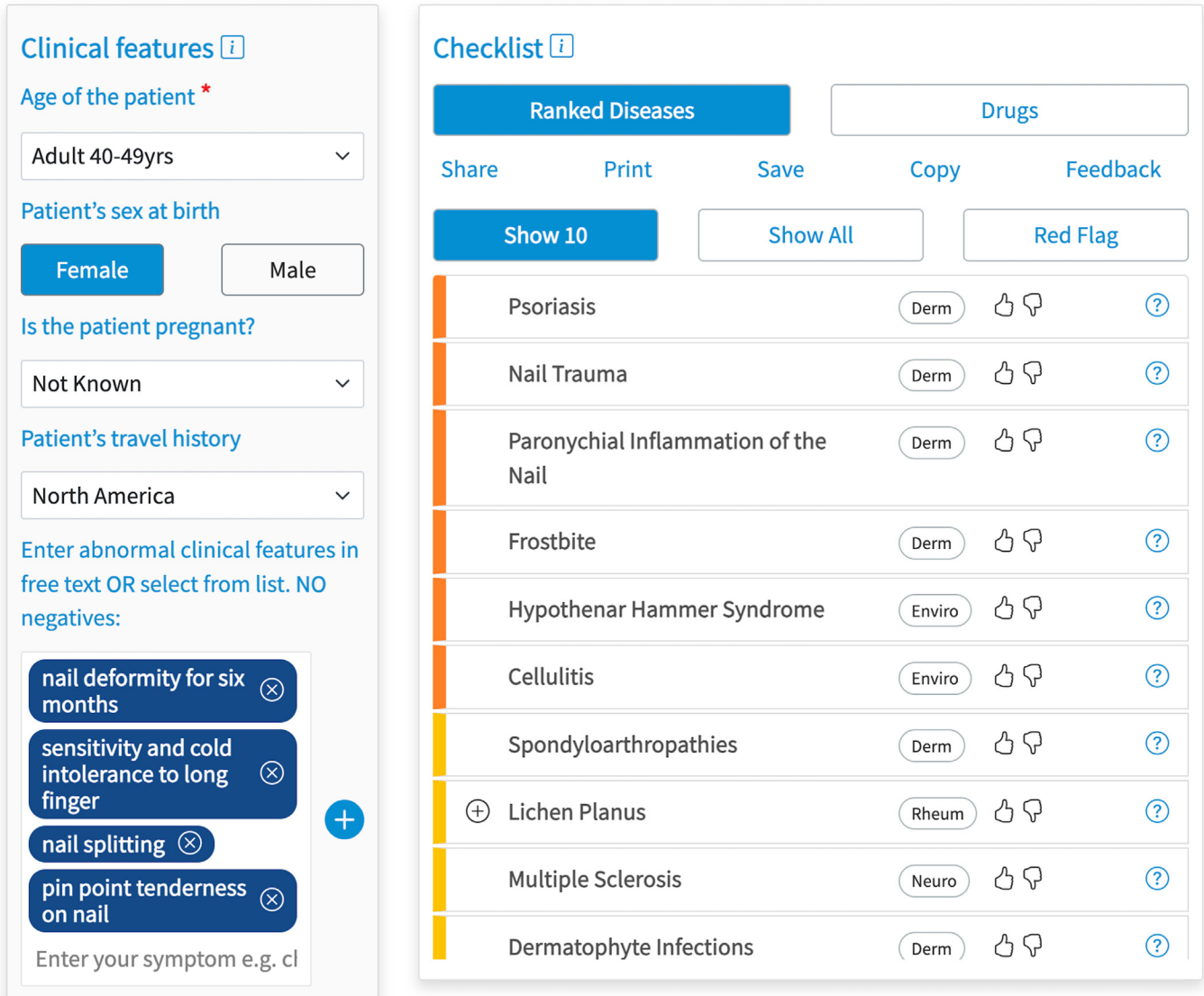

Figure 1. Example of a clinical case entered into Isabel.

Table 1 illustrates (1) the "Top 10" differential diagnoses lists generated by Isabel and ChatGPT-4 and (2) physician feedback. Correct diagnoses and partially correct diagnoses, if any, are labeled. Isabel correctly identified 7/16 (44%) cases as one of its "Top 10" differential diagnoses and was partially correct for 5/16 (31%). It did not correctly identify 4/16 (25%) of cases, namely diagnoses of a glomus tumor, a giant cell tumor of the tendon sheath, Kienbock disease, and Dupuytren disease. Meanwhile, ChatGPT correctly identified 14/16 (88%) of cases as one of its "Top 10" differential diagnoses and was partially correct for 1/16 (6%). It did not correctly identify 1/16 (6%) of cases, namely Hypothenar hammer syndrome. A summary of data measurements, including the median rank of the correct diagnoses, is outlined in Table 2. Regarding physician feedback, physician 1 preferred the lists generated by ChatGPT for 12/16 (75%) of cases, Isabel for 2/16 (13%) cases, and was indifferent for 2/16 (13%) cases. Meanwhile, physician 2 preferred the lists generated by ChatGPT for 13/16 (81%) of cases, Isabel for 1/16 (6%) of cases, and was similarly indifferent to 2/16 (13%) of cases. Only disagreements existed where one physician was indifferent while the other preferred ChatGPT or Isabel.

Using Isabel, the median rank of the correct diagnosis was 2 (IQR = 3), with a "Top 1" accuracy of 3/16 (19%) and "Top 3" accuracy of 5/16 (31%) (Table 2). Meanwhile, using ChatGPT, the median rank of the correct diagnosis was 1 (IQR = 2), with a "Top 1" accuracy of 9/16 (56%) and "Top 3" accuracy of 12/16 (75%). Among ChatGPT and Isabel's correctly identified diagnoses, most were within their "Top 3" ranking (n = 12 for ChatGPT; n = 5 for Isabel). Additionally, ChatGPT had significantly greater "Top 10," "Top 3," and "Top 1" accuracies compared with Isabel ($P < .05$). When including partially correct diagnoses, the median rank marginally decreased to 1.5 (IQR = 2) for Isabel and did not change for ChatGPT (median = 1, IQR = 2). The range of correct diagnoses ranged from 1 to 6 for Isabel compared with 1 to 7 for ChatGPT.

Discussion

Search engines are used by physicians across the world to help supplement daily clinical decision making.[30] With continual advancements in our understanding of medical diseases, treatments, and management pathways, it becomes difficult to keep up with the most up-to-date medical knowledge without the assistance of

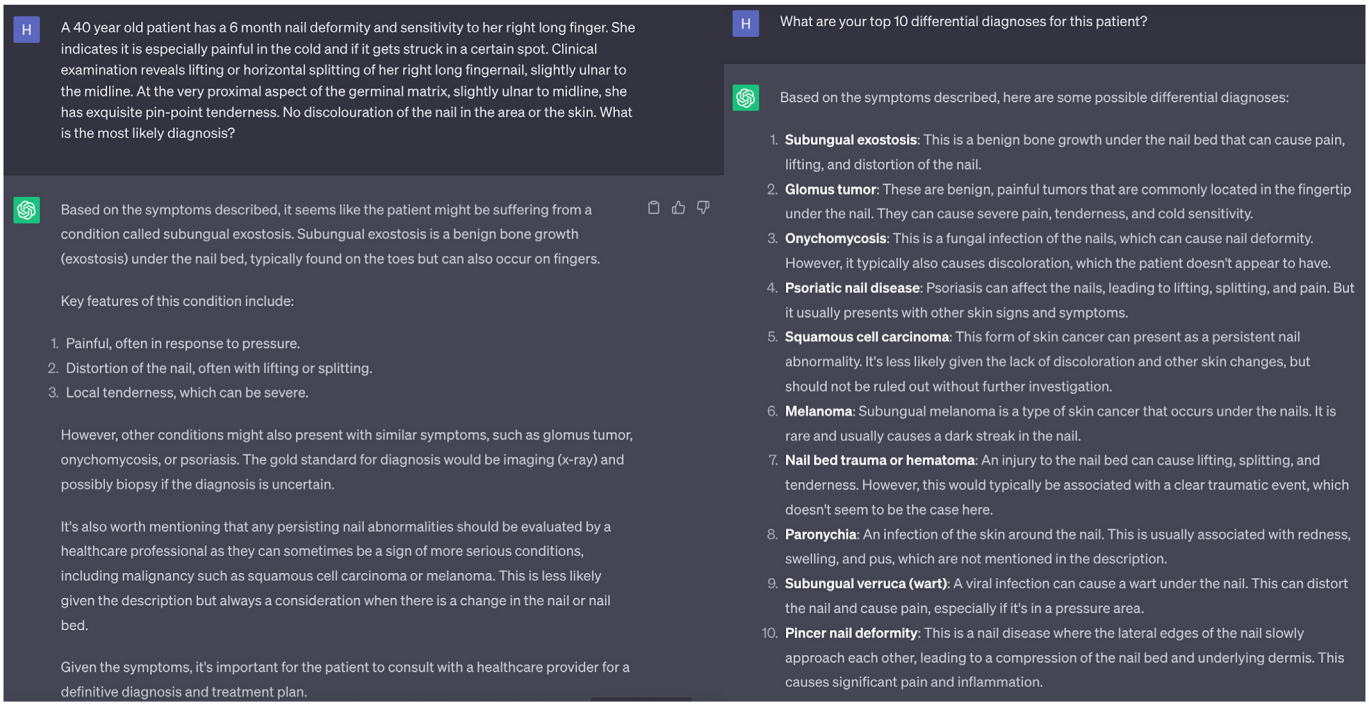

Figure 2. Example of a clinical case entered into ChatGPT-4.

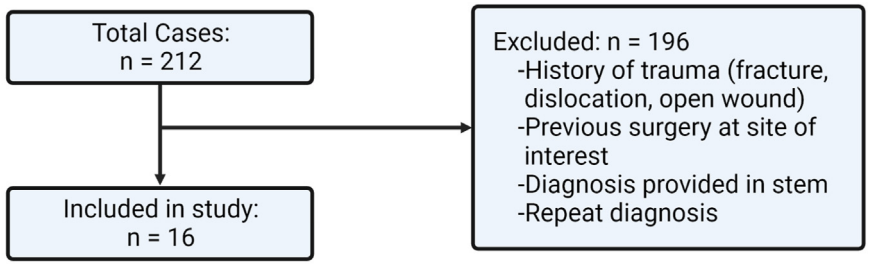

Figure 3. Selection procedure for clinical cases included in the study.

technology.[9] In this way, AI-based tools have the potential to improve diagnostic practice.[9–11] Our results demonstrate that ChatGPT-4 was highly accurate in diagnosing clinical cases within our sample ("Top 10" accuracy = 14/16 or 88%; median rank = 1) and generated significantly more accurate differential diagnoses when compared with Isabel ($P < .05$). The study shows promise in the use of LLMs like ChatGPT as potential diagnostic aids in the initial evaluation and referral of hand and peripheral nerve injuries. This is particularly relevant to primary care physicians as hand and peripheral nerve injuries can be challenging to diagnose and require extensive training, given there are overlapping symptoms, variable presentations, complex anatomy, and a potential need for specialized testing or imaging, which may not be readily available (eg, nerve conduction studies, EMG, and sensation testing).[3] Furthermore, hand and peripheral nerve injuries can lead to significant impairments in function, reinforcing that timely and accurate diagnosis is crucial.[3,4]

When given the classic presentation for hand surgery-related presentations, Isabel was successful in identifying the correct diagnosis for only 7/16 (44%) cases and was incorrect for 4/16 (25%) cases. Many of Isabel's identified diagnoses were broad and partially accurate (5/16 or 31%). This is in stark contrast with the literature, where Isabel's "Top 10" diagnostic accuracy varied from 65% to 91% in various clinical settings.[14,16,31,32] Our findings can be partly explained by Isabel's database being limited to only 6,000 medical conditions. As a result of its limited data set, it is not trained on specific hand-related conditions and lists imprecise, broad diagnoses. For instance, Dupuytren contracture was misdiagnosed and not listed as a differential for all 16 cases despite being a common hand condition. Additionally, Isabel listed many systemic diseases in its differential list and used broad terms for hand conditions. For example, it listed "tendon injuries of the hand" rather than specifying an extensor pollicis longus tendon rupture. In comparison, ChatGPT is trained on vastly larger amounts of text and medical content exceeding that of Isabel's. Beyond limitations in its data set, Isabel is limited in its natural language processing platform. It requires a structured input of clinical features in an itemized format, lacks contextual understanding, and does not support the inclusion of pertinent negatives. These limitations severely limit the form of input entered into Isabel and, subsequently, its diagnostic utility. For instance, specifying a tumor is "not painful" is clinically relevant; however, it would not be permitted as input, given it is a pertinent negative. Furthermore, Isabel produced several inappropriate differentials. For instance, it listed mandibular cysts, genital ulcer syndrome, vulva leiomyoma, and head and neck neoplasms as differentials in the case of a

Table 1
The "Top 10" Differential Diagnoses Lists Generated by Isabel and ChatGPT-4 With Physician Feedback[*,†]

Case Diagnosis	Isabel Results	ChatGPT Results	Physician Preference
Glomus tumor	1. Psoriasis	1. Subungual exostosis	Physician 1: ChatGPT
	2. Nail trauma	2. Glomus tumor[*]	
	3. Paronychial inflammation of the nail	3. Onychomycosis	
		4. Psoriatic nail disease	
	4. Frostbite	5. Squamous cell carcinoma	Physician 2: ChatGPT
	5. Hypothenar hammer syndrome	6. Melanoma	
	6. Cellulitis	7. Nail bed trauma or hematoma	
	7. Spondyloarthropathies	8. Paronychia	
	8. Lichen Planus	9. Subungual verruca (wart)	
	9. Multiple Sclerosis	10. Pincer nail deformity	
	10. Dermatophyte Infections		
Giant cell tumor of the tendon sheath	1. Epidermoid cyst	1. Giant cell tumor of the tendon sheath (Tenosynovial giant cell tumor)[*]	Physician 1: ChatGPT
	2. Hand epidermal inclusion cyst	2. Epidermoid cyst	
	3. Mandibular cysts	3. Ganglion	
	4. Dermatofibrosarcoma protuberans	4. Lipoma	
	5. Lipoma	5. Glomus tumor	Physician 2: ChatGPT
	6. Neoplasms of the jaw	6. Fibroma	
	7. Genital ulcer syndrome	7. Neuroma	
	8. Vulva leiomyoma	8. Enchondroma	
	9. Head and neck neoplasms	9. Leiomyoma	
	10. Nonmelanoma skin cancer	10. Soft tissue sarcoma	
Kienbock Disease	1. Hand and wrist ligament injuries	1. Kienböck disease[*]	Physician 1: ChatGPT
	2. Osteoarthritis	2. Wrist osteoarthritis	
	3. Osteomyelitis and septic arthritis	3. SLAC wrist	
	4. Tendinitis	4. Wrist ligament injury or sprain	
	5. Rheumatoid arthritis	5. Triangular fibrocartilage complex tear or degeneration	Physician 2: ChatGPT
	6. Leukemia	6. Scaphoid fracture or nonunion	
	7. Cellulitis	7. Wrist tendonitis	
	8. Elbow Injuries	8. Gout or pseudogout	
	9. Sclerodermatomyositis	9. Rheumatoid arthritis (RA)	
	10. Adiposis dolorosa	10. Carpal tunnel syndrome	
Carpal Tunnel Syndrome	1. Carpal tunnel syndrome[*]	1. Carpal tunnel syndrome[*]	Physician 1: ChatGPT
	2. Peptic ulcer disease	2. Cervical radiculopathy	
	3. Intervertebral disc herniation	3. Thoracic outlet syndrome	
	4. Complex regional pain syndrome	4. Peripheral neuropathy	
	5. Thoracic outlet syndrome	5. Ulnar nerve entrapment (cubital tunnel syndrome)	Physician 2: ChatGPT
	6. Elbow injuries	6. De Quervain tenosynovitis	
	7. Fibromyalgia	7. Raynaud's phenomenon	
	8. Cystitis/urethritis	8. Rheumatoid arthritis	
	9. Rotator cuff injury	9. Tendonitis	
	10. Diabetic neuropathy	10. Diabetes mellitus	
De Quervain Tenosynovitis	1. Tendinitis[†]	1. De Quervain tenosynovitis[*]	Physician 1: ChatGPT
	2. Osteoarthritis	2. Thumb carpometacarpal arthritis	
	3. Tendon injuries of the hand	3. Trigger thumb	
	4. Ulnar nerve injury	4. Thumb sprain	
	5. Trigger thumb and finger	5. Carpal tunnel syndrome	Physician 2: ChatGPT
	6. Hand and wrist ligament injuries	6. Thumb fracture	
	7. Fractures/dislocations of hand/wrist	7. Tendonitis or tendinopathy	
		8. Gout	
	8. Carpal tunnel syndrome	9. Scaphoid fracture	
	9. Osteomyelitis and septic arthritis	10. Radial nerve injury or entrapment	
	10. Herpetic whitlow		
Trigger finger	1. Trigger thumb and finger[*]	1. Dupuytren contracture	Physician 1: Indifferent
	2. Carpal tunnel syndrome	2. Trigger finger (stenosing tenosynovitis)[*]	
	3. Thoracic outlet syndrome	3. Rheumatoid arthritis	
	4. Tendon injuries of the hand	4. Osteoarthritis	
	5. Phalanx fracture/dislocation	5. Psoriatic arthritis	Physician 2: ChatGPT
	6. Fractures/dislocations of hand/wrist	6. Gout	
		7. Tendon injury or inflammation	
	7. Hypothenar hammer syndrome	8. Carpal tunnel syndrome	
	8. Osteoarthritis	9. Peripheral neuropathy	
	9. Cirrhosis	10. Hand infections or abscesses	
	10. Tendinitis		
Dupuytren disease	1. Carpal tunnel syndrome	1. Dupuytren contracture[*]	Physician 1: ChatGPT
	2. Osteoarthritis	2. Trigger finger (stenosing tenosynovitis)	
	3. Tendon injuries of the hand	3. Ganglion	
	4. Thoracic outlet syndrome	4. Giant cell tumor of tendon sheath	
	5. Rheumatoid arthritis	5. Rheumatoid arthritis	Physician 2: ChatGPT
	6. SLE	6. Collagen vascular diseases	
	7. Phalanx fracture/dislocation	7. Carpal tunnel syndrome	
	8. Trigger thumb and finger	8. Infectious tenosynovitis	
	9. Epidermoid cyst	9. Hand tumors (benign or malignant)	
	10. Endocarditis	10. Palmoplantar fibromatosis	

(*continued on next page*)

Table 1 (*continued*)

Case Diagnosis	Isabel Results	ChatGPT Results	Physician Preference
Cubital tunnel syndrome	1. Elbow injuries	1. Cubital tunnel syndrome or Ulnar neuropathy*	Physician 1: ChatGPT
	2. Cervical spondylosis	2. Cervical radiculopathy (C8-T1)	
	3. Intervertebral disc herniation	3. Thoracic outlet syndrome	
	4. Cubital tunnel syndrome*	4. Peripheral neuropathy	
	5. Repetitive strain injury	5. Carpal tunnel syndrome	Physician 2: ChatGPT
	6. Arm fracture	6. Medial epicondylitis (Golfer's elbow)	
	7. Osteoarthritis	7. Cardiac ischemia referred pain	
	8. Leukemia	8. Ulnar artery thrombosis	
	9. Lyme disease	9. Infections or tumors compressing the nerve	
	10. Ulnar nerve injury	10. Psychogenic pain or somatization disorder	
Anterior interosseus nerve syndrome	1. Median nerve injury†	1. FPL tendonitis or rupture	Physician 1: ChatGPT
	2. Hand and wrist ligament injuries	2. De Quervain tenosynovitis	
	3. Tendon injuries of the hand	3. Trigger thumb	
	4. Brachial plexus injury	4. Thumb arthritis	
	5. Carpal tunnel syndrome	5. Carpal tunnel syndrome	Physician 2: ChatGPT
	6. Dystonia disorders	6. Intersection syndrome	
	7. Tendinitis	7. Neuropathy or radiculopathy†	
	8. Trigger thumb and finger	8. Thumb fracture or post-traumatic stiffness	
	9. Vitamin C deficiency	9. Dupuytren contracture	
	10. Rheumatoid arthritis	10. Infectious or inflammatory arthritis	
Palmar arch aneurysm	1. Aortic aneurysm/dissection	1. Arteriovenous malformation	Physician 1: ChatGPT
	2. Hypothenar hammer syndrome	2. Glomus tumor	
	3. Keratoderma of palms and soles	3. Giant cell tumor of the tendon sheath	
	4. Hyperthyroidism	4. Lipoma	
	5. Dyshidrotic eczema	5. Aneurysm*	Physician 2: ChatGPT
	6. Arterial aneurysms*	6. Foreign body granuloma	
	7. Carpal tunnel syndrome	7. Epidermal inclusion cyst	
	8. Tendon injuries of the hand	8. Ganglion	
	9. Thoracic outlet syndrome	9. Dupuytren contracture	
	10. Renal failure	10. Soft tissue sarcoma	
Hypothenar hammer syndrome	1. Thoracic outlet syndrome	1. Ulnar nerve entrapment/cubital tunnel syndrome	Physician 1: Isabel
	2. Hypothenar hammer syndrome*	2. Carpal tunnel syndrome	
	3. Tendon injuries of the hand	3. Dupuytren contracture	
	4. Carpal tunnel syndrome	4. Trigger finger/thumb	
	5. Brain neoplasms	5. De Quervain tenosynovitis	Physician 2: Indifferent
	6. SLE	6. Rheumatoid arthritis	
	7. Systemic sclerosis	7. Osteoarthritis	
	8. Peripheral arterial disease	8. Ganglion	
	9. Verrucas/warts	9. Peripheral neuropathy	
	10. Asthma	10. Raynaud's disease	
Osteoarthritis	1. Phalanx fracture/dislocation	1. Boutonniere deformity	Physician 1: Isabel
	2. Osteoarthritis*	2. Rheumatoid arthritis	
	3. Fractures/dislocations of hand/wrist	3. Osteoarthritis*	
		4. Psoriatic arthritis	
	4. Rheumatoid arthritis	5. Gout	Physician 2: Isabel
	5. Tendon injuries of the hand	6. Septic arthritis	
	6. Systemic sclerosis	7. Dupuytren contracture	
	7. Carpal tunnel syndrome	8. Flexor tendon injury	
	8. Thoracic outlet syndrome	9. Mallet finger	
	9. Boutonniere deformity	10. SLE	
	10. Trigger thumb and finger		
Ganglion	1. Ganglion*	1. Ganglion*	Physician 1: Indifferent
	2. Fractures/dislocations of hand/wrist	2. Tendonitis	
		3. De Quervain tenosynovitis	
	3. Tendinitis	4. Intersection syndrome	
	4. Wrist bursitis	5. SLAC	Physician 2: Indifferent
	5. Hand and wrist ligament injuries	6. Scaphoid fracture or nonunion	
	6. Leukemia	7. Kienbock disease	
	7. Carpal tunnel syndrome	8. Infection	
	8. Osteomyelitis and septic arthritis	9. Rheumatoid arthritis	
	9. Distal radius ulna fractures	10. Gout or pseudogout	
	10. Benign bone tumors		
Flexor sheath infection (Space of Parona)	1. Trigger thumb and finger	1. Felon	Physician 1: ChatGPT
	2. Cellulitis	2. Paronychia	
	3. Tendinitis†	3. Cellulitis	
	4. Tendon injuries of the hand	4. Herpetic Whitlow	
	5. Carpal tunnel syndrome	5. Septic Arthritis	Physician 2: ChatGPT
	6. Thoracic outlet syndrome	6. Deep space infection	
	7. Erythromelalgia	7. Pyogenic flexor tenosynovitis*	
	8. Coronavirus	8. Osteomyelitis	
	9. Gestational diabetes	9. Foreign body reaction or infection	
	10. Furunculosis, furuncles and carbuncles	10. Gout or Pseudogout	

Table 1 (*continued*)

Case Diagnosis	Isabel Results	ChatGPT Results	Physician Preference
Extensor pollicis longus tendon rupture	1. Rheumatoid arthritis	1. Rheumatoid arthritis-associated tendon rupture*	Physician 1: ChatGPT
	2. Median nerve injury	2. Rheumatoid arthritis-associated tenosynovitis	
	3. Tendon injuries of the hand†	3. Rheumatoid arthritis-associated tendon subluxation or dislocation	
	4. Hand and wrist ligament injuries	4. Dupuytren contracture	Physician 2: ChatGPT
	5. Tendinitis	5. De Quervain tenosynovitis	
	6. Trigger thumb and finger	6. Carpal tunnel syndrome	
	7. Osteoarthritis	7. Trigger finger	
	8. Aortic arch syndrome	8. Ulnar collateral ligament injury	
	9. Spondyloarthropathies	9. Diabetic hand syndrome (cheiroarthropathy)	
	10. Boutonniere deformity	10. Osteoarthritis	
Scapholunate ligament injury	1. Hand and wrist ligament injuries†	1. Scapholunate ligament tear*	Physician 1: ChatGPT
	2. Thoracolumbar spinal injuries	2. SLAC wrist	
	3. Nephrogenic systemic fibrosis	3. Wrist osteoarthritis	
	4. Spinal infections	4. Distal radius fracture	
	5. Bacterial meningitis	5. Kienbock disease	Physician 2: ChatGPT
	6. Atlantoaxial instability	6. Ganglion	
	7. Carpal tunnel syndrome	7. Carpal tunnel syndrome	
	8. Boutonniere deformity	8. De Quervain tenosynovitis	
	9. Osteomyelitis and septic arthritis	9. Triangular fibrocartilage complex injury	
	10. Epidermolysis bullosa	10. Ulnar impaction syndrome	

SLAC, scapholunate advanced collapse; SLE, systemic lupus erythematosus; FPL, flexor pollicus longus.
* Correct diagnoses.
† Partially correct diagnoses.

Table 2
ChatGPT and Isabel's Accuracy in Capturing Correct Diagnoses When Excluding Versus Including Partially Correct Diagnoses, Within the Top 10, Top 3, and Top 1 Differential*

Accuracy Measurement	Excluding Partially Correct Diagnoses		Including Partially Correct Diagnoses	
	Isabel	ChatGPT	Isabel	ChatGPT
Top 1 accuracy (%)	3/16 (19%)*	9/16 (56%)*	5/16 (31%)	9/16 (56%)
Top 3 accuracy (%)	5/16 (31%)*	12/16 (75%)*	10/16 (63%)	12/16 (75%)
Top 10 accuracy (%)	7/16 (44%)*	14/16 (88%)*	12/16 (75%)	15/16 (94%)
Median rank of diagnosis (IQR)	2 (IQR = 3)	1 (IQR = 1)	1.5 (IQR = 2)	1 (IQR = 2)

Median rank of the diagnosis within the differential list is reported with an IQR.
* $P < .05$.

patient with giant cell tumor of the tendon sheath. This is a known and worrisome problem for EDS tools wherein multiple differential diagnoses are generated that have limited relevance to the actual disease.[14,15] If pursued by clinicians, it would result in over-testing or misuse of time and medical resources. Therefore, physicians must practice high caution with EDS tools and continue to exercise their clinical judgment to ensure that differential diagnoses are related and relevant.

On the other hand, ChatGPT-4 was successful in identifying the correct diagnosis in 14/16 (88%) of cases, partially correct in 1/16 (6%) of cases, and incorrect in 1/16 (6%) of cases. Most of the correct diagnoses were within the top three (n = 12) with a median rank of 1. These findings are supported by the current literature, with results of ChatGPT's "Top 10" accuracy varying from 75.6% to 93.3%.[19,21–24] Additionally, ChatGPT partially identified a diagnosis in a case of a 48-year-old patient presenting with a 7-month inability to flex their thumb's interphalangeal joint. The correct diagnosis was anterior interosseus nerve syndrome. ChatGPT listed "neuropathy" in the differential and "flexor pollicis longus (FPL) rupture" as the top diagnosis. This top differential is highly plausible, as FPL rupture is a more common cause of an inability to flex the interphalangeal joint; anterior interosseus nerve is rarer and more difficult to distinguish as it is an FPL palsy, requiring magnetic resonance imaging or electrodiagnostic studies that are not provided in the prompt.[33] This supports the importance of examining the quality of differentials beyond the correct diagnosis in our study, which was assessed by two independent hand surgeons. Physicians 1 and 2, respectively, preferred the differential diagnoses generated by ChatGPT for 12/16 and 13/16 cases and were indifferent for 2/16 cases. These findings can be explained by LLMs' strength in being trained with a vastly large data set and being able to integrate context in conversations, such as patient presentations.[34] Additionally, ChatGPT is continuously being improved with larger data sets and will be able to interpret images as input, which will likely improve its diagnostic accuracy and clinical utility with time. However, despite its strengths, ChatGPT has important limitations. For example, it is prone to "hallucination"—a phenomenon where LLMs output convincing responses that are not factual.[35] Hallucination is well-documented by the literature and places into question the system's reliability in clinical settings.[36] Additionally, ChatGPT requires a sufficiently detailed input of patient presentations and can produce varied responses when given the same input. Given LLMs are trained on a set of data, the data could also be biased, exacerbate health disparities, and not sufficiently account for rare and unique presentations that may be otherwise found in hand and peripheral nerve cases.[35,37] Therefore, despite demonstrating high diagnostic accuracy, outputs from ChatGPT should be interpreted with caution at present.

We acknowledge the limitations of our study. First, the inputs entered into Isabel are not analogous to those entered into ChatGPT. This is because of constraints in the inputs supported by Isabel, as previously described. Although this limitation cannot be controlled because of inherent limitations in Isabel's platform, it is important to consider, given there is an asymmetry in the data entered. However, comparing ChatGPT with Isabel despite its limited platform is relevant, as it evaluates and demonstrates the inherent advantage of LLMs similar to how they would be used to their full extent. Additionally, our data show that ChatGPT performed well

independent of Isabel's results. Second, although we compared ChatGPT with Isabel, we did not compare its ability to produce differential diagnoses against those produced by a physician. This comparison would allow us to assess their diagnostic utility compared with current clinical practice. A potential avenue of future research is to conduct a prospective study at a hand clinic. This would help include a representative sample of typical patient presentations and allow comparison to the physician's diagnostic decisions. However, this is currently difficult to conduct, given it is unclear if tools such as ChatGPT are ethical to use with patient data.

Our study demonstrates ChatGPT-4 had high diagnostic accuracy and generated higher quality DDx lists than Isabel within our sample. Isabel produced several inappropriate and imprecise differential diagnoses. These findings suggest LLMs may have better utility in the initial evaluation and referral of hand and peripheral nerve injuries. However, additional research is required to explore their use as a diagnostic aid in clinical practice. Although there are limitations in the use of AI tools, their accessibility and continuous development may show promise in improving patient care. Despite its potential utility in generating diagnoses, physicians must continue to exercise their clinical judgment when making diagnostic decisions.

Statements

- This study was approved by our institutional review board.
- This article does not contain any studies with human or animal subjects.
- The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
- The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest

No benefits in any form have been received or will be received related directly to this article.

References

1. Wojtkiewicz DM, Saunders J, Domeshek L, Novak CB, Kaskutas V, Mackinnon SE. Social impact of peripheral nerve injuries. *Hand (N Y)*. 2015;10(2):161–167.
2. Crowe CS, Massenburg BB, Morrison SD, et al. Global trends of hand and wrist trauma: a systematic analysis of fracture and digit amputation using the Global Burden of Disease 2017 Study. *Inj Prev*. 2020;26(suppl 1):i115–i124.
3. Griffin MF, Malahias M, Hindocha S, Khan WS. Peripheral nerve injury: principles for repair and regeneration. *Open Orthop J*. 2014;8:199–203.
4. Hile D, Hile L. The emergent evaluation and treatment of hand injuries. *Emerg Med Clin North Am*. 2015;33(2):397–408.
5. Meyerson J, Liechty A, Shields T, Netscher D. A national survey of hand surgeons: understanding the rural landscape. *Hand (N Y)*. 2023;18(4):686–691.
6. Bracey JW, Tait MA, Hollenberg SB, Wyrick TO. A novel telemedicine system for care of statewide hand trauma. *Hand (N Y)*. 2021;16(2):253–257.
7. Curtin CM, Yao J. Referring physicians' knowledge of hand surgery. *Hand (N Y)*. 2010;5(3):278–285.
8. Wildin C, Dias JJ, Heras-Palou C, Bradley MJ, Burke FD. Trends in elective hand surgery referrals from primary care. *Ann R Coll Surg Engl*. 2006;88(6):543–546.
9. Singh H, Graber ML. Improving diagnosis in health care–the next imperative for patient safety. *N Engl J Med*. 2015;373(26):2493–2495.
10. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. *BMJ*. 2016;353:i2139.
11. Shojania KG, Dixon-Woods M. Estimating deaths due to medical error: the ongoing controversy and why it matters. *BMJ Qual Saf*. 2017;26(5):423–428.
12. Isabel. Diagnose. Triage. Teach. Accessed September 4, 2023. https://www.isabelhealthcare.com
13. Zubiaga A. Natural language processing in the era of large language models. *Front Artif Intell*. 2024;6:1350306.
14. Riches N, Panagioti M, Alam R, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. *PLoS One*. 2016;11(3):e0148991.
15. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. *BMJ Qual Saf*. 2022;31(6):426–433.
16. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. *Arch Dis Child*. 2003;88(5):408–413.
17. School MM. Tool for physicians, residents: Isabel Pro helps diagnose complex cases. John P. and Kathrine G. McGovern Medical School at UTHealth. 2022. Accessed April 15, 2024. https://med.uth.edu/blog/2022/01/27/tool-for-physicians-residents-isabel-pro-helps-diagnose-complex-cases/
18. Ing EB, Balas M, Nassrallah G, DeAngelis D, Nijhawan N. The Isabel differential diagnosis generator for orbital diagnosis. *Ophthalmic Plast Reconstr Surg*. 2023;39(5):461–464.
19. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pre-trained transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. 2023;20(4):3378.
20. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224–226.
21. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. *medRxiv*. 2023:2023.01.30.23285067.
22. Mehnen L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. *medRxiv*. 2023. https://doi.org/10.1101/2023.04.20.23288859
23. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv*. 2023. https://doi.org/10.1101/2023.02.21.23285886
24. Benoit JRA. ChatGPT for clinical vignette generation, revision, and evaluation. *medRxiv*. 2023. https://doi.org/10.1101/2023.02.04.23285478
25. GPT-4. Accessed September 4, 2023. https://openai.com/gpt-4
26. Oh N, Choi G-S, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res*. 2023;104(5):269–273.
27. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare*. 2023:1357633X231181922.
28. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci*. 2023;192(6):3197–3200.
29. Hageman MGJS, Anderson J, Blok R, Bossen JKJ, Ring D. Internet self-diagnosis in hand surgery. *Hand (N Y)*. 2015;10(3):565–569.
30. Mikalef P, Kourouthanassis PE, Pateli AG. Online information search behaviour of physicians. *Health Info Libr J*. 2017;34(1):58–73.
31. Ramnarayan P, Cronje N, Brown R, et al. Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. *Emerg Med J*. 2007;24(9):619–624.
32. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*. 2015;351:h3480.
33. Ulrich D, Piatkowski A, Pallua N. Anterior interosseous nerve syndrome: retrospective analysis of 14 patients. *Arch Orthop Trauma Surg*. 2011;131(11):1561–1565.
34. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595.
35. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180.
36. *OpenAI*. GPT-4 Technical Report; 2023.
37. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res*. 2023;25:e48009.