Journal of
Artificial Intelligence and Soft Computing Techniques

Research Article Open Access

In Pursuit of an Expert Artificial Intelligence System: Reproducing Human Physicians Diagnostic Reasoning and Triage Decision Making

Azad Kabir*, Raeed Kabir and Jebun Nahar

Department of Research and Innovation, Doctor Ai LLC, 1120 Beach Blvd, Biloxi, MS 39530, United States of America

***Corresponding Author:** Azad Kabir, Department of Research and Innovation, Doctor Ai LLC, 1120 Beach Blvd, Biloxi, MS 39530, United States of America, Tel: 228-806-7777, E-mail: azad.kabir@ddxrx.net

Received Date: April 10, 2024 **Accepted Date:** May 10, 2024 **Published Date:** May 13, 2024

Citation: Azad Kabir, Raeed Kabir, Jebun Nahar (2024) In Pursuit of an Expert Artificial Intelligence System: Reproducing Human Physicians Diagnostic Reasoning and Triage Decision Making. J Artif Intel Soft Comp Tech 2: 1-14

Abstract

A primary objective in the development of an artificial intelligence (AI) system should be to replicate an expert physician's thought process. This study compares a hypothetico-deductive powered system (Doctor Ai) and a decision tree powered system (Babylon Health Ai) with human physicians to evaluate efficacy, measured by the time needed to (1) diagnose and (2) make triage decisions. In this study, both AI systems and the physicians evaluated a total of fifteen typical textbook presentations of clinical scenarios. The study found that both AI systems agreed on patient disposition decisions for 93% of cases (14 out of 15 cases; $p<0.08$) with no statistically significant difference from physicians, indicating that both AI systems are equally effective in patient triage decisions relative to the physicians. The Doctor Ai system agreed with the physicians on the final diagnosis for 73.3% (11 out of 15) of the cases, while Babylon Health Ai provided a final diagnosis in only 53% (8 out of 15) of cases. For the remaining cases, the diagnosis was either undisclosed or could not be determined. In this study, Doctor Ai used an average of 7.8 (±2.08) computer screens to reach diagnostic confirmation compared to Babylon Health's 21.5 (±9.63) screens ($p<0.001$). The number of screens utilized to reach a final disposition decision (triage) was 10.0 (±2.33) for Doctor Ai, whereas Babylon Health utilized 21.5 (±9.63) screens ($p<0.001$). Additionally, Doctor Ai used on average 13.9 (±6.54) Yes/No events to determine the final diagnosis compared to Babylon Health's 62.3 (±31.55) Yes/No events ($p<0.001$). In conclusion, the hypothetico-deductive system can diagnose more quickly and provide more accurate triage decisions compared to a decision tree powered system. However, both systems combined perform as well as physicians. It is important to evaluate whether such AI can be utilized to tackle pandemics in the U.S. healthcare system and in any developing country healthcare systems facing dire circumstances due to a scarcity of trained physicians.

Keywords: Artificial Intelligence; Hypothetico-Deductive Reasoning; Decision Tree Algorithm; Diagnostic Reasoning; Triage

Introduction

In healthcare systems today, medical errors are persistent, compromising patient safety and healthcare quality [1]. Despite advancements in medical technology and training, errors continue to create adverse events and prolonged hospital stays. The solution necessitates preemptively identifying, anticipating, and preventing errors before they occur. Artificial Intelligence (AI) stands alone in its ability to augment the clinical decision-making processes, improving patient outcomes, and allocating resources more efficiently [2].

Fully integrating AI in healthcare is a paradigm shift. AI-driven systems, equipped with great capability for prediction fueled by machine learning algorithms, can analyze expansive datasets of patient records, medical literature, and real-time clinical data to identify patterns more quickly than a human could [3]. Using the full extent of data available, AI can assist healthcare providers in making accurate diagnoses and minimizing the likelihood of errors throughout every step of patient care.

Moreover, AI technology has an incredible capability to improve patient triage, an essential component of healthcare delivery that influences treatment prioritization, resource utilization, and overall patient outcomes [4]. Leveraging sophisticated algorithms, AI-enabled triage systems can quickly assess the severity of patients' conditions based on symptoms, medical history, and calculate other relevant parameters; AI can also optimize the allocation of healthcare resources by suggesting an appropriate time to intervene in a patient's visit [5]. By streamlining the triage process, AI can simultaneously improve the efficiency of healthcare delivery but also increase the likelihood that patients receive appropriate care promptly, thereby mitigating risk of adverse events.

Several research studies have highlighted the efficacy of AI in preventing medical errors and optimizing patient triage across a diverse set of healthcare settings. For instance, Baker et al [5]. measure the expediting impact of AI-based triage systems with respect to assessment of patients with acute conditions; they find faster treatment initiation and improved clinical outcomes.

To develop an expert physician replica, it is important to understand pattern recognition and hypothetico-deductive reasoning in diagnostic decision making. Pattern recognition is widely used by machine learning algorithms to develop artificial intelligence systems. Therefore, it is crucial to have all key data points as available inputs for pattern recognition to work. This is the Achilles' Heel of machine learning process in aiding diagnostic reasoning. In real life, patients present with only a few symptoms and physicians are left to collect data points like a detective to find a diagnosis. The experts start data collection by generating hypotheses and using hypothetico-deductive reasoning. This process does not use pattern recognition until there is a critical mass of data available for analysis[6]. Expert clinical reasoning alternates between pattern recognition and hypothetico-deductive reasoning (systematically generating and testing hypotheses), depending on the difficulty of clinical cases [6]. Any artificial intelligence system that strives toward equivalent performance with expert clinical reasoning must also alternate between pattern recognition and hypothetico-deductive reasoning in a similar manner.

The DOCTOR Ai® system is a patent-protected software and includes a hypothetico-deductive powered algorithm which uses combination of automated chatbot functions with Natural Language Processing (NLP) algorithm and decision rules to both collect patient history using open-ended questions. Gathering all the pertinent information requires a conversational tool. To analyze the collected clinical information, the hypothetico-deductive system mimics the physician workflow, which includes high probability differential diagnosis generator based decision-making rules to collect medical history [7,8] (Patent granted in 2017 and 2024) and uses the natural laws of expert diagnostic thinking process. When a user clicks on the generated differential diagnosis, the system shows diagnostic confirmation pathways for each and every diagnosis. Based on a given patient clinical scenario, DOCTOR Ai confirms or denies a diagnosis and provides a specific treatment algorithm.

To the best of our knowledge, Babylon Health is developed based on decision tree powered algorithm [9], also one of the most advanced Ai tools that process patient symptoms and provide diagnosis and triage decisions. For this study, we compare both systems with physicians for

clinical efficacy. Babylon Health, a U.K.-based startup that developed Ai-based patient triage systems, using a chatbot that is currently used by the U.K.'s National Health Service to help diagnose ailments [10]. This algorithm currently has a low success rate for solving complex clinical problems [11]. Another system, IBM Dr. Watson, uses pattern recognition as a fundamental basic algorithm for diagnostic decision-making [12] and recently pulled away from the market after being tested at MD Anderson [13].

This study compares the efficacy of the two separate artificial intelligence algorithms with human physician performance in terms of the time needed to provide a final diagnosis and a disposition decision outpatient treatment under a general practitioner versus treatment in an emergency room for consideration of hospitalization, to solve clinical cases.

Method

A sample of fifteen clinical cases with typical clinical scenarios were chosen randomly. The clinical cases were drawn from a subset of cases that exclude physical exam findings or laboratory or radiological data. The final diagnosis for the following clinical scenarios was ultimately determined from the textbook: Chronic Obstructive Pulmonary Disease, Tuberculosis, Pneumonia, COVID-19, Influenza, Strep Throat, Pulmonary Embolism, Congestive Heart Failure, Myocardial Ischemia (MI), Diverticulitis, Urinary Tract Infection, Endometriosis, Hypothyroidism, Diabetes Mellitus (DM), Carpal Tunnel Syndrome, and Migraine. This sample size of fifteen in each group was considered adequate to compare the efficacy of each algorithm used to develop Doctor Ai and Babylon Health respectively. The study used the time needed to confirm a diagnosis or to provide a disposition decision as a surrogate measure for clinical efficiency during a clinical encounter, for both Ai systems. The following variables were utilized as surrogate measures to calculate the time needed to confirm a diagnosis or provide a final disposition decision: 1. The number of screens used to collect history with open-ended questions, from the start of the patient encounter until the differential diagnosis; 2. The number of screens used to find the final diagnosis, from the start to the end of the patient encounter; 3. The number of screens used to determine the final disposition, from the start to the end of the patient encounter; 4. The number of Yes/No events needed to determine the final diagnosis, from the point of differential diagnosis to the end of the patient encounter. The clinical history for each of the selected cases was entered into the artificial intelligence programs, one case at a time. Screenshots of the desktop computer were obtained and saved for data collection for each data entry point or click, to reflect how many computers screens any user needed to go through before finding the final diagnosis or coming to the final disposition decisions. As a part of the data collection, the investigator also collected the differential diagnoses, final diagnosis and disposition decision by each of the systems. And the Ai systems, Doctor Ai and Babylon Health were accessed using website address

https://ddxrx.com/ and https://www.babylonhealth.com respectively and also from smart phone apps (app store or google play).

All fifteen clinical scenarios were also evaluated by three internal medicine board certified physicians, who were working as hospitalists in the state of Alabama. A questionnaire was used to collect the physicians' decisions about the differential diagnosis, final diagnosis, and final disposition separately, for each case. As part of the data collection process, physicians were blinded from the Ai decisions. The investigator summed up all three physicians' diagnosis and disposition decisions. In general, the study demonstrates that an Ai system can generate the same results when repeatedly evaluating the same clinical case. The study further evaluates whether physicians can reproduce the same results while evaluating the same clinical cases. If physicians were unanimous in their decision, the decision was considered accurate.

A two-sided T-test was used to find a statistically significant difference between Doctor Ai and Babylon Health across the four surrogate measures of time to complete the clinical cases in terms of finding a diagnosis or final disposition decision. Both Ai systems were also compared with the expert physicians' final diagnoses and disposition decisions using a Chi-squared test.

Results

The physician's diagnostic accuracy was 73% (11 out of 15). Physicians agreed on disposition decisions for only 66.7% (10 out of 15) of the clinical cases. The Ai systems agreed upon 53% (8 out of 15) clinical cases; Babylon Health did not provide a final diagnosis for the remaining seven (7) clinical cases. Both Ai systems agreed on the disposition decisions for 93% (14 out of 15) of the clinical cases.

Table 1: Description of the clinical cases used to compare the diagnostic accuracy and disposition decisions of Doctor Ai, Babylon Health, and physicians

Clinical Cases	Final Diagnosis	Disposition
1. A 56-year-old male presents with chest pain for 6 hours. It is retrosternal, sharp, and rated 5/10 in intensity. The patient also complained of shortness of breath. Upon further questioning, the patient reported experiencing shortness of breath when lying flat and needing three pillows to sleep comfortably in the recent past. The patient further complained of lower extremity edema. Upon presentation, blood pressure was 125/89, heart rate 103, and pulse oximetry 92% on room air. The patient denied any other complaints.	Congestive Heart Failure **Doctor Ai:** Correct. **Babylon:** Not disclosed/critical. Physicians: 2 out of 3 correct	Hospital/Emergency Room **Doctor Ai:** Agree **Babylon**: Agree Physicians: 2 out of 3 agree
2. A 66-year-old female presents with a fever for 7 days. Her temperature was 100.4°F. The patient complained of chest pain and cough. She stated that her chest pain worsens when she takes a deep breath. The patient mentioned that she does not bring up any sputum while coughing and coughs all day long. She also reported worsening shortness of breath on walking or minor exertion for the last 4 days. Additionally, the patient complained of chills, nausea, and vomiting but denied any nasal congestion, headache, sore throat, or problems related to the throat, neck, ear, or face. Upon presentation, her blood pressure was 89/65, heart rate 110, and pulse oximetry 88% on room air. The patient denied any other complaints.	Pneumonia **Doctor Ai:** Correct. **Babylon**: Not disclosed/critical. Physicians: All correct	Hospital/Emergency Room **Doctor Ai**: Agree **Babylon**: Agree **Physicians**: All agree
3. A 45-year-old female presents with worsening shortness of breath for 12 hours while walking normally. The patient also complained of having stabbing chest pain on the left side of the chest when taking a deep breath, rated 3/10 in intensity. The chest pain started suddenly and was not radiating. She denied any cough, sore throat, nasal congestion, or headache. The patient had a heart rate of 110 and complained of palpitations. Upon further questioning, it was found that she had traveled a long distance and was in the hospital for 4 days prior to this event. The patient has right lower extremity swelling and a positive history of DVT in the past. Upon presentation, her blood pressure was 155/95, heart rate 110, and pulse oximetry 85% on room air. The patient denied any smoking, alcohol, or drug abuse.	Pulmonary Embolism **Doctor Ai**: Correct. **Babylon**: Not disclosed/critical Physicians: Allcorrect	Hospital/Emergency Room **Doctor Ai**: Agree **Babylon:** Agree Physicians: All agree

4. A 45-year-old female presents with chest tightness and shortness of breath during normal walking or at rest for the last 4 days, accompanied by a cough. She stated that she has had a nonproductive cough for more than 8 weeks, with symptoms persisting day and night. The shortness of breath is not associated with any allergens. Additionally, she noted having audible wheezes while exhaling. The patient denied any fever, chills, chest pain, palpitations, abdominal pain, heartburn, nasal congestion, headache, sore throat, leg pain, leg edema, anxiety or panic attacks, and weight loss. She usually smokes one pack per day but has not smoked in the past two days and complained of fatigue and tiredness. Furthermore, her pulse oximetry is 85% on room air. The patient denied any other complaints. Upon presentation, her blood pressure was 129/89, heart rate 87, and pulse oximetry 88% on room air.	Chronic Obstructive Pulmonary Disease **Doctor Ai:** Correct. **Babylon:** Correct. Physicians: All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree Physicians: 2 out of 3 disagree
5. A 35-year-old female presents with pain in her right wrist for the last 4 weeks. She also complained of right hand weakness. The patient has no history of wrist, hand, or elbow injuries, and her wrist is not swollen. She experiences pain in the hand, numbness, and a tingling sensation in the thumb, index, and middle fingers. There is no pain in any other part of the arm. Her symptoms worsen at night but are not associated with cold weather. Frequently, the patient finds it difficult to hold objects or turn keys and doorknobs. She works as a secretary where she mainly types all day. She denied having a fever, chills, joint pain, nausea, vomiting, chest pain, shortness of breath, palpitations, night sweats, dizziness, or lightheadedness, and she did not pass out. She has no past medical history of hypertension, hyperlipidemia, diabetes, no history of heart attack or stroke, and does not smoke. Upon presentation, her blood pressure was 155/95, heart rate 100, and pulse oximetry 98% on room air.	Carpal Tunnel **Doctor Ai:** Correct. **Babylon:** Correct. Physicians: All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree Physicians: All agree
6. A 55-year-old female presents with a burning and stinging sensation during urination. The patient reports increased frequency of urination at night and foul-smelling urine. She also complains of pain on the right side of her abdomen. She mentions a low-grade fever of 100.4°F for the last 2 days. She could not remember anything during the event. She denied any other complaints. Upon presentation, her blood pressure was 136/76, heart rate 78, and pulse oximetry 93% on room air.	Urinary Tract Infection **Doctor Ai:** Correct. **Babylon:** Correct. **Physicians:** All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree **Physicians:** All agree

7. A 75-year-old male has had a cough for the last 6 weeks accompanied by a low-grade fever. He decided to see a physician when he noticed blood in his sputum, predominantly at night. Recently, the patient traveled to Africa for a medical mission. He complained of night sweats and a persistent low-grade fever. Additionally, he has been losing weight, with a total loss of 20 pounds, and has complained of fatigue. He denied having coronary artery disease, hypertension, or heart failure. The patient also denied any smoking, secondary smoke exposure, and asbestos exposure. Upon presentation, his blood pressure was 125/85, heart rate 77, and pulse oximetry 98% on room air. He denied any other complaints.	Tuberculosis **Doctor Ai:** Correct. **Babylon:** Correct. **Physicians:** All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree. **Physicians:** 2 out of 3 agree
8. A 67-year-old male presents with chest pain and associated shortness of breath on exertion. The patient complained of retrosternal chest pain, rated 7/10 in intensity, radiating to the neck or jaw, and relieved with nitroglycerin. He stated that he has been progressively experiencing shortness of breath upon minor exertion for the last month. On presentation, his blood pressure was 187/95, pulse 115, respiratory rate 20, and temperature 98°F. He denied having any fever or cough but complained of nausea and vomiting. He denied any other complaints.	Myocardial Ischemia (MI) **Doctor Ai:** Correct. **Babylon**: Not disclosed/critical. **Physicians:** All correct	Hospital/Emergency Room **Doctor Ai:** Agree **Babylon**: Agree **Physicians:** All agree
9. A 24-year-old male presents with a fever of 101°F, chills, and a nonproductive cough for 3 days. He reports coughing both day and night, accompanied by a headache and sore throat. His headache started gradually, mainly in the forehead, and is rated 4/10 in intensity; it is constant and not associated with light sensitivity. The patient also complained of muscle aches, joint pain, fatigue, and lethargy. He has some skin rashes (petechiae) that become lighter when pressed. The patient denied experiencing nasal congestion, shortness of breath, chest pain, abdominal pain, nausea, vomiting, dysuria, diarrhea, and constipation. He also denied taking any medication that could cause immune deficiency. On presentation, his blood pressure was 120/65, heart rate 85, and pulse oximetry 95% on room air. He denied any other complaints.	Influenza/COVID-19 **DoctorAi:** Correct. **Babylon**: Correct. **Physicians:** 2 out of 3 correct	General Practitioner **Doctor Ai:** Agree **Babylon**: Agree **Physicians:** All agree
10. A 59-year-old female presents with a fever of 101°F. She complained of intractable vomiting since this morning and also mentioned having left lower quadrant abdominal pain. The patient reported constipation in the recent past. On presentation, her blood pressure was 135/65, heart rate 55, and pulse oximetry 94% on room air. All other findings and complaints were negative.	Diverticulitis **Doctor Ai:** Correct. **Babylon**: Correct. **Physicians:** All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree **Physicians:** 2 out of 3 agree

11. A 35-year-old female presents with extreme fatigue. She also complained of having heavy menstrual periods and constipation. She has experienced significant weight gain, cold intolerance, and hoarseness of voice. Additionally, she is losing hair, and her skin feels dry, thin, and brittle. She denied having any headaches, palpitations, shortness of breath, drowsiness, or tremors. The patient does not have any known heart or kidney diseases. She denied any other complaints. On presentation, her blood pressure was 145/85, heart rate 105, and pulse oximetry 98% on room air.	Hypothyroidism **Doctor Ai:** Correct. **Babylon**: Unable to determine. **Physicians:** All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree **Physicians:** All agree
12. A 55-year-old male presents with increased thirst and appetite for the past 6 weeks. He has also experienced an increased frequency of urination and has been passing more urine recently. He needs to wake up several times at night and has reported recent weight loss. The patient also feels tired and unwell but denied any confusion. He has a significant family history of type 2 diabetes mellitus but has never been diagnosed with diabetes. He denied any complaints of nausea and vomiting, numbness or loss of feeling, any skin lesions or rash, lightheadedness, dizziness, and has no history of immune deficiency. The patient has never been diagnosed with any autoimmune disease. On presentation, his blood pressure was 135/85, heart rate 102, and pulse oximetry 98% on room air.	Diabetes Mellitus (DM) **Doctor Ai:** Correct. **Babylon:** Correct. **Physicians:** All correct	General Practitioner **Doctor Ai:** Agree **Babylon:** Agree **Physicians:** All agree
13. A 51-year-old male presents with a constant headache on both sides of the head for 6 hours, which is precipitated by light and sounds. The headache started gradually, and he has sensitivity to light and sounds but denied any eye pain or changes in vision. He also complained of associated nausea and vomiting, though no blood was observed in the vomit. The patient reported seeing zig-zag lines and spots. Additionally, he mentioned that the headaches are precipitated by hunger and fatigue. He denied any head or neck injuries, neck pain, traffic accidents, nasal congestion, or fever. The patient also denied any numbness, weakness in any part of the body, or scalp tenderness. On presentation, his blood pressure was 136/85, heart rate 90, and pulse oximetry 96% on room air.	Migraine **Doctor Ai:** Correct. **Babylon**: Correct. **Physicians:** All correct	General Practitioner **Doctor Ai:** Agree **Babylon**: Agree **Physicians:** All agree
14. A 31-year-old female presents with abnormal vaginal bleeding. She complained of painful sexual intercourse and lower abdominal pain for the last 10 days. The patient is also experiencing constipation. In addition, she reports urinary frequency and dysuria. She denied any other complaints. On presentation, her blood pressure was 125/75, heart rate 110, and pulse oximetry 98% on room air.	Endometriosis **Doctor Ai:** Correct. **Babylon:** Unable to determine. **Physicians:** 1 out of 3 correct	General Practitioner **Doctor Ai:** Agree **Babylon**: Agree **Physicians:** All agree

15. A 35-year-old female presents with a fever of 102°F for 3 days. She complained of fatigue and a sore throat but denied any chills and shivering. The patient reported difficulty swallowing and speaking. She denied any cough, nasal congestion, and earache, but mentioned drooling, although she denied any difficulty breathing. She has a rash that resembles sandpaper. Her neck was found to have some tender, enlarged lymph nodes. She denied chest pain, abdominal pain, shortness of breath, abdominal discomfort, diarrhea, constipation, nausea, vomiting, dysuria, frequency, muscle ache, joint pain, and lethargy. The patient is not taking any medication that would suppress the immune system. On presentation, her blood pressure was 136/85, heart rate 90, and pulse oximetry 96% on room air.	Strep Throat **Doctor Ai:** Correct **Babylon**: Not disclosed/critical **Physicians:** 2 out of 3 correct	Hospital/Emergency Room **Doctor Ai:** General Practitioner **Babylon:** Agree **Physicians:** 2 out of 3 agree

Among the seven (7) cases where Babylon Health did not confirm a diagnosis, five (5) clinical cases were labelled as critical conditions. Babylon Health recommended for the patient to visit the emergency room immediately. Doctor Ai and Babylon Health both defined the remaining ten (10) clinical cases as non-critical. Among these 10 clinical cases, Babylon Health was unable to confirm a diagnosis for two (2) clinical scenarios, where the Babylon Health suggested those two patients to visit a general practitioner. Both artificial intelligence systems matched one or both of the differential diagnoses among the rest of the eight (8) non critical clinical cases. In addition, both the Ai systems agreed upon patient disposition decisions for 93% of the cases (14 out of 15) with no statistically significant difference ($P<0.08$). This indicates that there is no difference between both Ai systems in terms of the accuracy of provided triage decisions.

The time needed to complete the clinical cases was the primary focus of the study. This was measured using several surrogate variables to measure time needed to complete an encounter demonstrated in Table 2.

The Doctor Ai system accepts free text input of symptoms without validating errors and spelling mistakes. t can use these inputs to identify the nearest symptoms in the database for the Ai to analyze. In comparison, Babylon Health accepts data from a predefined selection of symptoms presented to the user in a drop-down menu, with limited options available. On average, Doctor Ai collects 2.8 (±0.86) symptoms using open-ended questions prior to generating differential diagnosis, while Babylon health only collects one symptom before collecting history and generating differentials.

Table 2 also shows the total number of screens Doctor Ai utilized to collect history with open- ended questions was 5.3 (±1.75); the total number of screens used to find the final diagnosis was 7.8 (±2.08); and the total number of screens to determine final disposition was 10.0 (±2.33); the total number of Yes/No events to find the final diagnosis 13.9 (±6.54). The corresponding numbers for Babylon Health were: 2.0 (±0), 21.5 (±9.63), 21.5 (±9.63) and 62.3 (±31.55) respectively. Each difference between Doctor Ai and Babylon Health was statistically significant at <0.001.

The study also compared the combined Ai system decisions against the physicians' diagnostic accuracy and disposition decisions. Only 46.7% (7 out of 15) of cases were accurately diagnosed and 53.3% (8 out of 15) cases received correct disposition decisions by physicians and the Ai system respectively.

Table 2: Comparison of Doctor Ai and Babylon Health Ai system in terms of time needed to find a diagnosis and provide triage decision

Surrogate measure for time needed to complete	DoctorAi	Babylon	P Value (two sided)
	(n=15)	(n=15)	
Number of symptoms collected using open-endedquestions to generate differential diagnosis:Number of symptoms collected using open-endedquestions to generate differential diagnosis:	2.8 (±0.86)	1.0 (±0)	<0.001
Number of screens used to collect history with open- ended questions (from the start to differentialdiagnosis):	5.3 (±1.75)	2.0 (±0)	<0.001
Number of screens used to find the final diagnosis (from the start to end):	7.8 (±2.08)	21.5 (±9.63)	<0.001
Number of screens to determine final disposition (from the start to end):	10.0 (±2.33)	21.5 (±9.63)	<0.001
Number of Yes/No events to find the final diagnosis (from the point of differential diagnosis to end):	13.9 (±6.54)	62.3 (±31.55)	<0.001
Did the Ai system provide a final diagnosis?	100.0% (±0)	53.3% (±0.13)	<0.001
Disposition decision to hospital (versus general practitioner)	26.7% (±0.11)	33.3% (±0.12)	<0.08

Table 3: Comparison between combined physicians' decisions to the Ai system in terms of diagnostic accuracy and disposition (triage) decisions

Combined Ai Decision			
Physicians Decision	**Correct Diagnosis**	**Wrong Diagnosis**	Chi-Square
Correct Diagnosis	7 (46.7%)	4 (26.7%)	
Wrong Diagnosis	1 (6.7%)	3 (20.0%)	<0.185
Total	8 (53.3%)	7 (46.7%)	
Correct Disposition Wrong Disposition			
Correct Disposition	8 (53.3%)	0 (0%)	
Wrong Disposition	6 (40.0%)	1 (6.7%)	<0.268
Total	14 (93.3%)	1 (6.7%)	

Discussion

Hypothetico-deductive powered system (Doctor Ai) provided a correct final diagnosis for 100% (15 out of 15) of clinical cases, whereas the physicians provided a correct final diagnosis for 73.3% (11 out of 15) of clinical cases. The decision tree powered system (Babylon Health) provided a final diagnosis for 53% (8 out of 15) of cases, reducing its efficiency in critical complex clinical cases. Both the system combined agreed with the disposition decision for 93% (14 out of 15) of cases to determine whether a patient should go to the general practitioner, the emergency room, or the hospital. In contrast, the physicians provided the correct disposition for 53.3% (8 out of 15) of clinical cases, with variation in disposition decision within the physicians. This study indicates that both of the Ai systems may provide cor-

rection disposition decisions like physicians.

Expert physician medical history collection depends on open-ended and unbiased questions that lead to improved diagnostic efficacy and accuracy. The Doctor Ai system used 2.8 (±0.86) symptoms on average (collected using open-ended text inputs) at the onset of the patient encounter. These symptoms were used to generate three to five differential diagnoses while Babylon Heath used only one (1) symptom (collected using an open-ended text input) to start collecting data. Babylon Health [10] utilizes decision tree powered algorithms, which possibly contribute to a decreased efficiency in the current study. This possibly resulted in a significant advantage for the Doctor Ai system in terms of time needed to complete any clinical encounter. The decision tree method utilizes yes and no based decision nodes, which leads to a different path for any yes or no answer in the decision tree. The questions asked may possibly require an expanded review of the system, or detailed specifications of the quality of symptoms, which may not pertain to the final diagnosis. Further, this may have no statistical importance in confirming a diagnosis and is rather used to rule out other, low probability diagnoses. This in turn may possibly lead to an increased time for Babylon Health to finish a clinical counter, resulting in a less efficient process to generate a differential diagnosis relative to Doctor Ai. In addition, the decision tree strategy reflects a faulty understanding of how an expert utilizes patient history and calculates its weight in the diagnosis process. During the initial stage of diagnostic reasoning, experts start by asking open-ended and unbiased questions in order to generate a high quality and diverse differential diagnosis. If a patient cannot provide any further history, then the experts ask questions relating to the presenting system, to avoid the use of leading questions. Thus, the most important strategy to develop an Ai system is to collect initial sets of history with open- ended and unbiased questions, with a focus on the presenting organ system, in order to generate a diverse, high probability differential diagnosis. The collection of further quantifying history from presenting symptoms does increase the probability of certain diagnoses but heterogeneous symptoms increase the likelihood of finding correct diagnosis among the generated differential diagnoses. Thus, these open-ended questions are the fundamental strength of Doctor Ai when compared to other decision tree powered systems.

The efficacy of hypothetico-deductive powered system in solving clinical scenarios is reflected in the comparison of the total number of screens any user has to go through before they can reach a final diagnosis or final disposition decision. The number of screens in this study is used as a surrogate measure for the time needed to reach the target endpoint. In this study, Doctor Ai used a 7.8 (±2.08) computer screen to reach diagnostic confirmation, compared to Babylon Health's 21.5 (±9.63) screen, which is statistically significant at <0.001. Even the number of screens utilized to reach a final disposition decision was 10.0 (±2.33) for Doctor Ai, in contrast to Babylon Health, which utilized a 21.5 (±9.63) screen, which is statistically significant at <0.001. The reason behind the difference in time needed to reach final disposition is the unique algorithm, which uses hypothetico-deductive reasoning as opposed to decision tree powered algorithm [7,8].

The decision to compare both systems by (1) the number of screens used to collect history with open-ended questions (from the start to differential diagnosis generation), (2) the number of screens used to find the final diagnosis (from the start to end) and (3) the number of screens to determine final disposition (from the start to end) was made as hypothetico-deductive system uses four separate stages to reach the final diagnosis: Stage 1 is the collection of patient history using open-ended questions; Stage 2 is the generation of high probability differential diagnoses Stage 3, utilizes a hypothetico-deductive reasoning-based diagnosis confirmation system; Stage 4, uses a mathematical equation-based cut-off point to confirm or deny a diagnosis [11]. Another problem with the common Ai algorithm is that it is fundamentally dependent on the Bayesian method and is unable to mimic an expert physician's thought process. The human, expert thinking process is simple in the sense that they cannot process complex math such as the calculation of diagnostic probability. If an Ai system use complex Bayesian weighting systems to solve clinical encounter, which any expert cannot compute as part of their workflow, does not follow physician clinical reasoning [7,8]. In stage 4, if a certain diagnosis is ruled out based on a unique cutoff points-based decision rules, then the system collects history about the second ranking diagnosis in the differential diag-

nosis list (which is generated from stage 2). This rank ordering of the high probability differential diagnosis list continues to get updated as clinical information is obtained during the diagnosis confirmation process. Whenever any pertinent new information is obtained even if such information is obtained at the end of the encounter, as would happen for an expert physician's workflow, the system updates the symptoms to find a new set of rank ordered differential diagnosis list and restarts the process from stage 3 again. Each of the above four stages are built to maximize efficacy and effectiveness of clinical encounters, considering how valuable of an asset the time needed to complete any clinical encounter is.

The other significant finding in the study was that the hypothetico-deductive powered system confirmed the presence of a diagnosis in 100% of cases. In contrast, Babylon Health could reach diagnostic confirmation only 53.3% of the time ($p<0.001$). To Babylon Health's credit, among 5 out of the 7 cases where decision tree powered was unable to find a diagnosis as the system refrained from collecting more clinical data when the Ai system determined that a patient's condition is critical and needs an emergency room visit. That's why the current study also compared the final disposition decision between the two Ai systems and found that there is no statistically significant difference between the system's respective final disposition decisions. This study found that hypothetico-deductive powered system recommends an emergency room visit for 26.6% of cases whereas Babylon Health recommends an emergency room visit for 33.3% of cases, with a p-value of <0.08. When the clinical cases were determined to be non-critical, Babylon Health completed the history taking process until they found a diagnosis (with the exception of 2 out of the 10 clinical cases) and recommended for the patient to visit a general practitioner. Overall, both the Ai system agreed on a patient's disposition decision of either the emergency room or general practitioner (primary care physician) in 93.3% of cases, indicating that the major difference between two systems is the time needed to complete any clinical encounter, measured as the number of screens utilized to complete the clinical encounter.

Hypothetico-deductive powered system (Doctor Ai) utilized on average 13.9 (±6.54) Yes/No events to find the final diagnosis compared to the decision tree powered system (Babylon Health) 62.3 (±31.55) Yes/No events, which were statistically significant at <0.001. It is understandable that the time taken to go through an average of 13.9 (±6.54) questions will be much less than 62.3 (±31.55) questions. That's why the current study measures screen number and the number of Yes/No events as surrogate measures of time to reflect clinical efficiency.

The future of healthcare belongs to artificial intelligence system-based healthcare delivery. Such a system can potentially address key challenges that healthcare systems in the U.S. and other countries are facing due to the COVID-19 pandemic. An Ai system can increase the number of critical patients that doctors and nurse practitioners can see by equipping physician assistants and nurses with the ability to diagnose more patients and triage them smoothly. If this is indeed the case, the following implications of the platform are significant in the time of COVID-19 as well as during future pandemics: (1) through better triage, the use of the platform can reduce the number of face-to-face interactions that healthcare providers must have with patients further reducing both patient and provider exposure as well as transmission risks and the exhaustion of scarce PPE equipment; (2) through improved allocation of services and reductions in unnecessary treatments and defensive medicine practices, the platform can reduce the costs of overburdened healthcare systems in an era when governments and banks are struggling to provide the loans required to keep such businesses afloat; and (3) through reducing wait times, the hassle and transmission risk of physically going to a healthcare center, and incorrect diagnoses and treatments due to time pressure and failure to attend to all symptoms by memory, the platform can improve overall patient welfare. Finally, such a artificial intelligence platform is desperately needed globally, where healthcare facilities are often scarce, poorly equipped, and physicians likewise scarce and poorly trained to manage the COVID-19 or any future pandemics.

However, there are challenges to develop such a perfect system that can mimic expert physician reasoning. Unfortunately, clinicians do not know all the sensitivities and specificities associated with each piece of a patient's history or physical exam that are pertinent to each medical di-

agnosis. This data is not available in the literature as well. But to solve the clinical scenario using Bayes' theorem, we need to know the sensitivity and specificity associated with each sign and symptom related to each diagnosis, to estimate their weight. Even using a supercomputer, Bayes theorem can't be solved as those numbers are virtually impossible to obtain. But using a simplified weight has shown to deliver accurate results compared to using accurate sensitivity and specificity for a known medical problem [13]. Thus, the incorporation of hypothetico-deductive reasoning to develop artificial intelligence software enables users to work as an expert by helping them generate a high-quality differential diagnosis at the beginning of the encounter, as well as allowing them to confirm the diagnosis if diagnostic criteria are met. Statistical models generate average overall predictions, not for an individual subject. Instead of complex statistical modeling of the different variables, artificial intelligence software needs an individual mathematical algorithm and not a statistical algorithm, to confirm or deny certain outcomes [14].

Limitations

The current study used typical presentation of clinical cases that are most commonly presents in physicians' clinics or in emergency room. In general, many attempts were failed because of the faulty expectation to solve atypical presentations of clinical cases rather than understanding that Ai is only good for simple textbook presentations of clinical cases. The problem with atypical clinical scenarios is that they do not present with commonly known signs and symptoms. Any attempt to provide diagnostic confirmation using atypical presentation will lead to diagnostic ineffectiveness and unnecessary testing and unnecessary resource utilization.

Conclusion

To solve complex critical clinical scenarios, the hypothetico-deductive powered system performs superior to decision tree powered system but both the system combined performs as well as human physicians, in terms of finding a correct diagnosis and provide disposition decisions (triage decisions). It is important to further evaluate whether such an Ai system can be used to tackle triage decisions during pandemics to reduce burden on the healthcare system in the U.S. and also in developing countries that are facing dire circumstances due to a scarcity of trained physicians.

Ethical Approval

The research related to human use has been complied with all the relevant national regulations, institutional policies, and in accordance the tenets of the Helsinki Declaration and has been approved by the institutional review board of Jackson Hospital, Montgomery, Alabama.

Conflict of Interest

The lead author owns financial interest that owns *DOCTOR Ai*® related patents and trademarks.

Acknowledgements

This study was inspired by Dr. Abul Hussam, Ph.D.

References

1. Choudhury A, Asan O (2020) Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform. 8: e18599.

2. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, Zhao J, Snowdon JL (2021) Precision Medicine, AI, and the Future of Personalized Health Care. Clin Transl Sci. 14: 86-93.

3. Bitkina OV, Park J, Kim HK. (2023) Application of artificial intelligence in medical technologies: A systematic review of main trends. Digit Health. 9: 20552076231189331.

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

5. Wolff P, Ríos S, Gonzales C (2023) Machine Learning Methods for Predicting Adverse Drug Reactions in Hospitalized Patients; Procedia Computer Science. 225: 22-31.

6. Elstein AS, A Schwartz (2002) "Clinical problem solving and diagnostic decision making: selective review of the cognitive literature." BMJ 324: 729-32.

7. Kabir A (2017) United States Patent No. US 9,536,051,B1; High probability differential diagnoses generator software; US Patent and Trademark Office; Alexandria, Virginia.

8. Kabir A (2024) United States Patent No. US11,972,865, B1; High probability differential diagnoses generator and smart electronic medical record; US Patent and Trademark Office; Alexandria, Virginia.

9. Khan RS, Zardar AA, Bhatti Z. Artificial Intelligence based Smart Doctor using Decision Tree Algorithm. Journal of Information & Communication Technology – JICT, 11: 1816-613X.

10. https://techcrunch.com/2019/08/02/babylon-health - confirms-550m-raise-to-expand-its-ai-based-health-services--to-the-us-and-asia/

11. https://www.forbes.com/sites/parmyolson/2018/12 /17/this-health-startup-won-big-government-dealsbut-inside-doctors-flagged-problems/#379eb58beabb

12. Cahan A, Cimino JJ. (2017) "A Learning Health Care System Using Computer-Aided Diagnosis." J Med Internet Res 19: e54

13. https://www.forbes.com/sites/matthewherper/ 2017/02/19/md-anderson-benches-ibm-watson-in-setback--for-artificial-intelligence-in-medicine/#131fd4683774

14. Kabir A (2017) Pending Patent Application No. 15/356,933 - Continuation of US Patent Number: 9,536,051 for High probability differential diagnoses generator software.