Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study

Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD.

Journal Club

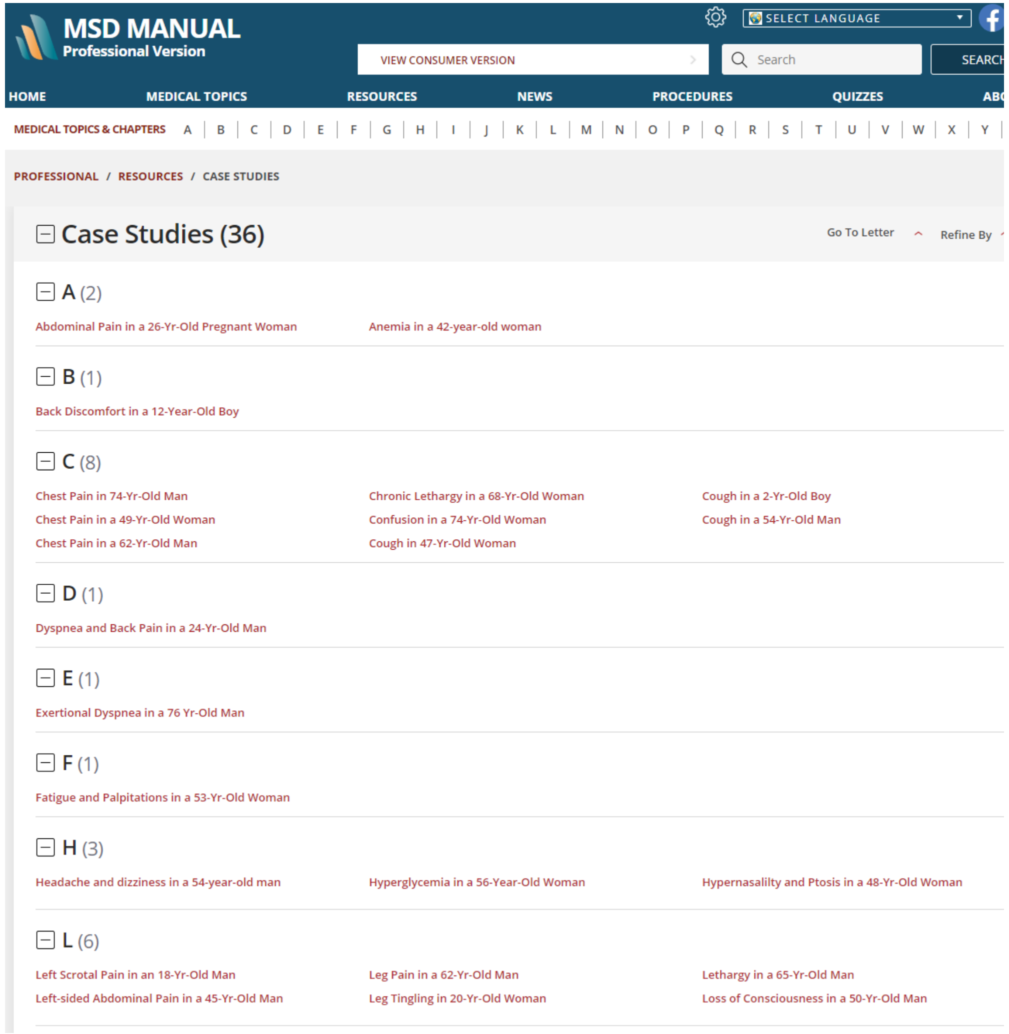
Nathan Yung 9/22/2023

Study design

- Hypothesis: When provided with clinical vignettes, ChatGPT would be able to recommend a diagnostic workup, decide the clinical management course, and ultimately make the diagnosis
- Assess accuracy in solving comprehensive clinical vignettes
- Clinical Vignettes provided by Merck Manual (MSD Manual)

MSD Manual

- Started in 1899 as a small reference book for physicians and pharmacists
- Started expanding in scope to a more comprehensive medical resource for professionals and consumers
- [Case studies - MSD Manual Professional Edition (msdmanuals.com)]


MSD MANUAL
Professional Version
SELECT LANGUAGE
VIEW CONSUMER VERSION
Search
SEARCH
HOME
MEDICAL TOPICS
RESOURCES
NEWS
PROCEDURES
QUIZZES
ABO
MEDICAL TOPICS & CHAPTERS
A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y |
PROFESSIONAL / RESOURCES / CASE STUDIES
Go To Letter
Refine By

⊟ Case Studies (36)

⊟ A (2)

Abdominal Pain in a 26-Yr-Old Pregnant Woman
Anemia in a 42-year-old woman

⊟ B (1)

Back Discomfort in a 12-Year-Old Boy

⊟ C (8)

Chest Pain in 74-Yr-Old Man
Chest Pain in a 49-Yr-Old Woman
Chest Pain in a 62-Yr-Old Man
Chronic Lethargy in a 68-Yr-Old Woman
Confusion in a 74-Yr-Old Woman
Cough in 47-Yr-Old Woman
Cough in a 2-Yr-Old Boy
Cough in a 54-Yr-Old Man

⊟ D (1)

Dyspnea and Back Pain in a 24-Yr-Old Man

⊟ E (1)

Exertional Dyspnea in a 76 Yr-Old Man

⊟ F (1)

Fatigue and Palpitations in a 53-Yr-Old Woman

⊟ H (3)

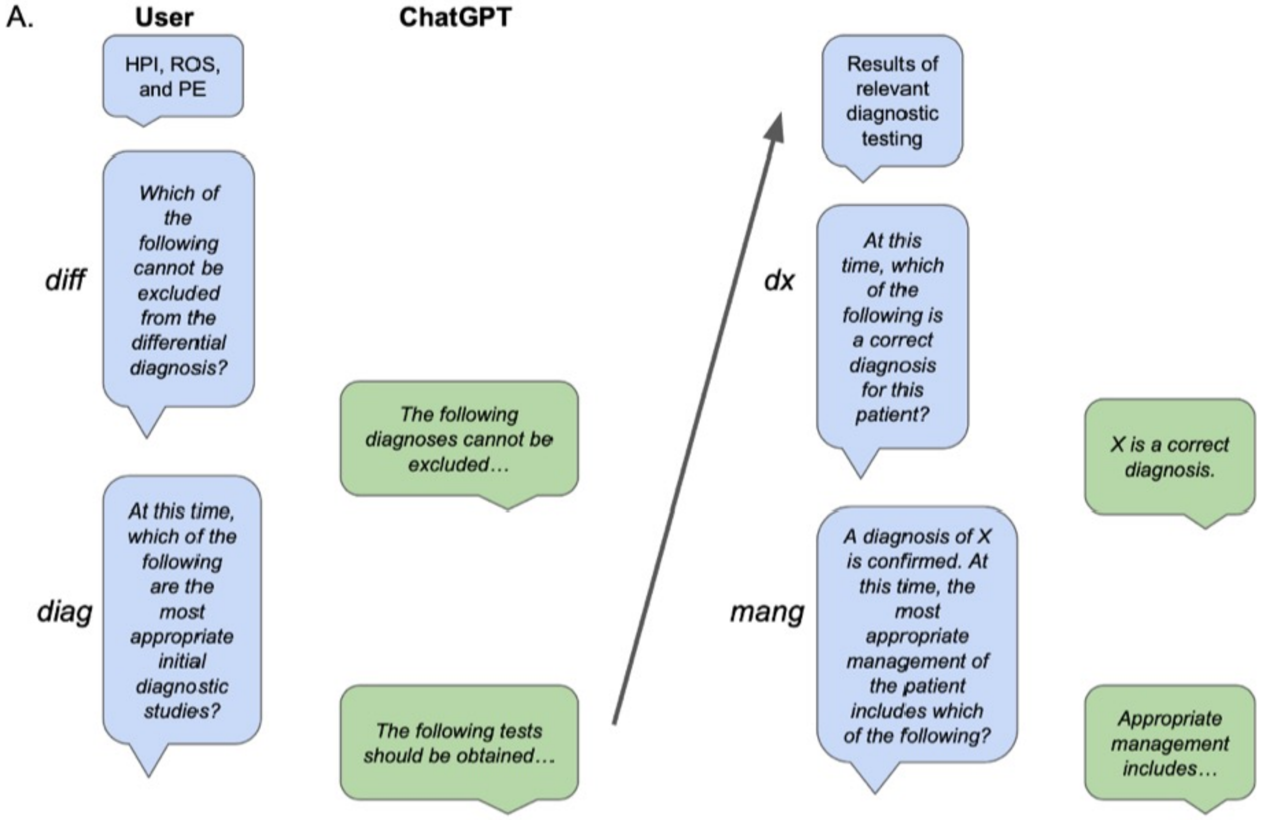
Headache and dizziness in a 54-year-old man
Hyperglycemia in a 56-Year-Old Woman
Hypernasalilty and Ptosis in a 48-Yr-Old Woman

⊟ L (6)

Left Scrotal Pain in an 18-Yr-Old Man
Left-sided Abdominal Pain in a 45-Yr-Old Man
Leg Pain in a 62-Yr-Old Man
Leg Tingling in 20-Yr-Old Woman
Lethargy in a 65-Yr-Old Man
Loss of Consciousness in a 50-Yr-Old Man

Figure1. Experimental workflow for determining ChatGPT accuracy in solving clinical vignettes.

- Panel A: Schematic of user interface with ChatGPT for this experiment.
- Blue boxes indicate prompts given to ChatGPT and green boxes indicate ChatGPT responses.
- Nonitalicized text indicates information given to ChatGPT without a specific question attached



A.

User

ChatGPT

HPI, ROS, and PE

diff

Which of the following cannot be excluded from the differential diagnosis?

The following diagnoses cannot be excluded…

diag

At this time, which of the following are the most appropriate initial diagnostic studies?

The following tests should be obtained…

Results of relevant diagnostic testing

dx

At this time, which of the following is a correct diagnosis for this patient?

X is a correct diagnosis.

mang

A diagnosis of X is confirmed. At this time, the most appropriate management of the patient includes which of the following?

Appropriate management includes…

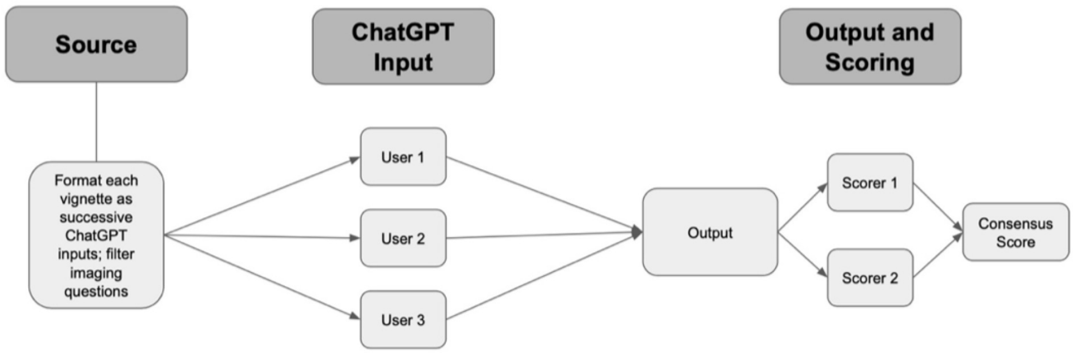
Interacting with the cases

- Which of the following cannot be excluded from the differential diagnosis
- At this time, which of the following are the most appropriate initial diagnostic studies
- At this time, which of the following is a correct diagnosis for this patient
- A Diagnosis of X if confirmed. At this time, the most appropriate management of the patient includes which of the following

Figure1. Experimental workflow for determining ChatGPT accuracy in solving clinical vignettes.

- Panel B: Schematic of experimental workflow. Prompts were developed from Merck Sharpe &Dohme (MSD) vignettes and converted to ChatGPT-compatible text input.
- Questions requiring the interpretation of images were removed. Three independent users tested each prompt. Two independent scorers calculated scores for all outputs; these were compared to generate a consensus score.
- *diag*: diagnostic questions
- *diff*: differential diagnoses
- *dx*: diagnosis questions;
- HPI: history of present illness
- *mang*: management questions
- Misc: Miscellaneous questions
- PE: physical exam
- ROS: review of systems.

B.



Clinical Acuity

- Cases were scored according to the Emergency Severity Index (ESI)
- ESI – Ordinal Scaled system 1-5.
- Scores were generated by ChatGPT and cross-validated with a human generated ESI score for validation

Supplemental

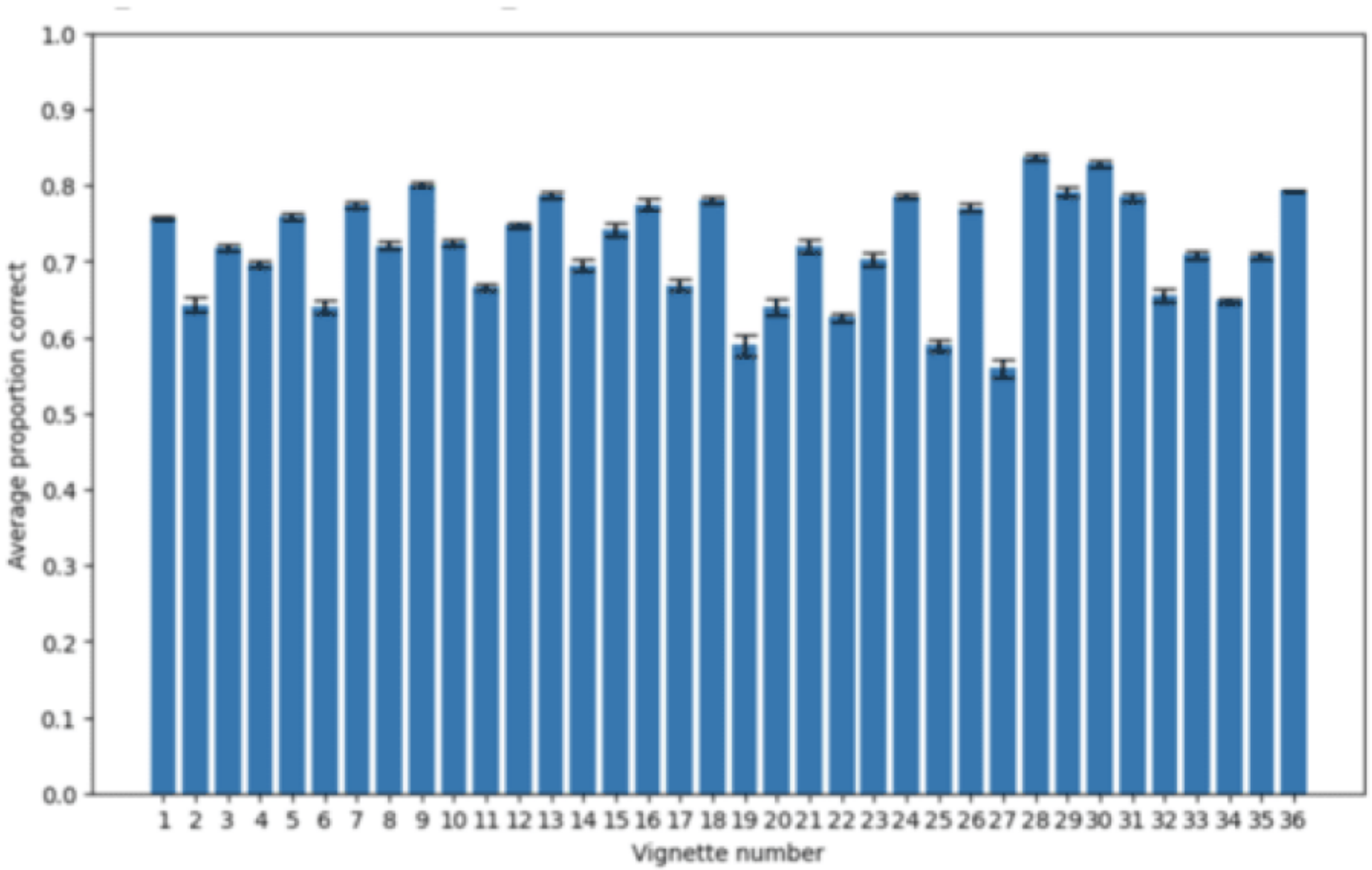
- **Table S1: Metadata for MSD Vignettes**
- Metadata for MSD vignettes. Age, gender, title, and final diagnosis were all provided within the vignettes themselves. ESI rating was calculated by ChatGPT and human scorers (see Methods). Vignette number was assigned by this research group in the order that the vignettes are published online.

Vignette Number	Title	Age	Gender	ESI	Final Diagnosis
1	Abdominal Pain in a 26-Yr-Old Pregnant Woman	26	F	3	Sepsis
2	Anemia in a 42-year-old woman	42	F	2	Transient pure red cell aplasia triggered by parvovirus B19 infection
3	Back Discomfort in a 12-Year-Old Boy	12	M	4	Adolescent idiopathic scoliosis
4	Chest Pain in 74-Yr-Old Man	74	M	2	Non-ST-segment elevation myocardial infarction (NSTEMI)
5	Chest Pain in a 49-Yr-Old Woman	49	F	2	Pulmonary embolism (PE)
6	Chest Pain in a 62-Yr-Old Man	62	M	4	Stable angina
7	Chronic Lethargy in a 68-Yr-Old Woman	68	F	3	Hashimoto thyroiditis
8	Confusion in a 74-Yr-Old Woman	74	F	2	Seizure due to meningioma
9	Cough in 47-Yr-Old Woman	47	F	2	Asthma
10	Cough in a 2-Yr-Old Boy	2	M	3	Foreign body aspiration, left lung
11	Cough in a 54-Yr-Old Man	54	M	3	Community-acquired pneumonia
12	Dyspnea and Back Pain in a 24-Yr-Old Man	24	M	3	Primary spontaneous pneumothorax
13	Exertional Dyspnea in a 76 Yr-Old Man	76	M	2	Severe aortic stenosis
14	Fatigue and Palpitations in a 53-Yr-Old Woman	53	F	3	Atrial fibrillation
15	Headache and dizziness in a 54-year-old man	54	M	4	Acoustic neuroma, right side
16	Hyperglycemia in a 56-Year-Old Woman	56	F	4	Type 2 diabetes
17	Hypernasalilty and Ptosis in a 48-Yr-Old Woman	48	F	2	Myasthenia gravis
18	Left Scrotal Pain in an 18-Yr-Old Man	18	M	3	Testicular torsion
19	Left-sided Abdominal Pain in a 45-Yr-Old Man	45	M	3	Ureteral lithiasis
20	Leg Pain in a 62-Yr-Old Man	62	M	3	Cellulitis

Results

- Case-by-case
- Question Category-by-Question Category

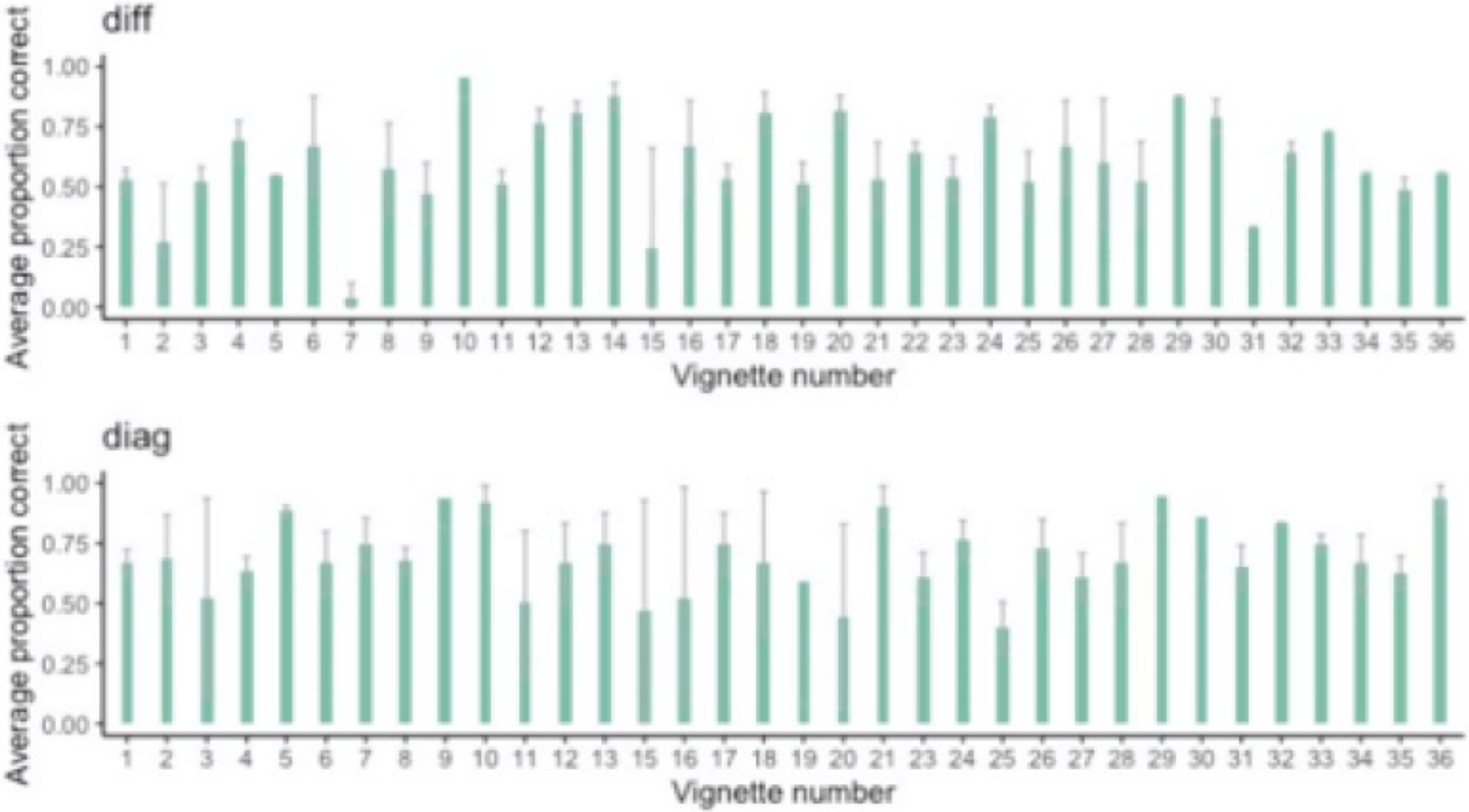
Results – Figure 2a



ChatGPT overall performance for each of the 36 Merck Sharpe & Dohme (MSD) vignettes; error bars are 1 SE of the mean.

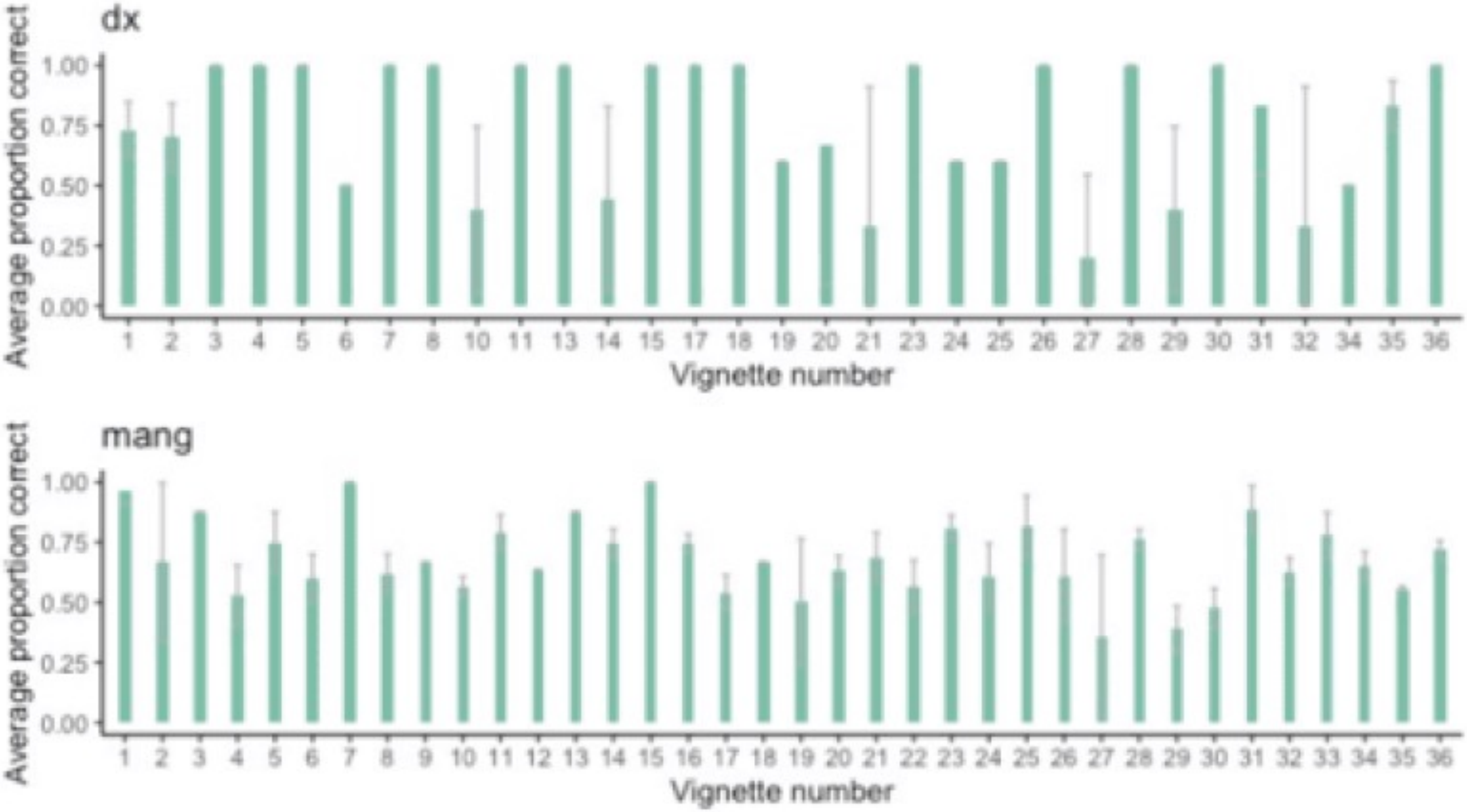
Average performance found to be 71.8%

Results – Figure 2c



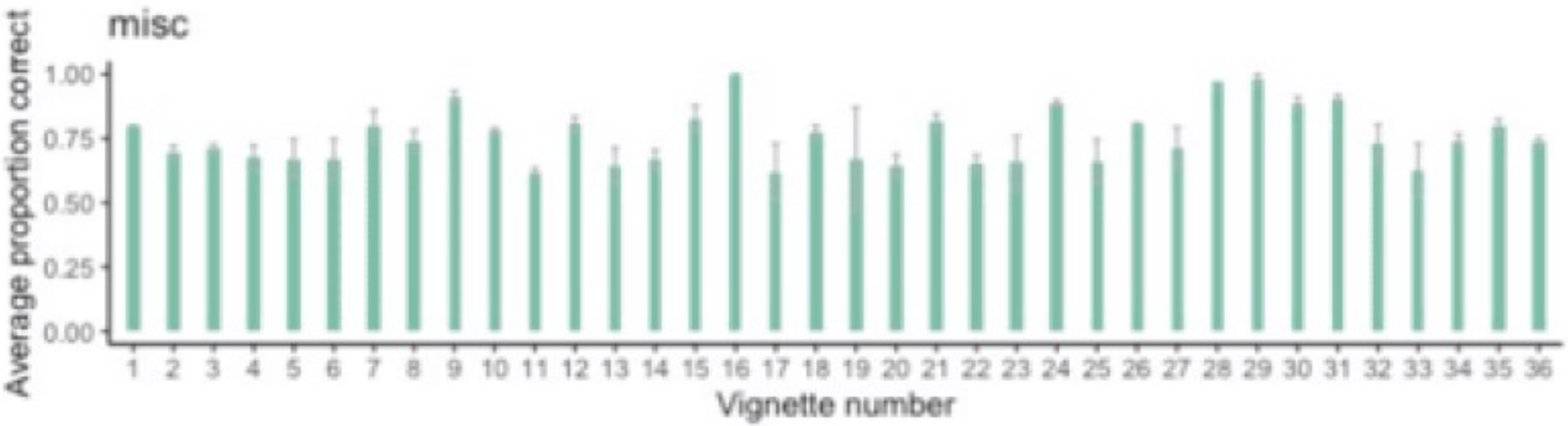
ChatGPT performance by question type for each of the 36 MSD vignettes; error bars are 1 SE of the mean. *diag*: diagnostic questions; *diff*: differential diagnoses; *dx*: diagnosis questions; *mang*: management questions; *misc*: miscellaneous question.

Results – Figure 2c



ChatGPT performance by question type for each of the 36 MSD vignettes; error bars are 1 SE of the mean. *diag*: diagnostic questions; *diff*: differential diagnoses; *dx*: diagnosis questions; *mang*: management questions; *misc*: miscellaneous question.

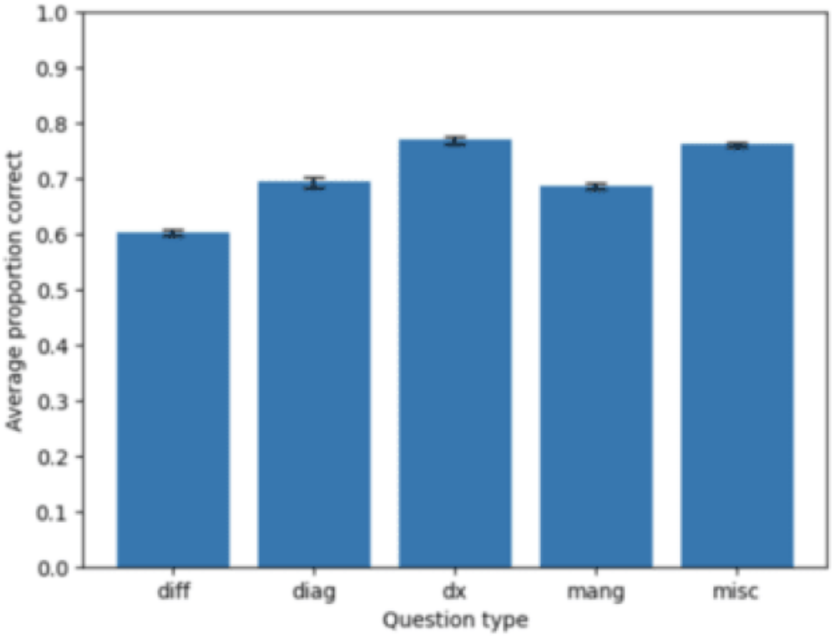
Results – Figure 2c



ChatGPT performance by question type for each of the 36 MSD vignettes; error bars are 1 SE of the mean. *diag*: diagnostic questions; *diff*: differential diagnoses; *dx*: diagnosis questions; *mang*: management questions; *misc*: miscellaneous question.

Table S2: ChatGPT Accuracy by Vignette

Vignette Number	Average Proportion Correct	Standard Error of the Mean
1	75.634921%	0.235725%
2	64.326599%	1.073817%
3	71.883267%	0.458918%
4	69.57294%	0.446449%
5	75.984848%	0.426538%
6	64.074074%	0.878604%
7	77.407407%	0.494539%
8	72.117304%	0.479085%
9	80.092593%	0.353047%
10	72.460317%	0.459257%
11	66.684704%	0.419683%
12	74.772727%	0.3634%
13	78.703704%	0.358782%
14	69.478438%	0.701748%
15	74.177489%	0.964328%
16	77.492151%	0.722724%
17	66.887755%	0.703032%
18	77.993197%	0.442806%
19	58.979107%	1.423041%
20	64.087302%	1.071508%
21	72.09127%	0.871775%
22	62.61992%	0.503332%
23	70.305217%	0.780044%
24	78.562031%	0.347551%
25	58.941799%	0.740824%
26	77.089947%	0.493125%
27	55.929705%	1.23026%
28	83.767952%	0.440924%
29	79.166667%	0.552734%
30	82.738095%	0.437194%
31	78.439153%	0.555443%
32	65.509259%	0.879226%
33	70.833333%	0.491073%
34	64.839958%	0.351188%
35	70.652054%	0.408724%
36	79.298777%	0.163206%



Question Type	Average Proportion Correct	Standard Error of the Mean
Differential	60.354963%	0.577072%
Diagnostic	69.336524%	0.969417%
Diagnosis	76.975309%	0.697913%
Management	68.579747%	0.548833%
misc	76.129811%	0.465247%

Results – Figure 2b

ChatGPT performance by question type; error bars are 1 SE of the mean

Table 1. Multivariable linear regression examining the relationship between ChatGPT accuracy and patient age, gender, and Emergency Severity Index (ESI), as well as question type.

Variable	B Coefficient (%, 95% CI)	P value
Age	-0.05 (-0.17 to 0.60)	.35
Male Gender	1.28 (-3.36 to 5.92)	.59
ESI	-0.98 (-4.15 to 2.96)	.55
Diagnostic Question	-6.62 (-13.42 to 0.18)	.06
Differential Questions	-15.80 (-22.90 to -8.70)	<.001
Diagnosis	-0.89 (-6.42 to 8.21)	.81
Management	-7.44 (-13.93 to -0.9)	.02

Authors’ Discussion

- The authors claimed to show that ChatGPT achieves an accuracy of 60.3% intervening differential diagnosis based on HPI, physical exam, and review of systems alone
- ChatGPT achieves an accuracy of 76.9 in narrowing towards a final diagnosis
- Report an average performance average performance of 71.8% across all vignettes and question types
- When comparing the performance of diagnosis accuracy and miscellaneous accuracy, the performance of ChatGPT approximates its accuracy in providing General Medical facts
- The authors note that the overall accuracy was lower in diagnostic questions and management questions because ChatGPT recommended extra or unnecessary diagnostic testing or clinical intervention respectively
- For several differential diagnosis questions and diagnosis questions, ChatGPT refused to provide a diagnosis altogether suggesting that ChatGPT is not always able to properly navigate clinical scenarios with a well established standard of care
- The authors highlighted that small errors in clinical judgment can result in adverse outcomes evidenced by the recommendation of feudal care or refusing to provide a final diagnosis

Limitations – Author suggested

- The adverse effects are hard to predict given that ChatGPT lacks reasoning capacity
- The authors feel that these cases could have been in the training set but felt that this fact was immaterial given that they were attempting to study the applications of current tools in clinical decision making

Impressions
Open Discussion

- Interesting study that works through standardized vignettes
- Different medical societies view standardized exams differently

Case Vingette 12 – Dyspnea and back pain in a 24yo Male

M12_DIFF	Based on the history and physical examination, the following differential diagnoses cannot be excluded: Pneumonia Pulmonary embolism Spontaneous pneumothorax Tension pneumothorax Thoracic aortic dissection	Asthma exacerbation, lung cancer, pneumonia, pulmonary embolism, spontaneous pneumothorax, and tension pneumothorax cannot be excluded based on the patient's history and physical examination.	Based on the history and physical examination, the following differential diagnoses cannot be excluded: Asthma exacerbation Lung cancer Musculoskeletal back pain Pneumonia Pulmonary embolism Spontaneous pneumothorax Tension pneumothorax Thoracic aortic dissection
M12_DIAG	Given the patient's current history and examination findings, the following steps are most appropriate to do at this time: Chest x-ray ECG D-dimer testing Oxygen 6 L by nasal cannula It is also likely that a CT angiography (CTA) of the chest or thoracic spine x-rays may be performed later in the evaluation, if indicated by the results of the initial tests.	Chest x-ray, oxygen 6 L by nasal cannula, and ECG are the most appropriate steps to do at this time based on the patient's history and examination findings.	Based on the patient's history and examination findings, the following steps are most appropriate to do at this time: Chest x-ray ECG Oxygen 6 L by nasal cannula D-dimer testing A CT angiography (CTA) of the chest or thoracic spine x-rays may be considered if the patient's condition worsens or if the results of initial tests are not conclusive.
M12_MANG	Given the patient's current history and examination findings, the following steps are essential to do immediately: Continue oxygen 6 L by nasal cannula Admit the patient to the hospital Other steps such as a CT scan of the chest with or without IV contrast, increasing supplemental oxygen, or chest tube placement may be necessary later in the evaluation and management, but these decisions should be made based on the results of initial tests and the patient's response to treatment. Discharging the patient, administering azithromycin, or scheduling follow-up chest x-ray in 1 day are not appropriate based on the limited information provided. Needle decompression, video-assisted thoracoscopic surgery, and chemical pleurodesis are more advanced management options that may be considered later in the evaluation, but are not essential at this time.	Continue oxygen 6 L by nasal cannula, and admit the patient to the hospital are essential steps to do immediately based on the patient's history and examination findings.	Based on the patient's history and examination findings, the following steps are essential to do immediately: Continue oxygen 6 L by nasal cannula Increase supplemental oxygen to 100% via nonrebreathing face mask Admit the patient to the hospital CT scan of the chest with IV contrast The other options listed (administering azithromycin, discharging the patient, needle decompression, chest tube placement, and scheduling video-assisted thoracoscopic surgery and chemical pleurodesis) may be considered based on the results of the patient's imaging studies and the progression of their condition.

Case Vingette 12 – Dyspnea and back pain in a 24yo Male

M12_Q1	Approximately 20% of pneumothorax patients have primary spontaneous pneumothorax.	20%	Approximately 20% of pneumothorax cases are classified as primary spontaneous pneumothorax.
M12_Q2	20s	20s	The peak age of patients who develop primary spontaneous pneumothorax is in the late teens to early 30s.
M12_Q3	Male sex, smoking, and family history are risk factors for primary spontaneous pneumothorax.	Male sex, Smoking, Family history.	Male sex and smoking are risk factors for primary spontaneous pneumothorax.