De nauwkeurigheid en volledigheid van ChatGPT in de Belgische huisartspraktijk: een descriptieve analyse van diagnostische casussen

Haio: Vanpoecke Sander , KUL
Promotor: AERTGEERTS Bert , KUL
Praktijkopleider: Jong Sigrid , MARIEN NELE , Van Duffel Marc , WOUTERS Raf

Context: Chat Generative Pre-Trained Transformer (ChatGPT) is artificiële intelligentie, gebaseerd op “advanced large language models”, dat sinds de introductie in 2022 een veelbesproken onderwerp is. De potentiële toepassingen in de geneeskunde worden momenteel onderzocht. In de huisartsgeneeskunde kan ChatGPT helpen bij het creëren van educatiemateriaal, de ontwikkeling van informatiefolders (bijv. plasdagboeken, hoofdpijndagboeken, vertaalde vragenlijsten) en het verbeteren van de communicatie met patiënten. Bovendien heeft het potentieel nut in klinische besluitvorming door te ondersteunen bij het uitbreiden van een diagnostisch landschap, triage en het verkennen van therapeutische opties. Er zijn echter beperkingen. Zo is er geen transparante bronvermelding. Daarnaast zijn de trainingsdata van de openbaar beschikbare versie van ChatGPT 3.5 beperkt tot september 2021, wat kan leiden tot verouderde medische informatie. In tegenstelling hiermee heeft ChatGPT 4.0, beschikbaar via abonnement, directe informatie via het internet, waardoor deze beperking mogelijks wordt verholpen. Gezien deze beperkingen heeft deze studie als doel de betrouwbaarheid en toepasbaarheid van ChatGPT in de huisartsgeneeskunde te evalueren, waar onnauwkeurigheden ernstige gevolgen kunnen hebben.

Onderzoeksvraag: Wat is de betrouwbaarheid en toepasbaarheid van ChatGPT als diagnostisch tool in de huisartsgeneeskunde?

Methode: Deze studie beoordeelde de diagnostische capaciteiten van ChatGPT 3.5 en ChatGPT 4.0 met behulp van 7 diagnostische casussen, uit de Belgische huisartsgeneeskunde. Aan elk model werden de casussen drie keer gepresenteerd. De door ChatGPT gegenereerde antwoorden werden systematisch vergeleken met door experts afgeleide antwoorden. Een Likertschaal werd gebruikt om de nauwkeurigheid van de antwoorden (beoordeeld op 5) en de volledigheid (beoordeeld op 3) te evalueren.

Resultaten: Wat betreft diagnostische nauwkeurigheid behaalde ChatGPT 3.5 een gemiddelde score van 3.524 (SD = 0.8373) op 5, wat wijst op een matige nauwkeurigheid met enige variabiliteit. ChatGPT 4.0 vertoonde een lagere gemiddelde score van 3.048 (SD = 0.218), wat duidt op grotere consistentie maar verminderde nauwkeurigheid. Wat betreft diagnostische volledigheid behaalde ChatGPT 3.5 een gemiddelde score van 2.333 (SD = 0.658), wat duidt op matige volledigheid met enige variabiliteit. ChatGPT 4.0 vertoonde een lagere gemiddelde score van 1.857 (SD = 0.727), wat wijst op minder uitgebreide diagnostische uitkomsten met iets hogere variabiliteit.

Conclusie: Deze studie benadrukt het potentieel en de beperkingen van ChatGPT in een huisartsenpraktijk. ChatGPT toonde goede diagnostische nauwkeurigheid voor de gegeven medische casussen. In deze studie presteerde ChatGPT 3.5 beter dan ChatGPT 4.0 op zowel diagnostische nauwkeurigheid als diagnostische volledigheid, met statistisch significante verschillen in beide categorieën. De kleine steekproefgrootte van deze studie vraagt om voorzichtigheid bij de interpretatie van deze resultaten, aangezien dit ruimte geeft voor bias. Toekomstig onderzoek moet gebeuren met grotere steekproeven om de betrouwbaarheid van deze bevindingen te verbeteren voordat het in de medische praktijk kan worden toegepast.