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

Haio: Devolder Nathalie , KUL
Promotor: AERTGEERTS Bert , KUL
Praktijkopleider: CALLENS Griet , DOUBEL Peter , Vandamme Hilde , VERBEKE Annelies

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 therapeutische tool in de huisartsgeneeskunde.

Methode: Deze studie beoordeelde de therapeutische capaciteiten van ChatGPT 3.5 en ChatGPT 4.0 aan de hand van vijf therapeutische casussen uit de Belgische huisartsgeneeskunde. Aan elk model werden de casussen drie keer gepresenteerd en klinische vragen werden gesteld. De door ChatGPT gegenereerde antwoorden werden systematisch vergeleken met richtlijnen uit klinische databases. Een Likertschaal werd gebruikt om de nauwkeurigheid (beoordeeld op 5) en de volledigheid (beoordeeld op 3) van de antwoorden te evalueren.

Resultaten: Wat betreft therapeutische nauwkeurigheid behaalde ChatGPT 3.5 een gemiddelde score van 3.867 (SD = 1.187), wat wijst op een redelijk accurate therapeutische nauwkeurigheid met aanzienlijke variabiliteit. In tegenstelling daarmee behaalde ChatGPT 4.0 een hogere score van 4.667 (SD = 0.816), wat wijst op betere nauwkeurigheid en minder variabiliteit ten opzichte van ChatGPT 3.5. Wat betreft therapeutische volledigheid behaalde ChatGPT 3.5 een gemiddelde score van 2.667 (SD = 0.617), wat aangeeft dat hoewel therapeutische aanbevelingen matig volledig waren, er enige inconsistentie bestaat. ChatGPT 4.0 vertoonde een lichte verbetering met een gemiddelde score van 2.733 (SD = 0.458), wat wijst op een betere volledigheid en betere consistentie ten opzichte van ChatGPT 3.5.

Conclusie: Deze studie benadrukt het potentieel en de beperkingen van ChatGPT in de huisartsenpraktijk. ChatGPT toonde een matige therapeutische nauwkeurigheid, waarbij ChatGPT 4.0 meer accurate antwoorden gaf. Deze verschillen waren echter statistisch niet significant. Beide modellen vertoonden goede vergelijkbare volledigheid. 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.