Minerva - Duiding - 15/09/2023

Vergelijking van antwoorden tussen artsen en artificiële intelligentie op medische vragen op een openbaar gezondheidsforum (ChatGPT-3.5)

Referentie
Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-96. DOI: 10.1001/jamainternmed.2023.1838

Duiding
Mohamed Ismail Saubry, assistant en médecine générale à l'UCLouvain
Geen belangenconflict met het onderwerp.

Klinische vraag

Hoe valide zijn de antwoorden van ChatGPT versus die van artsen op vragen gesteld op een gezondheidsforum?

Achtergrond

In de voorbije jaren heeft artificiële intelligentie (of AI) als onderzoeksgebied grote vooruitgang geboekt. Er wordt ook onderzoek gedaan naar mogelijke toepassingen op medisch vlak. ChatGPT is een door OpenAI ontwikkelde chatbot op basis van artificiële intelligentie. De chatbot is sinds juni 2020 beschikbaar voor het grote publiek. ChatGPT-3 is getraind om teksten te begrijpen en te genereren met behulp van een enorme en op verschillende talen gebaseerde gegevensverzameling afkomstig van het Internet. Het kan vragen beantwoorden en conversaties opstarten door de aangeboden context te analyseren en relevante en samenhangende antwoorden te genereren in natuurlijke taal (1). In principe zou ChatGPT-3 elke vraag moeten kunnen beantwoorden, dus ook medische vragen. Men kan deze tool gratis gebruiken op de OpenAI-website. De nog krachtigere versie 4 is beschikbaar tegen betaling. Volgens specialisten vertegenwoordigt ChatGPT-4 een aanzienlijke vooruitgang. Tot op heden publiceerde Minerva nog geen duiding met betrekking tot artificiële intelligentie (2).

Samenvatting

Bestudeerde populatie

- gebruikers van een openbaar forum (Reddit).

Onderzoeksopzet

- uit noodzaak en uit pragmatisme, en om snel een beschikbare en gemeenschappelijke databank van vragen van patiënten te kunnen ontwikkelen, verzamelde men vragen van het publiek en patiënten en door artsen gepubliceerde antwoorden op een online sociale media forum, Reddit's r/AskDocs; dit online forum is een sub-Reddit met ongeveer 474 000 leden die medische vragen kunnen posten en die door gescreende zorgverleners op vrijwillige basis beantwoord worden
- men selecteerde willekeurig 195 medische vragen, met het antwoord van een door het forum geverifieerde arts
- de oorspronkelijke vraag, inclusief titel en tekst, bewoorden men voor de analyse, en het antwoord van de arts werd gebruikt als referentieantwoord
- de vragen werden woordelijk ingevoerd in een nieuwe ChatGPT-3.5-sessie
- de antwoorden van de artsen en de antwoorden gegenereerd door ChatGPT-3.5 werden op geblindeerde wijze verzameld

- men onderzocht alleen de antwoorden van artsen omdat aangenomen werd dat die doorgaans beter zouden zijn dan die van andere zorgprofessionals of leken
- drie beoordelaars evalueerden de antwoorden; ze waren allemaal lid van een team van erkende zorgprofessionals werkzaam in de pediatrie, geriatrie, interne geneeskunde, oncologie, infectieziekten en preventieve geneeskunde.

Uitkomstmeting

- blinde beoordeling van het antwoord dat als beste van de twee werd beschouwd
    - kwaliteit van de verstrekte informatie: beoordeeld op een Likert-schaal variërend van 'zeer slecht' (1) tot 'zeer goed' (5)
    - graad van empathie in het antwoord: beoordeeld op een Likert-schaal variërend van 'helemaal niet empathisch' (1) tot 'zeer empathisch' (5)
- de scores van de verschillende beoordelaars werden gesynthetiseerd met behulp van het 'crowd scoring system'; dat systeem levert een waarde op die de consensus tussen beoordelaars weerspiegelt en kent tegelijkertijd een variantie aan deze score toe, hetgeen de mogelijke meningsverschillen tussen beoordelaars reflecteert
- verschil in aantal woorden van de reacties van ChatGPT versus van artsen
- percentage antwoorden waarbij ChatGPT als beter werd beschouwd
- met behulp van tweezijdige T-tests (t) werden de gemiddelde scores voor de kwaliteit van de antwoorden en de empathie in de antwoorden van de artsen vergeleken met die van de chatbots
- de mate van correlatie tussen de kwaliteit van de antwoorden en de empathie in de antwoorden werd gemeten aan de hand van de correlatiecoëfficiënt van Pearson (r).

Resultaten

- het gemiddelde antwoord van artsen was korter dan het antwoord gegenereerd door ChatGPT, met respectievelijk 52 woorden (17-62) versus 211 woorden (168-245); t=25,4; p<0,001
- de beoordelaars gaven in 78,6% van de gevallen de voorkeur aan het antwoord van de chatbot (met een 95% BI van 75,0 tot 81,8%); t=13,3, p<0,001
- de gemiddelde score voor kwaliteit van het antwoord bedroeg 3,256 (op 5) voor artsen en 4,132 (op 5) voor ChatGPT
- de gemiddelde score voor empathie in het antwoord bedroeg 2,147 (op 5) voor artsen en 3,655 (op 5) voor ChatGPT; t=18,9, p<0,001
- met betrekking tot de kwaliteit van het antwoord werd 22,1% (95% BI van 16,4 tot 28,2) van de antwoorden van artsen beoordeeld als goed tot zeer goed, versus 78,5% (95% BI van 7,3 tot 84,1) van ChatGPT
- van de antwoorden van artsen werd 4,6% beoordeeld als empathisch tot zeer empathisch (met een 95% BI van 2,2 tot 7,7) tegenover 45,1% van ChatGPT (met een 95% BI van 38,5 tot 51,8)
- Pearsons correlatiecoëfficiënt tussen de kwaliteit van het antwoord en de empathie in het antwoord lag hoger voor de artsen (r=0,59) dan de voor de AI gegenereerde antwoorden (r=0,32).

Besluit van de auteurs

Hoewel deze cross-sectionele studie gunstige resultaten toont in verband met het gebruik van artificiële intelligentie-wizards om vragen van patiënten te beantwoorden, besluiten de auteurs dat het cruciaal is om verder onderzoek te verrichten vooraleer definitieve conclusies te trekken over de potentiële impact ervan in een klinische context. Ondanks de beperkingen van dit onderzoek en het vaak overmatige enthousiasme rond nieuwe technologieën, biedt deze studie veelbelovende perspectieven aangaande de toevoeging van AI-wizards aan de flow met boodschappen voor patiënten, met de mogelijkheid om de resultaten voor zowel artsen als patiënten te verbeteren.

Financiering van de studie

Burroughs Wellcome Funds, Universiteit van Californië San Diego, Institut PREPARE, National Institute of Health.

Belangenconflicten van de auteurs
Verschillende auteurs melden banden met bedrijven die betrokken zijn bij telegeneeskunde of data-analyse; een van de auteurs geeft aan dat hij aandelen heeft in bedrijven die gespecialiseerd zijn in data-analyse; dezelfde auteur was tot juni 2018 CEO van Good Analytics, een consultancybedrijf actief op het vlak van wiskundige modellering en simulaties (3); twee andere auteurs hebben financiële banden met Good Analytics; één auteur zegt adviseur te zijn voor LifeLink, een ChatBot die wordt gebruikt in de gezondheidszorg; een andere beweert adviseur te zijn bij en aandelen te bezitten van Doximity, een online netwerk en dienst voor telegeneeskunde gericht op zorgprofessionals; andere verstrengelingen werden gemeld, maar zonder indicatie van mogelijk risico van bias voor deze studie.

Bespreking

Beoordeling van de methodologie
Deze originele cross-sectionele observationele studie, overigens zeer actueel gezien de populariteit van ChatGPT, vertoont een aantal relatief belangrijke methodologische zwakheden. Ten eerste is er weinig informatie beschikbaar over de willekeurige selectie van vragen en antwoorden voor deze studie. Dat laat ons niet toe om eventuele selectiebias uit te sluiten. We hebben evenmin duidelijke informatie gevonden over de inclusiecriteria voor de geselecteerde vragen. Een andere beperking betreft de steekproefgrootte van de onderzochte vragen/antwoorden, die erg klein blijft in verhouding tot het aantal geregistreerde gebruikers van het forum (iets minder dan 500 000 op het moment dat de studie werd uitgevoerd). We moeten opmerken dat de auteurs van de studie een steekproef van 200 vragen beoogden, uitgaande van een power van 80% om een verschil van 10 procentpunten tussen de antwoorden van de arts en die van de chatbot (45% versus 55%) te kunnen detecteren. Een mogelijk belangenconflict kan niet met zekerheid worden uitgesloten vermits sommige auteurs zeer nauwe banden hebben met bedrijven die actief zijn op het gebied van IT-toepassingen in de gezondheidszorg. Dit roept vragen op over de validiteit van de beoordeling van de verstrekte antwoorden. De structuur van de door ChatGPT gegenereerde antwoorden kan immers relatief makkelijk worden achterhaald. Op basis van de vaststelling dat het aantal beoordelaars beperkt is en dat er duidelijke banden met bedrijven vermeld zijn, kan wijzen op een zekere vertrouwdheid van dat de auteurs met het gebruik van deze nieuwe technologieën. Bijgevolg lijkt alleen blindering van de antwoorden geen afdoende maatregel te zijn om een echt blind onderzoeksproces te garanderen. Met betrekking tot de belangrijkste evaluatiecriteria (kwaliteit van en empathie in het antwoord) gebruikten de beoordelaars geen gestandaardiseerde criteria. Trouwens, de vraag "Welk antwoord lijkt beter?" is vrij vaag. De auteurs geven niet aan in hoeverre de antwoorden overeenkomen met praktijkrichtlijnen of aanbevelingen. Er wordt evenmin gerapporteerd of er maatregelen zijn genomen om overeenstemming tussen de waarnemers en consistentie in de beoordelingsmethode te garanderen. Een slechte interobserver agreement kan namelijk tot onbetrouwbare en slecht reproduceerbare resultaten geleid hebben. Ook de lengte van de antwoorden werd onder de loep genomen. Zo ging men na of langere antwoorden konden worden geïnterpreteerd als antwoorden van betere kwaliteit en met meer empathie. Andere confounders moeten in rekening gebracht worden, zoals de taalbeheersing van het Engels van de arts-redacteur, diens expertise in het domein waarop de vraag betrekking had of culturele factoren met betrekking tot de ervaren empathie. Het is overigens niet ondenkbaar dat de motivatie van mensen om vragen op een gezondheidsforum te beantwoorden kan slinken, wat niet het geval is met een machine. Voor het criterium 'empathie' was een jury met patiënten misschien geschikter geweest. Ten slotte is het moeilijk om de specifieke context van een openbaar forum te vergelijken met de meer professionele en intieme setting waarin een arts reageert op een elektronische vraag van een patiënt. De controlegroep van artsen die de vragen beantwoordden is daarom mogelijk niet representatief voor de werkelijkheid.

Interpretatie van de resultaten
Uit de resultaten van dit onderzoek blijken de beoordelaars de antwoorden van ChatGPT-3.5 statistisch significant positiever te evalueren op basis van de door de auteurs geselecteerde criteria.

Globaal worden langere antwoorden beschouwd als antwoorden van betere kwaliteit. Antwoorden door AI met gelijke lengte als de antwoorden van artsen blijven echter nog steeds de voorkeur hebben.

Wat zeggen de richtlijnen voor de klinische praktijk?
Het gebruik van artificiële intelligentie lijkt nog niet te zijn opgenomen in richtlijnen voor goede klinische praktijk.

Besluit van Minerva

Deze observationele studie vergeleek de antwoorden van artificiële intelligentie (ChatGPT-3.5) met de antwoorden van artsen op vragen gesteld op een openbaar forum. Door een aantal methodologische tekortkomingen kunnen we hieruit onmogelijk voldoende betrouwbare besluiten trekken. Deze studie heeft echter wel als meerwaarde dat ze het potentiële gebruik van artificiële intelligentie belicht als hulpmiddel voor de medische praktijk. Verdere studies zijn nog steeds nodig om het kader en de beperkingen te preciseren. Bovendien is het evenzeer noodzakelijk om stil te staan bij de ethische en deontologische implicaties met betrekking tot het gebruik van AI in een gezondheidszorgcontext.

Referenties

1. What is ChatGPT ? [Internet, website geconsulteerd op 24/08/2023]. Url: https://help.openai.com/en/articles/6783457-what-is-chatgpt
2. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-96. DOI: 10.1001/jamainternmed.2023.1838
3. Url: https://good-analytics.org/