Fül-orr-gégészeti tünetorientált kérdésekre adott ChatGPT-válaszok értékelése

Molnár Fiona Anna dr.*[1] ▪ Ambrus Andrea dr.*[1]
Sándor Médi dr.[1] ▪ Csanády Miklós Jr. dr.[1] ▪ Perényi-Csáthi Éva[2]
Rovó László dr.[1] ▪ Perényi Ádám dr.[1]

[1]Szegedi Tudományegyetem, Szent-Györgyi Albert Orvostudományi Kar, Fül-Orr-Gégészeti és Fej-Nyaksebészeti Klinika, Szeged
[2]Szegedi Tudományegyetem, Bölcsészet- és Társadalomtudományi Kar, Nyelvtudományi Doktori Iskola, Szeged

Bevezetés: A Chat Generative Pre-Trained Transformer (ChatGPT) egy újonnan kifejlesztett mesterségesintelligencia-alapú nyelvi modell, amely a betegek számára nyújtott könnyű hozzáférhetősége révén egyre gyakrabban jelenik meg egészségügyi információforrásként, azonban válaszainak szakmai pontosságáról és megbízhatóságáról jelenleg korlátozott mennyiségű evidencia áll rendelkezésre.
Célkitűzés: A jelen tanulmány célja az volt, hogy megállapítsuk a ChatGPT használhatóságát fül-orr-gégészeti panaszokra adott válaszai alapján.
Módszer: 10 fül-orr-gégészeti témakörben, a betegek tünetei alapján összesen 24 kérdést fogalmaztunk meg. A kérdéseket egymás után adtuk meg a ChatGPT 4.0 verziójának, amely minden esetben szöveges választ generált. Nyelvi értékelést követően a válaszokat 6 fül-orr-gégész szakorvos értékelte háromfokozatú skálán: helytelen (1 pont), helyes, de hiányos (2 pont), teljes mértékben elfogadható (3 pont).
Eredmények: A nyelvi értékelések alapján a generált válaszok közérthetősége, logikai felépítése és nyelvi szerkezete megfelelőnek bizonyult. A ChatGPT fül-orr-gégészeti kérdésekre adott válaszainak szakmai helyessége 2,00 és 2,83 közötti átlagértékeket mutatott a háromfokozatú skálán. Három kérdés (K4, K9 és K24) esetében azonban szignifikánsan gyengébb lett az eredmény. Négy szakorvos (R1, R2, R4 és R6) pontszámai jól korreláltak egymással, míg R3 és R5 válaszai között statisztikailag szignifikáns eltérés mutatkozott ($p<0,001$).
Következtetés: A ChatGPT által adott válaszok fül-orr-gégészeti témákban nyelvi szempontból alkalmasnak bizonyultak a további orvosszakmai értékelésre, ugyanakkor a szakmai tartalom helyessége változó képet mutatott. Eredményeink alapján a ChatGPT potenciálisan alkalmas lehet laikus felhasználók alapvető tájékoztatására, azonban jelenlegi formájában nem alkalmas a klinikai döntéshozatal támogatására. A jövőben elengedhetetlen a szakterület-specifikus, transzparens és validált mesterségesintelligencia-rendszerek fejlesztése, amelyek megbízható orvosszakmai forrásokon alapulnak, és biztonságosan integrálhatók az egészségügyi ellátórendszerbe.
Orv Hetil. 2025; 166(42): 1666–1674

Kulcsszavak: ChatGPT, fül-orr-gégészet, mesterséges intelligencia az egészségügyben, egészségügyi kommunikáció, orvosszakmai helyesség

Evaluation of ChatGPT's responses to symptom-oriented questions in otolaryngology

Introduction: Chat Generative Pre-Trained Transformer (ChatGPT) is a recently developed artificial intelligence (AI)-based language model that has become an increasingly common source of health-related information due to its accessibility. However, there is limited evidence regarding the accuracy and reliability of its responses.
Objective: This study aimed to assess ChatGPT's usability in otolaryngology by analyzing its answers to common patient questions.

*A két első szerző egyenlő arányban járult hozzá a publikációhoz

Method: 24 patient-oriented questions were created across 10 otolaryngological disease categories and submitted sequentially to ChatGPT version 4.0. The generated responses were evaluated from a linguistic perspective and 6 board-certified otolaryngologists using a three-point scale: incorrect (1), correct but incomplete (2), and correct (3).
Results: Language evaluations indicated that the responses were generally clear, well-structured, and of good quality for further medical evaluations. ChatGPT's medical accuracy scores ranged from 2.00 to 2.83. Three specific questions (Q4, Q9, Q24) received significantly lower ratings. Four raters (R1, R2, R4, R6) showed strong agreement in their evaluations, while significant differences emerged between the scores of R3 and R5 ($p<0.001$).
Conclusion: ChatGPT's responses in otolaryngology were coherent and well-structured, but the accuracy of medical content varied by topic. While the tool may be beneficial for basic patient education, it is not currently reliable enough to support clinical decision-making. Future development of validated, specialty-specific artificial intelligence systems based on trustworthy medical sources will be crucial for safe implementation in healthcare.

Keywords: ChatGPT, otolaryngology, artificial intelligence in healthcare, health communication, medical accuracy

Molnár FA, Ambrus A, Sándor M, Csanády M Jr., Perényi-Csáthi É, Rovó L, Perényi Á. [Evaluation of ChatGPT's responses to symptom-oriented questions in otolaryngology]. Orv Hetil. 2025; 166(42): 1666–1674.

(Beérkezett: 2025. július 23.; elfogadva: 2025. augusztus 23.)

Rövidítések

ChatGPT = (Chat Generative Pre-Trained Transformer) generatív betanított szövegátalakító; DL = (deep learning) mély gépi tanulás; COVID–19 = (coronavirus disease 2019) koronavírus-betegség 2019; K = Kérdés; R = Bíráló

A Chat Generative Pre-Trained Transformer (ChatGPT) egy mesterséges intelligencián alapuló nyelvi modell, amely mélytanulási algoritmusok alkalmazásával képes emberihez hasonló párbeszédek generálására szöveges bemenet alapján. A modell előrejelzéseit egy nagyméretű, többek között weboldalakat, könyveket és tudományos cikkeket is tartalmazó, nyilvánosan nem hozzáférhető szöveges adatbázison végzett betanításon alapuló tanulási folyamat révén végzi [1]. A ChatGPT a rendkívül gyors elterjedését nagyrészt annak köszönheti, hogy intuitív és felhasználóbarát felületet kínál, miközben képes számos különböző feladat ellátására, valamint az emberi nyelv finom árnyalatainak és kontextuális jelentéseinek magas szintű modellezésére. Ennek eredményeként a rendszer számos esetben képes koherens, a kérdéskörhöz illeszkedő válaszokat generálni [2].

A mesterséges intelligencia egészségügyi alkalmazása számos területen jelentősen növelheti a hatékonyságot és a pontosságot, többek között a potenciális kutatási témák azonosításában, valamint a klinikai és laboratóriumi diagnosztikai döntéshozatal támogatásában [3]. Ugyanakkor a ChatGPT alkalmazása több etikai kérdést is felvet. Különösen problematikus a plágium lehetősége, valamint az a tény, hogy a modell a jelenlegi formájában nem képes minden esetben megbízható módon megkülönböztetni a hiteles és a nem hiteles forrásokat [4–7].

Mindazonáltal a ChatGPT formailag jól strukturált, választékos szóhasználatú szövegeket képes generálni, és egyszerűbb kérdések esetén mintegy keresőmotorként is használható. E tulajdonságainak köszönhetően egyre több páciens fordul hozzá elsődleges információforrásként, különösen diagnózisra és lehetséges kezelési módokra vonatkozó panaszai kapcsán. Ugyanakkor a ChatGPT által adott, magyar nyelvű egészségügyi válaszok megbízhatóságát és szakmai pontosságát még nem vizsgálták részletesen. Tanulmányunk célja az volt, hogy megállapítsuk a ChatGPT használhatóságát fül-orr-gégészeti panaszokra adott válaszainak orvosszakmai helyessége alapján.

Módszer

Kérdések, adatforrás

A kérdéseket a klinikai gyakorlat során, ambuláns fül-orr-gégészeti rendelések alkalmával gyakran előforduló betegpanaszok alapján alkottuk meg. Ezek olyan panaszokat érintettek, amelyek rendszeresen előkerülnek az orvos–beteg kommunikáció során. A leggyakrabban felmerülő, tényleges betegségekhez köthető panaszköröket 10 fő témakörbe soroltunk. A végleges vizsgálati kérdéssort e leggyakoribb témakörök alapján állítottuk össze. Ezt követően két szerző (A. A. és M. F. A.) függetlenül átnézte, értékelte és jóváhagyta a panaszokból megalkotott kérdéseket a következő szempontok alapján: szakmailag relevánsak-e, gyakori fül-orr-gégészeti panaszokra vonatkoznak-e, laikus betegek által is közérthetően megfogalmazhatók-e, valamint a ChatGPT számára értelmezhető módon strukturáltak-e.

A statisztikai elemzéshez a kérdéseket 10, fül-orr-gégészeti szakterülethez kapcsolódó témakörbe soroltuk, hogy hatékonyan értékelhessük a ChatGPT teljesítményét: (1) heveny mandulagyulladás (tonsillitis acuta), mandula körüli tályog (abscessus peritonsillaris), garatgyulladás (pharyngitis); (2) fülzsírdugó (impact cerumen), külsőfül-gyulladás (otitis externa), középfülgyulladás (otitis media); (3) orrvérzés (epistaxis); (4) gégegyulladás (laryngitis), gégedaganat (tumor laryn-

gis); (5) nyálkövesség (sialolithiasis); (6) nyaki terimék; (7) orrnyálkahártya-gyulladás (rhinitis); (8) horkolás; (9) fülzúgás (tinnitus); (10) szédülés (vertigo). Végül összesen 24 kérdést használtunk fel a ChatGPT által generált válaszok elkészítéséhez.

Válaszok generálása

A kérdéseket magyar nyelven adtuk meg a ChatGPT 4.0 verziójának *(Kiegészítő az 1. ábrához)*. Minden kérdést külön-külön, egymástól függetlenül futtattunk, a platform „új beszélgetés” („new chat”) funkciójának használatával annak érdekében, hogy a válaszadást ne befolyásolják az előző kérdésekkel való összefüggések.

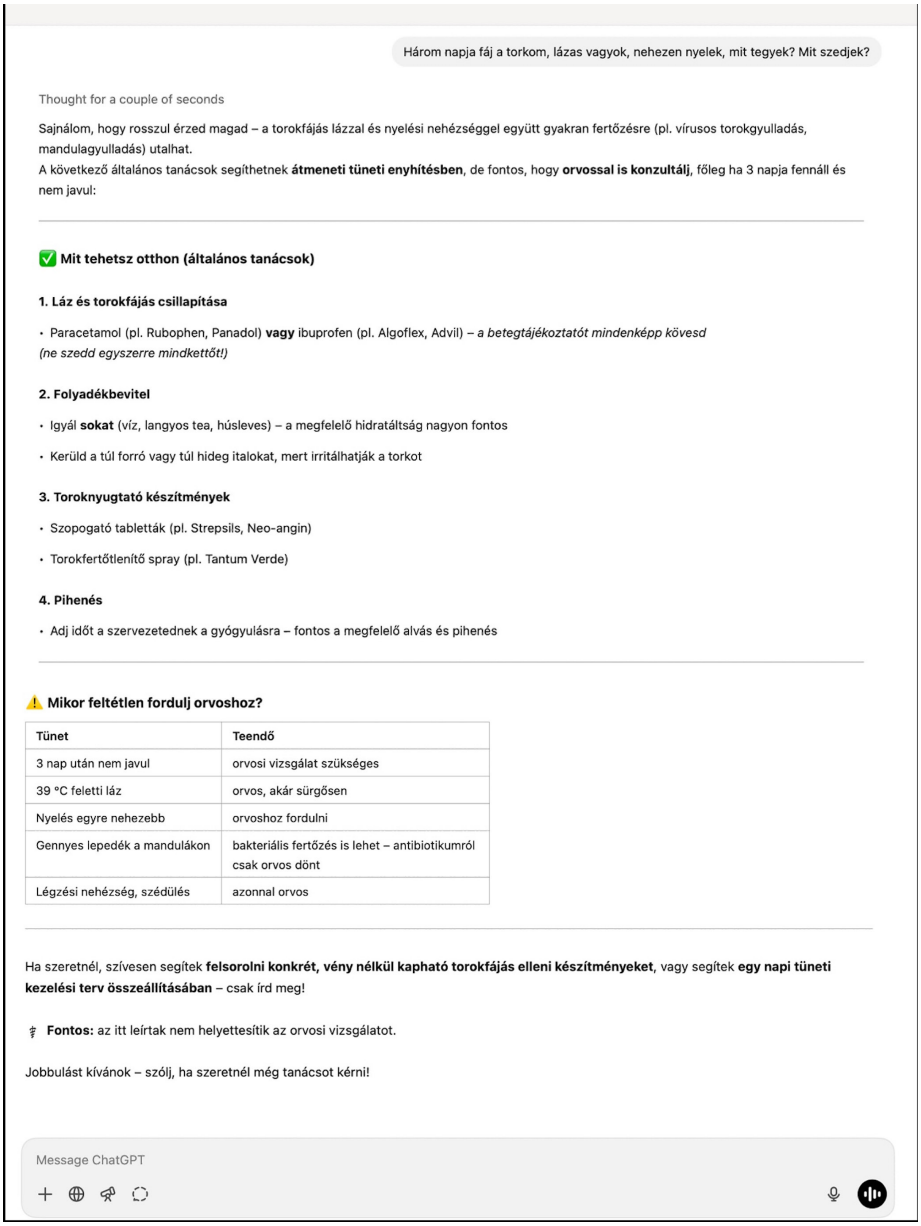
A válaszok értékelése

A ChatGPT által adott válaszokat 6, fül-orr-gégészeti és fej-, nyaksebészeti szakterületen szakvizsgával rendelkező orvos, valamint egy nyelvész szakértő értékelte.

Nyelvi értékelés

A válaszok nyelvi minőségét egy nyelvész végzettségű szerzőtárs értékelte három fő szempont alapján:

I. Közérthetőség: a válaszban használt kifejezések világosak és laikus olvasó számára is érthetők.
II. Szerkesztettség: a mondatok tömörek, logikus felépítésűek.



Kiegészítő 1. ábra Példa a ChatGPT válaszára az 1. tudásalapú kérdésre (K1), amelyet a jelen tanulmányban használtunk. A képernyőkép szó szerint mutatja be a felhasználó és a ChatGPT (OpenAI, GPT-4.0, 2024. márciusi verzió) közötti interakciót az 1. tudáskérdésre (K1) adott válasz esetében. Látható a modell számára megfogalmazott pontos kérdés (felül), valamint a ChatGPT teljes, változtatás és rövidítés nélküli szöveges válasza (alul). Ez a példa kvalitatív szemléltetésként szolgál a modell érvelési módjára és válaszstílusára, és a publikáció céljára nem módosítottuk

1. táblázat A ChatGPT-válaszok értékelése különböző fül-orr-gégészeti állapotok kapcsán. A vizsgált betegségek között heveny és idült kórképek szerepelnek; az eredmények a legalább 5 éves szakmai tapasztalattal rendelkező szakértők (R1–R6) egyéni értékeléseit mutatják. Az alkalmazott osztályozási skála a következő volt: helytelen (1 pont), helyes, de hiányos (2 pont), teljes mértékben elfogadható (3 pont). Minden kérdés nyelvi értékelését egy nyelvész (L1) végezte a következő szempontok alapján: I. közérthető kifejezések használata, II. tömör és logikus mondatszerkesztés, valamint III. a szöveg logikus szerkezete. Az értékelés háromfokozatú pontozási skála szerint történt, ahol 1 pont jelentette a kritériumnak való nem megfelelést, 2 pont a részleges megfelelést, míg 3 pont a teljes mértékű megfelelést. Megjegyzés: a kérdések után zárójelben megadott diagnózis nem szerepelt a keresésben

A kategóriák *dőlt* kiemeléssel szerepelnek *(1–10)* A témaköröket (K-24) az alábbi listában soroltuk fel. A kérdések után zárójelben megadott betegség a kérdéshez tartozó kórképet jelöli, amely alapján a szakértői értékelés tematikus bontása történt. Ezt a megjelölést elsősorban az adatelemzés és kategorizálás átláthatósága érdekében alkalmaztuk, azonban nem feltétlenül tükrözi a laikus kérdező pontos diagnózisát.	R1	R2	R3	R4	R5	R6	L1		
							I.	II.	III.
1. Akut torokpanaszok									
– K1. 3 napja fáj a torkom, lázas vagyok, nehezen nyelek, mit tegyek? Mit szedjek? (Betegség: torokgyulladás, mandula körüli tályog)	2	3	3	3	2	2	3	3	3
– K2. 1 hete fáj a torkom, lázas vagyok, nehezen nyelek, mit tegyek? (Betegség: mandula körüli tályog)	3	2	3	2	2	2	3	3	3
– K3. 2 hete fáj a torkom, nyelési nehezítettségem van, mit tegyek? (Betegség: tumorgyanú)	2	2	2	2	2	2	2	2	3
2. Fülpanaszok									
– K4. 3 napja dugul a fülem, mit tegyek? Mit szedjek? (Betegség: impaktált cerumen)	2	2	3	1	1	2	3	2	2
– K5. 1 hete dugul a fülem, és fáj a fülem, mit tegyek? Mit szedjek? (Betegség: otitis externa)	3	2	2	2	2	2	2	3	3
– K6. 1 hete dugul a fülem, folyik a fülem, fáj, és lázas vagyok, mit tegyek? Mit szedjek? Mi lehet? (Betegség: otitis externa, otitis media)	3	2	2	3	2	2	3	3	3
3. Orrvérzés									
– K7. 3 napja néhány percig vérzik az orrom, mit tegyek? Mit szedjek? (Betegség: epistaxis)	3	2	3	3	2	1	3	2	2
– K8. 2 hete néhány percig vérzik az orrom, mit tegyek? Mit szedjek? (Betegség: epistaxis)	2	2	3	2	2	2	3	2	3
– K9. 2 hete néhány percig vérzik és dugul az orrom. Mit tegyek? Mit szedjek? (Betegség: tumorgyanú)	2	1	3	1	2	2	3	3	3
4. Gégepanaszok									
– K10. Több mint 2 hete rekedt vagyok, mit tegyek? Mit szedjek? (Betegség: laryngitis acuta, tumorgyanú)	3	2	3	1	1	2	3	3	3
– K11. Több mint 2 hete rekedt vagyok, fáj a fülem, véres köpetem van, nyakamon csomót tapintok, mit tegyek? Mit szedjek? (Betegség: laryngitis chronica, tumorgyanú)	3	3	3	2	1	3	3	2	3
5. Nyálmirigykövesség									
– K12. 3 napja evéskor fájdalmat érzek a számban, csomót tapintok a nyakamon, mit tegyek? Mit szedjek? (Betegség: nyálmirigykövesség)	2	3	3	2	1	3	3	3	3
– K13. 2 hete fájdalmat érzek a számban, csomót tapintok a nyakamon, mit tegyek? Mit szedjek? (Betegség: nyálmirigykő/tumorgyanú)	2	2	3	2	1	3	3	3	3
6. Nyaki panaszok									
– K14. 3 napja nyakamon fájdalmas csomót tapintok, mit tegyek? Mit szedjek? (Betegség: infektív eredet)	3	3	3	2	1	3	3	3	2
– K15. 3 napja nyakamon fájdalmatlan csomót tapintok, mit tegyek? Mit szedjek? (Betegség: infektív eredet)	3	3	3	2	2	3	3	3	2
– K16. 1 hete nyakamon fájdalmas csomót tapintok, mit tegyek? Mit szedjek? (Betegség: infektív eredet/tumorgyanú)	3	3	3	2	1	2	3	3	2
– K17. 1 hete nyakamon fájdalmatlan csomót tapintok, mit tegyek? Mit szedjek? (Betegség: infektív eredet/tumorgyanú)	3	2	3	2	2	2	2	2	3
– K18. 2 hete nyakamon fájdalmas csomót tapintok, mit tegyek? Mit szedjek? (Betegség: infektív eredet/tumorgyanú)	1	3	3	2	2	2	3	3	3
– K19. 2 hete nyakamon fájdalmatlan csomót tapintok, mit tegyek? Mit szedjek? (Betegség: infektív eredet/tumorgyanú)	2	3	3	3	2	3	3	3	3
7. Orrpanaszok									
– K20. 3 napja folyik az orrom, dugul az orrom, mit tegyek? Mit szedjek? (Betegség: infektív/allergiás eredet)	3	3	3	3	2	3	3	3	3
– K21. 2 hete folyik az orrom, dugul az orrom, mit tegyek? Mit szedjek? (Betegség: infektív/ krónikus folyamat)	2	3	3	3	2	3	3	3	3
8. Horkolás									
– K22. több mint 2 hete horkolok, nappal fáradt vagyok, mit tegyek? Mit szedjek? (Betegség: horkolás)	3	2	3	3	3	2	3	3	3
9. Fülzúgás									
– K23. 3 napja zúg a fülem, rosszabbul hallok, szédülök, mit tegyek? Mit szedjek? (Betegség: infektív eredet)	1	3	2	3	2	2	3	3	3
10. Szédülés									
– K24. 3 napja szédülök, hányingerem van, mit tegyek? Mit szedjek? (Betegség: periferiás szédülés/stroke)	1	2	1	2	2	2	3	3	3

III. Szövegszerkezet: a válasz egészének szerkezete koherens, jól tagolt, logikusan felépített.

1 pont – nem felel meg a kritériumnak;
2 pont – részben megfelel a kritériumnak;
3 pont – teljes mértékben megfelel a kritériumnak.

A fül-orr-gégész és fej-, nyaksebész orvosokat arra kértük, hogy a ChatGPT által adott válaszokat tartalmilag értékeljék a szakmai pontosság és a teljesség szempontjából. Az értékelés háromfokozatú skálán történt:

- 1 – Helytelen: a ChatGPT válasza téves vagy potenciálisan veszélyes információt tartalmaz.
- 2 – Helyes, de hiányos: a megadott információk önmagukban helyesek, a válasz azonban nem teljes, lényeges elemek hiányoznak belőle.
- 3 – Teljes mértékben elfogadható: a válasz pontos, tartalmilag teljes, a kérdés minden releváns aspektusát lefedi. A szakértői értékelésekből kapott pontszámokat összesítettük, majd ezek alapján elemeztük a ChatGPT teljesítményét a fül-orr-gégészeti témájú kérdések megválaszolásában.

Az értékelésben külön figyelmet fordítottunk a nyelvi érthetőségre és az orvosszakmai helyességre *(1. táblázat)*.

Statisztikai elemzés

Az adatokat a SigmaStat 4.0 szoftverrel (Systat Software Inc., 13. verzió, Palo Alto, CA, Egyesült Államok) elemeztük. Kruskal–Wallis-féle nemparaméteres tesztet végeztünk több adathalmazon is:

1. a nyelvi szempontú értékelésre,
2. a 6 fül-orr-gégész szakorvostól érkező válaszok összehasonlítására,
3. valamint az egyes témák és az összes válasz kombinált eredményének összehasonlítására.

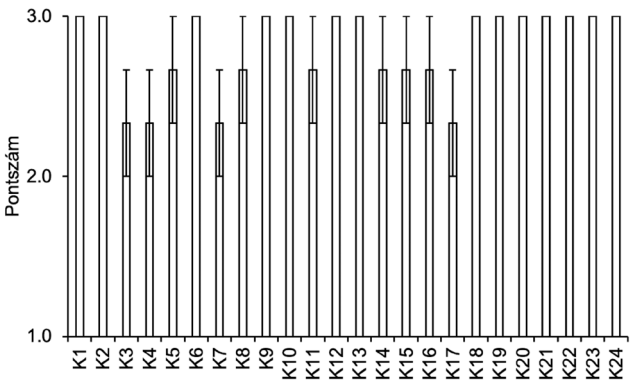
Eredmények

Összesen 24, fül-orr-gégészeti panaszkört érintő kérdést tettünk fel a ChatGPT-nek (*1. táblázat,* K1–K24). Az *1. táblázat* tartalmazza a nyelvész végzettségű szerzőtárs és a fül-orr-gégész szakorvosok által adott értékelések eloszlását. A ChatGPT minden kérdésre választ adott. A kérdéseket 10 különböző témakörbe soroltuk.

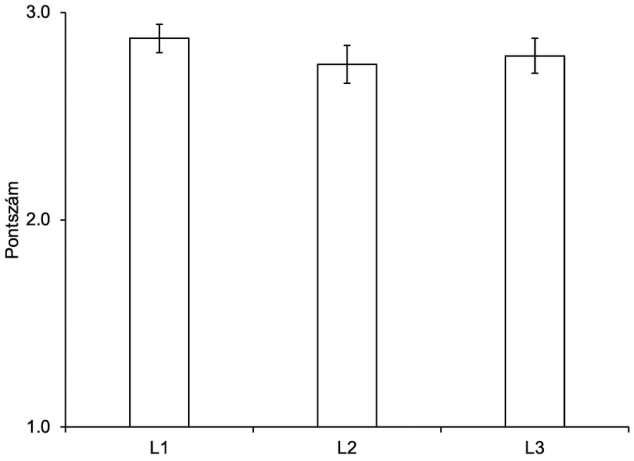
A nyelvi szempont – I. közérthető kifejezések használata (2,88 ± 0,07), II. tömör és logikus mondatok (2,75 ± 0,09), III. logikus szerkezet (2,79 ± 0,08) –, valamint az összesített átlagpontszám alapján sem mutatkozott szignifikáns különbség az egyes kategóriák között *(1. ábra)*. A Kruskal–Wallis-próba alapján egyik vizsgált szempontban sem volt statisztikailag szignifikáns eltérés, ami arra utal, hogy a ChatGPT nyelvhasználata következetesen magas szintű *(2/A és 2/B ábra)*.

A 6 fül-orr-gégész szakorvos által adott pontszámok átlaga alapján a ChatGPT válaszai a szakmai tartalom pontossága és részletessége tekintetében a legtöbb kérdés esetében 2 és 3 pont között mozgott. Kivételt képezett azonban a K4-es (1,83 ± 0,31), a K9-es (1,83 ± 0,31) és a K24-es (1,66 ± 0,21) kérdés, amelyek esetében a mesterséges intelligencia válaszai jelentősen pontatlanabbnak bizonyultak *(3. ábra)*. Különösen a K24-es kérdés kapta a legalacsonyabb átlagpontszámot, ami a tartalmi tévedések nagyobb arányára utal. Ezzel szemben a legmagasabb pontszámot a K20-as kérdés érte el (2,83 ± 0,17), ami a legnagyobb szakmai pontosságot tükrözi.

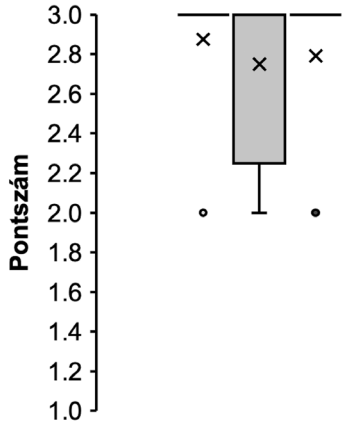
Ezután a kérdéseket témakörök szerint csoportosítottuk, és a ChatGPT-válaszokat értékelő valamennyi fül-orr-gégész szakorvos (R1–R6) által adott pontszám azt mutatta, hogy az átlagpontszám szinte minden témában 2,11–2,75 között alakult, kivéve a 10. témakört, ahol az



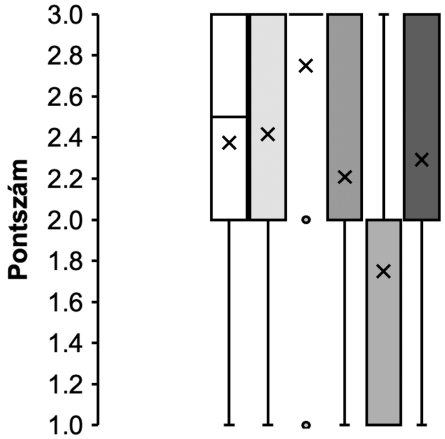
1. ábra A nyelvész szakértő értékelései a ChatGPT-válaszokra. A három skálán (I. közérthető kifejezések használata, II. tömör és logikus mondatok, III. logikus szerkezet) mért átlagpontszámokat és az átlagok standard hibáját számoltuk ki. A pontszámrendszer a következő volt: 1 pont – nem felel meg a kritériumnak, 2 pont – részben megfelel a kritériumnak, 3 pont – teljes mértékben megfelel a kritériumnak



2/A ábra A nyelvész szakértő pontszámai a ChatGPT összes válaszára a három skála alapján. A három szempont (L1–L3) átlagpontszámait és azok standard hibáját számítottuk ki. A pontszámok 1 és 3 között mozogtak: 1 pont – nem felel meg a kritériumnak, 2 pont – részben megfelel a kritériumnak, 3 pont – teljes mértékben megfelel a kritériumnak. A Kruskal–Wallis-próba alapján nem találtunk szignifikáns eltérést az egyes nyelvi szempontok között (L1 és L2: p = 0,28; L1 és L3: p = 0,47; L2 és L3: p = 0,72)



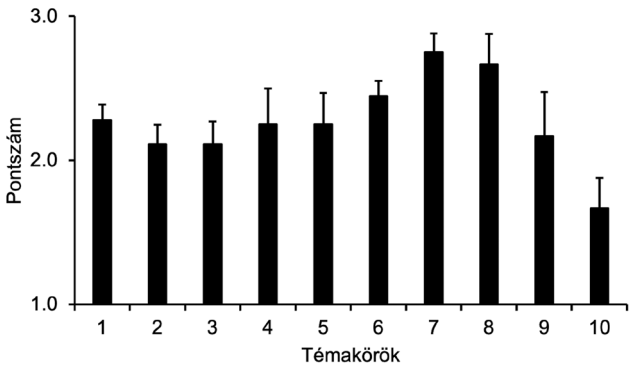
2/B ábra A nyelvész szakértő pontszámai a ChatGPT összes válaszára a három skála alapján (L1: közérthető kifejezések használata, L2: tömör és logikus mondatok, L3: logikus szerkezet). A doboz- és bajuszdiagramon az x az átlagértéket jelöli; a doboz az adatok középső 50%-át (interkvartilis tartomány) foglalja magában; a „bajuszok" a legkisebb és legnagyobb, nem kiugró értékekig nyúlnak; a kiugró értékeket külön pontokkal jelöltük



3. ábra A fül-orr-gégész szakorvosok (R1–R6) által adott értékelések átlagpontszámai és azok standard hibái a ChatGPT 24 válaszára (K1–K24) vonatkozóan. Az értékelés 1 és 3 közötti skálán történt, ahol 1 = helytelen, 2 = helyes, de hiányos, 3 = teljes mértékben elfogadható

átlagpontszám 1,66 ± 0,21 volt *(4. ábra)*. Eközben a legmagasabb átlagpontszámot a 7. témakör érte el (2,75 ± 0,13), ezt követte a 8. témakör (2,66 ± 0,21) és a 6. témakör (2,44 ± 0,11).

A fül-orr-gégészek értékeléseit összehasonlítva, a Kruskal–Wallis-próba alkalmazásával szignifikáns eltérés mutatkozott az R1–R5 (p = 8,38E-04), az R2–R5 (p = 6,26E-0.4), az R3–R4 (p = 2,57E-03) és az R3–R5 (p = 7,58E-08) között. A legmagasabb pontszámot az R3 jelzésű szakorvos adta, aki a 24 ChatGPT-választ átlagosan 2,75 ± 0,15 ponttal értékelte. Ezzel szemben a legalacsonyabb pontszám az R5-tel jelzett szakorvos bírálótól származott, 1,75 ± 0,11 átlagponttal a 24 ChatGPT-válasz esetében. A többi értékelő klinikus pontszámai viszonylag egységesek voltak (R1: 2,38 ± 0,15; R2: 2,42 ± 0,12; R4: 2,21 ± 0,13; R6: 2,29 ± 0,11).



4. ábra Az egyes témakörökre adott pontszámok összehasonlítása a fül-orr-gégész szakorvosok értékelései alapján. Az egyes témakörök átlagpontszámait és azok standard hibáit számítottuk ki. Szignifikáns különbséget figyeltünk meg a 7. és a 10. témakör között: a Kruskal–Wallis-teszt alapján a p-érték $9{,}95 \times 10^{-4}$ volt, amely alacsonyabb a Bonferroni-korrekcióval módosított α-szintnél ($p = 1{,}10 \times 10^{-3}$). A pontozási rendszer 1-től 3-ig terjedő skálán mozgott, a következő értelmezéssel: 1 = helytelen, 2 = helyes, de hiányos, 3 = teljes mértékben elfogadható

Megbeszélés

Napjainkban a betegek egészségügyi panaszok vagy tünetek észlelésekor gyakran nem orvoshoz fordulnak először, hanem online forrásokból próbálnak tájékozódni. E források között egyre nagyobb figyelmet kapnak a mesterséges intelligencia által vezérelt rendszerek, különösen a ChatGPT, amely gyorsan és széles körben vált elérhetővé az interneten. A ChatGPT ingyenesen hozzáférhető, mesterséges intelligencián alapuló, párbeszédorientált nyelvi modell, amely képes lehet egészségügyi jellegű információk közvetítésére is. A betegek különböző módokon próbálnak tájékozódni betegségeikről. A közelmúltig az internetes keresőmotorok voltak az elsődlegesen alkalmazott eszközök erre a célra. 2022. novemberi megjelenése óta a ChatGPT-t széles körben alkalmazzák betegtájékoztatásra, tünetek előzetes értékelésére és betegségmegelőzésre; különösen a COVID–19-járvány utáni időszakban terjedt el a használata, ami felgyorsította a telemedicina és a digitális egészségügyi eszközök elterjedését is [8]. Annak ellenére, hogy egyszerű felhasználhatósága révén rendkívül ígéretesnek tűnik, a ChatGPT klinikai alkalmazhatósága továbbra is heves vita tárgyát képezi, a leginkább azzal kapcsolatban, hogy mennyire pontos és átlátható az általa prezentált információ, valamint milyen súlyos kockázattal járhat az esetleges félretájékoztatás.

Nyelvi szempontú értékelés

A nyelvi szempontú értékelés a generált szöveg értelmezhetőségét kívánta tesztelni, vagyis azt, hogy a szöveg megértését biztosan nem annak nyelvi alapú feldolgozása korlátozza. A nyelvi értékelést végző szerzőtárs megállapította, hogy a 24 válaszból 14 általánosságban jól érthető és koherens volt, ugyanakkor a nyelvi feldolgo-

zás szempontjából a többi válasz is megfelelőnek bizonyult.

Az általunk is tapasztalt hiányosságok arra utalnak, hogy a ChatGPT válaszai gyakran olyan szaknyelvezeti szintet képviseltek, amely inkább orvosi ismeretekkel rendelkező felhasználóknak kedvezett, esetenként csak számukra volt értelmezhető annak ellenére is, hogy a kérdések megalkotásakor szigorúan ügyeltünk a laikus nyelv használatára [9].

Ennek ellenére a mesterséges intelligencián alapuló eszközök elősegíthetik a betegek aktív részvételét saját ellátásukban, maximalizálva azt, hogy orvosi segítséget a valóban indokolt esetekben kérjenek. A ChatGPT mindennapi, párbeszédalapú stílusa tehát segítheti a pácienseket abban, hogy felkészültebben jelenjenek meg orvosi konzultációkon [9], annak ellenére is, hogy a pontosság terén a jelen verziójának ismert – és általunk is észlelt – hiányosságai lehetnek. Egyes kutatások szerint, míg a ChatGPT nagy pontszámot ért el az empátia (4,28/5) tekintetében, a pontosság (3,76/5) és az alaposság (3,59/5) területén mérsékeltek voltak ezek a mutatók az általuk alkalmazott pontrendszer alapján [10]. Különösen aggasztó, hogy az általuk [10] vizsgált válaszok közül öt potenciálisan veszélyesnek bizonyult, ami komoly kockázatot jelenthet abban az esetben, ha a felhasználók kritikátlanul és felelősségteljes mérlegelés nélkül bíznak meg a rendszer válaszaiban.

A ChatGPT teljesítménye némelyik, rosszindulatú daganatos megbetegedésekre vonatkozó kérdéseinkben (például K9, K11) közepes volt, ami összhangban van más tanulmányokban foglalt eredményekkel. Egyes vizsgálatok szerint a ChatGPT 96,9%-os pontosságot ért el rosszindulatú daganatokkal kapcsolatos tévhitekre adott válaszokban, megközelítve a National Cancer Institute 100%-os eredményét [11]. Ugyanakkor más vizsgálatokban a fej-nyaki daganatok területén a modell diagnosztikai pontossága (88,9%) a legalacsonyabb értékű volt a vizsgált hat klinikai szakterület közül [12]. A szerzők rámutattak arra is, hogy a ChatGPT olykor nem ismerte fel megfelelően az egyértelműen alarmírozó, daganatra gyanús tüneteket. Más eredmények tovább erősítik ezt a képet: elemzésük szerint a megadott differenciáldiagnózisok 27–29%-a teljes mértékben valószínűtlennek bizonyult a rendelkezésre álló tünetek ismeretében [13]. Mindez kiemeli a szakemberek általi kritikus felügyelet szükségességét az ilyen rendszerek egészségügyi alkalmazásakor.

Más tanulmányok is alátámasztják eredményeinket: fül-orr-gégészeti kérdéseket vizsgálva a legnagyobb pontosságot allergológiai témákban tapasztalták (72%), ami összhangban áll a rinológiai kérdéseink magas pontszámaival [14]. Ezzel szemben egyes vizsgálatok szerint a ChatGPT nehezen boldogult bizonyos orrmelléküregi betegségekkel – például az odontogen sinusitisszel. A nem specifikus tünettan gyakran bizonytalan válaszokat idézett elő, emellett a modell által nyújtott kezelési tanácsadás sem bizonyult kielégítőnek [15].

A nyálmirigy-rendellenességek kezelésével kapcsolatban megállapították, hogy bár a ChatGPT klinikailag elfogadható terápiás opciókat javasolt, a szakértők elsődleges ajánlásaival való egyezés alacsony maradt [16]. Jelen tanulmányunkban a K15-ös kérdés (akut nyirokcsomógyulladás), amely a nagy nyálmirigyeket is érintheti, hasonló tendenciát mutatott, hiszen a diagnózis felállítása megfelelő volt, ugyanakkor a kezelési stratégiák bizonytalanok, esetenként helytelenek voltak.

Teljesítménybeli eltérések a fül-orr-gégészeti tünetcsoportok között

Eredményeink jelentős eltéréseket mutattak a ChatGPT válaszainak pontosságában a különböző tünetcsoportok között. Az olyan kérdések esetében, mint a K4 (fülduguláls), K9 (orrgarati daganat gyanúja) és K24 (heveny szédülés, esetleges stroke), a modell alacsony pontszámot ért el. Ez rávilágít arra, hogy a ChatGPT nehezen kezeli azokat az eseteket, amelyekben több diagnózis is felmerülhet (K4, K24), illetve korlátozottan képes felismerni azokat a tüneteket, amelyek egy szakember számára egyértelműen riasztóak (például K9). E megállapításaink összhangban állnak *Maksimovski és mtsai* eredményeivel, akik a ChatGPT válaszainak mindössze 45,5%-át találták helyesnek fül-orr-gégészeti témákban, míg a válaszok 21,6%-a egyértelműen hibásnak bizonyult [17]. Ez is alátámasztja, hogy a modell jelenleg korlátozott klinikai háttérismerettel rendelkezik, különösen olyan szakterületeken, ahol bizonyítékon alapuló, szaktudást igénylő értelmezésre van szükség [17]. Ezzel szemben a legjobb értékeléseket azok a viszonylag egyszerű, tünetek alapján könnyen azonosítható állapotok kapták, mint a K15 (akut nyirokcsomó-gyulladás), K20 (heveny nátha), K22 (horkolás) és K23 (akut fülzúgás). Ez összhangban áll *Mete és mtsai* megállapításaival, akik kedvezőbb válaszminőséget figyeltek meg fülészeti jellegű témákban – például fülzúgás esetében –, míg az otosclerosis gyógyszeres és sebészi kezelési lehetőségeivel kapcsolatos kérdésekre adott válaszok pontatlanabbnak (52,95%) bizonyultak [18].

Ez az inkonzisztencia – még az egyes fül-orr-gégészeti részterületeken belül is – arra utal, hogy a mesterséges intelligencián alapuló modelleket, mint a ChatGPT jelenleg használt verzióját, tovább kell finomítani, fejleszteni és szakmai háttértudással felruházni. Például *Radulesco és mtsai* megállapították, hogy a ChatGPT az esetek 62,5%-ában helyesen azonosította a rinológiai diagnózisokat, de gyakran túlzott mennyiségű diagnosztikai vizsgálatot javasolt [19]. Jelen tanulmányunk szintén megfelelő szakmai pontosságot mutatott a náthával kapcsolatos kérdésekben, azonban a modell legfőbb korlátait a részletes kezelési javaslatok tekintetében észleltük.

Etikai aggályok és a mesterséges intelligencia korlátai orvosi kontextusban

Számos lehetséges előnye [20, 21] ellenére a ChatGPT felhasználása komoly etikai és gyakorlati aggályokat vet fel. A „mesterségesintelligencia-hallucináció" – azaz hihető, de helytelen információk generálása – visszatérő probléma [22]. Ez különösen veszélyes lehet klinikai helyzetekben, ahol a félretájékoztatás késleltetheti a kritikus diagnózisok helyes megállapítását, és ezzel együtt a megfelelő kezelések elmaradásához vagy hibás alkalmazásához vezethet. Bár a ChatGPT által generált válaszok gyakran tartalmaznak figyelmeztetéseket – például: „Nem vagyok orvos" –, amelyek a leginkább a fejlesztők számára jelentenek jogi biztonságot, és bizonyos szintű etikai biztosítékként is szolgálnak, a hivatkozások hiánya – ahogy azt mi is megfigyeltük – aláássa a tanácsai iránti bizalmat [23].

Más tanulmányokban megállapították, hogy bár a ChatGPT gyakran javasolta szakember felkeresését, hiányoztak belőle az olyan források, mint az *UpToDate* mélyreható hivatkozási rendszere [22]. Hasonlóképpen, a mi vizsgálatunkban sem szerepelt egyetlen hivatkozás sem az általa generált válaszokban, ami korlátozza ezek felhasználhatóságát az evidenciákon alapuló orvoslásban. Vizsgálatok hangsúlyozzák, hogy a mesterséges intelligencia jelen verziójából hiányoznak azok a mély szakmai ismeretek, amelyek nélkülözhetetlenek a megbízható orvosi szövegalkotáshoz, ezért elengedhetetlen a klinikai szaktekintélyek általi szigorú ellenőrzés a klinikai gyakorlatban való alkalmazás előtt [4]. Az aggályok között szerepel a szerzői jog megsértésének lehetősége, az orvosjogi felelősség kérdése, valamint az elfogult vagy kitalált válaszok kockázata. Ezek a veszélyek kiemelik a gondosan válogatott, lektorált orvosi adatokon alapuló, szakterület-specifikus mesterségesintelligencia-rendszerek fejlesztésének fontosságát [9, 10]. Az ilyen rendszerek pontosabb és átláthatóbb működést, valamint alaposabb forrásellenőrzést tesznek lehetővé, ami kiemelten fontos a mindennapi gyakorlatban való alkalmazás előtt.

Következtetés

A ChatGPT mint nagy méretű nyelvi modell a betegek előzetes tájékozódásának és információgyűjtésének ígéretes eszköze lehet a fül-orr-gégészet területén. Átláthatóságának hiánya és az evidenciákra épülő, validált források mellőzése nem teszi lehetővé, hogy önálló klinikai döntéstámogatásként használjuk. A leginkább abban segítheti a betegeket, hogy tüneteik és kérdéseik alapján tájékozódjanak állapotukról, de nem helyettesítheti a képzett szakemberek klinikai tapasztalatát és terápiás javaslatait. A jövőben kulcsfontosságú a szakterület-specifikus, átlátható és validált mesterségesintelligencia-rendszerek fejlesztése (mint például a deep learning – DL), amelyek képesek érthető nyelvezetű, laikusok és szakemberek számára egyaránt hasznos, megbízható orvosi irányelvekre és naprakész tudományos adatokra épülő tájékoztatást adni. Mindezek megvalósítása még kihívást jelent. Addig is a ChatGPT és a hasonló eszközök használata körültekintést és kritikus hozzáállást igényel, és egészségügyi alkalmazásuk kizárólag képzett szakemberek felügyelete mellett javasolt.

Anyagi támogatás: A közlemény megírása, illetve a kapcsolódó kutatómunka nem részesült anyagi támogatásban.

Szerzői munkamegosztás: M. F. A.: A vizsgálati terv kidolgozása, a kézirat szövegezése, irodalmi áttekintés és statisztika. A. A.: A vizsgálati terv kidolgozása, a kézirat szövegezése, irodalmi áttekintés. S. M., Cs. M. Jr.: Szakmai véleményezés. P.-Cs. É.: Irodalmi áttekintés, szakmai tanácsadás. R. L.: Szakmai tanácsadás. P. Á.: Szakmai tanácsadás, a kézirat szövegezése. A közlemény végleges változatát valamennyi szerző elolvasta és jóváhagyta.

Érdekeltségek: A dolgozattal kapcsolatban a szerzőknek nincsenek érdekeltségeik.

Irodalom

[1] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. OpenAI, 2018. Available from: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed: July 1, 2025].
[2] OpenAI. ChatGPT: Optimizing language models for dialogue. 2022. Available from: https://openai.com/blog/chatgpt [accessed: July 1, 2025].
[3] Kharat A, Zhang L, Mehta N, et al. Applications of artificial intelligence in medical diagnostics and healthcare delivery. J Biomed Inform. 2022; 130: 104093.
[4] Biswas S. ChatGPT and the rise of medical misinformation: a critical review. BMJ Health Care Inform. 2023; 30: e100612.
[5] King MR, Chauhan V, Miller DD. Ethical concerns in large language model use in medicine. Nat Med. 2023; 29: 249–251.
[6] Gordijn B, Have H. ChatGPT: ethical issues in the era of generative artificial intelligence. Med Health Care Philos. 2023; 26: 1–3.
[7] Chatterjee K, Dethlefs N. Explainable AI in healthcare: a critical review and path forward. Artif Intell Med. 2023; 140: 102412.
[8] Angyal D, Molnár E, Tóth L. Role of the artificial intelligence in the digital healthcare in the post-COVID–19 era. [Az AI szerepe a digitális egészségügyben a COVID–19 utáni időszakban.] Egészségügyi Informatika 2024; 19(1): 33–41. [Hungarian]
[9] Zalzal GH, Park C, Moon J. Linguistic complexity and patient accessibility of ChatGPT responses in otolaryngology. AI Health Commun. 2024; 9(1): 45–52.
[10] Carnino JM, Chong NY, Bayly H, et al. AI-generated text in otolaryngology publications: a comparative analysis before and after the release of ChatGPT. Eur Arch Otorhinolaryngol. 2024; 281: 6141–6146.
[11] Johnson BC, Lee MJ, Anderson RK. Evaluating ChatGPT's ability to address cancer myths compared to NCI resources. J Cancer Educ. 2023; 38: 1023–1030.
[12] Oguz S, Demirtas B, Karamustafaoglu Y. Head and neck cancer evaluation by ChatGPT: a diagnostic accuracy study. Clin Oncol Inform. 2023; 14: 199–205.

[13] Lechien JR, Khalife M, Saussez S. Limitations of ChatGPT in differential diagnosis of head and neck symptoms. ENT J. 2023; 102: 412–419.
[14] Hoch S, Eisenhardt D, Lemke J. Assessment of ChatGPT's performance in answering ENT-related medical questions. Laryngo-Rhino-Otologie 2023; 102: 275–280.
[15] Saibene AM, Pipolo C, Felisati G, et al. Can ChatGPT assist with sinonasal disease management? Evaluation in odontogenic sinusitis. Acta Otorhinolaryngol Ital. 2024; 44:109–115.
[16] Chiesa M, Grandi C, Armaroli G. ChatGPT as a tool in salivary gland disorders: diagnostic and therapeutic performance. J Otolaryngol Res. 2023; 5: 88–94.
[17] Maksimovski V, Petrov T, Szabó D. ChatGPT accuracy in otolaryngology: an Eastern European multicenter analysis. Int J Otolaryngol AI. 2024; 3: 78–86.
[18] Mete O, Kuru T, Ayhan F. Evaluating AI-generated responses to neurotological patient questions. Otology Today 2025; 12: 13–20.
[19] Radulesco T, Laccourreye O, Bozec A. Use of ChatGPT in rhinology: accuracy and overdiagnosis concerns. Eur Arch Otorhinolaryngol. 2024; 281: 123–130.
[20] Angyal V, Bertalan Á, Domján P, et al. ScreenGPT – The opportunities and limitations of artificial intelligence in primary, secondary and tertiary prevention. [ScreenGPT – A mesterséges intelligencia alkalmazásának lehetőségei és korlátai a primer, szekunder és tercier prevencióban.] Orv Hetil. 2024; 165: 629–635. [Hungarian]
[21] Horváthné Kónya E, Virág A, Lajkó P, et al. Artificial intelligence in health education: blessing or curse? [Mesterséges intelligencia az egészségügyi oktatásban: áldás vagy átok?] Orv Hetil. 2024; 165: 2061–2064. [Hungarian]
[22] Athaluri V, Kodumuri N, Ponnaganti A. Artificial hallucinations and the ethical implications of ChatGPT in medicine. J Med Ethics AI. 2023; 12: 45–52.
[23] Karimov T, Aslan M, Chen Y. The missing link: source transparency and clinical trust in ChatGPT medical outputs. Health Inform Int. 2024; 21: 17–25.

(Molnár Fiona Anna dr.,
e-mail: drmolnarfiona@gmail.com)

„Sensus, non aetas, invenit sapientiam."
(Nem a kortól, az értelemtől függ a bölcsesség.)