ARTYKUŁ ORYGINALNY / ORIGINAL PAPER
Otrzymano/Submitted: 07.11.2025 • Zaakceptowano/Accepted: 09.03.2026

© *Akademia Medycyny*

Nowe spojrzenie na kompetencje specjalisty geriatrii w erze sztucznej inteligencji. Zastosowanie modeli językowych w analizie treści Państwowego Egzaminu Specjalizacyjnego (PES)

A new perspective on the competencies of geriatric specialists in the era of artificial intelligence: the application of language models in the analysis of the State Specialization Exam (PES)

Joanna Monika Najbar[1], Oskar Makuch[1], Małgorzata Stompór[2]

[1] Studenckie Koło Naukowe z Dziedziny Geriatrii i Gerontologii, Unwersytet Warmińsko Mazurski w Olsztynie, Olsztyn, Polska

[2] Katedra Medycyny Rodzinnej i Chorób Zakaźnych, Uniwersytet Warmińsko-Mazurski w Olsztynie, Olsztyn, Polska

Streszczenie

Wstęp. Geriatria stanowi jedną z najbardziej wymagających dziedzin medycyny, ze względu na wielochorobowość oraz konieczność uwzględniania aspektów funkcjonalnych, poznawczych i społecznych. W ostatnich latach rośnie zainteresowanie wykorzystaniem sztucznej inteligencji (AI) jako potencjalnego wsparcia w opiece nad seniorami. Pojawia się pytanie, czy modele językowe AI mogą osiągnąć poziom kompetencji specjalisty geriatrii. ***Cel pracy.*** Celem niniejszego badania było zbadanie zdolności współczesnych modeli AI do posługiwania się wiedzą specjalistyczną w geriatrii na podstawie analizy odpowiedzi na pytania z Państwowego Egzaminu Specjalizacyjnego (PES) oraz identyfikacja obszarów, w których przewyższają, a w których ustępują lekarzom geriatrom. ***Materiały i metody.*** Przeanalizowano pytania z 10 egzaminów specjalizacyjnych PES z różnych dziedzin medycyny w sesji jesiennej 2024 roku. Wyniki trzech modeli językowych AI (DeepSeek-V3, GPT-4o oraz Gemini 2.0 Flash-Lite) zestawiono z osiągnięciami lekarzy przystępujących do egzaminu. Dodatkowo zbadano mechanizmy popełniania błędów przez wszystkie modele, skupiając się przede wszystkim na PES z Geriatrii. ***Wyniki i omówienie.*** W PES z geriatrii AI osiągnęła wyraźnie lepsze wyniki niż lekarze – średnia wszystkich modeli wyniosła 98,7 pkt, przewyższając średni wynik lekarzy rezydentów o 9,5 pkt. Modele AI bardzo dobrze radziły sobie z zadaniami strukturyzowanymi. Największe ograniczenia AI dotyczyły: wykrywania nietypowych prezentacji chorób u seniorów, podejmowania decyzji w przypadku pacjentów z wielochorobowością, i przewidywania kaskad lekowych. ***Wnioski.*** AI może efektywnie wspierać lekarzy geriatrów w zadaniach strukturyzowanych i opartych na wytycznych, jednak nie zastąpi klinicznej intuicji potrzebnej przy nietypowych prezentacjach i złożonych decyzjach terapeutycznych. Największą wartość przynosi model współpracy lekarz-AI, zwiększający bezpieczeństwo farmakoterapii i usprawniający proces diagnostyczny w geriatrii. *Geriatria 2026;20:24-31. doi: 10.53139/G.20262001*

Słowa kluczowe: sztuczna inteligencja, geriatria, Państwowy Egzamin Specjalizacyjny, wspomaganie decyzji klinicznych

Abstract

Introduction. Geriatrics is one of the most challenging fields in medicine due to the high prevalence of multimorbidity and the need to integrate functional, cognitive, and social factors into patient care. Recently, artificial intelligence (AI), particularly large language models (LLMs), has emerged as a potential tool to support the management of older adults. This study examines whether AI LLMs can demonstrate competency comparable to board-certified geriatric specialists. ***Objective.*** To evaluate the ability of AI LLMs to apply specialized geriatric knowledge by analyzing their performance on the Polish National Specialty Examination (PES) and to identify areas where AI excels

ORCID: Joanna Monika Najbar 0009-0003-2667-5700, Oskar Makuch 0009-0008-9738-5045, Małgorzata Stompór 0000-0001-7746-4506

or falls short compared with human geriatricians. ***Materials and methods.*** 10 PES exams from the Autumn 2024 session were analyzed. Three AI LLMs (DeepSeek-V3, GPT-4o, and Gemini 2.0 Flash-Lite) were tested and compared with resident physicians. Incorrect responses were analyzed to identify specific limitations of AI in geriatric practice. ***Results.*** AI models outperformed physicians in the PES geriatrics section, achieving an average score of 98.7 points, 9.5 points higher than residents. Models excelled in structured, guideline-based tasks but struggled with atypical disease presentations, decision-making in multimorbidity, and predicting medication cascades. ***Conclusions.*** AI LLMs can effectively support geriatricians in structured clinical tasks but cannot replace the intuition required for complex cases. Their greatest value lies in a collaborative model, enhancing diagnostic efficiency, guiding medication safety, and complementing human decision-making in geriatrics. *Geriatria 2026;20:24-31. doi: 10.53139/G.20262001*

Keywords: artificial intelligence, geriatrics, National Specialty Exam, clinical decision suport

Wstęp

Geriatria jako specjalizacja medyczna koncentrująca się na holistycznym podejściu do leczenia i opieki nad osobami starszymi, wymaga uwzględnienia wielochorobowości, polipragmazji oraz aspektów psychospołecznych pacjentów, co jest dużym wyzwaniem dla lekarzy innych dziedzin medycyny. Starzenie się społeczeństw na całym świecie stawia przed systemami opieki zdrowotnej bezprecedensowe wyzwania. W Polsce widać to szczególnie wyraźnie, gdyż osoby starsze stanowią ponad 18% populacji, a do połowy XXI wieku odsetek ten może przekroczyć nawet 30% [1]. W tym kontekście sztuczna inteligencja (AI, ang. *Artificial Intelligence*), szczególnie modele uczenia maszynowego i duże modele językowe (LLM, ang. *Large Language Model*), wyłaniają się jako potencjalnie przełomowe narzędzia, które mogą wspierać klinicystów w podejmowaniu decyzji diagnostycznych i terapeutycznych [2].

LLM to zaawansowane modele uczenia maszynowego trenowane na ogromnych zbiorach tekstowych, które potrafią generować i rozumieć język, co umożliwia im interakcję z użytkownikami w sposób zbliżony do człowieka [2]. Wdrożenie różnych modeli sztucznej inteligencji, w tym LLM, oraz innych inteligentnych technologii wspomagających (IAT, ang. *Intelligent Assistive Technology*) dają obiecujące możliwości w poprawie zarządzania chorobami przewlekłymi osób starszych [3-5]. Technologie te mogą poprawiać wykorzystanie zasobów medycznych, wspierać pracę zespołów w zakładach opiekuńczych, wzmacniać edukację zdrowotną, usprawniać zarządzanie lekami oraz zapewniać wsparcie psychologiczne [3]. Jednakże zastosowanie AI i IAT w opiece nad seniorami wiąże się zarówno z barierami, jak i czynnikami sprzyjającymi implementacji na poziomie indywidualnym, społecznym, organizacyjnym, ekonomicznym i etycznymi [2]. AI w medycynie ma ogromny potencjał i wydaje się rozwiązywać wiele problemów związanych ze starzejącym się społeczeństwem oraz brakami w zasobach ludzkich, choć w praktyce wdrażanie jej rozwiązań często pozostaje suboptymalne. Przeniesienie rozwiązań AI bądź IAT z laboratoriów do codziennej opieki nad pacjentem bywa trudne, a kluczowe wyzwania związane z ich wdrożeniem obejmują m.in. akceptację przez użytkowników, przestrzeganie wytycznych, zachowanie zasad etyki, jakość zbieranych danych i ochronę prywatności [3,5].

Cele

Celem niniejszego badania było zbadanie zdolności współczesnych modeli sztucznej inteligencji (AI) do posługiwania się wiedzą specjalistyczną w geriatrii na podstawie analizy odpowiedzi na pytania z Państwowego Egzaminu Specjalizacyjnego (PES) Badanie miało charakter wielowymiarowy i obejmowało: ocenę kompetencji AI w geriatrii, porównanie wyników AI i lekarzy, analiza stopnia trudności pytań dla AI w odpowiedziach na PES z geriatrii w porównaniu z pytaniami z egzaminu PES dla z innych specjalizacji lekarskich, identyfikacja ograniczeń i typowych błędów modeli oraz próba określenia potencjału AI w wsparciu decyzji klinicznych.

Materiały i metody

Badanie miało charakter retrospektywno-porównawczy i obejmowało analizę wyników Państwowego Egzaminu Specjalizacyjnego (PES) w sesji jesiennej 2024 roku w dziesięciu specjalizacjach lekarskich, ze szczególnym uwzględnieniem geriatrii – ze względu na jej wyjątkową złożoność, wymagającą całościowej oceny pacjenta, stanowiącej jedno z największych wyzwań dla systemów AI.

W badaniu wykorzystano trzy nowoczesne modele generatywne AI – GPT-4o, Deepseek-V2, Gemini 2.0, które zostały poddane ocenie pod kątem zdolności do

rozwiązywania pytań egzaminacyjnych o charakterze teoretycznym. W analizie skoncentrowano się na porównaniu wyników AI z wynikami uzyskanymi przez lekarzy na tym samym egzaminie, co pozwoliło na ocenę czy modele te są w stanie dorównać w zakresie analizy danych wiedzy lekarza rezydenta lub nawet otrzymać wynik wyższy.

Egzamin PES składa się z pytań jednokrotnego wyboru, dotyczących szerokiego spektrum problemów zdrowotnych w poszczególnych specjalizacjach lekarskich. Każdy z modeli AI został poddany tej samej procedurze egzaminacyjnej, co lekarze, a odpowiedzi były rejestrowane i oceniane w systemie punktacji binarnej (0 – odpowiedź nieprawidłowa, 1 – odpowiedź prawidłowa) zgodnie z obowiązującym na PES schematem punktacji. Łącznie modele odpowiedziały na 1200 pytań (10 x 120).

Dodatkowo przeprowadzono szczegółową analizę pytań, które sprawiały trudności AI, w celu identyfikacji dominujących trybów błędów oraz ograniczeń modeli. Uwzględniono przy tym różne aspekty: złożoność rozumowania klinicznego, rzadkie prezentacje chorób, dylematy etyczne zależne od kontekstu oraz fragmentaryczne rozumienie pytań wynikające z nadmiernego skupienia na słowach kluczowych.

W trakcie przygotowywania tekstu zastosowano generatywną sztuczną inteligencję w ograniczonym zakresie, wyłącznie w celu poprawy spójności stylistycznej i językowej. Wykorzystany model to GPT-4o. AI nie ingerowało w treść merytoryczną ani interpretację wyników. Wszystkie dane, analizy i wnioski zostały opracowane niezależnie przez autorów.

Wyniki

Analiza przeprowadzona w oparciu o wyniki Państwowego Egzaminu Specjalizacyjnego (PES) z jesieni 2024 roku objęła dziesięć różnych dziedzin medycyny, ze szczególnym uwzględnieniem geriatrii. Wyniki lekarzy wskazują, że najwyższe średnie wyniki osiągano w dziedzinie endokrynologii, natomiast najniższe w patomorfologii. W przypadku geriatrii średni wynik lekarzy wyniósł 89,2 punktu na 120 możliwych. Dla porównania, średni wynik wszystkich analizowanych modeli AI we wszystkich specjalnościach osiągnął 91,2 punktu, natomiast w samej geriatrii AI zdobyło średnio 98,7 punktu, wyraźnie przewyższając wyniki lekarzy. Wśród badanych modeli najwyższą skuteczność wykazał DeepSeek-V3, natomiast GPT-4o uzyskał najniższe rezultaty.

Próg zaliczenia Polskiego Egzaminu Specjalizacyjnego (PES) w części testowej wynosił 60% maksymalnej liczby punktów. Co istotne, żaden z modeli AI nie uzyskał wyniku poniżej progu zaliczenia egzaminu, co już samo w sobie stanowi istotny sygnał o rosnącym potencjale tych systemów w dziedzinie medycyny. Wszystkie wyniki zestawiono w tabelach I i II.

Najwyższe średnie wyniki egzaminu osiągnęli lekarze w dziedzinie endokrynologii (98,4 pkt), a najniższe w patomorfologii (83,3 pkt) – tabela. I. W geriatrii średni wynik lekarzy wyniósł 89,2 punktu, co plasuje ich w środkowej części rozkładu wyników.

Zestawiono także średnie wyniki uzyskane przez lekarzy z wynikami trzech modeli AI (DeepSeek-V3, GPT-4o, Gemini 2.0) w dziesięciu specjalizacjach lekarskich. Analiza wskazuje, że wynik osiągnięty przez modele AI przewyższał wyniki lekarzy w większości specjalizacji, a największa przewaga została odnotowana w geriatrii (+9,5 pkt) (tabela II). Najmniejsza różnica wystąpiła w chirurgii naczyniowej (+1,2 pkt). Wyniki te sugerują, że nowoczesne modele AI osiągają wysoką skuteczność w teoretycznych testach medycznych, przy czym największy potencjał przewyższania wiedzy lekarzy może występować w dziedzinach medycyny wymagających złożonego przetwarzania wiedzy, takich jak geriatria.

Pomimo imponujących wyników AI, analiza szczegółowa ujawniła obszary, w których systemy te nadal napotykają trudności. Spośród 74 pytań, które sprawiały problemy wszystkim trzem modelom, tylko 4 dotyczyły bezpośrednio PES z geriatrii (tabela III). W wyniku analizy wyłoniono cztery główne typy błędów. Najczęściej występowały problemy z wielostopniowym rozumowaniem klinicznym (41%), które wymagało sekwencyjnej analizy przypadków oraz wyciągania wniosków na podstawie złożonych informacji. Kolejnym znaczącym problemem były rzadkie prezentacje chorób (28%), czyli sytuacje, w których objawy pacjenta były nietypowe i wymagały doświadczenia klinicznego do prawidłowej diagnozy. Trzecim obszarem trudności były dylematy etyczne zależne od kontekstu (19%), pokazujące, że AI nadal nie potrafi w pełni uwzględniać wartości kulturowych, preferencji pacjenta czy zasad autonomii w procesie decyzyjnym. Ostatnią, choć mniej liczną grupą problemów (12%), było fragmentaryczne rozumienie pytań, polegające na nadmiernym skupieniu się na pojedynczych słowach kluczowych, np. *„ból”,* bez uwzględnienia szerszego kontekstu klinicznego. Analiza wzorców błędów w PES

Tabela I. Wyniki uzyskane przez lekarzy na Państwowym Egzaminie Specjalizacyjnym (PES) w sesji jesiennej 2024 roku w dziesięciu dziedzinach medycyny. Zawarto w niej liczbę wszystkich uczestników egzaminu, liczbę osób, które ukończyły egzamin w pełnym zakresie, średnią punktację, odchylenie standardowe, medianę, wartości minimalne i maksymalne oraz wskaźnik trudności pytań dla każdej specjalności [*Źródłem wyników są oficjalne statystyki PES 2024 publikowane przez Centrum Egzaminów Medycznych (CEM)*]

Table I. Results achieved by physicians in the Polish Specialty Exam (PES) in Autumn 2024 session, across ten medical specialties. The table includes the total number of exam participants, the number of candidates who completed the full exam, the average score, standard deviation, median, minimum and maximum values, as well as the item difficulty index for each specialty [*The data come from official PES 2024 statistics published by the Medical Examination Centre (CEM)*]

Specjalizacja	Przystąpiło	Zdało	Średni wynik	Odchylenie standardowe	Mediana	Maks.	Min.	Wskaźnik trudności
Anestezjologia i intensywna terapia	85	82	92,3	9,81	92,0	113	63	0,766
Chirurgia naczyniowa	12	9	93,8	6,04	93,0	103	82	0,781
Endokrynologia	31	31	98,4	10,81	98,5	114	64	0,811
Geriatria	15	15	89,2	11,40	89,5	109	64	0,738
Kardiologia	89	83	90,4	12,67	93,0	108	37	0,732
Neurologia	55	52	85,8	13,00	87,5	106	47	0,706
Onkologia kliniczna	26	26	88,7	9,90	90,0	103	64	0,733
Patomorfologia	10	9	83,3	23,67	88,0	103	15	0,610
Pediatria	162	155	88,6	10,72	90,0	110	57	0,733
Psychiatria	112	105	92,3	13,51	96,0	110	49	0,751

Tabela II. Zestawienie średnich wyników uzyskanych przez lekarzy oraz przez trzy modele AI (DeepSeek-V3, GPT-4o, Gemini 2.0) w dziesięciu dziedzinach medycyny.
Kolumna „Średni wynik AI" przedstawia średnią wyników wszystkich trzech modeli, natomiast kolumna „Różnica AI-lekarze" pokazuje różnicę między średnim wynikiem AI a średnim wynikiem lekarzy w danej specjalizacji

Table II. Comparison of average scores achieved by physicians and three AI models (DeepSeek-V3, GPT-4o, Gemini 2.0) across ten medical specialties.
The column shows the mean score of all three models, while the column indicates the difference between the AI average score and the average score of physicians in each specialty

Specjalizacja	Średni wynik lekarzy	Średni wynik AI	DeepSeek-V3	GPT-4o	Gemini 2.0	Różnica AI-lekarze
Anestezjologia i intensywna terapia	92,3	94,5	96,0	92,1	95,3	+2,2
Endokrynologia	98,4	97,1	98,9	95,3	97,2	-1,3
Geriatria	89,2	98,7	99,5	97,1	99,4	+9,5
Kardiologia	90,4	93,8	95,2	91,0	95,1	+3,4
Neurologia	85,8	89,3	91,5	86,1	90,2	+3,5
Onkologia kliniczna	88,7	94,2	95,9	91,5	95,1	+5,5
Patomorfologia	83,3	86,7	88,1	84,3	87,6	+3,4
Pediatria	88,6	92,4	94,0	89,7	93,4	+3,8
Psychiatria	92,3	95,1	96,7	92,8	95,8	+2,8
Chirurgia naczyniowa	93,8	95,0	96,4	93,2	95,3	+1,2

Tabela III. Tabela przedstawia zestawienie czterech pytań testowych z PES z Geriatrii, w których wszystkie modele AI popełniły błąd, wraz z możliwymi odpowiedziami, odpowiedzią prawidłową, odpowiedzią AI oraz analizą mechanizmu popełnionych błędów. Pytania pochodzą z archiwalnego testu PES z Geriatrii, jesień 2024, udostępnionego na platformie Remedium

Table III. The table presents a summary of four test questions from PES in Geriatrics in which all AI models made errors, including the possible answers, the correct answer, the AI-generated answer, and an analysis of the underlying error mechanisms. The questions are drawn from the archival PES Geriatrics exam, Autumn 2024, available on the Remedium platform

Treść pytania	Odpowiedź poprawna	Odpowiedź AI	Omówienie
78-letni pacjent (waga 66 kg, wzrost 181 cm, BMI 21,1 kg/m², obwód łydki 28 cm, masa ciała sprzed 6 miesięcy 76 kg), leczony z powodu zaawansowanej niewydolności serca i ciężkiej POChP (grupa E), dotychczas sprawny, skierowany z powodu utraty apetytu >1 miesiąca. Na podstawie zgromadzonych informacji można podejrzewać: A. niedożywienie, które nie wymaga pogłębionej oceny dietetycznej B. łagodne niedożywienie z potwierdzeniem 3 kryteriów fenotypowych i 3 kryteriów etiologicznych wg GLIM C. umiarkowane niedożywienie z potwierdzeniem 2 kryteriów fenotypowych wg GLIM D. umiarkowane niedożywienie z potwierdzeniem 3 kryteriów fenotypowych i 3 kryteriów etiologicznych wg GLIM E. ciężkie niedożywienie z potwierdzeniem 3 kryteriów fenotypowych i 2 kryteriów etiologicznych wg GLIM	E	C	AI uwzględniło jedynie kryteria fenotypowe, nie łącząc ich z kontekstem etiologicznym. Brak pełnej analizy danych doprowadził do niepełnej i błędnej klasyfikacji niedożywienia.
W przypadku przygotowania do planowego zabiegu operacyjnego osoby starszej z nadciśnieniem tętniczym należy pamiętać: 1) pacjent może być skierowany, jeśli ciśnienie <160/100 mmHg w ciągu 12 miesięcy; 2) przy zabiegach naczyniowych lub nerki należy mierzyć ciśnienie na obu ramionach, różnica >20 mmHg → powtórzyć i uwzględnić wyższy wynik; 3) pacjent może być skierowany, jeśli ciśnienie <130/80 mmHg w ciągu 12 miesięcy; 4) nadciśnienie >180/100 mmHg zwiększa ryzyko sercowo-naczyniowe i wymaga unormowania przed planową operacją; 5) wartości >180/100 mmHg wymagają odroczenia planowej operacji. A. 2,3,5 B. 1,3,4,5 C. 1,2,4,5 D. 1,3,4 E. wszystkie wymienione	C	E	AI analizowało tylko fragment informacji (ogólne zalecenia dotyczące ciśnienia) zamiast pełnego kontekstu klinicznego i celu pytania (czy pacjent może zostać skierowany na operację planową przy określonych wartościach ciśnienia). Fragmentaryczne podejście doprowadziło do nadmiernego uproszczenia i błędnej odpowiedzi.
Które leki należy odstawić u starszych pacjentów u kresu życia (przewidywana długość życia <2 miesiące)? A. leki przeciwkrzepliwe przy migotaniu przedsionków B. klopidogrel przed upływem 12 miesięcy od wszczepienia stentu C. kwas acetylosalicylowy w prewencji pierwotnej D. beta-blokery w zaburzeniach rytmu E. heparyna w prewencji wtórnej po incydencie zakrzepowym	C	A	AI błędnie wybrało odstawienie leków przeciwkrzepliwych przy migotaniu przedsionków zamiast odstawienie ASA. Nie uwzględniło sekwencyjnej analizy korzyści i ryzyka leczenia w kontekście przewidywanej krótkiej długości życia pacjenta.
Wskaż prawdziwe stwierdzenie dotyczące pacjentów starszych z cukrzycą u kresu życia (perspektywa przeżycia kilka tygodni, placówki opiekuńcze): A. wymagają regularnej kontroli HbA1c w celu dostosowania leków B. wymagają kontroli glikemii, aby unikać hipoglikemii i objawowej hiperglikemii C. jeśli leczeni insuliną, wymagają wielokrotnych pomiarów glikemii w ciągu dnia D. jeśli leczeni insuliną, wymagają kontynuacji insuliny okołoposiłkowo E. jeśli leczeni insuliną, wymagają odstawienia insuliny	B	E	Poprawna odpowiedź koncentruje się na utrzymaniu komfortu i bezpieczeństwa metabolicznego, czyli unikaniu hipoglikemii i objawowej hiperglikemii. AI dążyło do uproszczenia schematu insulinoterapii, ignorując kontekst kliniczny, krótki przewidywany czas przeżycia i priorytetu opieki paliatywnej.

wykazała, że wieloetapowe rozumowanie kliniczne stanowiło najczęstszy tryb niepowodzenia dla modeli AI, odpowiadając za 41% błędów w pytaniach z wszystkich analizowanych dziedzin.

Omówienie

Ostatnia dekada przyniosła rewolucyjne postępy w dziedzinie AI, szczególnie w kontekście generatywnych modeli sztucznej inteligencji, które wykazują zdolność do przetwarzania złożonych danych medycznych, generowania diagnoz różnicowych oraz wspierania procesów decyzyjnych co do podejmowanej terapii, stąd nadzieja, że ChatGPT spowoduje przełomowe zmiany w opiece i leczeniu osób starszych. Liczne badania wykazały, że AI może osiągać dokładność diagnostyczną porównywalną z lekarzami, a w niektórych przypadkach nawet przewyższać ją, osiągając poziom zbliżony do ekspertów w danej dziedzinie [6-8]. Badania przeprowadzone na dużej kohorcie, obejmującej ponad 2 000 przypadków medycznych, wykazały, że hybrydowe zespoły lekarzy i LLM osiągają wyższą dokładność diagnostyczną niż zarówno pojedynczy specjaliści, jak i same systemy AI [9].

W kontekście egzaminów medycznych, duże modele językowe takie jak GPT-4 i Claude 3.5 Sonnet konsekwentnie osiągają wyniki przewyższające progi zdawalności egzaminów lekarskich, takie jak United States Medical Licensing Examination (USMLE), ze skutecznością sięgającą 80-95% [10]. Podobne wyniki przedstawiono w niniejszej pracy w przypadku polskiego PES. Jednakże należy pamiętać, że same wyniki egzaminów nie odzwierciedlają wszystkich kompetencji niezbędnych w rzeczywistej praktyce klinicznej. Zastosowanie AI w geriatrii niesie ze sobą unikalne wyzwania wynikające z wielowymiarowego charakteru opieki nad osobami starszymi. Całościowa Ocena Geriatryczna (CGA) stanowi złoty standard w ocenie pacjentów geriatrycznych, obejmując nie tylko aspekty biomedyczne, ale również funkcjonalne, poznawcze, emocjonalne i społeczne. Integracja AI w ten proces wymaga nie tylko wysokiej dokładności diagnostycznej, ale również zdolności do interpretacji uzyskanych informacji w danym kontekście klinicznym, uwzględniania wartości i preferencji pacjentów oraz podejmowania decyzji etycznych [11].

Główne wyzwanie polega na tym, że AI opiera się głównie na rozpoznawaniu wzorców zamiast głębokiego zrozumienia mechanizmów zdrowia i choroby osób starszych. Rozumowanie hipotetyczno-dedukcyjne, często wykorzystujące skrypty chorób *(*ang. *illness scripts),* różni się od rozumowania indukcyjnego opartego na głębszym zrozumieniu mechanizmów. Chociaż wykorzystane także w niniejszym badaniu modele AI wykazały wysoką dokładność w zadaniach związanych z rozpoznawaniem wzorców, ich zdolność modyfikowania podejścia w odpowiedzi na nowe, nieoczekiwane sytuacje kliniczne – pozostaje ograniczona [12].

Wieloetapowe rozumowanie kliniczne wymaga uwzględnienia różnorodnych informacji, a skuteczna diagnoza opiera się na logicznym powiązaniu objawów, wyników badań oraz historii medycznej pacjenta. Badania wykazały, że choć duże LLM osiągały znakomite wyniki na egzaminach lekarskich, testy te nie oceniały wielu umiejętności podejmowania decyzji niezbędnych do pracy w realnym środowisku, takich jak interpretacja informacji, przestrzeganie wytycznych czy integracja z procesami klinicznymi. W wynikach badań naukowych nad zastosowaniem AI w medycynie wykazano też zależność od przestarzałych lub nie dopasowanych do danej populacji danych, co podważa jej kliniczną trafność. Badanie Hager i współpracowników wykazało, że LLM nie są obecnie gotowe do autonomicznego podejmowania decyzji klinicznych, szczególnie w przypadkach wymagających sekwencyjnej diagnozy i iteracyjnego wyboru pytań diagnostycznych i testów [1,2]. Badania porównawcze na zestawach pytań z New England Journal of Medicine wykazały, że DeepSeek-R1 osiągał 35% dokładności w ustaleniu ostatecznej diagnozy i 48% w umieszczaniu prawidłowej diagnozy na liście różnicowej, podczas gdy GPT-4 uzyskał odpowiednio 39% w ustaleniu ostatecznej diagnozy i 64% w umieszczaniu prawidłowej diagnozy na liście różnicowej [3]. DeepSeek-R1 generował przy tym dłuższe listy diagnoz średnio o długości 11,9 w porównaniu do 9,0 dla GPT-4. DeepSeek-R1 generuje bardziej rozbudowane listy diagnostyczne, odzwierciedlające szersze rozumowanie kliniczne, podczas gdy GPT-4 wykazuje większą precyzję w identyfikacji właściwej diagnozy [3].

Halucynacje AI, czyli generowanie wiarygodnie brzmiących, lecz nieprawdziwych informacji, stanowią istotne ryzyko w medycynie i jedną z głównych barier we wdrażaniu AI. Badania szacują, że wskaźniki halucynacji w modelach AI wykorzystywanych do systemów wspomagania decyzji klinicznych wahają się od 8% do 20%, w zależności od złożoności modelu i jakości danych treningowych [13]. Badanie przeprowadzone przez startup AI Mendel i University of Massachusetts Amherst

oceniło podsumowania medyczne wygenerowane przez dwa duże modele językowe i wykazało, że najczęstsze nieprawidłowości dotyczyły danych pacjenta: historii choroby, diagnoz, procedur i zaleceń lekowych (GPT-4o miał 21 podsumowań z nieprawidłowymi informacjami i 50 podsumowań z uogólnionymi informacjami), podkreślając potrzebę ostrożności przy stosowaniu AI w diagnostyce i dokumentacji geriatrycznej [14]. Z uwagi na powyższe ograniczenia nieustannie prowadzone są badania nad zwiększaniem specyficzności i zaufania do uzyskiwanych danych przy użyciu coraz bardziej złożonych systemów cybernetycznych, takich jak generatywna sztuczna inteligencja (GenAI – ang. *generative AI*), która tworzy nowe, oryginalne treści na podstawie danych, na których została wytrenowana. W odróżnieniu od tradycyjnej AI, która analizuje istniejące dane, GenAI jest w stanie generować treści, które nie istniały wcześniej, naśladując ludzką kreatywność [15]. Takie modele wykorzystano na przykład do badań nad przewidywaniem ryzyka zgonu i ustalenia rokowania, aby ułatwiać podejmowanie decyzji terapeutycznych. Zhang i jego zespół użył modelu uczenia maszynowego do przewidywania śmiertelności w demencji, używając danych 45 tys. pacjentów z U.S. National Alzheimer's Coordinating Center (NACC) [16]. Badacze skonstruowali wieloczynnikowy model XGBoost (wzmacnianie gradientowe, ang. *extreme gradient boosting*) – otwarto-źródłowej biblioteki uczenia maszynowego z wykorzystaniem sekwencyjnego budowania drzew decyzyjnych, gdzie każde nowe drzewo poprawia błędy poprzednich. Uzyskane dane pozwoliły z dużą dokładnością wyodrębnić pacjentów zagrożonych największą śmiertelnością w poszczególnych typach demencji [16].

Biorąc pod uwagę wszystkie te ograniczenia integracja AI w medycynie, a szczególnie w opiece geriatrycznej wymaga szczególnej ostrożności. Należy zapewnić, aby zarówno lekarze, jak i beneficjenci użycia technologii mieli do niej wystarczające zaufanie, a technologie muszą być używane w sposób transparentny, uwzględniający autonomię pacjenta, jego wartości i preferencje, a także respektujący zasady sprawiedliwości i niedyskryminacji. Modele językowe nie powinny zastępować lekarza, lecz funkcjonować jako jego rozszerzenie, wspierając procesy decyzyjne i minimalizując ryzyko błędów medycznych, przy jednoczesnym zachowaniu pełnej odpowiedzialności człowieka za podejmowane decyzje.

Przyszłość AI w geriatrii powinna opierać się na modelach współpracy między lekarzami a dużymi modelami językowymi. W tego rodzaju modelu lekarz integruje doświadczenie kliniczne z możliwościami analitycznymi AI. Dzięki temu powstaje system diagnostyczny o charakterze komplementarnym, w którym doświadczenie kliniczne lekarzy koryguje typowe dla AI błędy, a AI uzupełnia ewentualne ograniczenia ludzkiej pamięci i wiedzy, zwiększając precyzję, skuteczność i zorientowanie na pacjenta.

Wnioski

Analiza wyników uzyskanych przy użyciu modeli AI w odpowiedziach na pytania Polskiego Egzaminu Specjalizacyjnego w dziedzinie geriatrii ukazuje rosnący potencjał AI jako wsparcia lekarza w procesie diagnostycznym i terapeutycznym. Modele takie jak GPT-4o, DeepSeek-V3 czy Gemini 2.0 wykazują już na ten moment imponującą zdolność do syntetyzowania wiedzy, generowania propozycji diagnoz i leczenia, często przewyższającą możliwości lekarzy przystępujących do egzaminu specjalizacyjnego. Analiza wskazała jednak na przeszkody w zastosowaniu AI w momencie konieczności podjęcia wieloetapowego rozumowania klinicznego, a niewątpliwie efektywne wykorzystanie AI w praktyce klinicznej nie może ograniczać się do mechanicznego przetwarzania danych. Na podstawie wyników niniejszego badania możemy wysnuć wniosek, że na obecnym etapie kluczowe jest zachowanie synergii między pracą lekarza a możliwościami – modeli językowych AI. Mogą one służyć do analizy dokumentacji, ułatwiać diagnostykę różnicową i wspierać planowanie terapii, lecz ostateczne decyzje muszą pozostać w rękach doświadczonego klinicysty.

Konflikt interesów / Conflict of interest
Brak/None

Adres do korespondencji / Correspondence address
Joanna Monika Najbar
Wydział Lekarski, Uniwersytet Warmińsko-Mazurski w Olsztynie
Aleja Warszawska 30, 11-041 Olsztyn
☎ (+48) 785 409 889
najbarjm@gmail.com

Piśmiennictwo/References

1. Główny Urząd Statystyczny (GUS). Prognoza ludności Polski do 2050 roku. Warszawa: GUS; 2023. n.d. https://stat.gov.pl/obszary-tematyczne/ludnosc/prognoza-ludnosci/prognoza-ludnosci-na-lata-2023-2060,11,1.html n.d. https://stat.gov.pl/obszary-tematyczne/ludnosc/prognoza-ludnosci/prognoza-ludnosci-na-lata-2023-2060,11,1.html (accessed December 1, 2025).
2. Shankar R, Bundele A, Mukhopadhyay A. Barriers and enablers for the deployment of large language model-based conversational robots for older adults: A protocol for a systematic review of qualitative studies. PLoS One 2025;20:e0321093. https://doi.org/10.1371/JOURNAL.PONE.0321093.
3. Feng G, Weng F, Lu W, et al. Artificial Intelligence in Chronic Disease Management for Aging Populations: A Systematic Review of Machine Learning and NLP Applications. Int J Gen Med 2025;18:3105. https://doi.org/10.2147/IJGM.S516247.
4. Ma B, Yang J, Wong FKY, et al. Artificial intelligence in elderly healthcare: A scoping review. Ageing Res Rev 2023;83. https://doi.org/10.1016/J.ARR.2022.101808.
5. Wangmo T, Lipps M, Kressig RW, Ienca M. Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders. BMC Med Ethics 2019;20. https://doi.org/10.1186/S12910-019-0437-Z.
6. Gao S, Xu Z, Kang W, et al. Artificial intelligence-driven computer aided diagnosis system provides similar diagnosis value compared with doctors' evaluation in lung cancer screening. BMC Med Imaging 2024;24:141-. https://doi.org/10.1186/S12880-024-01288-3/TABLES/3.
7. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. Npj Digital Medicine 2025 8:1 2025;8:175-. https://doi.org/10.1038/s41746-025-01543-z.
8. Hu W, Zhang J, Zhou D, et al. A comparison study of artificial intelligence performance against physicians in benign-malignant classification of pulmonary nodules. Oncologie 2024;26:581-6. https://doi.org/10.1515/ONCOLOGIE-2023-0319/DOWNLOADASSET/SUPPL/J_ONCOLOGIE-2023-0319_SUPPL_001.XLSX.
9. Zöller N, Berger J, Lin I, et al. Human-AI collectives produce the most accurate differential diagnoses 2024.
10. Lin SY, Hsu YY, Ju SW, et al. Assessing AI efficacy in medical knowledge tests: A study using Taiwan's internal medicine exam questions from 2020 to 2023. Digit Health 2024;10. https://doi.org/10.1177/20552076241291404/SUPPL_FILE/SJ-DOCX-1-DHJ-10.1177_20552076241291404.DOCX.
11. Cesario A, D'oria M, Calvani R, et al. The Role of Artificial Intelligence in Managing Multimorbidity and Cancer. J Pers Med 2021;11:314. https://doi.org/10.3390/JPM11040314.
12. Xu H, Wang Y, Xun Y, Shao R, Jiao Y. Artificial intelligence for clinical reasoning: the reliability challenge and path to evidence-based practice. QJM 2025. https://doi.org/10.1093/QJMED/HCAF114.
13. Chen F, Li Y, Chen Y, Bian Z, et al. Strategies for the Analysis and Elimination of Hallucinations in Artificial Intelligence Generated Medical Knowledge. J Evid Based Med 2025;18. https://doi.org/10.1111/JEBM.70075.
14. Rumale P, Tiwari S, Naik TG, et al. Faithfulness Hallucination Detection in Healthcare AI. Proceedings of KDD 2024 Workshop – Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare (KDD-AIDSH 2024) 2024;1.
15. Baig MM, Hobson C, GholamHosseini H, Ullah E, Afifi S. Generative AI in Improving Personalized Patient Care Plans: Opportunities and Barriers Towards Its Wider Adoption. Applied Sciences 2024, Vol 14, Page 10899 2024;14:10899. https://doi.org/10.3390/APP142310899.
16. Zhang J, Song L, Miller Z, Chan KCG, Huang KL. Machine learning models identify predictive features of patient mortality across dementia types. Communications Medicine 2024 4:1 2024;4:1–13. https://doi.org/10.1038/s43856-024-00437-7.