Now Audiofonol, 2025; 14(4): 46–53
DOI: 10.17431/na/213882
CC BY-NC-ND 3.0 PL, © Autorzy

Nowa Audiofonologia®
Prace badawcze

Przesłano do redakcji: 06.08.2025
Zaakceptowano po recenzji: 03.11.2025
Opublikowano: 08.01.2026

Zgodność ocen progów słyszenia między ekspertami w badaniu słuchowych potencjałów wywołanych pnia mózgu (ABR) u dzieci według procedury szeregu natężeniowego

Agreement of hearing threshold ratings among experts in auditory brainstem responses (ABR) in children using the intensity series procedure

Elżbieta Gos[1A-F], Małgorzata Pastucha[2BEF], Krzysztof Kochanek[3ACE-G], Adam Piłka[3ACE-G], Elżbieta Włodarczyk[4CEF], Henryk Skarżyński[5AEG]

[1] Instytut Fizjologii i Patologii Słuchu, Światowe Centrum Słuchu, Zakład Teleaudiologii i Badań Przesiewowych, Warszawa/Kajetany

[2] Instytut Fizjologii i Patologii Słuchu, Światowe Centrum Słuchu, Zakład Audiologii Eksperymentalnej, Warszawa/Kajetany

[3] Instytut Fizjologii i Patologii Słuchu, Światowe Centrum Słuchu, Warszawa/Kajetany

[4] Instytut Fizjologii i Patologii Słuchu, Światowe Centrum Słuchu, Klinika Rehabilitacji, Warszawa/Kajetany

[5] Instytut Fizjologii i Patologii Słuchu, Światowe Centrum Słuchu, Klinika Oto-Ryno-Laryngochirurgii, Warszawa/Kajetany

Wkład autorów:
A Projekt badania
B Gromadzenie danych
C Analiza danych
D Interpretacja danych
E Przygotowanie pracy
F Przegląd literatury
G Gromadzenie funduszy

Streszczenie

Wprowadzenie: Badanie słuchu za pomocą potencjałów wywołanych pnia mózgu (ABR) odgrywa podstawową rolę w diagnostyce audiologicznej dzieci, szczególnie tych, które nie są w stanie współpracować w badaniach behawioralnych. Mimo dużego znaczenia tej metody, jest ona obarczona pewnym subiektywizmem interpretacji i oceny wyniku badania, co może niekorzystnie wpływać na trafność diagnozy. Celem pracy była analiza zgodności ocen progu słyszenia na podstawie zapisów ABR u dzieci, dokonanych niezależnie przez czterech ekspertów dla różnych bodźców akustycznych (trzask, 500, 1000, 2000, 4000 Hz).

Materiał i metody: Materiał badawczy stanowiły 322 zapisy ABR uzyskane od 54 dzieci. Zapisy ABR pozyskano z repozytorium wyników zgromadzonych w Instytucie Fizjologii i Patologii Słuchu (IFPS). Badania były wykonywane przy zastosowaniu standardowej procedury szeregu natężeniowego. Zapisy ABR były analizowane przez czterech ekspertów z wieloletnim doświadczeniem klinicznym, którzy w niezależny sposób dokonywali oceny progów słyszenia u dzieci.

Wyniki: W 53,5% analizowanych zapisów czterech ekspertów osiągnęło całkowitą zgodność w ocenie progów słyszenia. W 93,2% przypadków rozbieżność ocen była akceptowalna klinicznie, wynosiła do ±10 dB. Korelacja między ocenami ekspertów była bardzo wysoka, $ICC = 0{,}98$; $p < 0{,}001$. Rozbieżności ocen między ekspertami zaobserwowano głównie w przypadku bodźca tonalnego o częstotliwości 500 Hz, co prawdopodobnie wynika z właściwości propagacji fali wędrownej o niskiej częstotliwości w ślimaku.

Wnioski: Doświadczeni eksperci klinicyści osiągają dużą zgodność w ocenie progów słyszenia u dzieci na podstawie fali V w badaniu ABR. Ta wysoka zgodność ocen eksperckich jest podstawą do opracowania automatycznych metod detekcji fali V, które mogą skutecznie wspierać proces diagnostyczny, zapewniając większą obiektywność interpretacji zapisów ABR, oszczędzając jednocześnie czas i koszty diagnostyki audiologicznej.

Słowa kluczowe: słuchowe potencjały wywołane pnia mózgu • ABR • progi słyszenia • zgodność ocen ekspertów

Autor korespondencyjny: Małgorzata Pastucha, Zakład Audiologii Eksperymentalnej, Światowe Centrum Słuchu, Instytut Fizjologii i Patologii Słuchu, ul. Mochnackiego 10, 02-042 Warszawa; email: m.pastucha@ifps.org.pl

Abstract

Introduction: Hearing testing using auditory brainstem responses (ABR) plays a key role in the audiological diagnosis of children, especially those who are unable to cooperate in behavioural testing. Despite the great importance of this method, it involves a certain degree of subjectivity in the interpretation and evaluation of the test results, which may negatively affect the accuracy of the diagnosis. The aim of this study was to analyse the agreement of hearing threshold ratings based on ABR recordings in children made independently by four experts for different acoustic stimuli (click, 500, 1000, 2000, 4000 Hz).

Material and methods: The study material consisted of 322 ABR recordings from 54 children. ABR recordings were sourced from the repository of results collected at the Institute of Physiology and Pathology of Hearing (IFPS). The tests were performed using a standard intensity series procedure. The ABR recordings were analysed by four experts with many years of clinical experience, who independently assessed the children's hearing thresholds.

Results: In 53.5% of the records analysed, four experts reached complete agreement in the assessment of hearing thresholds. In 93.2% of cases, the discrepancy in assessments was clinically acceptable, up to ±10 dB. The correlation between expert ratings was very high, ICC = 0.98; $p < 0.001$. Discrepancies in ratings between experts were mainly observed for the 500 Hz tonal stimulus, which is probably due to the low-frequency travelling wave propagation properties of the cochlea.

Conclusions: Experienced expert clinicians reach a high level of agreement in the assessment of hearing thresholds in children based on wave V in the ABR test. This high agreement of expert assessments is the basis for the development of automated V-wave detection methods that can effectively support the diagnostic process, providing greater objectivity in the interpretation of ABR recordings while saving audiological diagnostic time and cost.

Keywords: auditory brainstem responses • ABR • hearing thresholds • agreement of experts' ratings

Wprowadzenie

Słuchowe potencjały wywołane pnia mózgu (ang. *auditory brainstem responses*, ABR) stanowią obecnie jedną z podstawowych obiektywnych metod badania słuchu. Znajdują zastosowanie m.in. w szacowaniu progu słyszenia [1], badaniach przesiewowych słuchu u noworodków, diagnostyce różnicowej zaburzeń słuchu [2], monitoringu funkcji nerwu słuchowego i pnia mózgu w czasie zabiegów otoneurochirurgicznych [3,4]. Z uwagi na obiektywny charakter, metoda ta jest szeroko stosowana w diagnostyce dzieci, osób z niepełnosprawnościami i pacjentów niewspółpracujących, a także w przypadkach podejrzenia symulacji oraz w celach orzeczniczych [5,6].

Odpowiedzi rejestrowane w badaniu ABR tworzą charakterystyczny układ fal (I–VII), spośród których fala V uznawana jest za najważniejszy wskaźnik kliniczny wykorzystywany do estymacji progu słyszenia [7]. Badanie progu słyszenia polega na rejestracji odpowiedzi dla bodźców o zmiennej intensywności (procedura szeregu natężeniowego) i wyznaczeniu progu fali V – najniższego poziomu, przy którym fala ta jest obecna. Badanie wykonuje się z użyciem bodźców typu trzask oraz impulsów tonalnych o częstotliwościach 500, 1000, 2000 i 4000 Hz.

Pomimo obiektywnego charakteru badania i standaryzowanych procedur rejestracji odpowiedzi, końcowa interpretacja wyniku opiera się na subiektywnej, wzrokowej analizie zapisu, co może skutkować istotnymi różnicami w ocenie między ekspertami [8]. Ze względu na niewielką amplitudę rzędu dziesiątych części μV, fala V może być trudna do jednoznacznej identyfikacji, szczególnie przy niskich poziomach natężenia bodźca [7]. Na wiarygodność oceny wpływają m.in. rodzaj zastosowanego bodźca (np. 500 Hz vs trzask), obecność artefaktów, jakość zapisu, zmienność odpowiedzi osobniczej oraz doświadczenie osoby oceniającej wynik [9,10]. Nieprawidłowa identyfikacja fali V może skutkować zarówno zawyżeniem, jak i zaniżeniem progu słyszenia. Wśród najczęstszych błędów wymienia się mylne uznanie odpowiedzi sonomotorycznej lub artefaktu za falę V, nieuwzględnienie charakterystyki bodźca (np. wpływu częstotliwości na latencję odpowiedzi) oraz niewłaściwe ustawienie czułości prezentacji zapisu na ekranie [10]. W efekcie, nawet wśród doświadczonych klinicystów, występują istotne różnice w wyznaczaniu progu, sięgające nawet 60 dB [11], choć klinicznie akceptowalne rozbieżności między ekspertami nie powinny przekraczać 10 dB [12].

W związku z trudnościami w jednoznacznej interpretacji rejestracji ABR podejmowano liczne próby automatyzacji detekcji fali V i obiektywnej estymacji progu słyszenia [13]. Stosowane metody obejmują m.in. analizę korelacji krzyżowej [14,15], statystyki detekcyjne [16], przekształcenia falkowe [17], uczenie maszynowe [18] oraz rozpoznawanie wzorców [19]. Choć systemy te mogą wspierać pracę klinicysty, żaden z dotychczasowych algorytmów nie osiąga pełnej niezawodności w warunkach klinicznych. Dodatkowym ograniczeniem jest fakt, że wiele badań nad automatyczną detekcją opiera się na sztucznie wyselekcjonowanych zapisach dobrej jakości lub danych uzyskanych w warunkach eksperymentalnych, a nie na rzeczywistych, zróżnicowanych danych klinicznych.

W rezultacie wartość diagnostyczna ABR nadal opiera się głównie na wiedzy i doświadczeniu osoby interpretującej wynik, a tym samym jest podatna na subiektywizm. W praktyce klinicznej mogą występować istotne rozbieżności w interpretacji zapisów – różni diagności, analizując ten sam wynik badania ABR, mogą wyznaczyć różne wartości progu słyszenia. Niespójność ocen może prowadzić do błędów diagnostycznych, a w konsekwencji – do niewłaściwych decyzji terapeutycznych.

Z tego względu istotnym elementem oceny jakości diagnostyki ABR jest analiza zgodności międzyeksperckiej (ang. *inter-rater agreement*), pozwalająca oszacować stopień spójności ocen dokonywanych niezależnie przez różnych diagnostów. Może ona obejmować zarówno przypadki pełnej zgodności (identycznie wyznaczone progi), jak i akceptowalne rozbieżności, np. w granicach ±5, ±10, czy ±20 dB. W literaturze opisano różne metody ilościowej oceny tej zgodności, w tym współczynnik *ICC*

(ang. *intraclass correlation coefficient*), analizę Blanda–Altmana, współczynnik kappa Cohena [20].

Pomimo istotnego znaczenia tego zagadnienia, brakuje badań dotyczących zgodności ocen progów słyszenia na podstawie rzeczywistych zapisów ABR. Jak dotąd jedynie Vidler i Parker [11] ocenili rozbieżności między ekspertami w wyznaczaniu progów ABR, jednak wykorzystali nie zapisy rzeczywiste ABR, a symulowane (wygenerowane w sztuczny sposób). Pozostałe prace [21,22] ograniczały się do oceny obecności lub braku odpowiedzi, natomiast Naves i wsp. [8] skupiali się na analizie latencji, nie odnosząc się do estymacji progu.

Celem niniejszej pracy była analiza zgodności ocen progu słyszenia na podstawie zapisów ABR u dzieci, dokonanych niezależnie przez czterech ekspertów dla różnych bodźców akustycznych (trzask, 500, 1000, 2000, 4000 Hz). Oceniono zgodność idealną (0 dB różnicy), zgodność w granicach tolerancji ±10 dB oraz maksymalną rozpiętość ocen. Analiza ta pozwoliła wskazać obszary trudności diagnostycznych, istotne dla dalszego rozwoju standardów oceny wyników badania ABR oraz narzędzi wspomagających automatyzację procesu tej oceny.

Materiał i metody

Praca została zrealizowana w ramach grantu: ABR.AI – system oparty o technologię AI do szacowania progu słyszenia na podstawie szeregów natężeniowych rejestracji odpowiedzi fali V; 2024/ABM/03/KPO/KPOD.07.07-IW.07-0222/24-00.

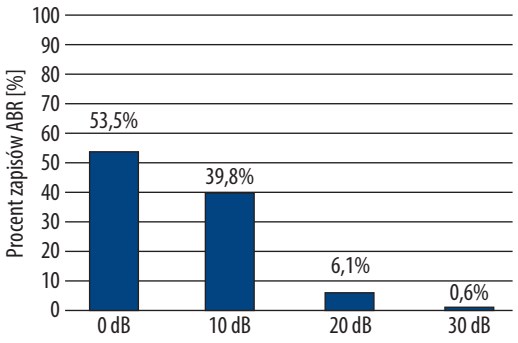
Procedura badania

Badanie miało charakter retrospektywny. Zapisy ABR pozyskano z repozytorium wyników zgromadzonych w Instytucie Fizjologii i Patologii Słuchu. Badania były wykonywane przez wykwalifikowaną kadrę, przy zastosowaniu standardowej procedury szeregu natężeniowego. Wszystkie badania przeprowadzono na tym samym typie sprzętu diagnostycznego. Zakres natężeń bodźców oraz poziom początkowy bodźca był dobierany przez osobę wykonującą badania w zależności od sytuacji klinicznej. Zakres natężeń bodźców stosowanych podczas badań zawierał się w granicach od 0 do 100 dB nHL, a krok zmiany poziomu bodźca wynosił najczęściej 10 dB.

O liczbie uśrednień, zakresie intensywności bodźców oraz liczbie powtarzanych zapisów odpowiedzi ABR dla każdej intensywności decydował osoba wykonująca badanie w jego trakcie. Zapisy ABR uzyskane w procedurze szeregu natężeniowego zawierały kolejne odpowiedzi od najwyższej intensywności do najniższej. Próg odpowiedzi ABR to najniższa intensywność bodźca, przy której w zapisie ABR była widoczna fala V.

Eksperci

Zapisy ABR były analizowane przez czterech ekspertów. Trzech z nich było lekarzami ze specjalizacją audiologia i foniatria. Ich staż pracy w zawodzie lekarza wynosił od 14 do 23 lat, zaś analizą zapisów ABR zajmowali się od 12 do 20 lat. Wszyscy lekarze w pracy klinicznej pracowali zarówno z dziećmi, jak i dorosłymi. Czwartym ekspertem był inżynier biomedyczny, który wykonywał badania ABR przez 15 lat, a ich analizą zajmował się ponad 40 lat. Eksperci dokonywali ocen niezależnie, żaden z nich nie miał informacji na temat ocen pozostałych ekspertów.



Rycina 1. Rozbieżności (dB, wartości bezwzględne) w ocenach 312 zapisów ABR między czterema ekspertami
Figure 1. Discrepancies (dB, absolute values) in the ratings of 312 ABR records between four experts

Materiał badawczy

Uzyskano 322 zapisy ABR od 54 dzieci. Do analizy zakwalifikowano 312 zapisów, natomiast 10 odrzucono – w każdym z tych 10 przypadków przynajmniej jeden ekspert uznał, że ze względu na niską jakość wyniku nie jest możliwa wiarygodna ocena progu słyszenia. Liczba zapisów dla poszczególnych bodźców przedstawiała się następująco: trzask – 71 zapisów, 500 Hz – 91, 1000 Hz – 105; 2000 Hz – 23, 4000 Hz – 22 zapisy.

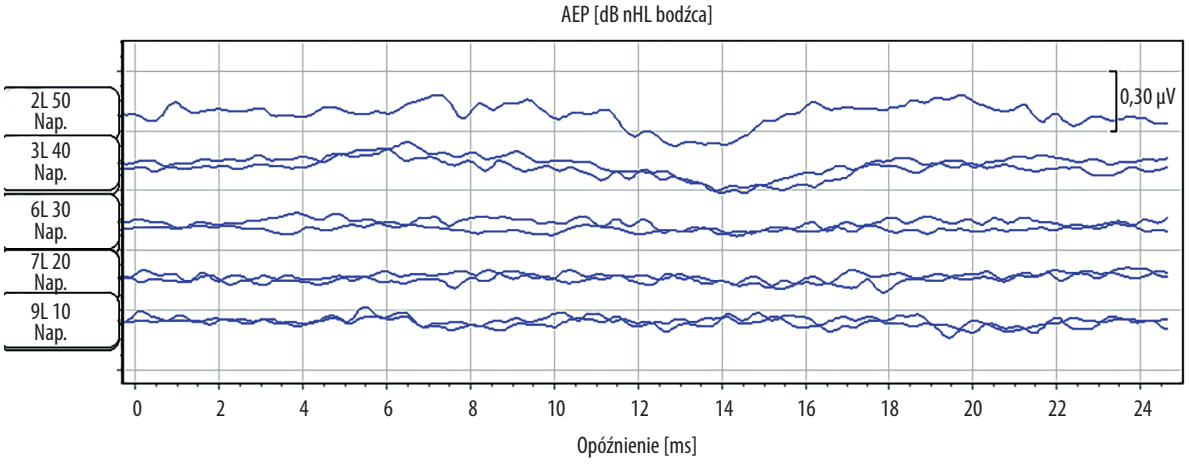
Wyniki

Zgodność ocen ekspertów

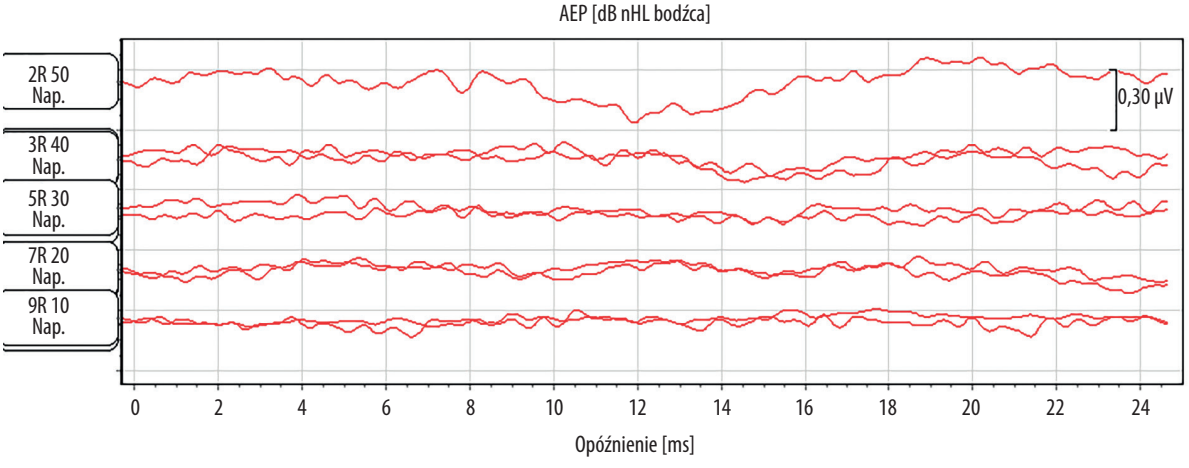
Różnice ocen ekspertów zawierały się w przedziale od 0 do 30 dB, średnia różnica w ocenach wynosiła 5,4 dB ($SD = 6{,}4$).

W 167 na 312 badanych zapisów odnotowano pełną zgodność między czterema ekspertami. Oznacza to, że w 53,5% przypadków eksperci identycznie oszacowali próg słyszenia, różnica między ich ocenami wynosiła 0 dB. W 124 przypadkach (39,7%) rozbieżność ocen wynosiła 10 dB, w 19 zapisach (6,1%) rozbieżność ocen wynosiła ±20 dB, a w dwóch zapisach rozbieżność ocen wynosiła ±30 dB. Na **rycinie 1** przedstawiono powyższe dane w ujęciu procentowym.

Zidentyfikowano dwa zapisy ABR, w których oceny ekspertów różniły się w największym stopniu – rozbieżności wynosiły do 30 dB. Oba te zapisy zaprezentowano poniżej (**ryciny 2 i 3**). Rozbieżności z właściwą identyfikacją fali V w ocenach poszczególnych mogły wynikać przede wszystkich z braku powtórzenia zapisów ABR dla natężenia 50 dB nHL oraz zignorowania faktu braku fali V w zapisach odpowiedzi dla 30 dB nHL.



Rycina 2. Zapis ABR dla bodźca 500 Hz (ucho lewe); oceny ekspertów: 50–20–40–30 dB HL
Figure 2. ABR recording for a 500 Hz stimulus (left ear); expert ratings: 50–20–40–30 dB HL



Rycina 3. Zapis ABR dla bodźca 500 Hz (ucho prawe); oceny ekspertów: 50–20–30–40 dB HL
Figure 3. ABR recording for a 500 Hz stimulus (right ear); expert ratings: 50–20–30–40 dB HL

Korelacja między ocenami ekspertów

Korelacja między ocenami progów słyszenia, dokonanymi niezależnie przez czterech ekspertów, została oszacowana za pomocą współczynnika korelacji wewnątrzklasowej – ICC (2,k) wyniósł 0,98 i był istotny statystycznie ($p < 0{,}001$). Wynik ten świadczy o bardzo dużej zgodności ocen ekspertów w szacowaniu progów słyszenia. Na **rycinie 4** przedstawiono graficznie zależności między ocenami poszczególnych ekspertów. W każdej z sześciu kombinacji ($\binom{4}{2} = 6$) widoczna jest bardzo silna, pozytywna korelacja ocen w danej parze ekspertów.

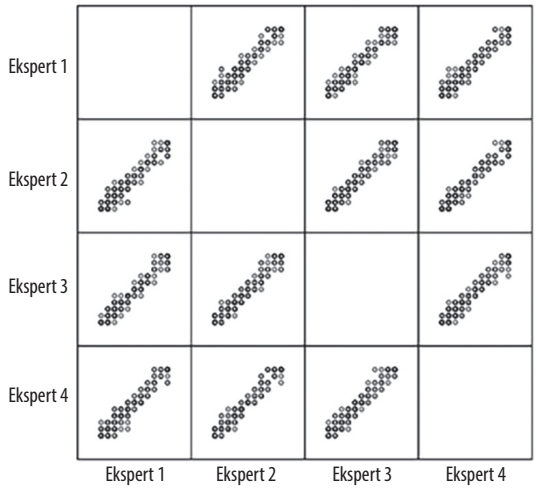
Zgodność ocen ekspertów a rodzaj bodźca

Zbadano dodatkowo, czy zgodność ocen progów słyszenia pomiędzy ekspertami różni się w zależności od rodzaju bodźca. W **tabeli 1** zestawiono rozbieżności ocen dla bodźca typu trzask oraz tonów o częstotliwościach 500, 1000, 2000 i 4000 Hz.

Z danych przedstawionych w **tabeli 1** wynika, że w ponad 90% przypadków zapisów uzyskanych zarówno dla trzasku, jak i krótkich tonów, rozbieżności między ekspertami nie przekraczały 10 dB, były zatem klinicznie akceptowalne. Największy odsetek przypadków, w których rozbieżności ocen przekraczały 10 dB, zaobserwowano dla bodźca 500 Hz (9,9%) oraz dla bodźca 4000 Hz (9,1%). W przypadku bodźca 4000 Hz możliwym wyjaśnieniem może być stosunkowo niewielka liczba analizowanych zapisów, co zwiększa podatność wyników na przypadkowe wahania. Z kolei zaobserwowane rozbieżności dla bodźca 500 Hz mogą być związane z właściwościami propagacji fali wędrownej o niskiej częstotliwości w ślimaku, co zostanie omówione w dalszej części pracy.

Dyskusja

Celem pracy była ocena zgodności ocen progów słyszenia u dzieci, dokonywanych na podstawie zapisów słuchowych potencjałów wywołanych pnia mózgu (ABR) przez kilku



Rycina 4. Korelacje między ocenami progów słyszenia dokonanymi przez czterech ekspertów
Figure 4. Correlations between hearing threshold ratings made by four experts

Tabela 1. Rozbieżności (dB, wartości bezwzględne) w ocenach 312 zapisów ABR między czterema ekspertami
Table 1. Discrepancies (dB, absolute values) in the ratings of 312 ABR records between four experts

Bodziec [Hz]	Różnica bezwzględna [dB]	*n*	%
Trzask	0	36	50,7
	10	32	45,1
	20	3	4,2
500	0	51	56,0
	10	31	34,1
	20	7	7,7
	30	2	2,2
1000	0	54	51,4
	10	45	42,9
	20	6	5,7
2000	0	12	52,2
	10	10	43,5
	20	1	4,3
4000	0	14	63,6
	10	6	27,3
	20	2	9,1

niezależnych ekspertów. Choć badanie ABR odgrywa kluczową rolę w diagnostyce audiologicznej dzieci – zwłaszcza tych, u których nie można uzyskać wiarygodnych odpowiedzi behawioralnych – problem subiektywności oceny progów słyszenia w procedurze szeregu natężeniowego jest wciąż rzadko podejmowany w literaturze naukowej. Tymczasem zagadnienie to ma nader istotne znaczenie kliniczne, np. zaopatrzenie w aparaty słuchowe czy kwalifikacja do implantacji. Są to decyzje o fundamentalnym znaczeniu dla dalszego rozwoju dziecka – zarówno w wymiarze poznawczym, jak również emocjonalnym i społecznym. Dlatego konieczne jest pogłębianie wiedzy na temat rzetelności ocen dokonywanych przez osoby intepretujące wyniki badań ABR oraz identyfikacja ewentualnych źródeł rozbieżności.

Pierwsze doniesienia dotyczące zgodności ocen ekspertów analizujących zapisy ABR ukazały się w latach 70. XX wieku. Rose i wsp. [22] zgromadzili 613 zapisów słuchowych potencjałów wywołanych, pochodzących od 50 pacjentów. Sześciu doświadczonych klinicystów oceniało, czy odpowiedź występuje czy nie, jednak w publikacji nie sprecyzowano, co dokładnie rozumiano pod słowem „odpowiedź". Można przypuszczać, że chodziło o identyfikację charakterystycznych komponentów odpowiedzi korowej (P1, N1, P2, N2), nie wskazano jednak jakie parametry (latencja, amplituda lub inne) podlegały ocenie. W 333 przypadkach na 613 (54%) sędziowie byli całkowicie zgodni (tzn. albo wszyscy zgadzali się, że odpowiedź występuje, albo wszyscy zgadzali się, że odpowiedź nie występuje). Zgodność częściowa (między pięcioma, czterema lub trzema 3 ekspertami) wystąpiła odpowiednio w 20%, 18% i 8% przypadków. Dodatkowo, eksperci oceniali każdy zapis trzykrotnie, z co najmniej tygodniową przerwą między kolejnymi rundami ocen. Powtarzalność ich własnych ocen (ang. *intra-rater reliability*) mieściła się w zakresie od 65% do 88%. Badanie Rose i wsp. [22], choć cenne, to jest obarczone istotnymi ograniczeniami metodologicznymi, przede wszystkim niejednoznacznością samego przedmiotu oceny oraz jej kryteriów.

Mało precyzyjna była także metodologia zastosowana przez Cohena i wsp. [21]. W badaniu analizowano oceny ekspertów dla 464 zapisów słuchowych potencjałów wywołanych, pozyskanych od 20 dzieci; mediana ich wieku wyniosła 28,5 miesiąca. Trzech doświadczonych ekspertów miało za zadanie wskazać, czy odpowiedź jest obecna, czy jest jej brak, mogli również stwierdzić, że nie są pewni obecności lub braku odpowiedzi. Zgodność między trzema ekspertami wynosiła 49%. Każdy ekspert dokonał również powtórnej oceny tych samych zapisów po co najmniej 2 tygodniach. Powtarzalność wynosiła 59–75%. Niestety nie sposób wprost odnieść wyników obu wspomnianych powyżej prac do wyników badań własnych, ponieważ nie dotyczą one wyznaczania progów słyszenia na podstawie fali V. Jednak ich wartością jest pokazanie, że już na wczesnym etapie rozwoju metod elektrofizjologicznych oceny słuchu dostrzegano problem dużej zmienności subiektywnych ocen klinicznych.

Kolejne badanie przeprowadzili Gans i wsp. [23]. Ekspertami było 9 studentów audiologii, kontynuujących naukę na studiach magisterskich lub doktoranckich (ang. *audiology gratuate students*), posiadających kilka miesięcy doświadczenia klinicznego. Analizowali oni zapisy ABR pochodzące od 50 dzieci z różnymi niepełnosprawnościami, w tym wszystkie miały zdiagnozowaną niepełnosprawność intelektualną. Badanie ABR wykonano dla bodźców typu trzask o intensywnościach 30, 50, 70 i 90 dB nHL. Analizie poddano 900 zapisów. Wszystkie zapisy, zanim zostały poddane ocenie 9 ekspertów, zostały

ocenione przez doświadczonego klinicystę z ponad 10-letnim stażem w analizie ABR. Metodyka badania była nietypowa, bowiem ekspertom podano uprzednio informację, jaki próg został stwierdzony przez doświadczonego klinicystę, jednak w połowie przypadków podano progi prawdziwe, a w połowie – progi fałszywe, przesunięte o ±20 dB. W przypadku zapisów, gdy podany był prawdziwy próg, w 71% osiągnięto zgodność ekspertów z wzorcową oceną. Natomiast gdy podany próg był fałszywy, zgodność ekspertów z wzorcową oceną wynosiła 56%. Ogółem zgodność ocen ekspertów z prawdziwym progiem osiągnięto w 64% zapisów.

Przedstawione powyżej wyniki badań Gansa i wsp. są bardzo cenne w kontekście badań własnych, ponieważ odnoszą się do oceny progów słyszenia i bazują na analizie szeregów natężeniowych. Jednak w badaniu Gansa i wsp. ekspertami nie byli w pełni wykwalifikowani specjaliści, a ich kompetencje kliniczne były ograniczone. Mogło to mieć wpływ na trafność i powtarzalność ocen, a tym samym na ogólną interpretację wyników badania. Z drugiej strony, badanie ukazuje interesujące zjawisko podatności na sugestię: gdy ekspertom przekazano nieprawdziwe informacje o progach, ich zgodność z rzeczywistą oceną spadła. Może to świadczyć o dużym wpływie oczekiwań poznawczych na analizę sygnałów ABR, co jest istotne także w kontekście praktyki klinicznej. W polskich warunkach osoba wykonująca badanie ABR nieraz samodzielnie oznacza falę V i taki wynik przekazuje lekarzowi. W świetle wyników Gansa i wsp. zasadne jednak staje się pytanie – czy powinna to robić? Oznaczanie fali V to kluczowy etap interpretacji zapisu ABR, który wymaga: wiedzy teoretycznej o neurofizjologii słuchu, doświadczenia klinicznego, umiejętności odróżniania sygnału od artefaktu, uwzględnienia kontekstu pacjenta (wiek, stan neurologiczny, współistniejące zaburzenia). Wyniki przytoczonego badania pokazują, że nawet osoby z wiedzą akademicką i z pewnym doświadczeniem mogą być podatne na sugestię i błędy w oznaczaniu fali V. Wydaje się więc zasadne, by w praktyce klinicznej technicy ograniczyli się do poprawnego wykonania badania i zapewnienia dobrej jakości zapisu, natomiast interpretacja – w tym oznaczanie fali V – powinna należeć do lekarza specjalisty, ewentualnie do doświadczonego audiologa w krajach, gdzie ma on uprawnienia diagnostyczne.

Powtarzalności ocen ABR dotyczyło także badanie Navesa i wsp. [8]. Od 10 osób dorosłych (po 5 kobiet i mężczyzn) ze słuchem w normie, w wieku od 20 do 52 lat zebrano 160 zapisów ABR uzyskanych dla bodźca typu trzask o czterech poziomach intensywności: 20, 40, 60 i 80 dB HL. Zapisy były oceniane przez czterech ekspertów klinicystów ze stażem audiologicznym trwającym od 6 do 15 lat, przy czym zajmowali się analizą wyników ABR od co najmniej 3 lat. Zadaniem ekspertów była ocena latencji fal I–V (w milisekundach). Okazało się, że oceny ekspertów były najbardziej zbieżne przy dużych intensywnościach bodźca, natomiast wariancja ocen między ekspertami rosła, gdy bodźce były słabsze. Dla kluczowej w diagnostyce słuchu fali V średnia wartość latencji (dla bodźca o intensywności 80 dB) wynosiła 5,53 ms, a rozbieżności między ekspertami wynosiły od 0,023 ms do 0,080 ms. Badacze uznali, że różnica między ocenami ekspertów wynosząca poniżej 0,1 ms jest klinicznie akceptowalna. Taka różnica została osiągnięta w 87% przypadków i na tej podstawie badacze uznali, że eksperci osiągali wysoką zgodność w ocenie latencji fal I–V. Jednak ta konkluzja może budzić wątpliwości. Po pierwsze, przyjęcie jednej wartości granicznej (0,1 ms) jako kryterium akceptowalności może nie być adekwatne dla wszystkich fal – rozbieżność o tej samej wartości ma relatywnie większe znaczenie przy krótszych latencjach (np. fali I) niż przy dłuższych (np. fali V). Po drugie, w literaturze spotyka się również wyższe progi akceptowalności, np. 0,2 ms [24,25], co wskazuje na niejednoznaczność standardów w tym zakresie. Wreszcie fakt, że w 13% przypadków rozbieżność ocen przekraczała ustalony próg, może wskazywać na potrzebę dalszej standaryzacji procedur oceny zapisów ABR, zwłaszcza przy niższych poziomach intensywności bodźca, w przypadku których zgodność między ekspertami była zauważalnie niższa.

Najbardziej interesujące badanie, z którego wynikami można porównać wyniki badań własnych, przeprowadzili Vidler i Parker [11]. Zebrali oni grupę 16 ekspertów, każdy z nich miał doświadczenie audiologiczne (od 3,5 do 25 lat) oraz doświadczenie kliniczne w wykonywaniu i interpretacji badań ABR (od 1,5 do 25 lat). Każdy z ekspertów oceniał progi słyszenia na podstawie fali V dla 12 przypadków. Eksperci mieli do dyspozycji pełne szeregi natężeniowe w zakresie od 75 do 15 dB nHL, ze skokiem 5 dB. Dla każdej intensywności bodźca wykonywano po dwa przebiegi. Istotny w metodologii tego badania jest fakt, że oceniane zapisy nie pochodziły od rzeczywistych pacjentów, ale były wynikiem symulacji komputerowej.

Z badania Vidlera i Parkera wynika, że rozbieżności między ocenami ekspertów były znaczne – wynosiły od 10 do nawet 60 dB. Kiedy grupa liczy 16 ekspertów, trudno o wyliczenie zgodności w ujęciu procentowym, posłużono się w tym celu dodatkową miarą, jaką był próg rzeczywisty, tzn. taki, jaki oszacowali sami badacze. Okazało się, że rozbieżności między ekspertami nie zależały od tego progu. Na przykład przy progu rzeczywistym ocenionym na 20 dB HL zdarzały się rozbieżności zarówno stosunkowo niewielkie (15 dB), jak i znaczne, wynoszące 40 dB. Natomiast w przypadku szeregów, gdy próg wynosił 80 dB Hz, rozbieżność między ekspertami wyniosła raz 20 dB, a innym razem – 35 dB. Zgodność ekspertów z rzeczywistym progiem wynosiła od 81% (tzn. 81% ekspertów wskazało próg tożsamy z tym określonym przez badaczy) do zaledwie 6%.

W tym kontekście wyniki uzyskane w badaniach własnych wydają się cechować wysoką zgodnością. Zgodność między czterema ekspertami była idealna w 53,5% analizowanych zapisów. Natomiast jeśli dołączyć do tego zapisy, w których rozbieżność wyniosła ±10 dB, to wskaźnik zgodności wzrastał do 93,2%. Pozostało zatem 6,8% przypadków, w których różnice ocen między ekspertami można uznać za znaczne – wynoszące ±20 dB i więcej. W kontekście przytoczonych powyżej wyników osiągniętych przez innych badaczy, te rezultaty można uznać za więcej niż zadowalające. Niemniej nie można bagatelizować tych przypadków, w których wystąpiły znaczne rozbieżności w ocenie wyniku badania, trudne do zaakceptowania z klinicznego punktu widzenia. Warto przy tym podkreślić, że w przypadku wykonywania badań

rzeczywistych ABR u dzieci, przeprowadzenie pełnego szeregu natężeniowego – jak miało to miejsce w badaniu Vidlera i Parkera (choć szereg był generowany komputerowo) – jest utrudnione. Ograniczenia wynikają przede wszystkim z czasu trwania badania – dzieci badane są podczas snu, a zbyt długi protokół może prowadzić do ich wybudzenia. Dodatkowo, przy bodźcach o najwyższych intensywnościach, istnieje ryzyko, że dziecko wybudzi się i badanie będzie trzeba przerwać. W rzeczywistych warunkach klinicznych konieczne jest zatem dążenie do możliwie wiernego przestrzegania protokołu badania przy jednoczesnym uwzględnieniu praktycznych ograniczeń związanych z diagnostyką małego dziecka.

W badaniach własnych stwierdzono, że szczególne trudności interpretacyjne może sprawiać bodziec tonalny o częstotliwości 500 Hz. Jest to o tyle zrozumiałe, że w przypadku bodźca o częstotliwości 500 Hz w przedziale intensywności od 60 do 40 dB HL następuje zmiana miejsca maksymalnego pobudzenia błony podstawnej ślimaka. Dla wyższych intensywności (60–100 dB HL), na skutek ograniczonej specyficzności częstotliwościowej, odpowiedź jest generowana głównie w zakręcie podstawnym ślimaka, natomiast dla niższych intensywności (40–20 dB HL) – zarówno w zakręcie podstawnym, jak i w zakręcie szczytowym [26]. A ponieważ latencje odpowiedzi powstających w dwóch odległych miejscach ślimaka różnią się znacznie między sobą, to w procesie uśredniania może nastąpić ich zniesienie. Wynika stąd praktyczne zalecenie dla klinicysty intepretującego wynik badania o zachowanie ostrożności w szacowaniu progu w przypadku bodźca 500 Hz. Z kolei dla osoby wykonującej badanie za pomocą bodźca 500 Hz można sformułować zalecenie, by procedury nie kończyć zbyt wcześnie, ale kontynuować ją przynajmniej dla bodźca o intensywności o 20 dB niższej od intensywności, przy której nie zarejestrowano fali V.

Wyniki uzyskane w badaniach własnych mają duże znaczenie w kontekście opracowania algorytmów służących do automatycznej detekcji progu słyszenia na podstawie fali V. Próby automatyzacji tego procesu podejmowane są od wielu lat, jednak w ostatnim czasie nabrały szybkiego tempa dzięki rosnącej dostępności zaawansowanych metod obliczeniowych, takich jak np. głębokie sieci neuronowe, maszyny wektorów nośnych czy modele zespołowe. Opracowywane algorytmy bazują przede wszystkim na metodach uczenia nadzorowanego (ang. *supervised learning*), w którym model uczy się na podstawie danych uprzednio oznaczonych – w tym przypadku na wartościach progów słyszenia oszacowanych przez eksperta [13,27–30]. Model taki uczy się przewidywać próg słyszenia na podstawie wybranych cech sygnału ABR, takich jak latencja, amplituda czy kształt odpowiedzi. Skuteczne trenowanie tego typu algorytmów wymaga jednak dostępu do dużych, dobrze opisanych zbiorów wiarygodnych danych klinicznych. W tym kontekście szczególnie istotna jest wysoka zgodność ocen progów słyszenia dokonywanych przez doświadczonych klinicystów. Ich spójne decyzje stanowią wiarygodny punkt odniesienia (ang. *ground truth*) dla algorytmów uczących się. Wyniki badań własnych wskazują, że ogólny poziom zgodności między doświadczonymi specjalistami jest wysoki, a rozbieżności przekraczające dopuszczalny klinicznie poziom (10 dB) zdarzają się w kilku procentach przypadków. Ta wysoka zgodność ocen eksperckich jest podstawą do opracowania, trenowania i walidacji algorytmów automatycznych, które mogą skutecznie wspierać proces diagnostyczny, zapewniając większą obiektywność interpretacji zapisów ABR.

Wnioski

Doświadczeni klinicyści, posiadający długi staż w interpretacji wyników badań ABR, osiągają wysoką zgodność w ocenie progów słyszenia na podstawie fali V. Niemniej zdarzają się przypadki, w których rozbieżność ocen nawet między doświadczonymi ekspertami przekracza ±10 dB. Nieprawidłowa ocena progu słyszenia na podstawie fali V w badaniu ABR może prowadzić do wyników zarówno fałszywie dodatnich, jak i fałszywie ujemnych, co skutkuje nietrafną diagnozą. W celu zwiększenia zgodności ocen i tym samym poprawienia trafności diagnoz, pomocne mogą być m.in.: weryfikacja wyników badania przez drugiego specjalistę, walidacja krzyżowa (ang. *cross-checking*) za pomocą badań behawioralnych lub audiometrycznych, a także opracowanie lokalnych protokołów, standaryzacja procedur wykonywania badania oraz zasad interpretacji odpowiedzi. Nowoczesne automatyczne narzędzia wspomagające decyzje kliniczne mogą dodatkowo zwiększyć trafność diagnozy, a jednocześnie zmniejszyć koszty i skrócić czas diagnostyki audiologicznej.

Finansowanie

Badania finansowane ze środków Agencji Badań Medycznych przyznanych na projekt ABR.AI – system oparty o technologię AI do szacowania progu słyszenia na podstawie szeregów natężeniowych rejestracji odpowiedzi fali V; 2024/ABM/03/KPO/KPOD.07.07-IW.07-0222/24-00.

Piśmiennictwo

1. Stapells DR. Threshold estimation by the tone-evoked auditory brainstem response: a literature meta-analysis. J Speech Lang Pathol Audiol, 2000; 24(2): 74–83.
2. O-Uchi T, Kanzaki J. Diagnostic value of auditory brain stem responses (ABR) in differential diagnosis between cochlear deafness and acoustic neuromas. Acta Otolaryngol Suppl, 1983; 393: 85–91; https://10.3109/00016488309129582.
3. Sato S, Yamada M, Koizumi H, Onozawa Y, Shimokawa N, Kawashima E i wsp. Neurophysiological mechanisms of conduction impairment of the auditory nerve during cerebellopontine angle surgery. Clin Neurophysiol, 2009; 120(2): 329–35; https://doi.org/10.1016/j.clinph.2008.11.005.
4. Tokimura H, Asakura T, Tokimura Y, Atsuchi M, Kimotsuki K, Sato E i wsp. [Intraoperative ABR monitoring during cerebellopontine angle surgery]. No Shinkei Geka, 1990; 18(11): 1023–7.
5. Jacobson JT, Jacobson CA, Spahr RC. Automated and conventional ABR screening techniques in high-risk infants. J Am Acad Audiol, 1990; 1(4): 187–95.
6. Serban S, Rusescu A, Ionescu MT. Usefulness of auditory brainstem audiometry in forensic medicine and expertise of auditory disabilities. Clinical Aspects, 2016.
7. Hall J. New Handbook of Auditory Evoked Responses. Boston: Pearson, 2007.

8. Naves KFP, Pereira AA, Nasuto SJ, Russo IPC, Andrade AO. Assessment of inter-examiner agreement and variability in the manual classification of auditory brainstem response. Biomed Eng Online, 2012; 11: 86; https://doi.org/10.1186/1475-925X-11-86.
9. Hildesheimer M, Muchnik C, Rubinstein M. Problems in interpretation of brainstem-evoked response audiometry results. Audiology, 1985; 24(5): 374–9; https://doi.org/10.3109/00206098509078355.
10. Kochanek K. Przyczyny błędów w progowych badaniach ABR. Audiofonologia, 1998; XIII: 139–50.
11. Vidler M, Parker D. Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. Int J Audiol, 2004; 43(7): 417–29; https://doi.org/10.1080/14992020400050053.
12. Kochanek K. Zasady stosowane w progowych badaniach ABR. Audiofonologia, 1998; XIII: 131–7.
13. Trzaskowski B. Porównanie algorytmów detekcji fali V w badaniu słuchowych potencjałów wywołanych pnia mózgu. Now Audiofonol, 2015; 4(2): 43–52; https://doi.org/10.17431/894416.
14. Berninger E, Olofsson A, Leijon A. Analysis of click-evoked auditory brainstem responses using time domain cross-correlations between interleaved responses. Ear Hear, 2014; 35(3): 318–29; https://doi.org/10.1097/01.aud.0000441035.40169.f2.
15. Wang H, Li B, Lu Y, Han K, Sheng H, Zhou J i wsp. Real-time threshold determination of auditory brainstem responses by cross-correlation analysis. iScience, 2021; 24(11): 103285; https://doi.org/10.1016/j.isci.2021.103285.
16. McKearney RM, Bell SL, Chesnaye MA, Simpson DM. Auditory brainstem response detection using machine learning: a comparison with statistical detection methods. Ear Hear, 2022; 43(3): 949–60; https://doi.org/10.1097/AUD.0000000000001151.
17. Zhang R, McAllister G, Scotney B, McClean S, Houston G. Coupling wavelet transform with bayesian network to classify auditory brainstem responses. Conf Proc IEEE Eng Med Biol Soc, 2005; 2005: 7568–71; https://doi.org/10.1109/IEMBS.2005.1616263.
18. Wimalarathna H, Ankmnal-Veeranna S, Allan C, Agrawal SK, Samarabandu J, Ladak HM i wsp. Machine learning approaches used to analyze auditory evoked responses from the human auditory brainstem: a systematic review. Comput Methods Programs Biomed, 2022; 226: 107118; https://doi.org/10.1016/j.cmpb.2022.107118.
19. Naït-Ali A, Adam O, Motsch JF. Modelling and recognition of brainstem auditory evoked potentials using Symlet wavelet. ITBM-RBM, 2000; 21(3): 150–7; https://doi.org/10.1016/S1297-9562(00)90029-9.
20. Zidan M, Thomas RL, Slovis TL. What you need to know about statistics, part II: reliability of diagnostic and screening tests. Pediatr Radiol, 2015; 45(3): 317–28; https://doi.org/10.1007/s00247-014-2944-x.
21. Cohen MM, Rapin I, Lyttle M, Schimmel H. Auditory evoked response (AER): consistency of detection in young sleeping children. Arch Otolaryngol, 1971; 94(3): 214–9; https://doi.org/10.1001/archotol.1971.00770070208006.
22. Rose DE, Keating LW, Hedgecock LD, Schreurs KK, Miller KE. Aspects of acoustically evoked responses: inter-judge and intra-judge reliability. Arch Otolaryngol, 1971; 94(4): 347–50; https://doi.org/10.1001/archotol.1971.00770070539010.
23. Gans D, Del Zotto D, Gans KD. Bias in scoring auditory brainstem responses. Br J Audiol, 1992; 26(6): 363–8; https://doi.org/10.3109/03005369209076660.
24. Don M. Quantitative approaches for defining the quality and threshold of auditory brainstem responses. Images of the Twenty-First Century Proceedings of the Annual International Engineering in Medicine and Biology Society, 1989; 2: 761–2; https://doi.org/10.1109/IEMBS.1989.95969.
25. McKearney RM, Simpson DM, Bell SL. Automated wave labelling of the auditory brainstem response using machine learning. Int J Audiol, 2025; 64(7): 766–71; https://doi.org/10.1080/14992027.2024.2404537.
26. Kochanek K, Skarżyński H, Janczewski G, Grzanka A, Piłka A. Specyficzność częstotliwościowa słuchowych potencjałów pnia mózgu wywołanych impulsem tonu o obwiedni Gaussa o częstotliwościach 500 Hz. I: Uszy normalnie słyszące. Audiofonologia, 2001; XX: 113–23.
27. Chen C, Zhan L, Pan X, Wang Z, Guo X, Qin H i wsp. Automatic recognition of auditory brainstem response characteristic waveform based on bidirectional long short-term memory. Front Med (Lausanne), 2020; 7: 613708; https://doi.org/10.3389/fmed.2020.613708.
28. Chesnaye MA, Simpson DM, Schlittenlacher J, Laugesen S, Bell SL. Audiogram estimation performance using auditory evoked potentials and gaussian processes. Ear Hear, 2025; 46(1): 230–41; https://doi.org/10.1097/AUD.0000000000001570.
29. Liu Y, Xiang L, Li Q, Li K, Yang Y, Wang T i wsp. Comparison of deep learning models for objective auditory brainstem response detection: a multicenter validation study. Trends Hear, 2025; 29: 23312165251347773; https://doi.org/10.1177/23312165251347773.
30. Thalmeier D, Miller G, Schneltzer E, Hurt A, Hrabě de Angelis M, Becker L i wsp. Objective hearing threshold identification from auditory brainstem response measurements using supervised and self-supervised approaches. BMC Neuroscience, 2022; 23(1): 81; https://doi.org/10.1186/s12868-022-00758-0.

Dr hab. n. med. i n. o zdr. Elżbieta Gos, email: e.gos@ifps.org.pl • 0000-0003-3173-3867
Mgr Małgorzata Pastucha, email: m.pastucha@ifps.org.pl • 0000-0002-4866-1866
Prof. dr hab. n. med. i n. o zdr. Krzysztof Kochanek, email: k.kochanek@ifps.org.pl • 0000-0003-2860-5784
Mgr inż. Adam Piłka, email: a.pilka@ifps.org.pl • 0000-0002-3382-1126
Dr hab. n. med. i n. o zdr. Elżbieta Włodarczyk, email: e.wlodarczyk@ifps.org.pl • 0000-0003-3737-5730
Prof. dr hab. med. dr h.c. multi Henryk Skarżyński, email: skarzynski.henryk@ifps.org.pl • 0000-0001-7141-9851