PRACA POGLĄDOWA

PRZEGLĄD WŁAŚCIWOŚCI PSYCHOMETRYCZNYCH KWESTIONARIUSZY OCENY JAKOŚCI ŻYCIA ZWIĄZANYCH ZE ZDROWIEM (HRQOL)

THE REVIEW OF PSYCHOMETRIC PROPERTIES OF HEALTH RELATED QUALITY OF LIFE (HRQOL) INSTRUMENTS

Katarzyna Młyńczak[1], Dominik Golicki[2]

[1] HealthQuest sp. z o.o., sp. k. z siedzibą w Warszawie
[2] Katedra i Zakład Farmakologii Doświadczalnej i Klinicznej, Warszawski Uniwersytet Medyczny

DOI: https://doi.org/10.20883/ppnoz.2016.30

STRESZCZENIE

Wiarygodność kwestionariuszy oceny jakości życia jest niezbędna do prawidłowej interpretacji ich wyników. Walidacja pozwala określić, na ile przydatne będzie testowane narzędzie w konkretnym kontekście klinicznym. Do takiej oceny służą parametry określane mianem właściwości psychometrycznych, na które składają się przede wszystkim trafność (ang. *validity*), rzetelność (ang. *reliability*) oraz czułość testu. Celem pracy jest usystematyzowanie wiedzy w zakresie psychometrii wykorzystywanej w badaniu wiarygodności kwestionariuszy oceny jakości życia.

Trafność określa, czy kwestionariusz mierzy to, co powinien mierzyć. Najczęściej wyróżniamy jej trzy aspekty: trafność treściową, kryterialną oraz teoretyczną. Trafność treściowa określa, na ile instrument pokrywa zamierzony obszar zainteresowania. Trafność kryterialna to analiza zgodności pomiarów otrzymanych za pomocą narzędzia z określonym kryterium o udowodnionej relacji z pierwotnie mierzoną wartością wyjściową (walidacja poprzez porównanie wyników z testem referencyjnym). Trafność teoretyczna określa stopień zależności pomiędzy wynikami kwestionariusza a teoretycznym modelem zjawiska, zbudowanym na podstawie osobnych badań jego zależności.

Rzetelność w świetle klasycznej teorii testów ocenia stopień losowej rozbieżności w wynikach pomiarów, które z założenia powinny być jednakowe oraz określa zakres, w którym kwestionariusz będzie dawał spójne i powtarzalne wyniki. Rozróżnia się rzetelność rozumianą jako spójność wewnętrzną oraz rzetelność związaną z powtarzalnością pomiarów (testu–retestu, pod względem zgodności ocen).

Czułość kwestionariusza to jego zdolność do pomiaru różnic w jakości życia między pacjentami lub grupami pacjentów. Wyróżnia się również ocenę czułości na zmiany kliniczne (responsywność), wykonywaną u tego samego pacjenta.

Ocena właściwości psychometrycznych kwestionariuszy jakości życia pozwala określić ich wiarygodność w warunkach konkretnego kontekstu klinicznego. Istotna jest ocena wszystkich jej aspektów, gdyż tylko wtedy można mówić o pozytywnym wyniku walidacji.

Słowa kluczowe: walidacja, trafność, rzetelność, czułość, responsywność.

ABSTRACT

Health-related quality of life (HRQoL) questionnaires need to be validated for their proper interpretation and determining, how useful the test will be in the specific clinical context. Validity, reliability and sensitivity are usually used for it. The aim of this work is to summarize psychometric properties, used in assessing questionnaires' validity.

Validity determines whether the instrument measures what it is intended to measure. Its three aspects: content, criterion and construct, need to be considered. Content validity concerns the extent to which the questionnaire reflects the intended domain of interest. Criterion validity is the analysis of the agreement between measurements obtained using the test and the specified criterion with a proven relationship with the originally measured value (quality of life). Construct validity determines the extent of correlation between the test and the theory built on modelling the clinical context structure.

According to the classical test theory reliability assesses the impact of a random variability in the measurement results, which are supposed to be similar, and determines the range of internal consistency and repeatability (inter-rater reliability, and test-retest reliability) of questionnaire's results.

The sensitivity of the test is its ability to measure differences in quality of life between patients or groups of patients. It also assesses the sensitivity to clinical changes (responsiveness), performed on the same patient.

The evaluation of psychometric properties of the HRQoL questionnaires determines their validity in the clinical context. It is important to assess all their aspects, because only then we can talk about the positive validation result.

Keywords: validation, validity, reliability, sensitivity, responsiveness.

Wstęp

Jakość życia związana ze zdrowiem (ang. *health related quality of life*) charakteryzuje odczuwanie przez pacjenta jego stanu zdrowia, a jej zmiana może stanowić wyznacznik skuteczności działania zastosowanej terapii medycznej. Ponieważ jednak jakość życia jest cechą latentną, tzn. niepoddającą się obserwacji, nie jest możliwe jej zmierzenie w sposób bezpośredni [2, 3]. Pośrednie metody pomiaru to różnego rodzaju kwestionariusze, których celem jest wnioskowanie o jakości życia na podstawie charakterystyki zakresu ograniczeń aktywności fizycznej, psychicznej czy społecznej pacjenta.

Ocena jakości życia związanej ze zdrowiem stanowi istotny etap procesów wnioskowania o efektywności kosztowej technologii medycznej. Aby zapewnić poprawność takiego wnioskowania, należy być przekonanym o poprawności oceny jakości życia, którą może zapewnić walidacja narzędzi pomiarowych. Walidacja to proces badania właściwości psychometrycznych kwestionariusza i weryfikacji jego wiarygodności [3].

Wiarygodność w kontekście właściwości psychometrycznych może być rozumiana w dwóch aspektach. Po pierwsze, wiarygodny kwestionariusz powinien być użyteczny w konkretnej sytuacji klinicznej, tzn. poprawnie rozpoznawać różnice w jakości życia wynikające z różnych, możliwych do wystąpienia w danej chorobie stanów zdrowia. Ten rodzaj wiarygodności charakteryzuje pojęcie trafności testu (ang. *validity*). Z kolei wyniki kwestionariusza, uzyskiwane przy założeniu stałości oceny jakości życia, powinny być spójne, powtarzalne i wolne od błędów pomiarowych, co nazywamy rzetelnością testu (ang. *reliability*). Dopiero równowaga w ocenie trafności oraz rzetelności kwestionariusza pozwala twierdzić o jego wiarygodności [1].

Trafność instrumentu określa, czy kwestionariusz mierzy to, co powinien mierzyć [2]. Trafność należy do tej grupy pojęć z zakresu psychometrii, według których wiarygodność instrumentu związana jest nie tylko z poprawnym jego działaniem, ale z działaniem w obrębie konkretnego kontekstu klinicznego. Nie jest zatem cechą samego narzędzia, tylko wniosków, jakie będą wyciągane na podstawie jego wyników. Rzetelność z kolei określa cechę narzędzia, ale może być zmienna w zależności od kontekstu. Ocenia, czy wyniki określające tę samą cechę będą powtarzalne dla kolejnych pomiarów [2].

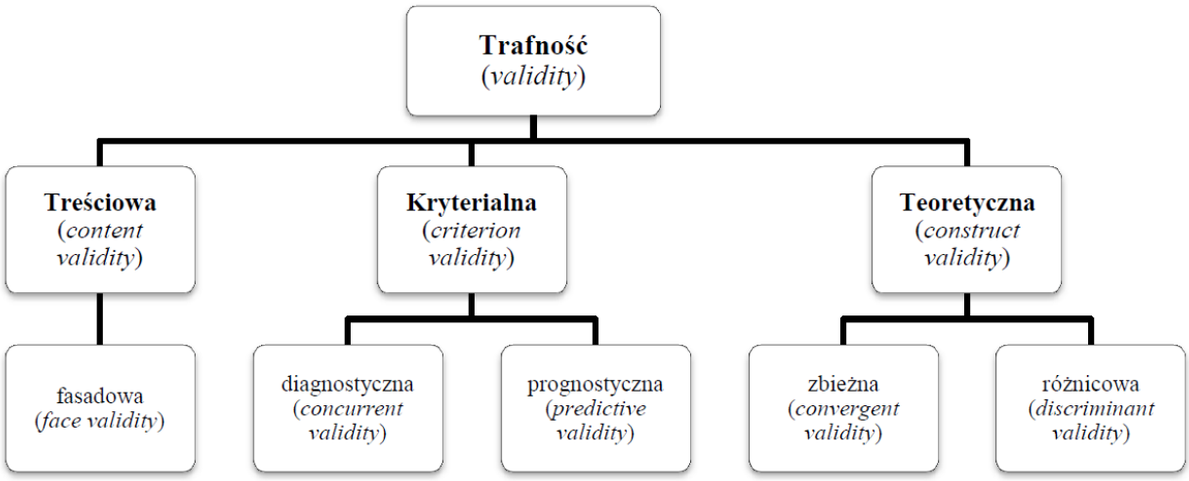
Na **rycinie 1** w graficzny sposób scharakteryzowano pojęcia trafności i rzetelności. Przyjęto założenie, że środek tarczy oznacza wartość prawdziwą, zaś kropki oznaczają wyniki kolejnych pomiarów tej wartości. Pomiary na tarczy A są teoretycznie trafne – średnia z pomiarów jest zgodna z wartością prawdziwą – ale cechują się brakiem powtarzalności (mała rzetelność). Na tarczy B pomiary są powtarzalne (rzetelne), ale nie mierzą rzeczywistej wartości (mała trafność). Dopiero wyniki pomiarów na tarczy C są powtarzalne i mierzą wartość prawdziwą (są rzetelne i trafne). W ocenie rzetelności badamy różnice pomiędzy pomiarami tej samej cechy. Sytuacją pożądaną jest, aby te różnice dążyły do zera. W przypadku badania trafności istotna jest zbieżność wyników poszczególnych pomiarów oraz wartości prawdziwej. O ile w ramach oceny rzetelności różnica między pomiarami jest stosunkowo łatwa do określenia, o tyle w ramach oceny trafności nie jest znana wartość prawdziwa, ponieważ jakość życia jest cechą ukrytą. Możemy ją jednak oszacować na podstawie wyników z obserwacji klinicznej lub przyjąć zgodnie z wartością określoną jako referencja.

Celem niniejszej pracy jest usystematyzowanie wiedzy o właściwościach psychometrycznych kwestionariuszy oceny jakości życia. Oprócz wspomnianych wyżej pojęć związanych z trafnością oraz rzetelnością omówione zostaną również pojęcia czułości oraz czułości na zmiany kliniczne (responsywność).

Trafność

Najbardziej złożone z opisywanych pojęć to pojęcie trafności. Najczęściej wyróżnia się trzy główne aspekty trafności: treściową, kryterialną oraz teoretyczną (**Rycina 2**) [1, 2]. Mówiąc o trafności treściowej (ang. *content validity*) instrumentu, zależy nam na tym, aby narzędzie w pełni pokrywało zamierzony obszar zainteresowania, tzn. uwzględniało wszystkie możliwe do wystąpienia w danej chorobie stany zdrowia i związaną z nimi jakość

Rycina 1. Pojęcia trafności i rzetelności przedstawione graficznie w postaci tarczy z wynikami



Rycina 2. Zestawienie pojęć związanych z trafnością (*validity*)

życia. Trafność tę określa się, badając proces tworzenia kwestionariusza, przeglądając procedury stosowane przy jego opracowywaniu oraz analizując jego strukturę. Przykładem oceny trafności treściowej może być opis budowy kwestionariusza Kidscreen.

Związana z trafnością treściową jest trafność fasadowa (ang. *face validity*). Określa ona subiektywny sposób postrzegania i rozumienia kwestionariusza przez respondentów i jest oceniana po zaprojektowaniu testu [2]. Jest to jednak koncepcja postrzegana przez naukowców za nazbyt subiektywną i rozmytą, przez co jest rzadziej wykorzystywana i raportowana w badaniach walidacyjnych [3].

Trafność kryterialna (ang. *criterion validity*) to analiza zgodności pomiarów otrzymanych za pomocą testu z pomiarami otrzymanymi w wyniku oceny zmiennej pozatestowej (kryterium), o udowodnionej relacji z pierwotnie mierzoną wartością wyjściową, tj. z jakością życia [2]. Na ogół przybiera postać badania zgodności między otrzymanym wynikiem a wynikiem innego, już zwalidowanego kwestionariusza, stanowiącego „złoty standard" postępowania [3]. W ramach trafności kryterialnej ocenia się trafność diagnostyczną (ang. *concurrent validity*) oraz prognostyczną (ang. *predictive validity*). Trafność diagnostyczna opiera się na zgodności wyników ze stanem faktycznym, z kryterium aktualnie występującym u badanej jednostki. Trafność prognostyczna z kolei określa zdolność kwestionariusza do przewidywania przyszłych stanów zdrowia czy zdarzeń. Określa zgodność z kryterium, według którego można przewidzieć możliwą do wystąpienia zmianę stanu zdrowia, np. kiedy na podstawie jakości życia może być dokonywana predykcja czasu przeżycia chorych na nowotwory [2].

Ostatnia i najważniejsza odmiana trafności to trafność teoretyczna (ang. *construct validity*). Określa ona stopień zależności pomiędzy wynikami testu a teorią zbudowaną na podstawie modelowania struktury i zależności badanego zjawiska, np. stawiamy hipotezę, że jakość życia będzie się zmieniać w określony sposób w określonych grupach/stanach zdrowia, następnie badamy korelację wyników kwestionariusza w poszczególnych, analizowanych grupach i określamy stopień ich zgodności [2]. Do oszacowania trafności teoretycznej niezbędne jest przeprowadzenie badań i sprawdzenie ich rezultatów z założeniami teoretycznymi. W pierwszym kroku budujemy model zjawiska, względem którego będą porównywane wyniki testów. Następnie stawiamy hipotezy co do kierunku i siły zależności wyników oraz modelu. Ostatecznie weryfikujemy postawione uprzednio hipotezy. W praktyce nie można jednoznacznie wskazać, czy pozycje kwestionariusza są wiarygodne, a jedynie można zebrać wystarczającą ilość dowodów na potwierdzenie ich zależności z modelem [2]. Jest to najważniejszy rodzaj trafności.

W ramach trafności teoretycznej określa się trafność zbieżną (ang. *convergent validity*), kiedy różne miary mające w zamierzeniu mierzyć tę samą cechę korelują między sobą, oraz trafność różnicową (ang. *discriminant validity*), kiedy miara nie powinna korelować zbyt silnie z miarami, które w zamierzeniu powinny mierzyć inne cechy [2]. Należy zwrócić uwagę na fakt, że często wiele domen może być ze sobą zależnych i należy brać pod uwagę przede wszystkim wysoki lub niski stopień zależności.

Jednym ze sposobów realizacji sprawdzenia trafności teoretycznej jest metoda znana jako walidacja znanych grup (ang. *known groups validation*), która polega na ba-

daniu zależności wyników testu w grupach o znanych właściwościach i badaniu różnic pomiędzy grupami o znanych różnicach [2]. Na przykład: można oczekiwać, że pacjenci z zaawansowanym stadium nowotworu będą mieli gorszą jakość życia niż pacjenci w jego wczesnym stadium. Wówczas wykazanie tej zależności pomiędzy wynikami jakości życia a zdiagnozowanym stanem zdrowia potwierdzi trafność teoretyczną testu. Metoda *known groups validation* jest kombinacją testu trafności oraz czułości.

Narzędziem, dzięki któremu można zobrazować wyniki zarówno dotyczące oceny trafności zbieżnej, różnicowej, jak i rzetelności, jest tzw. macierz wielu cech–wielu metod (ang. *multitrait–multimethod analysis*) [2, 7]. W **tabeli 1** zestawiono porównanie współczynników korelacji dwóch testów (1, 2), każdy z trzema domenami (fizyczna, emocjonalna, społeczna). Komórki C określają wartości współczynników korelacji opisujących trafność zbieżną, komórki D – trafność różnicową, a R – rzetelność [2].

Rzetelność

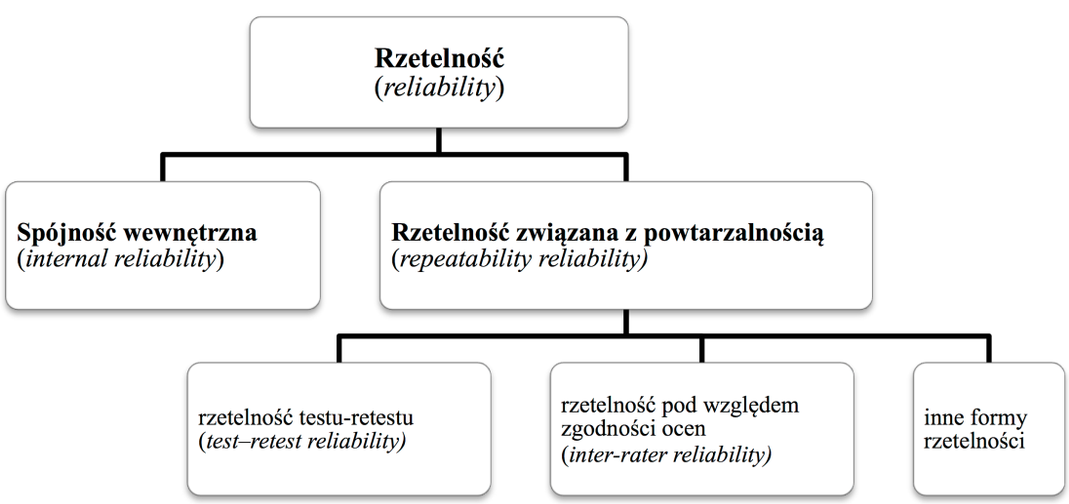
Drugą z istotnych właściwości psychometrycznych jest rzetelność, która określa cechę samego narzędzia. W świetle klasycznej teorii testów rzetelność ocenia wpływ losowej rozbieżności w wynikach pomiarów, które z założenia powinny być jednakowe, oraz określa zakres, w którym kwestionariusz będzie dawał spójne i powtarzalne wyniki [6].

Rozróżnia się rzetelność rozumianą jako spójność wewnętrzną oraz rzetelność związaną z powtarzalnością (**Rycina 3**). Spójność wewnętrzna (ang. *internal reliability*) ocenia, czy pozycje kwestionariusza, które w zamierzeniu mierzą tę samą cechę, dają spójne wyniki. Określana jest za pomocą korelacji pozycji kwestionariusza między sobą [2].

Rzetelność związana z powtarzalnością określa, czy wyniki pomiarów oceniających w założeniu niezmieniony stan zdrowia pozostaną spójne [2, 3]. Ten niezmieniony stan zdrowia może być interpretowany jako stan niezmie-

Tabela 1. Macierz wielu cech–wielu metod zestawia w ramach jednej macierzy wyniki związane z oceną trafności oraz rzetelności (zaadaptowane z [2])

	Instrument	Funkcjonowanie fizyczne		Funkcjonowanie emocjonalne		Funkcjonowanie społeczne	
		1	2	1	2	1	2
Funkcjonowanie fizyczne	1	R					
	2	C	R				
Funkcjonowanie emocjonalne	1	D		R			
	2		D	C	R		
Funkcjonowanie społeczne	1	D		D		R	
	2		D		D	C	R



Rycina 3. Zestawienie pojęć związanych z rzetelnością (*reliability*)

niony w czasie, wówczas mamy do czynienia z oceną rzetelności testu–retestu (ang. *test–retest reliability*), a więc badamy powtarzalność wyników pomiarów oddalonych od siebie nieznacznie w czasie (w założeniu stałych); lub też niezmieniony stan rozumiany jako ocena stanu zdrowia u tego samego badanego wykonywana przez różnych badaczy – jest to wówczas rzetelność pod względem zgodności ocen (ang. *inter-rater reliability*). W przypadku rzetelności testu–retestu czas przerwy pomiędzy pomiarami powinien być na tyle krótki, żeby stan zdrowia nie uległ zmianie, ale też na tyle długi, by respondenci nie pamiętali swoich poprzednich odpowiedzi [3]. Rzetelność pod względem zgodności ocen służy do wykazania spójności między samodzielną oceną pacjenta a oceną wykonaną przez krewnych lub personel medyczny, dlatego też jest ważna w kontekście decyzji o zastosowaniu oceny proxy (ang. *proxy assessment*, tł. ocena przez przedstawiciela/pełnomocnika/respondenta zastępczego) [2].

Do pomiarów rzetelności wykorzystuje się miary współczynników stosowanych w statystyce w ramach testów zgodności. Do najczęściej stosowanych zalicza się współczynniki kappa oraz kappa ważone, a także współczynnik korelacji wewnątrz klasy (ang. *intraclass correlation coefficient* – ICC) [2].

Kappa to unormowany współczynnik obrazujący wpływ przypadkowej zgodności na wynik zgodności obserwowanej (**Tabela 2**). Kappa ważone z kolei uwzględnia wagi zgodności pomiędzy poszczególnymi domenami, a sposób obliczeń i interpretacji jest analogiczny [2]. Znajduje to zastosowanie w przypadku porównania kwestionariuszy o hierarchicznej strukturze domen, wówczas bezwzględna różnica pomiędzy odpowiedziami w obu pomiarach będzie miała wpływ na ostateczny wynik rzetelności. Wartość kappa zbliżona do 0 oznacza zgodność na poziomie zgodności przypadkowej, zaś bliska 1 charakteryzuje test o dużej rzetelności.

Współczynnik ICC stosowany jest w przypadku skali interwałowej lub ilorazowej i określa stosunek wariancji prawdziwej do wariancji otrzymanej, zgodnie z klasyczną teorią testów [2]. Jeśli ICC jest bliskie 1, wówczas losowa rozbieżność w wynikach jest mała, a rzetelność testu duża.

Czułość/responsywność

Oprócz terminów związanych z trafnością i rzetelnością w ramach oceny wiarygodności kwestionariusza określa się także jego czułość oraz czułość na zmiany istotne kliniczne (responsywność). Czułość testu to jego zdolność do pomiaru różnic w jakości życia między pacjentami lub grupami pacjentów. Test powinien potrafić wskazać na różnice w grupach przy stosunkowo umiarkowanej (możliwej do przeprowadzenia badania) liczebności badanej populacji [2]. Im bardziej czuły test, tym mniejsza próba potrzebna do przeprowadzenia poprawnego w sensie statystycznym badania. W odróżnieniu od *known-group validation*, w którym oczekujemy spodziewanej różnicy, w ocenie czułości chcemy wykazać, że liczebność badanej próby jest wystarczająca do wykazania istotnej statystycznie i klinicznie różnicy, jaka może wystąpić między grupami [2]. Czułość na zmiany (responsywność) to zdolność testu do wykrywania istotnych klinicznie zmian jakości życia u tego samego pacjenta, co odróżnia ją od klasycznego pojęcia czułości. Wykazano, że instrumenty specyficzne chorobowo są z reguły bardziej czułe na zmiany niż testy generyczne (ogólne) [2].

Wyróżnia się kilka parametrów służących do oceny czułości na zmiany. *Effect size* (ES) to współczynnik średniej zmiany wyników porównywanych kwestionariuszy do wartości odchylenia standardowego (SD) początkowego pomiaru. *Standardised response mean* (SRM) jest podobnym parametrem, ale uwzględnia w mianowniku odchylenie standardowe zmiany, a nie pojedynczego pomiaru. Jest to miara częściej stosowana niż *effect size* i rekomendowana w ocenie rzetelności. Bardzo podobny parametr do SRM, ale stosujący zamiast odchylenia standardowego błąd standardowy dla średniej, to statystyka sparowanego testu T (ang. *paired T-statistic*). Efektywność względna/względna wydajność to współczynnik kwadratów wyników testów T (lub F dla porównań więcej niż dwóch skali/domen). Przy porównaniu większej liczby skali/domen

Tabela 2. Wytyczne wskazujące na poziom zgodności wyników w zależności od wartości parametru kappa (κ)

κ	Zgodność
< 0,20	brak zgodności
0,21–0,40	bardzo słaba
0,41–0,60	średnia
0,61–0,80	dobra
0,81–1,00	bardzo dobra

Tabela 3. Zestawienie najczęściej stosowanych parametrów służących ocenie rzetelności kwestionariusza [2]

Nazwa parametru	Wzór
wielkość efektu (ang. *effect size* – ES)	$\frac{(\bar{x}_2 - \bar{x}_1)}{SD(x_1)}$
Standaryzowana średnia odpowiedź (ang. *standardised response mean* – SRM)	$\frac{(\bar{x}_2 - \bar{x}_1)}{SD(x_2 - x_1)}$
Statystyka sparowanego testu T (ang. *paired T-statistic*)	$\frac{(\bar{x}_2 - \bar{x}_1)}{SE(x_2 - x_1)}$
Efektywność względna/względna wydajność (ang. *relative efficiency/relative validity*)	współczynnik kwadratów statystyk T (lub F) dwóch skali

SD – odchylenie standardowe, SE – błąd standardowy

przyjmuje się w mianowniku najmniejszą wartość testu-T lub F [2]. Opisane odpowiednimi wzorami parametry zestawiono w **tabeli 3**.

Nie zawsze zdarza się, że otrzymane wyniki kwestionariuszy będą rozkładać się zgodnie z rozkładem normalnym, a zatem nie będzie możliwe wyznaczenie średnich różnic w pomiarach i odchyleń standardowych. Kiedy znaczna część wyników odpowiada maksymalnej/minimalnej wartości skali, mamy do czynienia z efektem zwanym efektem sufitu/efektem górnego pułapu (ang. *ceiling effect*) lub efektem podłogi/efektem dolnego pułapu (ang. *floor effect*). Występowanie któregoś z tych zjawisk wiąże się na ogół z niską czułością kwestionariusza [2].

Podsumowanie

Istotne jest, aby przed zastosowaniem kwestionariusza oceny jakości życia w badaniu bądź praktyce klinicznej przeprowadzić jego walidację. Ocena właściwości psychometrycznych kwestionariuszy jakości życia pozwala określić ich wiarygodność w warunkach konkretnego kontekstu klinicznego. Ważna jest ocena wszystkich jej aspektów, gdyż tylko wtedy można mówić o pozytywnym wyniku walidacji. Kwestionariusz powinien cechować się odpowiednią czułością oraz rzetelnością, jednak są to jedynie warunki konieczne, ale nie wystarczające, by mówić o jego pełnej wiarygodności. Walidacja testu w populacji docelowej, w konkretnej sytuacji klinicznej umożliwia oszacowanie jego trafności, a w dalszych krokach pozwala wnioskować o przydatności jego stosowania w badaniach bądź praktyce klinicznej.

Oświadczenia

Oświadczenie dotyczące konfliktu interesów
Autorzy deklarują brak konfliktu interesów.

Źródła finansowania
Autorzy deklarują brak źródeł finansowania.

Piśmiennictwo

1. Brzyski P. Trafność i rzetelność skal pomiarowych. W: Kawecka-Jaszcz K, Klocek M, Tobiasz-Adamczyk B (red.). Jakość życia w chorobach układu sercowo-naczyniowego. Metody pomiaru i znaczenie kliniczne. Poznań:Termedia; 2006. 43–54.
2. Fayers PM, Machin D. Quality of Life. The assessment, analysis and interpretation of patient-reported outcomes. Second edition. Chichestrer: John Wiley & Sons, Ltd.; 2007. 77–108.
3. Lis J, Jahnz-Różyk K, Hermanowski T, Czech M. Koszty, jakość i wyniki w ochronie zdrowia. Leksykon podstawowych pojęć ISPOR. Warszawa: Polskie Towarzystwo Farmakoekonomiczne; 2009. Pojęcia: Jakość życia związana ze zdrowiem, Stan zdrowia, Rzetelność, Zdolność do uchwycenia zmian klinicznych, Trafność.
4. KIDSCREEN Opis budowania struktury kwestionariusza, http://www.kidscreen.org/english/measurement-model/ [data dostępu: 12.09.2016].
5. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52:281–302.
6. Brzeziński J. Metodologia badań psychologicznych. Warszawa: PWN; 2002.
7. Brzeziński J. Trafność i rzetelność testów psychologicznych. Wybór tekstów. Gdańsk: GWP; 2005.
8. Golicki D, Farkowski MF, Grabowska-Krawiec P, Szkultecka-Dębek M, Zawodnik S, Borowiec Ł, Czech M. Słownik jakości życia, angielsko-polski, polsko-angielski. Warszawa: Farmapress; 2014.
9. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420–428.
10. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33:613–619.

11. Norman GR, Stratford PW, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50:869–879.

Zaakceptowano do edycji: 2016–04–12
Zaakceptowano do publikacji: 2016–04–28

Adres do korespondencji:
Katarzyna Młyńczak
HealthQuest sp. z o.o., sp. k.
ul. Adama Mickiewicza 63, budynek Megadex A
01-625 Warszawa
tel. kom.: 698 414 630
e-mail: katarzyna.mlynczak@healthquest.pl