Biostatystyka w badaniach medycznych i praktyce klinicznej

Agata Smoleń

Katedra i Zakład Epidemiologii i Metodologii Badań Klinicznych, Uniwersytet Medyczny w Lublinie

Wprowadzenie Biostatystyka odgrywa bardzo istotną rolę na każdym etapie naukowego badania biomedycznego. Najważniejszym etapem jest planowanie i projektowanie badania. Podstawą jest sformułowanie precyzyjnego i uzasadnionego pytania klinicznego, na które chcemy uzyskać jednoznaczną odpowiedź. Bardzo często nie mamy już możliwości poprawienia błędów popełnionych na tym etapie. Nawet najbardziej skomplikowane analizy statystyczne nie pozwolą na ich wyeliminowanie i prowadzą najczęściej do nieprawidłowych wniosków.

Natomiast prawidłowa interpretacja i rozumienie publikowanych w czasopismach naukowych wyników badań i odnoszenie się do nich w praktyce klinicznej nie jest w pełni możliwe bez znajomości podstaw metodologii i prawidłowego rozumienia koncepcji statystycznych.

Nie zawsze badacze rozpoczynający pracę naukową, a nawet lekarze uczestniczący w konferencjach naukowych i dyskusjach dotyczących nowych procedur mają świadomość istotności i konieczności stosowania właściwych metod statystycznych w badaniach medycznych. Należy podkreślić, że opracowywanie wyników badań naukowych, opartych na analizach statystycznych jest standardem postępowania w publikacjach naukowych. Każdy etap planowanego lub prowadzonego badania w dziedzinie biomedycyny wymaga konieczności opracowania zarówno klinicznego, jak i statystycznego.

Wiadomo, że badanie naukowe powinno odpowiadać na istotne i interesujące pytania i pozwala uzyskać precyzyjną odpowiedź. Z tego względu bardzo ważne jest dokładne zdefiniowanie problemu badawczego oraz sformułowanie celów i hipotez badawczych. Według International Conference of Harmonization „badanie potwierdzające jest badaniem, w którym postawiona na początku hipoteza jest następnie weryfikowana”.[1] Schemat postępowania w planowaniu badania nie jest więc dowolny (TABELA 1).

Badania w zakresie nauk biomedycznych dzieli się na **podstawowe** – przeprowadzane najczęściej na podstawie modelu badawczego oraz **populacyjne**. Z kolei badania populacyjne dzieli się na retrospektywne (kliniczno-kontrolne) lub prospektywne (kohortowe), które mają charakter obserwacyjny. Ponadto wyróżnić należy badania kliniczne (eksperymentalne), w których często bardzo ważnym zagadnieniem jest uwzględnienie grupy kontrolnej. Celem użycia grupy kontrolnej jest oszacowanie, co może się zdarzyć, jeśli nie zastosuje się np. leczenia. Prawdziwy efekt leczniczy określa się na podstawie porównywania wyników grupy leczonej i grupy kontrolnej. Należy tutaj pamiętać o losowym doborze grup, czyli randomizacji, bardzo istotnej w badaniach klinicznych. Randomizacja powinna być stosowana wszędzie tam, gdzie jest to możliwe w badaniu klinicznym. Badania eksperymentalne są na ogół prospektywne, gdyż analizują wpływ interwencji na wynik, który pojawi się w przyszłości.

Badania obserwacyjne mogą dostarczyć mniej informacji niż badania eksperymentalne, ponieważ często nie mamy możliwości kontrolowania wszystkich czynników zakłócających analizowane wyniki. Jednakże, w pewnych sytuacjach, mogą one być jedynym możliwym sposobem uzyskania potrzebnej informacji. Przykładem takich badań są badania epidemiologiczne, określające związek między czynnikami ryzyka a występowaniem choroby w populacji.

Wyniki doświadczeń naukowych zbieramy zawsze i opracowujemy z myślą o udowodnieniu postawionej wcześniej hipotezy. Wymaga to określenia pytania badawczego i populacji (próby) badanej, kryteriów doboru jednostek badanych (pacjentów), zdefiniowania głównych wielkości, które poddane zostaną pomiarom (parametry). Bardzo ważnym zagadnieniem jest odpowiedni dobór próby do badania, sposób pozyskiwania danych oraz czas trwania badania. Na tym etapie statystyka odgrywa ważną rolę zarówno w wyborze istotnych parametrów, poddawanych pomiarom, jak

Adres do korespondencji:
dr hab. n. med. Agata Smoleń, Katedra i Zakład Epidemiologii i Metodologii Badań Klinicznych, Uniwersytet Medyczny w Lublinie, ul. Chodźki 1, 20-093 Lublin, tel.: 81 448 63 70, e-mail: absmolen@wp.pl
Praca wpłynęła: 15.02.2016.
Przyjęta do druku: 15.02.2016.
Publikacja *online*: 14.04.2016.
Nie zgłoszono sprzeczności interesów.
Pol Arch Med Wewn. 2016; 126 (Special Issue 1): 1-24
doi: 10.20452/pamw.3370

TABELA 1 Schemat badania naukowego

1. Sformułuj pytanie kliniczne (problem badawczy, cel).
2. Określ stawianą hipotezę badawczą.
3. Określ, jakie parametry należy badać i w jakich skalach pomiarowych.
4. Dobierz odpowiednio próbę badaną i określ czas trwania badania.
5. Zbierz dane.
6. Zastosuj odpowiednie testy statystyczne.
7. Zinterpretuj wyniki i wyciągnij wnioski.

i określeniu skali ich pomiaru, w wyznaczaniu wymaganej minimalnej wielkości próby, czy w planowaniu rodzaju analiz koniecznych do przeprowadzenia w opracowywaniu uzyskanych danych. Pozwala to uzyskać odpowiedź na stawiane pytania badawcze oraz na wyciągnięcie istotnych klinicznie i wiarygodnych wniosków, zgodnych z głównym celem badania.[2-4]

Podsumowując, należy przytoczyć słowa Altmana: „... Mówiąc krótko, jest rzeczą nieetyczną przeprowadzanie błędnego doświadczenia naukowego. Jednym z aspektów błędnych doświadczeń są niewłaściwie stosowane metody statystyczne [...]. Statystyka wpływa na etykę w sposób znacznie bardziej specyficzny: jest rzeczą nieetyczną publikowanie niepoprawnych lub błędnych rezultatów".[2,5]

Populacja i próba badana Przedmiotem analiz statystycznych są grupy złożone z badanych jednostek, np. osób, obiektów lub zjawisk, które określamy jednostkami statystycznymi. Analiza statystyczna danych dotyczy metod opisu i analizy występujących prawidłowości. Jest to możliwe tylko wtedy, kiedy badaniu poddana jest duża liczba jednostek, a to pozwala zaobserwować występujące prawidłowości. Nie jest to więc możliwe na podstawie pojedynczej jednostki lub niewielkiego zbioru jednostek badanych.

Stosowanie metod statystycznych w badaniach biomedycznych umożliwia więc obiektywne wnioskowanie na podstawie wyników serii badań przeprowadzanych w badanej populacji. W medycynie populacja to grupa badanych, interesująca nas pod pewnym względem. Ustalenie co stanowi populację zależy od celu badania. Czasami możemy przebadać całą populację, np. spis powszechny. Najczęściej jednak w medycynie badanie całej populacji nie jest możliwe lub jest zbyt kosztowne czy czasochłonne. W takiej sytuacji w praktyce stosuje się badania częściowe na podstawie podzbioru pobranego z tej populacji, określanego jako próba statystyczna. Aby móc wnioskować o całej populacji na podstawie wyników badania próby, musi ona być reprezentatywna, czyli odpowiednio liczna i dobrana w sposób losowy. Dobór losowy polega na całkowicie przypadkowym włączeniu do próby jednostek badanych. Oznacza to, że każdy element populacji musi mieć jednakowe szanse wejścia do próby. W przypadku próby losowej, z określonym prawdopodobieństwem, istnieje jej podobieństwo do populacji, z której została wylosowana. Dlatego zbieramy dane na temat grupy przypadków, która musi być reprezentatywna dla populacji. W takiej sytuacji ma podobną charakterystykę do jednostek w populacji i dlatego wnioski prawdziwe dla próby są również prawdziwe dla całej populacji z dużym prawdopodobieństwem. Aby w doborze próby uniknąć manipulacji, stosuje się metody oparte na rachunku prawdopodobieństwa, np. tablice liczb losowych bądź zwykle losowanie.

Zasadniczą kwestią we wszystkich zastosowaniach statystyki jest więc odpowiedni dobór badanej próby.

Należy zwrócić uwagę, że czasami postępujemy w jakiś sposób celowy, czyli nielosowy. Dobór próby nielosowej odbywa się według ustalonych kryteriów. Jeśli znamy rozkład interesujących nas parametrów w populacji, możemy tak wybrać z niej taką próbę, aby jej struktura ze względu na te parametry dokładnie odzwierciedlała ich rozkłady w populacji.[2,6-8]

Rodzaje zmiennych i skal pomiarowych Każdy przypadek (jednostka badana) wchodzący w skład badanej próby czy populacji jest scharakteryzowany za pomocą różnych właściwości (zwanych cechami, zmiennymi czy parametrami). Parametrem w populacji opisującym badane przypadki jest cecha, która nie jest jednakowa u wszystkich badanych. Różnią się one pod względem informacji wynikającej ze skali pomiarowej. Ustalając sposób pomiaru cechy, ustalamy jej reprezentację w postaci zmiennej.

Najbardziej uproszczony podział mający znaczenie praktyczne pozwala wyróżnić **zmienne mierzalne (ilościowe) oraz niemierzalne (jakościowe)**. Niemierzalne, to takie które wyrażamy pewnymi kategoriami (np. nazwy, symbole, kody...). Natomiast mierzalne, to takie które można scharakteryzować wartością liczbową, często w danych jednostkach miary. Mogą one być **ciągłe**, czyli przyjmują każdą wartość z określonego przedziału (tzn. mierzone z dowolną dokładnością, np. wiek można zmierzyć w latach, miesiącach, dniach...; masę ciała – w kg, g, mg...). Ponadto wyróżniamy cechy skokowe – nieciągłe, które przyjmują tylko określone wartości z danego przedziału (można je uporządkować oraz określić różnicę między nimi [np. stopnie w indeksie 2; 3; 3,5; 4; 4,5; 5, liczba łóżek w szpitalu – przyjmuje tylko określone wartości 0, 1, 2, ale nie 1,5]).

Z kolei ze względu na skalę pomiaru parametry niemierzalne dzielą się na mierzone w skali nominalnej i porządkowej. Najsłabszą skalą pomiarową jest **skala nominalna** (pomiary dostarczają wyłącznie informacji do jakiej grupy kategorii należy badany obiekt i nie pozwalają na ich uporządkowanie, np. płeć, rasa). Cechy takie nazywa się również grupującymi, kategoryzującymi. Jeśli dana cecha ma tylko 2 warianty, nazywamy ją dychotomiczną, np. płeć.

Kolejną skalą pomiarową jest **skala porządkowa** (tzn. można uporządkować badaną zmienną pod względem jej natężenia, np. wykształcenie). Często wyniki pomiaru w skali porządkowej są

TABELA 2 Typy danych

Parametry			
niemierzalne (jakościowe)		mierzalne (ilościowe)	
nominalne	porządkowe	skokowe	ciągłe
kategorie różne, nie można uporządkować	kategorie różne, można uporządkować	przyjmują tylko określone wartości z danego przedziału, można je uporządkować oraz określić różnicę między nimi	przyjmują każdą wartość z określonego przedziału, można je uporządkować i wykonywać działania arytmetyczne
przykład: płeć, kolor oczu	przykład: wykształcenie, stopień zaawansowania choroby	przykład: stopnie w indeksie, liczba łóżek w szpitalu	przykład: wiek, wzrost

wyrażane symbolicznie liczbami, co staje się źródłem błędów interpretacyjnych. Parametry niemierzalne można zakodować, jednak należy pamiętać, że dana wartość liczbowa określa tylko kod, a nie jest wartością mierzalną i nie należy stosować tu operacji dostępnych tylko dla danych mierzalnych. Natomiast ze względu na skalę pomiaru parametry mierzalne dzielą się na mierzone w **skali ilorazowej lub interwałowej** (przedziałowej). W skali interwałowej punkt zero wybrany jest umownie, np. temperatura w stopniach Celsjusza, nie ma tu jednak sensu dzielenie dwóch wartości przez siebie.

Rozróżnienie typów danych jest ważne ze względu na dopuszczalne działania arytmetyczne tylko dla danych ilościowych. Ze skali mocniejszej można przejść do słabszej, ale nie na odwrót. Analizując, np. wiek można porównywać grupy, wykonywać dla tego parametru działania arytmetyczne (dodawanie, odejmowanie, mnożenie, dzielenie), można te dane uporządkować w kolejności rosnącej lub malejącej czy podzielić na kategorie (młodsi, starsi – wtedy informacja jest mniej dokładna). Mając natomiast kategorie wieku, można przydzielić badanego do danej kategorii, nie można jednak dokładnie określić wieku. Mocniejsze skale pomiaru dotyczą więc parametrów mierzalnych.

Rozróżnienie rodzaju danych i stosowanych skal pomiarowych jest istotne, gdyż każda wymaga stosowania innych metod statystycznych.[2,7,8] Podsumowanie typów danych przedstawia TABELA 2.

Punktem wyjścia do obliczeń statystycznych jest zawsze zebranie danych tzw. surowych (oryginalnych wyników badania dla poszczególnych osób) i odpowiednie przygotowanie ich do analizy. Po sprawdzeniu, że wszystkie dane i zmienne zdefiniowano jednakowo oraz sprawdzono poprawność wprowadzonych wartości (ten etap jest szczególnie ważny, gdy kilka osób jest odpowiedzialnych za zbieranie i wpisywanie danych) można rozpocząć analizę statystyczną.

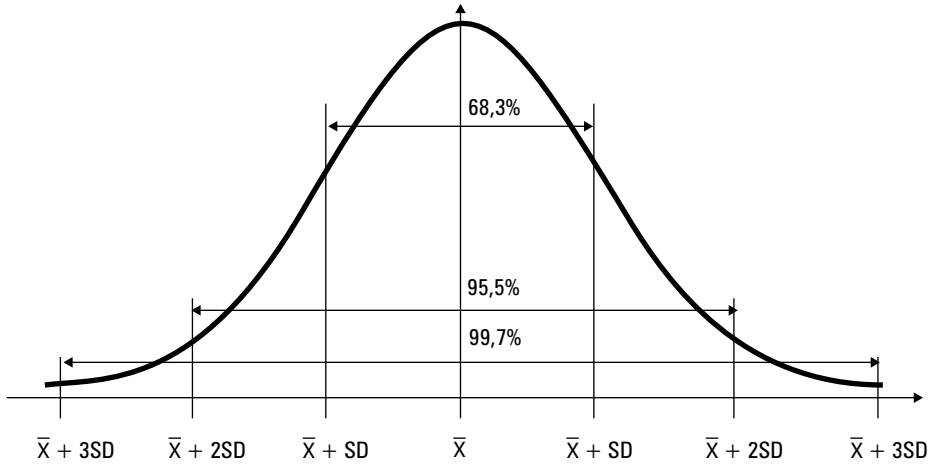
Rozkład empiryczny i teoretyczny Dla każdego analizowanego parametru na wstępie określamy empiryczny (obserwowany w badanej próbie) rozkład zmiennej. Istnieją również rozkłady teoretyczne (np. dwumianowy, Poissona, normalny...) opisane przez ściśle określone wzory matematyczne. Rozkłady teoretyczne stanowią model badanych przez nas zjawisk. Pełną informację o zmiennej losowej zawiera rozkład prawdopodobieństwa. Określa on prawdopodobieństwo rozkładu poszczególnych wartości lub przedziałów zmienności zmiennej losowej. Rozkład prawdopodobieństwa można przedstawić w formie tabelarycznej, graficznej lub analitycznej.

Bardzo ważna jest zgodność rozkładu wielu cech w rzeczywistych populacjach z rozkładami teoretycznymi, których właściwości są dobrze znane. Jednym z najistotniejszych rozkładów w statystyce jest **rozkład normalny**, charakteryzujący się dzwonowatym kształtem krzywej symetrycznej (krzywej Gaussa) względem wartości średniej, który reprezentuje wiele obserwowanych rozkładów. Ponadto należy podkreślić, że dokładny rozkład normalny jest osiągalny dopiero dla nieskończonej liczby przypadków. Wynika z tego, że w miarę wzrostu liczby analizowanych przypadków ich rozkład jest coraz bardziej zbliżony do rozkładu normalnego. Oparte jest to na bardzo ważnym w statystyce centralnym twierdzeniu granicznym. Mówi ono również, że średnia wszystkich możliwych średnich prób pobranych z populacji ogólnej jest równa średniej tej populacji, a odchylenie standardowe wszystkich możliwych średnich prób pobranych z populacji ogólnej jest równe odchyleniu standardowemu tej populacji podzielonemu przez pierwiastek z liczebności próby.

W praktyce najczęściej badana próba ma ograniczoną liczbę przypadków, ale należy pamiętać, że im będzie liczniejsza, tym lepiej zostanie odwzorowany rzeczywisty rozkład badanego parametru w populacji, z której została pobrana próba. Rozkład normalny jest ważny nie tylko dlatego, że jest dobrym doświadczalnym odwzorowaniem wielu zmiennych spotykanych w przyrodzie, ale także dlatego, że spełnienie założenia normalności rozkładu zmiennych jest wymagane dla większości stosowanych testów statystycznych. Cechą charakterystyczną wynikającą z kształtu i symetryczności rozkładu normalnego jest to, że ok. 68,3% wszystkich przypadków zawiera się w przedziale: średnia ± odchylenie standardowe (±SD), a 95,5% przypadków zawiera się w przedziale średnia ± 2odchylenia standardowe, natomiast średnia ±3 odchylenia standardowe zawiera 99,7% przypadków (RYCINA 1).

Do oceny zgodności rozkładu badanego parametru z rozkładem normalnym stosuje się najczęściej testy normalności, tj. test zgodności χ^2, test Kołmogorowa i Smirnowa, Lillieforsa czy test Shapiro i Wilka. Ostatni z wymienionych testów jest

RYCINA 1 Krzywa Gaussa



stosowany najczęściej ze względu na jego większą moc w porównaniu z pozostałymi wymienionymi testami.[2-4,7,9-11]

Statystyka opisowa Ze względu na fakt, iż bezpośrednie wyniki pomiarów badanych parametrów są trudne do interpretacji, a w przypadku bardzo licznych prób interpretacja badanych parametrów nie jest możliwa, konieczny jest syntetyczny opis badanych parametrów. Uproszczonym sposobem charakteryzowania zmiennej losowej jest podanie jej wartości parametrów. Mierzą one różne właściwości rozkładów: położenie, zmienność i asymetrię.

W praktyce stosuje się charakterystyki opisowe umożliwiające analizę struktury badanych parametrów. Pozwala to odpowiedzieć na pytanie, jak często badana cecha występuje w populacji, jaki jest jej rozkład, jaka wartość parametru jest najbardziej reprezentatywna w analizowanej próbie itp.

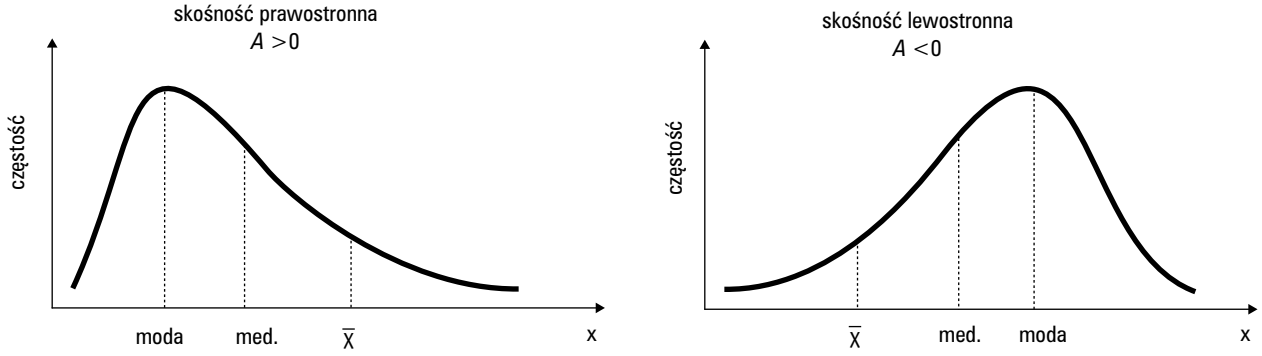
Dla parametrów niemierzalnych charakterystyka ich obejmuje liczność i najczęściej odsetek poszczególnych kategorii w badanej grupie przypadków. Natomiast dla parametrów mierzalnych niosących więcej informacji największe znaczenie mają miary położenia oraz miary rozproszenia.

Miary położenia oszacowują wartość charakteryzującą badany parametr, lecz każda z nich skupia się na innym aspekcie jego rozkładu.

Do najczęściej stosowanych miar położenia należą: **średnia arytmetyczna** (najczęściej określana jako średnia), **mediana** (wartość przeciętna) i **modalna** (wartość najczęściej występująca). Średnia arytmetyczna jest zarówno najczęściej używaną, jak i najczęściej nadużywaną miarą położenia. Wynika to z prostych obliczeń i powszechnego jej stosowania. Jej wadą jest fakt, że duży wpływ na jej wartość mają skrajne wyniki pomiarów odbiegające znacznie od pozostałych i dobrze opisuje tylko parametry o rozkładzie symetrycznym. Podstawowe założenie stosowania średniej arytmetycznej stanowi używanie ilorazowej lub interwałowej skali pomiaru dla badanych parametrów.

Mediana jest zbliżona do średniej, jeżeli dane mają rozkład symetryczny, jest mniejsza niż średnia, gdy dane mają rozkład prawoskośny, natomiast większa niż średnia przy rozkładzie lewoskośnym. Mediany nie można stosować dla parametrów mierzonych w skali nominalnej. Jest po uporządkowaniu wartości badanego parametru w kolejności rosnącej lub malejącej wartością środkową w przypadku nieparzystej liczby pomiarów lub średnią arytmetyczną dwóch wartości środkowych w przypadku parzystej liczby pomiarów. Nie mają więc na nią wpływu skrajne wartości badanego parametru, dlatego stosuje się ją jako miarę położenia dla silnie skośnych rozkładów. Inną miarą położenia jest wartość modalna (inaczej zwana modą lub dominantą). Może być wyznaczona w każdej skali pomiarowej. Czasami występuje więcej niż jedna wartość modalna, gdy dwie lub więcej wartości pojawiają się taką samą liczbę razy, a częstość występowania każdej z nich jest większa niż częstość występowania każdej innej wartości. Jeśli rozkład jest jednomodalny i względnie symetryczny, zbliżony do rozkładu normalnego, najlepszą charakterystyką jest średnia arytmetyczna. Jeśli rozkład jest jednomodalny, ale niesymetryczny – mediana. Natomiast jeśli rozkład jest wielomodalny – wykorzystujemy wartości modalne. Często pojedynczy parametr miary położenia nie wystarcza, gdy dysponujemy dużą liczbą pomiarów. Dokładniejszej charakterystyki można dokonać, stosując kwartyle (podział na 4 części), decyle (podział na 10 części) czy centyle (podział na 100 części). Kwartyl pierwszy (Q1) to wartość, poniżej której znajduje się 25% obserwacji, a 75% powyżej. Kwartyl drugi (Q2) jest równy medianie. Natomiast kwartyl trzeci (Q3) to wartość, poniżej której znajduje się 75% obserwacji, a 25% powyżej. Natomiast przy równie często stosowanym podziale, np. na 100 części – centyle, (percentyle) 25. centyl jest równy dolnemu kwartylowi, 50. centyl medianie, a 75. centyl górnemu kwartylowi.

Miary położenia nie wystarczą do charakterystyki badanego parametru, gdyż może się zdarzyć, że są one równe dla różnych zbiorów danych w różnych grupach, co może sugerować podobieństwo analizowanego parametru. Często prowadzi to do błędu, gdy nie ma dokładnych danych. Pomocne w ocenie podobieństwa są równie ważne miary charakteryzujące stopień rozproszenia wartości badanej zmiennej. Najprostszą miarą rozproszenia jest **rozstęp**,



RYCINA 2 Asymetria rozkładu

czyli różnica między maksymalną i minimalną wartością zmiennej, często podaje się te dwie wartości zamiast ich różnicy. Nie daje to jednak informacji o zróżnicowaniu wszystkich jednostek i podaje się w celu wstępnej informacji. Należy zwrócić uwagę, że rozstęp jest nieprawidłową miarą rozproszenia, gdy w danych znajdują się wartości odstające.

Kolejną bardzo istotną miarą jest **wariancja**. Jest ona znormalizowaną sumą kwadratów odległości pomiędzy każdym z pomiarów a wyznaczoną dla nich średnią arytmetyczną. Problemem w przypadku tej miary jest to, że mianem jej jest kwadrat jednostki w jakiej mierzona jest dana zmienna. Z tego powodu wprowadzono najpowszechniejszą miarę rozproszenia zwaną **odchyleniem standardowym**, aby uzyskać miano zgodne z mianem zmiennej, obliczane jako pierwiastek kwadratowy z wariancji. Jest zatem wyrażone w tych samych jednostkach, co analizowany parametr. Odchylenie standardowe określa o ile wszystkie jednostki badanej zmiennej różnią się przeciętnie od średniej arytmetycznej. Odchylenia standardowe stosuje się czasami do eliminacji grubych błędów. Stosuje się je, jeśli dane pochodzą z populacji o rozkładzie normalnym i różnią się od średniej o więcej niż 3 odchylenia standardowe.

Oprócz scharakteryzowanych miar rozproszenia często się używa również **rozstępu międzykwartylowego**. Obliczany jest jako różnica górnego i dolnego kwartyla. W przypadku zmiennej wyrażonej w skali porządkowej lub takiej, która nie podlega rozkładowi normalnemu miarą zmienności jest odchylenie ćwiartkowe, obliczane na podstawie średniej różnicy kwartyla górnego i dolnego. Wykorzystuje się je do konstrukcji typowego obszaru zmienności zmiennej w próbie, a w obszarze tym mieści się połowa wszystkich jednostek (Q1:Q3). Podobnie często się używa rozstępu zawierającego 95% centralnie położonych przypadków. Zakres ten nazywany jest zakresem normalnym lub zakresem odniesienia.

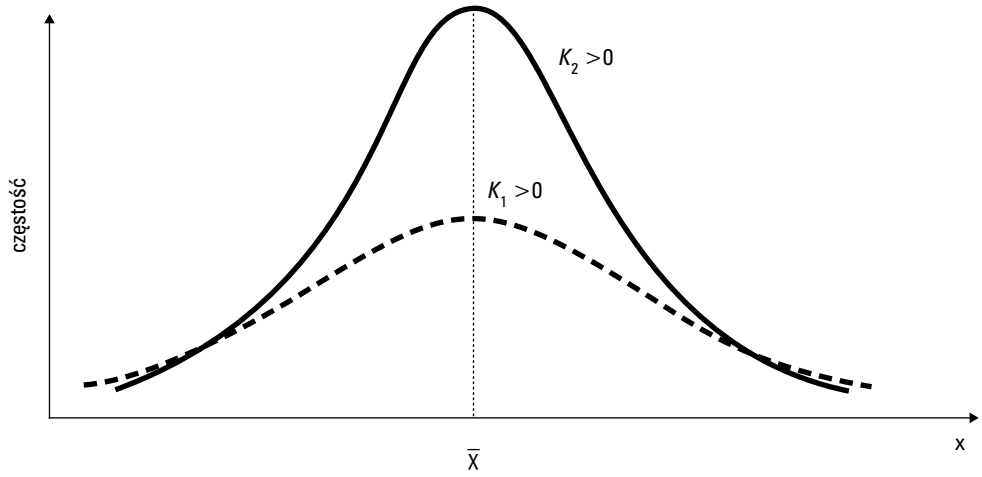
Należy pamiętać, że miary rozproszenia istnieją jedynie dla skali pomiarowej ilorazowej i interwałowej.

Inną miarą zmienności przydatną do porównania różnych zmiennych jest współczynnik zmienności. Jest to iloraz odchylenia standardowego i wartości średniej, często wyrażany w procentach. Pozwala on na porównanie rozproszenia parametrów mierzonych w różnych jednostkach miary.

Jeszcze inną miarą rozproszenia jest **błąd standardowy średniej**. Błąd standardowy średniej jest miarą precyzji wyznaczenia wartości średniej arytmetycznej w populacji ogólnej na podstawie określenia średniej dla badanej próby. Należy wytłumaczyć to w sposób następujący: jeżeli wybieramy kilka grup z danej populacji ogólnej i dla każdej z nich obliczymy wartości średnie, to błąd standardowy średniej jest to błąd szacowania średniej dla populacji ogólnej, tzn. określa rozrzut między kilkoma średnimi obliczanymi dla kilku grup, wybranych z danej populacji ogólnej. Gdy badane grupy są bardzo mało liczne, czyli nie są reprezentatywne dla całej populacji wówczas oszacowanie średniej z populacji ogólnej na podstawie średniej i odchylenia standardowego próby nie jest wskazane. Podawanie w takiej sytuacji błędu standardowego średniej ma większy sens statystyczny niż podawanie odchylenia standardowego.

Podsumowując najogólniej, jeżeli mamy określoną wartość średnią i jeżeli wiemy, jak rozproszone są wokół niej wartości, uzyskujemy już pewien obraz danych.

Innymi rzadziej stosowanymi miarami dotyczącymi kształtu rozkładu są miary skośności i koncentracji (kurtoza). **Skośność** charakteryzuje odchylenie rozkładu od symetrii. Jeżeli więc współczynnik asymetrii jest różny od 0 oznacza to asymetrię rozkładu (różnicę od rozkładu normalnego). Dla rozkładu lewoskośnego współczynnik asymetrii jest mniejszy od 0 ($A < 0$), natomiast dla prawoskośnego współczynnik asymetrii jest większy od 0 ($A > 0$). W przypadku rozkładu normalnego wartości średniej arytmetycznej, mediany i modalnej są równe ($\overline{X}$ =Med.=Moda). W przypadku rozkładów lewoskośnych średnia jest mniejsza od mediany, a mediana mniejsza od modalnej ($\overline{X}$ <Med. <Moda). W przypadku rozkładu prawoskośnego średnia jest większa od mediany, a mediana większa od modalnej ($\overline{X}$ >Med.>Moda) (RYCINA 2; źródło http://pqstat.pl/?mod_f=opisowa).



RYCINA 3 Koncentracja rozkładu (K_1 – rozkład platykurtyczny, K_2 – rozkład leptokurtyczny)

Natomiast kurtoza jest miarą spłaszczenia rozkładu. Dla rozkładu normalnego równa jest 0 (rozkład mezokurtyczny, K = 0). Wartości mniejsze od 0 oznaczają rozkład wysmukły (leptokurtyczny, K >0), a mniejsze od 0 rozkład spłaszczony (platykurtyczny, K <0) (RYCINA 3; źródło http://pqstat.pl/?mod_f=opisowa).

Wartości uzyskane w wyniku zastosowania statystyki opisowej są podstawą dla dalszych etapów wnioskowania statystycznego, takich jak testowanie różnic między grupami czy też badanie związków między wybranymi parametrami.[2-10]

Szacowanie liczności próby Szacowanie minimalnej wielkości próby jest głównym problemem i bardzo istotnym zagadnieniem w pracy naukowo-badawczej. Zależy od wielu czynników i opiera się często na skomplikowanych obliczeniach. Szacowanie optymalnej wielkości próby przeprowadzamy w odniesieniu do głównego celu badania i danego testu. Jeżeli natomiast nasze badanie zawiera większą liczbę testów, opieramy się na najważniejszym lub wyznaczamy wielkość próby wymaganą dla każdego z nich i wybieramy szacowaną największą liczność. Często popełniany błąd w badaniach to zbyt mała liczność próby. Wynika to z braku jednej uniwersalnej odpowiedzi na pytanie, ile należy zbadać przypadków, aby wnioski z badania były wiarygodne.

Często nie możemy stosować jednego schematu doboru próby do badania. Określenie liczności próby może być czasami niemożliwe. Zdarzyć się też może, że podział na grupy nie jest losowy. Należy pamiętać, że nielosowy czy niewłaściwy dobór próby może prowadzić do uzyskania wyników różnych od wyników uzyskiwanych dla całej populacji. W takiej sytuacji należy ocenić podobieństwo porównywanych grup pod względem parametrów istotnych dla wyników badania. Wyniki badania można uznać za wiarygodne, jeżeli nie stwierdza się różnic między analizowanymi grupami, tzn. są one jednorodne. Z kolei gdy liczność próby jest od nas niezależna, należy stosować odpowiednie testy diagnostyczne i odpowiednio duże przedziały ufności. Ponadto wyniki stosowanych testów statystycznych muszą być powiązane z logiczną analizą problemu.

Pomiary w naukach biomedycznych obciążone są pewnym nieuniknionym błędem, a z wpływem na liczność próby łączy się bardzo wiele zagadnień. Głównym celem badań empirycznych jest poznanie prawidłowości zachodzących w badanych sytuacjach. W każdej sytuacji należy uwzględnić wpływ tzw. przyczyn głównych, które powodują powstawanie prawidłowości, działając w sposób trwały i ukierunkowany. Natomiast zakłócenia w prawidłowościach powodują przyczyny przypadkowe, działając w sposób różny i nietrwały. W statystyce rozróżnienie tych dwóch rodzajów przyczyn możliwe jest dzięki prawu wielkich liczb. Wynika z niego, że wraz ze wzrostem liczby doświadczeń efekty oddziaływania przyczyn przypadkowych wzajemnie się znoszą, a uwidaczniają się efekty oddziaływania przyczyn głównych. Ponadto na liczność próby wpływa również fakt, że dane biomedyczne dotyczące najczęściej osób, charakteryzują się dużą zmiennością wewnątrz- i międzyosobniczą (np. reakcja na ten sam bodziec u tego samego pacjenta może być inna w innym czasie oraz reakcja dwóch różnych pacjentów na ten sam bodziec może być różna).

Ważne jest również to, czy badany parametr na tle losowych czynników zakłócających jego działanie uwidacznia się ewidentnie ze znaczną powtarzalnością, a czynniki przypadkowe nie występują lub nie są istotne, wtedy można opierać się na próbie o niezbyt dużej liczności. Gdy jednak badany parametr uwidacznia się w sposób bardzo subtelny i na wyniki obserwacji badanych pacjentów bardzo duży wpływ mają ich cechy osobnicze oraz inne czynniki losowe, wtedy konieczne jest oparcie się na znacznie liczniejszej próbie. W takiej sytuacji mała liczność próby powoduje to, że badanie może mieć za małą dokładność albo dawać wnioski ze zbyt dużym błędem. Natomiast wykonanie zbyt dużej liczby badań powoduje, że stają się one czasochłonne i wymagają większych nakładów finansowych, zazwyczaj przy minimalnym zysku statystycznym.

Podsumowując, należy podkreślić, że im większa liczność próby, tym mniejszy błąd. Natomiast próba powinna być wystarczająco liczna, aby otrzymać wynik o przyjętym poziomie dokładności. Należy pamiętać, aby dobierać najmniejszą

próbę, zapewniającą ustalony poziom wiarygodności wyników.

W prostych sytuacjach ważne jest zwrócenie uwagi, aby liczność próby nie była zbyt mała i dla porównywanych grup powinna wynosić co najmniej 50, grupy zaś o liczności mniejszej niż 30 powodują, że wnioskowanie nie jest wiarygodne. Dlatego też jako prostą wskazówkę można przyjąć jako **minimalną liczbę 30 elementów w próbie**. Czasami jednak nawet 50-elementowe grupy są zbyt małe. Powodem jest istotny wpływ rozproszenia wyników w małej próbie na wartości odpowiednich statystyk. Ponadto liczności porównywanych grup nie muszą być równe, ale duże różnice nie są wskazane. Jeśli natomiast porównujemy parametry niemierzalne z użyciem testu χ^2, to liczność oczekiwana w każdej komórce tabeli wielodzielczej powinna wynosić co najmniej 5. W przypadku szacowania liczności próby potrzebnej do określenia, jaką część populacji stanowią badani z wyróżnioną pewną cechą największą liczność próby należy przyjąć, gdy dana wartość cechy występuje w połowie populacji. Jeśli więc nic nie wiemy o populacji i nie możemy przeprowadzić badań wstępnych, pilotażowych, to możemy w odpowiednim wzorze przyjąć częstość występowania interesującej nas cechy równą 0,5 i wówczas uzyskamy oszacowanie liczności próby, które na pewno spełnia nasze wymagania.

Nie zawsze można oprzeć się na takich prostych założeniach i wymaga to bardziej precyzyjnego szacowania. Wyznaczenie liczności próby wymaga znajomości parametrów populacji, które oceniamy na podstawie badanej próby. W praktycznych zastosowaniach przy szacowaniu optymalnej liczności próby należy na wstępie określić pewne parametry lub wykorzystać wyniki wcześniej przeprowadzonych badań. W niektórych sytuacjach wymaga to określenia zmienności obserwacji, czyli np. przyjęcia pewnej wartości odchylenia standardowego, aby podstawić do równania na liczność. Często wymaga to oparcia się na danych z badań powadzonych w przeszłości w tej populacji lub wykonania badania pilotażowego na niewielkiej próbie w celu uzyskania przybliżonych wartości parametrów populacji.

Metody szacowania liczności próby opierają się na pewnych założeniach. Pierwsze z nich to, że parametry ilościowe w badanych próbach mają rozkład normalny (gdy liczność próby bardzo wzrasta, to średnie prób podlegają rozkładowi normalnemu, nawet kiedy badany parametr nie ma rozkładu normalnego w populacji). Następnie konieczne jest określenie z jakim prawdopodobieństwem wnioskujemy, np. o występowaniu różnic. Jest to związane z **poziomem istotności** α, poniżej którego odrzucamy **hipotezę zerową**. Oznacza to maksymalne prawdopodobieństwo błędnego stwierdzenia, że efekt rzeczywiście zachodzi. Zwykle przyjmujemy na poziomie 0,05 i odrzucamy hipotezę zerową, gdy wartość p jest mniejsza od tej wartości. Również konieczne jest określenie **mocy stosowanego testu** statystycznego (1 – błąd II rodzaju). Ryzyko błędu II rodzaju jest to prawdopodobieństwo nieodrzucenia hipotezy zerowej, gdy jest ona fałszywa. Innymi słowy jest to szansa wykrycia statystycznie istotnego efektu, jeżeli on rzeczywiście zachodzi. Zwykle przyjmuje się moc co najmniej 80%, a zatem błąd II rodzaju wynosi 20%. Należy również zwrócić uwagę, że obliczenie mocy testu jest również możliwe na podstawie uzyskanych wyników w próbach o danej liczbie badanych. Analiza mocy testu daje więc możliwość udzielenia odpowiedzi na pytanie: jaka jest moc naszego wnioskowania?

Ponadto **oszacowanie wielkości próby** bardzo często wymaga precyzyjnego określenia minimalnej wielkości efektu badania, który może zostać potraktowany jako klinicznie istotny. Często jest to różnica w średnich lub odsetkach. Czasami określona jest jako różnica standaryzowana, czyli wielokrotność odchylenia standardowego analizowanych wartości.

Optymalną wielkość próby można wyznaczyć kilkoma sposobami. Przydatne są np. wzory Lehra, specjalne tablice, nomogram Altmana, który można stosować dla różnych testów oraz różne programy komputerowe.

Podsumowując, jeszcze raz należy podkreślić, że zwiększenie dokładności szacowania można uzyskać, zwiększając liczność próby, ale tylko do pewnej wartości, gdy przyrost dokładności jest bardzo mały i nie osiągamy już istotnej korzyści. Natomiast próby o małej liczności charakteryzuje niska moc testu, niewystarczająca do wykrycia istotnych efektów. Obliczana liczność jest szacunkowa i nie określamy jej ściśle z dokładnością co do jedności.[2,3,8,10]

Wnioskowanie statystyczne Jednym z najistotniejszych zagadnień w analizie danych jest wnioskowanie statystyczne. Obejmuje dwa działy: estymację oraz weryfikację hipotez statystycznych. Estymacja jest to szacowanie pewnych wartości (np. średniej, odchylenia standardowego, współczynnika korelacji, odsetka wyróżnionej cechy) w całej populacji na podstawie wyników uzyskanych dla próby pobranej z tej populacji. Nie otrzymujemy więc wyniku dla populacji, tylko dla próby. Jednakże wnioski z tego wyniku chcemy uogólnić na całą populację. Oszacowanie charakterystyki cechy dla populacji dokonujemy na podstawie tej charakterystyki dla próby. Zatem otrzymany wynik uznajemy jako miarodajny dla populacji. Obliczana wartość dla próby jest więc estymatorem, czyli oszacowaniem tej wartości dla całej populacji, a estymację taką nazywamy estymacją punktową. Wyniki badania próby pozwalają obliczyć wartość danej cechy. Zwykle nie możemy na tej podstawie wnioskować, że w całej populacji jest taka sama wartość tej cechy. Najczęściej nie zadowala nas ta jedna wartość estymatora i potrzebne jest oszacowanie przedziału wartości, które może przybierać z założonym prawdopodobieństwem ten estymator. Przedział ten nazywamy przydziałem ufności, a metodę określania takiego przedziału estymacją przedziałową. Innymi słowy przedział ufności dla danej

miary statystycznej określa, w jakim przedziale mieści się poszukiwana przez nas rzeczywista wartość z określonym prawdopodobieństwem. Prawdopodobieństwo to określane jest poziomem ufności (1-a) i zwykle przyjmuje się 0,95. Poziom ufności wskazuje prawdopodobieństwo znalezienia rzeczywistej wartości poszukiwanego parametru w przedziale ufności. Natomiast prawdopodobieństwo znalezienia rzeczywistej wartości parametru poza przedziałem ufności, nazywamy poziomem istotności. Zatem są to prawdopodobieństwa ściśle ze sobą związane, a związek między nimi określa równanie: Poziom ufności = 1 – poziom istotności. Poziom istotności jest to wartość, względem której oceniamy z jakim prawdopodobieństwem różnice, które obserwujemy są przypadkowe.

Im węższy jest przedział ufności, tym precyzyjniej wartość estymatora przybliża wartość rzeczywistą. Z kolei szeroki przedział ufności świadczy o dużych różnicach wartości dla próby i dla populacji, a zatem małą wiarygodność uzyskanych wyników. W takiej sytuacji prawidłowe wnioskowanie na podstawie przedziałów ufności nie jest możliwe.

Jest oczywiste, że liczność próby ma istotne znaczenie dla oszacowania błędu określanej wielkości. Im liczność próby jest większa, tym większe prawdopodobieństwo, że uzyskane wyniki dla próby przybliżają wyniki dla populacji. Jeśli więc badanie przeprowadzane jest na większej liczbie osób, tym bardziej przedział ufności maleje. Należy pamiętać, że każde szacowanie obarczone jest pewnym błędem. Założone prawdopodobieństwo 95% oznacza, że mamy 5% szans na popełnienie błędu, czyli że prawdziwa wartość mieści się poza wyznaczonym przedziałem. Jeżeli zwiększymy prawdopodobieństwo, np. na 99%, to wyznaczony zakres ulegnie rozszerzeniu i na odwrót, jeżeli zmniejszymy prawdopodobieństwo na 90%, to zakres ulegnie zmniejszeniu.

Przedziały ufności i testowanie hipotez są ze sobą ściśle powiązane. Podstawowym celem estymacji jest uogólnienie uzyskanych wyników z próby na całą populację. Natomiast celem testowania hipotez jest stwierdzenie z określonym prawdopodobieństwem prawdziwości interesującego nas zdarzenia w populacji, na podstawie analizy badanych cech w próbie. Weryfikacja hipotez statystycznych po estymacji parametrów jest drugim zasadniczym działem wnioskowania statystycznego. Hipotezą statystyczną nazywamy każde przypuszczenie dotyczące badanej populacji, które testujemy, konfrontując wyniki uzyskane dla próby z treścią danej hipotezy. Hipotezę, którą sprawdzamy, nazywamy hipotezą zerową i oznaczamy symbolem $\mathbf{H}_0$. Natomiast to, co nas zazwyczaj interesuje hipotezą alternatywną – $\mathbf{H}_1$, czyli tą, którą przyjmujemy w przypadku odrzucenia hipotezy zerowej. Narzędziem służących do weryfikacji hipotez statystycznych jest test statystyczny. Należy pamiętać, że testy statystyczne zakładają, że próba została wybrana w sposób losowy i pozwalają uzyskać poprawne wyniki tylko, gdy jest ona reprezentatywna dla populacji. Test statystyczny jest to reguła postępowania, która każdej próbie losowej przypisuje decyzję przyjęcia lub odrzucenia sprawdzanej hipotezy. Żeby mieć podstawy do stwierdzenia o prawdziwości hipotezy alternatywnej, przyjmujemy ją tylko wtedy, gdy prawdopodobieństwo prawdziwości hipotezy zerowej będzie małe, zwykle równe 0,05. Każdy wynik testu statystycznego ma określony poziom istotności. Wartość założonego poziomu istotności jest porównywana z wyliczoną z testu statystycznego wartością p. Jeżeli obliczona przez program wartość p, tzw. prawdopodobieństwo testowe, jest mniejsza od przyjętego poziomu istotności testu to sprawdzaną hipotezę zerową odrzucamy i przyjmujemy alternatywną. Im mniejsza jest wartość p, tym mamy większą pewność o nieprawdziwości hipotezie zerowej i możemy powiedzieć, że wyniki są istotne na poziomie 5%. Jeśli wartość $p \geq 0,05$, oznacza to, iż nie ma podstaw do odrzucenia hipotezy zerowej, która zwykle stwierdza, że obserwowany efekt jest wynikiem przypadku. Mówimy wtedy, że uzyskane efekty nie są istotne na poziomie 5%. Należy podkreślić, że stwierdzenie: brak podstaw do odrzucenia hipotezy zerowej nie oznacza, że hipoteza zerowa jest prawdziwa. Natomiast wartość p nie oznacza prawdopodobieństwa, że hipoteza, zerowa jest prawdziwa czy fałszywa, a określa jedynie siłę dowodu do odrzucenia hipotezy zerowej. Ponadto opisywanie wyników tylko jako istotnych lub nie, na poziomie $p < 0,05$ nie daje dokładnych informacji. Na przykład, jeżeli $p = 0,04$, odrzucimy H_0; natomiast dla $p = 0,06$ oraz $p = 0,6$ nie mamy podstaw do jej odrzucenia. Podawanie dokładnej wartości p, otrzymanych w wynikach komputerowych pozwala na dokładniejsze zobrazowanie istotności statystycznej i jej braku.

Z uwagi na to, że testowanie hipotez statystycznych opiera się na wynikach próby losowej, decyzja o przyjęciu lub odrzuceniu sprawdzanej hipotezy podjęta w wyniku zastosowania testu zawsze obarczona jest pewnym błędem. Możliwe jest więc popełnienie jednego z dwóch rodzajów błędów, o których już wspomniano ze względu na ich bardzo istotną rolę w planowaniu badania i wyznaczaniu optymalnej liczności próby. O przyjęciu konkretnych ich wartości musimy zdecydować przed rozpoczęciem zbierania danych. Błąd I rodzaju oznacza, że odrzucamy hipotezę zerową, gdy w rzeczywistości jest ona prawdziwa i stwierdzamy istnienie różnicy, gdy w rzeczywistości jej nie ma. Jest to opisywany poziom istotności testu (α). Błąd II rodzaju oznacza, że nie odrzucamy hipotezy zerowej, gdy jest ona fałszywa, i stwierdzamy brak efektu, gdy w rzeczywistości on istnieje (β). Natomiast moc testu oznaczana ($1 - \beta$), jest prawdopodobieństwem odrzucenia hipotezy zerowej, gdy jest ona fałszywa; tzn. jest to szansa wykrycia jako statystycznie istotnego rzeczywistego efektu leczenia o określonej wielkości. Optymalna wielkość próby zapewnia równowagę

TABELA 3 Błędy przy testowaniu hipotez statystycznych (I i II rodzaju)

Hipoteza / Decyzja	Odrzucić H_0	Nie odrzucać H_0
H_0 prawdziwa	błąd I rodzaju (α)	decyzja prawidłowa (1- α)
H_0 fałszywa	decyzja prawidłowa (1- β)	błąd II-go rodzaju (β)

pomiędzy skutkami błędów I i II rodzaju. Określenie błędów pierwszego i drugiego rodzaju przedstawia **TABELA 3**.

Podsumowując przydatność praktyczną przedziałów ufności, należy podkreślić, że mogą być one użyte do podejmowania decyzji, informując o zakresie wiarygodnych wartości wyniku dla całej populacji, z przyjętym 95% prawdopodobieństwem. Umożliwiają ogólną ocenę czy wartość p <0,05, jeśli wartość wyniku jest poza 95% przedziałem ufności. Nie jest możliwe jednak na ich podstawie określenie dokładnej wartości p. W praktyce użycie przedziałów ufności pozwala na interpretowanie wartości statystycznych w odniesieniu do istotności klinicznej. Wytłumaczyć należy to tym, iż poziom p <0,05 w interpretacji oznacza efekt istotny statystycznie, ale nie musi być to równoważne z istotnością kliniczną. Należy pamiętać, że na wartość p ma wpływ wielkość badanej próby, czyli uzyskany bardzo mały efekt przy dużej próbie może być istotny statystycznie, czasami nie mając żadnego znaczenia klinicznego. Odwrotnie również poziom p >0,05 oznacza efekt nieistotny statystycznie, może być jednak istotny klinicznie. Zobrazować to można w bardzo prosty sposób, biorąc pod uwagę np. pacjentów z chromaniem przestankowym i możliwością przejścia przez nich dodatkowo odległości 2 m po zastosowanym leczeniu. Może się okazać, że uzyskany efekt leczenia jest efektem nieistotnym statystycznie, ale w praktyce klinicznej jest efektem istotnym klinicznie. Istotność kliniczna świadczy o wykryciu czegoś znaczącego w praktyce dla lekarza i pacjenta.

Z kolei przy weryfikacji hipotez statystycznych, bez których nie jest możliwe dokładne i poprawne wnioskowanie problemstanowi zawsze dobór odpowiednich metod opracowywania danych w zależności od celu badania. Najczęściej stosowane i opisywane w publikacjach metody dotyczą modeli prostych. Na przykład test istotności pozwala na ocenę, np. czy uzyskane wyniki różnią się czy nie od pewnej wartości hipotetycznej, tzn. czy ta wartość znajdzie się poza wyznaczonym przedziałem ufności lub czy znajdzie się w zakresie wyznaczonego przedziału ufności. Taka wiedza jest konieczna i wystarczająca dla każdego lekarza do prawidłowego rozumienia i interpretowania większości wyników, przedstawianych w oryginalnych artykułach, publikowanych w renomowanych czasopismach naukowych.

Skupiając swoją uwagę na analizach dotyczących porównywania grup, należy w pierwszej kolejności odróżnić **modele niepowiązane** (grupy niezależne) i **modele powiązane** (grupy zależne). Porównanie tego samego parametru dla całkowicie różnych grup, np. chorzy i zdrowi oznacza porównanie grup niezależnych. Natomiast porównanie tego samego parametru mierzonego kilka razy u tych samych pacjentów (np. w różnym czasie) lub różnymi metodami pomiarowymi (np. ultrasonografia i tomografia) oznacza porównanie grup zależnych. Przed przystąpieniem do weryfikacji określonej hipotezy statystycznej należy wybrać odpowiedni do tego celu test statystyczny. Prawidłowy wybór odpowiedniego testu nie jest dowolny i wymaga przestrzegania pewnych reguł. Znajomość tych reguł jest podstawą metodologii analizy statystycznej danych. Na początku należy stosowane testy podzielić na **testy parametryczne i testy nieparametryczne**. Testy parametryczne dotyczą porównywania pewnych parametrów populacji (np. średnia, wariancja). Natomiast testy nieparametryczne dotyczą porównywania pewnych opisów parametrów, czyli ich rozkładów, które nie są parametrami. Testy parametryczne charakteryzują się większą mocą, czyli zdolnością do wykrywania różnic rzeczywiście występujących, jednak wymagają spełnienia pewnych założeń. Brak spełnienia wymaganych założeń najczęściej prowadzi do uzyskania błędnych wyników. Warunkiem stosowania testów parametrycznych jest normalność rozkładów badanych parametrów lub zbliżonych do rozkładu normalnego. Jeśli rozkład danego parametru znacznie odbiega od rozkładu normalnego, testy parametryczne mogą prowadzić do niepoprawnych wyników i nieprawidłowych wniosków. Testy nieparametryczne można zastosować zawsze w powyższych sytuacjach, jednak należy pamiętać, że ich moc wynosi ok. 95% mocy testów parametrycznych. Dlatego też, kiedy spełnione są wymagane założenia, należy zastosować mocniejsze testy parametryczne.

Należy zwrócić uwagę na wpływ obserwacji odstających na uzyskiwane wyniki, np. średniej arytmetycznej czy odchylenia standardowego, co często jest powodem błędnej interpretacji uzyskiwanych wyników. Jeżeli obserwacje odstające są wynikiem błędów, należy je, jeśli to możliwe, skorygować lub usunąć albo zastąpić średnią. Znanym testem statystycznym pozwalającym na wykrywanie obserwacji odstających jest np. test Grubbsa lub reguła „czterech sigma". Jeśli występują znaczne rozrzuty danych, niebędące wynikiem błędu, można dokonać transformacji, co często pozwala uzyskać rozkład normalny. Najczęściej stosowaną w medycynie transformacją jest **transformacja logarytmiczna** w przypadku rozkładów prawoskośnych i różnych wariancji dla porównywanych grup.

Udowodniono jednak, że konsekwencje niespełnienia założenia o normalności rozkładu nie

są tak bardzo istotne, jak niespełnienie warunków jednorodności wariancji i braku skorelowania średnich i odchyleń standardowych. W ocenie jednorodności (homogeniczności) wariancji najczęściej stosowany jest test Fishera, test Levene'a, czy często polecany jako bardziej odporny na skośność danych i różnice w liczności grup test Browna i Forsythe'a. Uzyskując na podstawie tych testów istotną różnicę w wariancjach, oznacza to niejednorodność wariancji.

Z przedstawionych względów częściej używamy testów nieparametrycznych, pamiętając że są one słabsze. Jednak są niewrażliwe na typ rozkładu i przy rozkładach znacznie odbiegających od rozkładu normalnego cechują się wyższą wiarygodnością wyników.

Ponadto bardzo istotna jest liczba badanych grup. Oczywiście najprostsza sytuacja to porównanie dla jednej grupy badawczej danego parametru, np. z pewną wartością referencyjną, czy ocena zgodności rozkładu, np. z rozkładem normalnym. Również prosta sytuacja to porównanie wyników uzyskanych w dwóch grupach badawczych. Często jednak konieczne jest porównanie trzech i większej liczby grup. W takich sytuacjach często zdążają się błędy metodologiczne, polegające na porównywaniu każdej grupy z każdą, przez intuicyjne stosowanie metod statystycznych, bez znajomości podstaw metodologii. Wytłumaczenie dlaczego nie można tak postąpić jest proste i wynika z rachunku prawdopodobieństwa. Na przykład, porównując ze sobą dwie grupy niezależne czy zależne na poziomie istotności $\alpha = 0,05$, mamy możliwość uzyskania prawidłowego wyniku z prawdopodobieństwem 0,95, czyli popełnimy 5 błędów na 100. Następnie, przeprowadzając kolejne porównanie na tym samym poziomie istotności, mamy taką samą sytuację. Łącznie więc, prawidłowe wyniki to $(0,95)^2 \approx 0,90$. I analogicznie dla kolejnego porównania szansa uzyskania prawidłowych wyników wynosi już tylko $(0,95)^3 \approx 0,86$. Wynika z tego, że zbyt często stwierdzamy istnienie różnic, gdy w rzeczywistości różnice te nie występują. Jednym z najprostszych rozwiązań w takiej sytuacji jest użycie najbardziej popularnej poprawki Bonferroniego. Polega na tym, że dzielimy błąd przez liczbę porównań i takie alfa stosujemy do pojedynczego testu, np. porównując 3 grupy poziom istotności $\alpha = 0,05$ zastępuje się poziomem istotności $\alpha/3 = 0,0167$. Wtedy w sumie nie przekraczamy błędu 0,05. Należy pamiętać, że powoduje to zmniejszenie mocy testu, czyli zdolności do wykrywania rzeczywistych różnic. Ze względu na łatwy dostęp do programów komputerowych poleca się wykorzystanie bardziej skomplikowanych testów zawierających poprawkę na wielokrotne porównania.

Poruszyć należy również zagadnienie jedno- i obustronnych testów. Należy podkreślić, że hipoteza zerowa (H_0) to hipoteza weryfikowana, natomiast hipoteza alternatywna (H_1) to ta, którą można przyjąć, gdy zostanie odrzucona hipoteza zerowa. Weryfikując hipotezę, uzyskujemy dwa wnioski. Pierwszy z nich to odrzucenie hipotezy zerowej i uznanie za prawdziwą hipotezę alternatywną, z określonym prawdopodobieństwem. Drugi natomiast brzmi następująco: nie mamy podstaw do odrzucenia hipotezy zerowej i nie oznacza to jej przyjęcia, czyli stwierdzenia jej prawdziwości. Zawsze hipoteza alternatywna to ta, którą w wyniku testowania możemy przyjąć. Kiedy np. chcemy wykazać istnienie różnicy między dwiema średnimi dla porównywanych grup, to hipoteza zerowa zakłada równość tych średnich. Natomiast hipoteza alternatywna może dotyczyć różnicy między tymi średnimi bez określania kierunku tej różnicy, czyli czy średnia jednej grupy jest większa lub mniejsza od drugiej. Wtedy stosujemy test dwustronny. Jeżeli natomiast określimy kierunek różnicy, czyli hipotezę alternatywną, że wartość średnia jedna jest większa lub mniejsza od drugiej, stosujemy test jednostronny. Prawdopodobieństwo wykrycia różnicy dla testu jednostronnego jest zawsze dwukrotnie mniejsze niż dla testu dwustronnego. Jeżeli mylnie odrzucamy prawdziwą hipotezę zerową, popełniamy błąd pierwszego rodzaju (prawdopodobieństwo = istotność), jeżeli natomiast mylnie nie odrzucamy fałszywej hipotezy zerowej to popełniamy błąd drugiego rodzaju (prawdopodobieństwo = 1-moc testu). Należy podkreślić, że wybór testu jedno- lub obustronnego musi być oparty na wiedzy i doświadczeniu, a nie na potrzebie wykazania wyższych istotności różnic, co pociąga za sobą większe ryzyko fałszywego wnioskowania.[2-13]

Porównywanie parametrów ilościowych Dla parametrów mierzonych w skali ilorazowej i interwałowej lub porządkowej testy istotności dla pojedynczej próby umożliwiają odpowiedź na pytanie czy średnia lub mediana dla badanej próby jest zgodna z wartością hipotetyczną. Stosuje się w takiej sytuacji test t dla pojedynczej próby lub test Wilcoxona dla pojedynczej próby. Natomiast testy istotności dla dwóch lub więcej prób umożliwiają odpowiedź na pytanie czy średnia lub mediana dla porównywanych grup się różnią.

Przy porównywaniu parametrów mierzalnych dla kilku grup kryteria wyboru testów statystycznych zależą od następujących czynników:

- czy grupy są niezależne czy zależne
- jaka jest liczba porównywanych grup; 2 czy więcej niż 2
- czy rozkład zmiennej w każdej z analizowanych grup jest zbliżony do rozkładu normalnego czy inny (odpowiedź wymaga zastosowana testów normalności)
- czy wariancje w porównywanych grupach są jednorodne czy niejednorodne (odpowiedź wymaga zastosowana testów jednorodności wariancji).

Grupy niezależne Test t-Studenta dla grup niezależnych stosujemy porównując dwie średnie, gdy rozkład danych w każdej analizowanej grupie jest zbliżony do rozkładu normalnego i gdy wariancje parametrów w porównywanych grupach są jednorodne. Gdy wariancje nie są jednorodne

TABELA 4 Wybór testu statystycznego (porównanie grup niezależnych)

Porównywane grupy	Charakter rozkładu	Jednorodność wariancji	Stosowany test
dwie grupy niezależne	rozkład normalny	tak	test t-Studenta dla grup niezależnych
		nie	test Cochrana i Coxa lub test Welcha
	brak rozkładu normalnego lub porządkowa skala pomiaru		test Manna i Whitney'a
więcej niż dwie grupy niezależne	rozkład normalny	tak	analiza wariancji (*post hoc* Tukeya)
		nie	test Kruskala i Wallisa (*post hoc* Dunna)
	brak rozkładu normalnego lub porządkowa skala pomiaru		

do porównań średnich należy użyć testu Cochrana i Coxa lub testu Welcha. W przypadku braku rozkładu normalnego lub porządkowej skali pomiaru (w skali porządkowej nie bada się normalności rozkładu) do porównania mediany w dwóch grupach niezależnych należy użyć testu Manna i Whitneya (ok. 95% mocy testu t-Studenta dla grup niezależnych).

Do jednoczesnego porównania kilku średnich dla więcej niż dwóch grup niezależnych, gdy spełnione jest założenie normalności rozkładu i jednorodności wariacji stosuje się **jednoczynnikową analizę wariancji (ANOVA)**. Jeśli różnice między średnimi okazują się nieistotne, oznacza to brak różnic między analizowanymi grupami. Natomiast jeśli analiza wariancji wykaże istnienie różnic, oznacza to że, przynajmniej jedna średnia różni się od pozostałych. Analiza ta nie pozwala jednak na ocenę, które średnie grupowe różnią się między sobą. W takiej sytuacji stosuje się testy *post hoc*. Do wyboru jest wiele testów, np. test NIR Fishera, Bonferroniego, Scheffego, Tukeya, Newmana i Keulsa, Duncana oraz porównania z grupą kontrolną Dunneta i zależy to od stopnia ich konserwatywności. Niska konserwatywność oznacza duże prawdopodobieństwo popełnienia błędu pierwszego rodzaju przy minimalizacji błędu drugiego rodzaju. Natomiast wysoka konserwatywność oznacza minimalizację błędu pierwszego rodzaju przy dużym prawdopodobieństwie popełnienia błędu drugiego rodzaju. Niską konserwatywnością cechuje się test NIR Fishera, a wysoką test Scheffego. Często stosowane są więc testy o średniej konserwatywności i niewrażliwości na naruszenie założeń normalności rozkładu i jednorodności wariancji, tj. Newmana i Keulsa i Tukeya. Najczęściej jest to test Tukeya, który ma największą moc.

W przypadku braku rozkładu normalnego lub porządkowej skali pomiaru do porównania mediany dla więcej niż dwóch grup niezależnych należy użyć testu Kruskala i Wallisa. Dla testów nieparametrycznych w analizach *post hoc* najczęściej stosowany jest test o umiarkowanym stopniu konserwatywności, tj. test wielokrotnych porównań Dunna.[2,4,6,8,10-14]

Grupy zależne Test t-Studenta dla grup zależnych stosujemy, porównując dwie średnie, gdy rozkład różnicy pomiarów analizowanych grup jest zbliżony do rozkładu normalnego. W przypadku braku rozkładu normalnego lub porządkowej skali pomiaru do porównania uporządkowania danych należy użyć testu kolejności par Wilcoxona (ok. 95% mocy testu t-Studenta dla grup zależnych). Do jednoczesnego porównania kilku grup zależnych stosuje się analizę wariancji z powtarzanymi pomiarami. Warunkiem koniecznym i wystarczającym poprawności wyników jednoczynnikowej analizy wariancji z powtarzanymi pomiarami jest warunek sferyczności, co oznacza, że po obliczeniu różnic dla wszystkich par wyników z powtarzanych pomiarów, wariancje tych różnic będą do siebie zbliżone. Sferyczność oceniamy na podstawie testu Mauchlaya. Po spełnieniu tego warunku, jeśli okaże się w analizie wariancji, że różnice między średnimi są nieistotne, oznacza to brak różnic między analizowanymi grupami. Natomiast jeśli analiza wariancji wykaże istnienie różnic oznacza to, że przynajmniej jedna średnia różni się od pozostałych. Analiza ta nie pozwala jednak na ocenę, które średnie grupowe różnią się między sobą. W takiej sytuacji stosuje się testy *post hoc*, najczęściej test Tukeya.

W przypadku niespełnienia powyższych warunków lub porządkowej skali pomiaru do porównania mediany dla więcej niż dwóch grup zależnych należy użyć **testu Friedmana**. Dla testów nieparametrycznych w analizach *post hoc* najczęściej stosuje się test wielokrotnych porównań Dunna.[2,4,6,8,10-14]

W analizie wariancji jednoczynnikowy model można często rozbudowywać, tworząc modele, złożone tj. model dwuczynnikowy czy model trójczynnikowy itd. Są to wtedy modele wieloczynnikowej analizy wariancji (MANOVA), których bardziej zaawansowana metodologia wykracza poza zakres koniecznych podstaw wiedzy statystycznej.[14]

Interpretując uzyskane wyniki dla wszystkich opisanych testów, porównuje się przyjęty przed testem poziom istotności α z obliczonym przez komputer prawdopodobieństwem testowym p.

Podsumowanie wyboru testów porównujące grupy niezależne oraz zależne przedstawiają **TABELE 4** i **5**.

Porównywanie cech nominalnych Skala nominalna jest tak słabą skalą pomiarową, iż zarówno ocenę istotności różnic, jak i ocenę zależności zmiennych można wykonać z użyciem statystyki χ^2. Najpowszechniej używany jest test χ^2 Pearsona. Wskazuje on na istnienie lub brak różnic między grupami bądź zależności lub jej brak między badanymi parametrami. Do prawidłowego stosowania testów χ^2 liczność próby musi wynosić co najmniej

TABELA 5 Wybór testu statystycznego (porównanie grup zależnych)

Porównywane grupy	Charakter rozkładu		Stosowany test
dwie grupy zależne	rozkład różnicy normalny		test t-Studenta dla grup zależnych
	brak rozkładu normalnego lub porządkowa skala pomiaru		test kolejności par Wilcoxona
więcej niż dwie grupy zależne	rozkład normalny	sferyczność	analiza wariancji z powtarzanymi pomiarami (*post hoc* Tukeya)
		brak sferyczności	test Friedmana (*post hoc* Dunna)
	brak rozkładu normalnego lub porządkowa skala pomiaru		

30 jednostek badanych (czasami nawet 50). Najczęstszym sposobem przedstawiania danych mierzonych w skali nominalnej są tabele wielodzielcze (kontyngencji). Są to tabele, w których w odpowiednich wierszach i kolumnach wpisujemy częstości występowania badanych cech. Najprostszą formą tabeli wielodzielczej jest tabela 2 × 2. Możliwość poprawnego stosowania testu χ^2 jest związana z koniecznością każdej liczności oczekiwanej równej co najmniej 5, która jest obliczana na podstawie wartości obserwowanych w tabeli wielodzielczej. Małe wartości oczekiwane mogą doprowadzić do niewłaściwych wniosków. Dlatego stosuje się regułę Cochrana, która brzmi: jeżeli przynajmniej jedna z wartości oczekiwanych jest mniejsza od 1 lub jeżeli więcej niż 20% wartości oczekiwanych jest mniejsza niż 5 szacowana wartość statystyki χ^2 jest niepoprawna. Jeśli liczności są małe, to zawsze należy się zastanowić nad możliwością połączenia kategorii tak, aby warunek powyższy był spełniony. Zauważmy również, że warunek ten wyznacza minimalną wielkość próby wymaganej do stosowania testu. Jeśli zatem każda komórka tabeli zawierać ma co najmniej 5 jednostek, to liczba jednostek w próbie musi być co najmniej 5 razy większa niż liczba komórek. Na przykład, w tabeli 3 × 3, która zawiera 9 komórek, wymagana liczność próby wynosi co najmniej 45 (5 × 9) jednostek badanych. W przypadku niespełnienia tego warunku konieczne jest stosowanie poprawek. Przyjęta przez statystyków empiryczna reguła mówi: jeżeli w tablicy wartości obserwowanych 2 × 2 (nie mylić z warunkami Cochrana, które dotyczą tablicy wartości oczekiwanych) występują w niektórych komórkach liczebności poniżej 10, należy użyć poprawki Yatesa. Kiedy całkowita liczność jest mniejsza od 20 lub między 20 a 40, ale najmniejsza z liczności jest niższa niż 5 stosuje się dokładny test Fishera. Ponadto należy jeszcze sprawdzić, czy są spełnione warunki Cochrana. W przypadku ich niespełnienia musimy użyć do analizy dokładnego testu Fishera (dla tabel 2 × 2) lub testu Fishera, Freemana i Haltona (poprawnego również dla tabel o wymiarach większych). Dla tabel o większych wymiarach nie ma możliwości uwzględnienia poprawki na ciągłość Yatesa.

W analizach niezależności z użyciem testu χ^2 i stwierdzeniu istnienia zależności cech, to znaczy po odrzuceniu hipotezy zerowej, można określić siłę związku między analizowanymi cechami. Współczynnikami opartymi na wartości statystyki χ^2 są współczynniki Q Yula, phi-kwadrat lub V-kwadrat, które określają siłę zależności między dwoma parametrami jakościowymi w tabeli 2 × 2. Przyjmują one wartości od 0 (brak związku między parametrami) do 1 (całkowita zależność między badanymi parametrami). Dla tabel 2 × 2 najczęściej stosowaną miarą zależności cech jest współczynnik Q Kendalla. Jest wykorzystywany, jeśli zmienne są zdychotomizowane (np. wartości powyżej mediany przyjmują wartość 1, a poniżej 0). Wartość współczynnika równa 1 (lub –1) oznacza całkowitą zależność cech, tzn. jednemu wariantowi jednej cechy odpowiada zawsze jeden wariant drugiej cechy, a drugiemu – drugi. Im mniejsza wartość bezwzględna Q, tym słabsza zależność cech, a Q = 0 oznacza całkowitą niezależność cech. Natomiast dla tabel wielodzielczych o większych wymiarach posłużymy się współczynnikiem Cramera. Współczynnik ten również przyjmuje wartości od 0 do 1.

W badaniach medycznych mamy często do czynienia ze specyficznymi układami danych dla powtarzanych pomiarów. Kiedy porównujemy dwie grupy zależne dla parametrów mierzonych w skali nominalnej, wykorzystujemy **test McNemary**. Natomiast dla więcej niż dwóch grup zależnych **test Q Cochrana**. Testem wielokrotnych porównań *post hoc* są testy Fleissa lub Marascuilo i McSweeneya. Innym wyjściem może być stosowanie testu McNemary i poprawki Bonferroniego.[2,4,6,8,10-13]

Należy tutaj również wspomnieć o istnieniu testów dla pojedynczego czy dwóch wskaźników struktury (odsetka). Jednak wymagana liczność grup w takiej sytuacji to co najmniej 100 jednostek.

Należy również wspomnieć, że szczególnym przypadkiem zastosowania tabel wielodzielczych, gdy chcemy testować wpływ różnych czynników i ich interakcji jest analiza log-liniowa, która należy do grupy analiz wieloczynnikowych.[15]

Podsumowanie wyboru testów, analizując parametry mierzone w skali nominalnej, przedstawia **TABELA 6**.

Podsumowując testy istotności, należy przypomnieć regułę praktycznego postępowania w ocenie, która z dwóch analizowanych hipotez jest bardziej prawdopodobna. Przypomnijmy, że opieramy się na obliczonej z określonego testu statystycznego wartości prawdopodobieństwa – p. Wartości prawdopodobieństwa p, poniżej założonego poziomu progowego a, oznaczają odrzucenie hipotezy zerowej, wartości zaś powyżej a nie pozwalają na odrzucenie hipotezy

TABELA 6 Wybór testu statystycznego dla skali nominalnej

Analizowane grupy	Stosowany test
dwie grupy niezależne	chi-kwadrat Pearsona dokładny test Fishera test Fishera, Freemana i Haltona
dwie grupy zależne	test McNemary
więcej niż dwie grupy niezależne	analiza log-liniowa
więcej niż dwie grupy niezależne	test Q Cochrana

TABELA 7 Tablica kontyngencji przedstawiająca związek ocenianego testu (parametru) z rzeczywistą sytuacją chorobową

Wynik testu	Rzeczywistość		Razem
	choroba	brak choroby	
pozytywny	TP	FP	TP+FP
negatywny	FN	TN	FN+TN
razem	TP+FN	FP+TN	TP+FP+FN+TN

Skróty: TP – wyniki prawdziwie pozytywne; FP – wyniki fałszywie pozytywne; FN – wyniki fałszywie negatywne; TP – wyniki prawdziwie negatywne

zerowej. Gdy p = a, należałoby rozstrzygnąć poprzez zwiększenie liczności badanej próby.

Przy analizie parametrów mierzonych w skali nominalnej należy również wspomnieć o ocenie jakości wykonywanych testów lub badanych parametrów. Pozwala na to sytuacja przedstawiona poniżej, dotycząca liczby przypadków prawdziwie pozytywnych, prawdziwie negatywnych, fałszywie pozytywnych i fałszywie negatywnych (TABELA 7).

Odsetki wyników prawdziwie dodatnich, prawdziwie ujemnych, fałszywie dodatnich i fałszywie ujemnych to atrybuty testu, na podstawie których decyduje się o zastosowaniu danego testu.

Podstawowe miary dokładności testu to czułość i specyficzność.

Czułość (*sensitivity* – SENS) to prawdopodobieństwo zarejestrowania testu pozytywnego (nieprawidłowy wynik) u osób rzeczywiście chorych:

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%$$

Test czuły powinien wykazać małą liczbę osób z wynikami fałszywie ujemnymi.

Specyficzność (*specifity* – SPEC) – inaczej swoistość, czyli prawdopodobieństwo uzyskania wyniku negatywnego (prawidłowy wynik) u osób zdrowych:

$$\text{SPEC} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\%$$

Test swoisty powinien wykazać małą liczbę osób z wynikami fałszywie dodatnimi.

Zatem czułość określa zdolność testu do wykrywania choroby, natomiast swoistość zdolność testu do wykluczania choroby.

Innym przydatnym wskaźnikiem jest **wskaźnik wiarygodności** (*likelihood ratio* – LR), który odzwierciedla wartość predykcji wyniku testu. Określa on stopień, w jakim wynik testu zmienia prawdopodobieństwo choroby.

Wskaźnik wiarygodności dodatniego wyniku testu (LR+) wskazuje na ile wynik dodatni zwiększa szansę wystąpienia choroby w porównaniu z szansą określoną przed zastosowaniem testu:

$$\text{LR+} = \frac{\text{SENS}}{1 - \text{SPEC}}$$

Wskaźnik wiarygodności ujemnego wyniku testu (LR-) wskazuje, na ile wynik ujemny zmniejsza szansę wystąpienia choroby w porównaniu z szansą określoną przed zastosowaniem testu:

$$\text{LR–} = \frac{1 - \text{SENS}}{\text{SPEC}}$$

Przyjmuje się, że test ma rzeczywistą wartość diagnostyczną, jeżeli LR wynosi około 10 i więcej lub około 0,1 i mniej. Wyniki między 5 a 10 oraz między 0,1 a 0,2 wskazują, że test jest przydatny. LR między 0,5 a 2 oznacza, że test nie wpływa w sposób istotny na ocenę prawdopodobieństwa choroby. Porównanie LR dla różnych testów umożliwia również szybką ocenę użyteczności badanych parametrów (testów).

Ponieważ ani czułość, ani swoistość nie odpowiadają na pytanie: jeśli test jest dodatni, to jaka jest szansa na występowanie choroby oraz jeśli test jest ujemny, to jaka jest szansa, że chorobę można wykluczyć, stosuje się wtedy dodatnią i ujemną wartość predykcyjną.

Dodatnia wartość predykcyjna (*positive predictive value* – PPV), czyli wartośc predykcji wyniku dodatniego testu, określa prawdopodobieństwo stwierdzenia choroby na podstawie pozytywnego wyniku testu

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%$$

czyli odpowiada na pytanie: jeżeli wynik testu był nieprawidłowy, to jakie jest prawdopodobieństwo, że choroba rzeczywiście występuje.

Ujemna wartość predykcyjna (*negative predictive value* – NPV), czyli wartość predykcji wyniku ujemnego testu, określa prawdopodobieństwo wykluczenia choroby przy negatywnym wyniku testu

$$NPV = \frac{TN}{FN + TN} \times 100\%$$

czyli odpowiada na pytanie: jeżeli wynik testu wskazuje na istnienie prawidłowości, to jakie jest prawdopodobieństwo, że chorobę można wykluczyć.

SENS, SPEC i LR są niezależne od częstości występowania choroby w populacji, dlatego przyjmuje się je często za wartości stałe dla danego testu. Natomiast PPV i NPV zmieniają się w zależności od częstości występowania choroby w populacji.

Z kolei **dokładność** (*accuracy* – ACC) oznacza odsetek pozytywnego wyniku testu u chorych oraz negatywnego wyniku testu u zdrowych:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \times 100\%$$

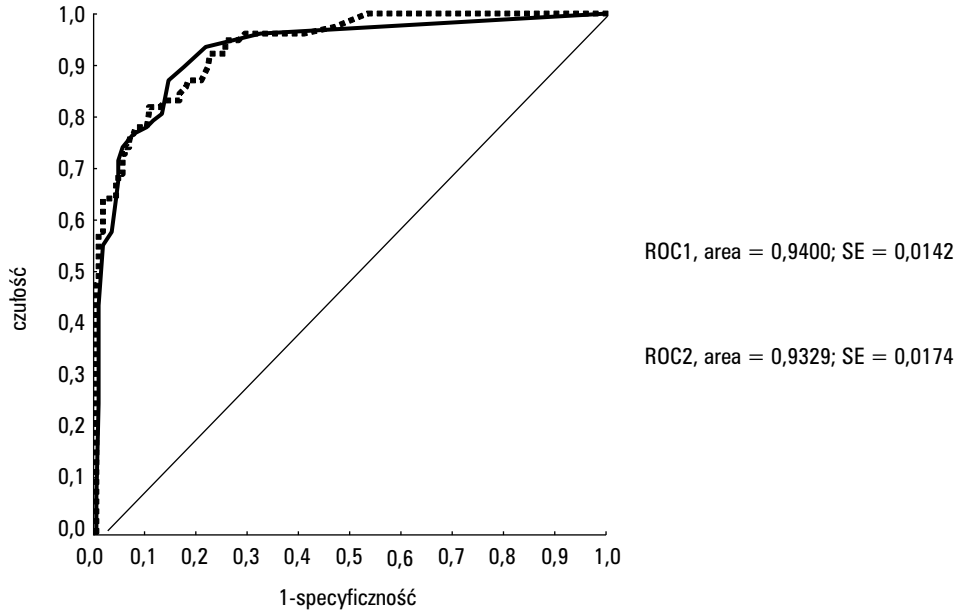
Należy zwrócić uwagę na fakt, że złagodzenie kryteriów dla testu pozytywnego zwiększa czułość, ale jednocześnie zmniejsza specyficzność, podczas gdy zaostrzenie kryteriów oceny zwiększa specyficzność, ale obniża czułość. W praktyce klinicznej często stosuje się w związku z tym dwa różne testy jeden po drugim – pierwszy o dużej czułości i drugi o dużej specyficzności.[3,16]

Dokładniejszą analizę możemy przeprowadzić, kiedy wartości testu diagnostycznego są ciągłe lub mają wiele kategorii. Wtedy oblicza się wartości predykcyjne dla najlepiej różnicujących występowanie choroby **wartości granicznych** (*cutoff*). Wartości decyzyjne przyjęte dla mierzonych parametrów to wartości graniczne, których przekroczenie pozwala na dokonanie kwalifikacji osoby badanej, jako tej, u której istnieje podejrzenie choroby. Wartość graniczna obliczana dla parametrów mierzalnych ma więc oddzielać zdrowych od chorych. Wybór „najlepszej" wartości progowej wyniku testu jest często kompromisem pomiędzy największą czułością testu (najmniejszym odsetkiem wyników fałszywie ujemnych) i największą specyficznością (najmniejszym odsetkiem wyników fałszywie dodatnich). Dla parametrów mierzalnych zmiany czułości i specyficzność testów przy przesuwaniu wartości granicznej przedstawia się graficznie w formie krzywych **ROC (Receiver Operating Characteristics)** oraz wykorzystano dodatkowe informacje, porównując pola pod krzywą ROC. Konstrukcja krzywych ROC jest efektywną metodą pomocną w ustaleniu wartości progowej dla testu diagnostycznego. Krzywe te zajmują centralną i unifikującą pozycję w procesie szacowania i stosowania różnych testów diagnostycznych. Dostarczają one współczynnika dokładności, pokazując ograniczenia zdolności testu do różnicowania alternatywnych stanów zdrowia i choroby na tle całego spektrum rozpatrywanych wyników badań. Krzywa jest konstruowana na podstawie zmieniających się progów decyzji w całym obszarze obserwowanych wyników. Krzywa ROC opisuje nakładanie się dwóch rozkładów poprzez wykreślenie czułości *vs* 1-specyficzność dla całego zakresu progów decyzji. Na osi Oy przedstawiona jest czułość (frakcja prawdziwie pozytywna), obliczana tylko na podstawie wyników testu w podgrupie chorych. Natomiast na osi Ox przedstawiona jest wartość 1-specyficzność (frakcja fałszywie pozytywna), obliczana wyłącznie w grupie zdrowych. Test dający perfekcyjne różnicowanie obu grup przy braku pokrywania się w dwóch rozkładach wyników reprezentuje krzywa, która przechodzi przez górny lewy róg, gdzie prawdziwie pozytywna frakcja równa się 1 lub 100% (perfekcyjna czułość), a fałszywie pozytywna frakcja równa się 0 (perfekcyjna specyficzność). Teoretyczny wykres dla testu, który nie daje żadnego różnicowania, to znaczy istnieje w nim identyczny rozkład wyników dla dwóch grup przedstawia prosta nachylona do osi Ox pod kątem 45°. Większość krzywych przebiega między tymi dwoma ekstremami. Obrazowo oznacza to, że im bliżej górnego lewego rogu przechodzi krzywa ROC, tym większa jest dokładność testu diagnostycznego. Wartości te mieszczą się w zakresie między 1,0, która odpowiada perfekcyjnej separacji zmiennych przez test w obu grupach i 0,5, co oznacza, że nie ma widocznej różnicy rozkładu wartości między dwiema grupami. Możliwe jest nie tylko porównanie testów ze sobą, ale też zbadanie czy test diagnostyczny jest w ogóle efektywny w rozróżnianiu dwóch populacji. Najbardziej użyteczną klinicznie metodą przedstawiania właściwości testu jest opisanie jego wyników za pomocą LR i/lub pola powierzchni pod krzywą ROC.[3,17,18] Przykład wykreślonych dwóch krzywych ROC przedstawia **RYCINA 4**.

Ocena zależności między badanymi parametrami mierzalnymi Ostatnim zagadnieniem najczęstszych zastosowań w badaniach medycznych dla parametrów mierzalnych po statystyce opisowej i testowaniu hipotez dotyczących różnic jest ocena istnienia i siły związku statystycznego. Bardzo często zasadniczym celem jest badanie współzależności między dwoma parametrami. Analiza pozwala odpowiedzieć na pytanie czy między badanymi parametrami istnieją jakieś zależności oraz jaki jest ich kształt, siła i kierunek.

Należy zwrócić uwagę, że współzależność między badanymi parametrami może być funkcyjna lub probabilistyczna. Zależność funkcyjna polega na tym, że zmiana wartości jednego parametru powoduje konkretną zmianę drugiego parametru i jest określona danym wzorem matematycznym. Zależność probabilistyczna ma miejsce wtedy, gdy zmienia się rozkład prawdopodobieństwa jednego parametru wraz ze zmianą drugiego parametru. Szczególnym przypadkiem zależności

RYCINA 4 Przykład hipotetycznych krzywych ROC



probabilistycznej jest oceniana metodami statystycznymi zależność korelacyjna. Dotyczy ona sytuacji, kiedy określonym wartościom jednego parametru odpowiadają konkretne średnie wartości drugiego parametru.

Należy również zwrócić również uwagę na fakt, iż liczbowe stwierdzenie występowania współzależności nie zawsze oznacza występowanie związku przyczynowo-skutkowego między badanymi parametrami. Hipoteza o tego typu związku musi być formułowana na podstawie wiedzy merytorycznej, pozwalającej wytłumaczyć sposób kształtowania się badanej zależności przyczynowej. Techniczne wykonywanie analizy, bez wiedzy merytorycznej dotyczącej możliwości występowania określonej zależności przyczynowo-skutkowej może prowadzić do wykrycia związków pozornych i być powodem często nieprawidłowej interpretacji otrzymanych wyników. Znane są przytaczane często przykłady występowania istotnych zależności, ocenione na podstawie metod statystycznych, np. liczby urodzeń i liczby gniazd bocianich.

Do oceny zależności między badanymi parametrami używa się **analizy korelacji i regresji**. Korelacja pozwala na ocenę siły związku liniowego, a regresja umożliwia określenie kształtu tego związku.

Najprostszy i najczęściej analizowany związek to związek liniowy. W medycynie rzadko mamy do czynienia z prostymi związkami, jednak zawsze należy rozpoczynać analizy od oceny istnienia związku liniowego. Wyjaśnić należy to możliwością prostej interpretacji oraz częstym odnoszeniem się do braku możliwości prostej interpretacji ze względu na brak zależności liniowych.

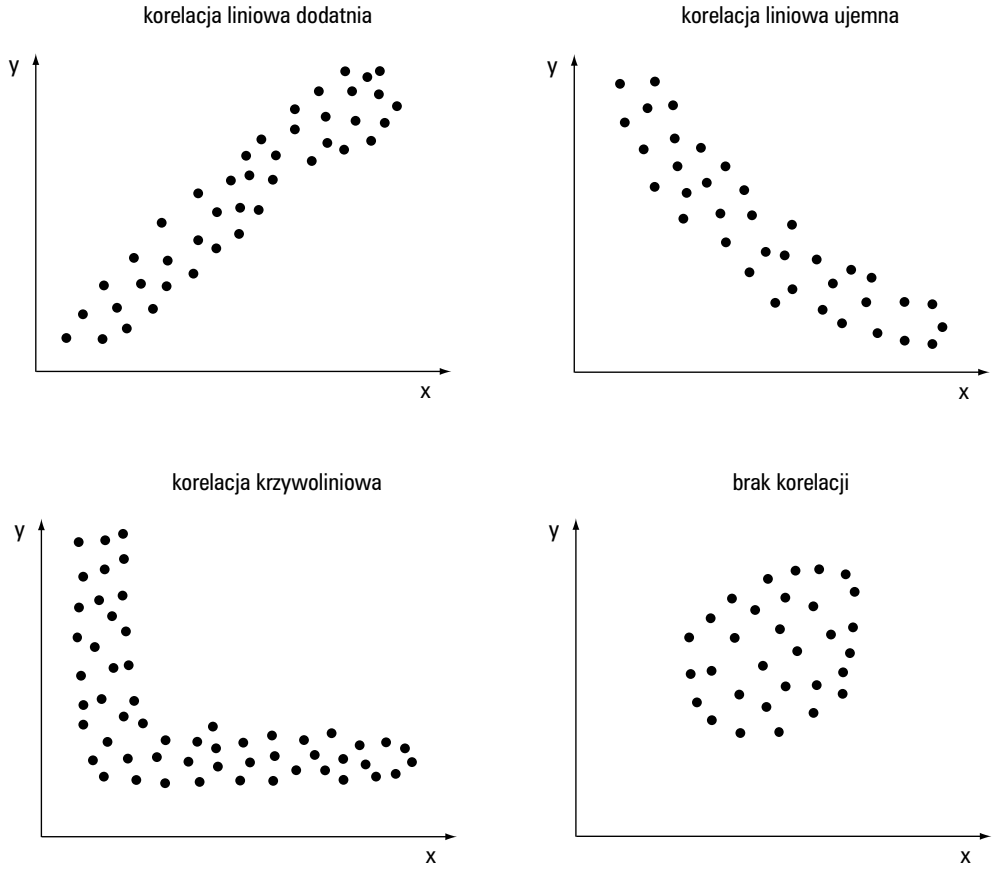
Analizę istnienia związku korelacyjnego między badanymi parametrami należy rozpocząć od sporządzenia wykresu rozrzutu punktów o współrzędnych x (pierwszy parametr) i y (drugi parametr) i na jego podstawie dokonać ogólnej oceny siły, kierunku i rodzaju zależności. Pozwala to więc na rozstrzygnięcie problemu czy zastosować regresję liniową czy nieliniową.

Gdy wyznaczone punkty układają się wzdłuż prostej, oznacza to istnienie związku liniowego. W praktyce niezwykle rzadko się zdarza, aby punkty układały się idealnie na prostej, co oznaczałoby całkowitą zależność. Im bliżej te punkty są prostej, tym silniejszy związek. Gdy punkty są chaotycznie rozproszone w układzie współrzędnych, oznacza to brak zależności liniowej. Jeżeli natomiast tworzą inny kształt, świadczy to o istnieniu zależności nieliniowej. W regresji krzywoliniowej możemy określić również typ krzywej reprezentującej związek, np. wykładniczy, logarytmiczny, hiperboliczny, potęgowy, wielomianowy czy logistyczny.

Zależność liniowa jest szczególnym przypadkiem związku monotonicznego. Związek monotonicznie rosnący (korelacja dodatnia) występuje wtedy, gdy wraz ze wzrostem wartości jednego parametru rosną wartości drugiego parametru, a związek monotonicznie malejący (korelacja ujemna) występuje wtedy, gdy wraz ze wzrostem wartości jednego parametru maleją wartości drugiego parametru. Graficzną prezentację opisanych sytuacji przedstawia RYCINA 5 (źródło: http://slideplayer.pl/slide/61 882/#).

Analiza korelacji pozwala więc na wykrycie związku liniowego badanych parametrów mierzonych w skali ilorazowej lub interwałowej. W takim przypadku najczęściej stosowaną miarą siły i kierunku zależności jest **współczynnik korelacji liniowej Pearsona**. Należy zwrócić uwagę na to, że współczynnik korelacji nie jest odporny na nietypowe obserwacje odstające, co wpływa na końcowe wyniki i w konsekwencji może się przyczynić do błędnej ich interpretacji. Pojedynczy pomiar odstający może nawet wpłynąć na nieprawidłowo stwierdzony istotny statystycznie związek liniowy.

Współczynnik korelacji przyjmuje wartości z przedziału od –1 do 1 i oznaczany jest zwykle symbolem r. Im jego wartość jest bliższa –1 lub 1, tym silniejszy jest związek między badanymi parametrami. Wartość współczynnika równa –1 lub 1



RYCINA 5 Wykresy rozrzutu prezentujące korelację liniową dodatnią, korelację liniową ujemną, korelację krzywoliniową i brak korelacji

oznacza występowanie związku funkcyjnego. Natomiast wartość 0 oznacza brak zależności liniowej między badanymi parametrami.

Jednak sama wartość r jest niewystarczająca do stwierdzenia istotnego związku. Może się czasami okazać, że nawet współczynnik r bliski 1 jest nieistotny statystycznie. Do stwierdzenia czy związek jest istotny służy test, w którym stawia się dwie hipotezy: hipotezę zerową mówiącą, że współczynnik r równy jest zero ($p > \alpha$), czyli oznacza brak zależności liniowej oraz hipotezę alternatywną mówiącą, że r różni się od zera ($p < \alpha$), czyli istnieje istotna statystycznie liniowa zależność między badanymi parametrami.

Często stosowana interpretacja wartości współczynnika korelacji jest następująca:

$-1 < r < -0{,}70$	bardzo silna korelacja ujemna
$-0{,}7 < r < -0{,}5$	silna korelacja ujemna
$-0{,}5 < r < -0{,}3$	korelacja ujemna o średnim natężeniu
$-0{,}3 < r < -0{,}2$	słaba korelacja ujemna
$-0{,}2 < r < 0{,}2$	brak korelacji
$0{,}2 < r < 0{,}3$	słaba korelacja dodatnia
$0{,}3 < r < 0{,}5$	korelacja dodatnia o średnim natężeniu
$0{,}5 < r < 0{,}7$	silna korelacja dodatnia
$0{,}7 < r < 1$	bardzo silna korelacja dodatnia.

Zwykle sam współczynnik korelacji nie wystarcza i chcemy wiedzieć więcej. Wtedy do danych możemy również dopasować najlepszą prostą zwaną prostą regresji, opisującą zależność między badanymi parametrami. Wyznaczenie prostej regresji polega na określeniu współczynników a i b w równaniu prostej $y = ax + b$.

W tym miejscu ważne jest podjęcie decyzji, która zmienna w równaniu jest zmienną niezależną, a która zależną. Zmienną wpływającą na zmiany nazywamy **zmienną niezależną**. Natomiast zmienną, której zmiany mogą być powodowane przez zmienną niezależną, nazywamy **zmienną zależną**. Zgodnie z podanym wcześniej określeniem zależności typu statystycznego równanie regresji jest więc ilościowym zapisem zależności, zachodzącej pomiędzy określonymi wartościami zmiennej niezależnej – x i odpowiadającymi im wartościami zmiennej zależnej – y.

Analiza regresji jest zatem opisem i oceną zależności między zmienną zależną będącą efektem wpływu jednej lub kilku innych zmiennych niezależnych, które są traktowane jako przyczyny. Badanie zależności tego typu można wykorzystywać do:

1. oceny efektów decyzji związanych ze zmianą x;
2. przewidywania wartości *y* na podstawie znanych wartości x;
3. badania czy zmienna x ma istoty wpływ na y.

Dalszym etapem analizy może być wyznaczenie również jednej z najczęściej stosowanych miar, którą jest **współczynnik determinacji**. Definiowany jest jako wartość r^2. Współczynnik ten określa procent wariancji zmiennej zależnej, który można wyjaśnić z użyciem modelu liniowego przez zmiany zmiennej niezależnej. Służy więc do oceny dopasowania modelu do danych empirycznych. Przyjmuje się, że modele dla których wartość $r^2 > 0{,}75$ to modele dobre, natomiast gdy $r^2 \leq 0{,}50$ są modelami słabej jakości.

Gdy założenie o ilorazowej lub interwałowej skali pomiaru badanych parametrów nie jest spełnione, musimy posłużyć się innymi współczynnikami. Nieparametrycznym odpowiednikiem współczynnika korelacji Pearsona jest współczynnik korelacji rangowej Spearmana. Jest on stosowany, gdy spełniony jest przynajmniej jeden z następujących warunków:

1. co najmniej jedna zmienna, x lub y, mierzona jest w skali porządkowej;
2. obie zmienne są wyrażone w skali ilorazowej lub interwałowej i nie interesuje nas wyłącznie związek liniowy, lecz jakikolwiek związek monotoniczny;
3. ani x, ani y nie mają rozkładu normalnego;
4. liczebność próby jest mała; zwykle nie przekracza 30 jednostek;
5. potrzebujemy miary związku między dwoma parametrami, gdy związek ten jest nieliniowy.

Współczynnik korelacji rangowej R Spearmana interpretuje się podobnie jak współczynnik korelacji liniowej r Pearsona. Jest on miarą zgodności uporządkowania dwóch szeregów. Do stwierdzenia czy związek jest istotny służy test, w którym hipoteza zerowa mówi, że współczynnik R nieistotnie różni się od zera i oznacza brak związku monotonicznego oraz hipotezę alternatywną mówiącą, że R istotnie różni się od zera, czyli istnieje istotny statystycznie związek monotoniczny między badanymi parametrami. Należy pamiętać, że w przypadku korelacji rangowej nie ma możliwości wyznaczenia współczynnika determinacji ani zbudowania modelu regresyjnego.

Z kolei najbardziej odpowiednim współczynnikiem do oceny miary zależności między parametrami porządkowymi pogrupowanymi na kilka kategorii jest współczynnik gamma. Jest to współczynnik korelacji rangowej uwzględniający przewagę rang wiązanych w tabeli (takich samych rang). Przy jego stosowaniu nie musimy ograniczać się do małych prób, przy tym ma on również zastosowanie w przypadku rang niewiązanych, jednak w takiej sytuacji lepszy jest **współczynnik tau Kendalla**. Współczynnik gamma ma również podobną interpretację. Im jego wartość jest bliższa 1 lub –1, tym silniejszy jest związek między parametrami porządkowymi. Wartość dodatnia tego współczynnika oznacza, że wzrost wartości obu parametrów jest zgodny, czyli obydwa rosną lub maleją. W przypadku współczynnika gamma nie można mówić o zależności liniowej, tylko rozpatrywać ten współczynnik jako miarę zgodności badanych parametrów.

Przy obliczaniu współczynników opisanych powyżej powinno się wyznaczyć dla nich przedziały ufności, aby znać rozpiętość przedziałów ich wartości, co pozwala określić ich przydatność w odniesieniu do całej populacji, z której pobrano próbę.[2-4,7,8,10,11,13,14]

Podsumowanie Statystyka medyczna obejmuje bardzo wiele zagadnień i bardziej zaawansowanych metod statystycznych, jednak wykracza to poza zagadnienia podstaw metodologii biostatystyki opisywanych w tym artykule. Nie przedstawiono zaawansowanych technik statystycznych, takich jak analizy wielowymiarowe, czy analiza szeregów czasowych lub analiza szeregów zdarzeń. Osoby zainteresowane tą tematyką mogą znaleźć informacje na ten temat w specjalistycznej literaturze.[6,9,13-15]

Praktyczne stosowanie statystyki jest możliwe dzięki rozwojowi technik informatycznych. Ze względu na powszechność komputerów i programów statystycznych wykonywanie skomplikowanych obliczeń jest stosunkowo proste i nie wymaga znajomości wzorów matematycznych. Prawidłowa interpretacja wymaga jednak znajomości metodologii analizy danych, a raport i prezentacja graficzna często ułatwiają merytoryczną interpretację i rozumienie uzyskanych wyników, ale nie zwalnia z wykonywania odpowiednich analiz. Często popełniane błędy wynikają przede wszystkim z nieznajomości podstaw metodologii, gdyż każda procedura wymaga spełnienia pewnych założeń. Niewłaściwie stosowane metody statystyczne mogą prowadzić do nieprawdziwych wniosków. Gore i Altman twierdzą, że „...jest rzeczą nieetyczną przeprowadzenie błędnego doświadczenia naukowego. Jednym z aspektów błędnych doświadczeń są niewłaściwie stosowane metody statystyczne”.[2,5] Należy jednak podkreślić, że w interpretowaniu uzyskanych wyników i wyciąganiu wiarygodnych wniosków nie może badacza zastąpić statystyk, gdyż tylko badacz może poprawnie zinterpretować wyniki zgodnie z niezbędną i aktualną wiedzą merytoryczną na dany temat. Wiedza ta jest konieczna również do określenia, który błąd należy minimalizować. Ponadto bardzo istotne jest, aby decyzja o poziomie istotności statystycznej została podjęta przed rozpoczęciem badania, a nie w trakcie jego trwania. Odpowiednia wiedza oparta na praktyce pozwala rozstrzygać, czy uzyskana w wyniku przeprowadzonych analiz istotność statystyczna ma znaczenie istotne klinicznie.

Nieznajomość podstaw metodologii biostatystyki i nieprawidłowe stosowanie testów statystycznych prowadzi do błędnych opinii, że z pomocą statystyki można wszystko udowodnić. W takich sytuacjach aktualne staje się znane stopniowanie kłamstwa, chętnie cytowane przy omawianiu najczęściej nieprawidłowych wniosków, będących właśnie często wynikiem źle zaplanowanego badania, niepoprawnej metodologii lub błędnie stosowanych analiz statystycznych. Dlatego tak bardzo istotna jest podstawowa wiedza przedstawiona w tym opracowaniu.

PIŚMIENNICTWO

1 International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceutical for Human Use. Guidelines: Statistical Principals for Clinical Trials (ICH E-9). http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.

2 Moczko JA, Bręborowicz GH. Nie samą biostatystyką. Poznań: Ośrodek Wydawnictw Naukowych; 2010.

3 Petrie A, Sabin C. Statystyka medyczna w zarysie. Warszawa: Wydawnictwo Lekarskie PZWL; 2006.

4 Watała C. Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych. Bielsko-Biała: a-Medica Press; 2002.

5 Gore SM, Altman DG. Statystyka w praktyce lekarskiej. Warszawa: Wydawnictwo Naukowe PWN; 1997.

6 Armitage P, Berry G, Matthews JNS. Statistical methods in medical reasearch. Blackwell Science Ltd.; 2002.

7 Moczko JA, Bręborowicz GH, Tadeusiewicz R. Statystyka w badaniach medycznych. Warszawa: Springer PWN; 2004.

8 Zieliński T. Jak pokochać statystykę czyli *STATISTICA* do poduszki. StatSoft. Kraków; 1999.

9 Dawson B, Trapp RG. Basic & clinical biostatistics. McGraw-Hill Companies, Inc. 2004.

10 Stanisz A. Biostatystyka. Podręcznik dla studentów medycyny i lekarzy. Kraków: Wydawnictwo UJ; 2005.

11 Stanisz A. Przystępny kurs statystyki z zastosowaniem *STATISTICA PL* na przykładach z medycyny. Tom 1. Statystyki podstawowe. StatSoft Polska. Kraków; 2006.

12 Conover WJ. Practical nonparametric statistics. John Wiley & Sons, 1999.

13 Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. Chapman&Hall/CRC 2007.

14 Stanisz A. Przystępny kurs statystyki z zastosowaniem *STATISTICA PL* na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe. StatSoft Polska. Kraków; 2007.

15 Stanisz A. Przystępny kurs statystyki z zastosowaniem *STATISTICA PL* na przykładach z medycyny. Tom 3. Analizy wielowymiarowe. StatSoft Polska. Kraków; 2007.

16 Jedrychowski W. Zasady planowania i prowadzenia badań naukowych w medycynie. Kraków: Wydawnictwo UJ; 2004.

17 Hanley JA, McNeil B. A method of comparing of the areas under the receiver operating characteristic curve derived from the same cases. Radiology. 1983; 148: 839-843.

18 Jaesche R, Cook D, Guyatt G. Ocena artykułów na temat testów diagnostycznych. Część II – metody określania przydatności testu. Kraków: Medycyna Praktyczna; 1998; 11: 184-191.