Centrum Medyczne Kształcenia Podyplomowego w Warszawie

ELEMENTY OCENY ORGANIZACJI I WYNIKÓW BADAŃ KLINICZNYCH

Redaktor naukowy
dr n. ekon. Michał Jakubczyk
dr n. med. Maciej Niewada

ELEMENTY OCENY ORGANIZACJI I WYNIKÓW BADAŃ KLINICZNYCH

dr n. ekon. Michał JAKUBCZYK
dr n. med. Maciej NIEWADA

Warszawa 2011

Przygotowanie i druk podręcznika współfinansowany przez Unię Europejską
z Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

AUTORZY
Michał Jakubczyk – Instytut Ekonometrii, Szkoła Główna Handlowa
Maciej Niewada – Katedra i Zakład Farmakologii Doświadczalnej i Klinicznej, Warszawski Uniwersytet Medyczny
Wojciech Masełbas – Katedra i Zakład Farmakologii Doświadczalnej i Klinicznej, Warszawski Uniwersytet Medyczny
Łukasz Borowiec

WYDAWCA
Centrum Medyczne Kształcenia Podyplomowego
01-813 Warszawa, ul. Marymoncka 99/103
tel. 22 56 93 700
fax 22 56 93 712
www.cmkp.edu.pl

ISBN 978-83-62110-29-2

Skład, przygotowanie do druku, druk i oprawa
Agencja Reklamowo-Wydawnicza
A. Grzegorczyk
www.grzeg.com.pl

Redaktor techniczny
Grażyna Dziubińska

Spis treści

I. Wprowadzenie

Michał JAKUBCZYK, Maciej NIEWADA

Niniejszy podręcznik powstał przede wszystkim z myślą o lekarzach, którzy czytając publikacje naukowe, a w szczególności opublikowane wyniki badań klinicznych, napotykają wiele sformułowań z obszaru zarówno metodologii organizacji badań klinicznych, jak i statystyki. Przykładowe pojęcia, występujące w części opisującej metody w takich publikacjach, to np.: randomizacja, zaślepienie, skala Jadad, badanie typu *superiority* vs. *non-inferiority*, kalkulacja liczebności próby, ukrycie kody randomizacji, analiza zgodna z zaplanowanym leczeniem, błąd I rodzaju, 90% moc testu, model ryzyka proporcjonalnego Coksa, 95% przedział ufności dla współczynnika ryzyka, analiza wieloczynnikowa, regresja logistyczna itd.

Celem tej książki jest przybliżenie Czytelnikowi najczęściej pojawiających się terminów i zaznajomienie ze sposobem organizacji i interpretacją wyników badań klinicznych. Niniejszy podręcznik należy traktować jako umiejscowiony w obszarze medycyny opartej na aktualnych i wiarygodnych publikacjach (ang. *EBM – evidence based medicine)* oraz oceny technologii medycznych (ang. *HTA – health technology assessment*), tj. w obszarze zbioru technik pozwalających na porównywanie technologii medycznych między sobą, w celu dostarczenia decydentowi informacji do wyboru technologii, które powinny być stosowane.

Nie chodzi przy tym o to, żeby zasypać Czytelnika definicjami wszystkich możliwych terminów i wzorami na wszystkie możliwe testy statystyczne. Nie chodzi też o to, aby zadość uczynić wszelkim formalnościom związanym z naukami statystycznymi. Niektóre tematy wymagałyby znacznie dłuższego i bardziej technicznego opracowania. Za cel stawiamy sobie raczej przedstawienie idei stojących za wybranymi, najczęściej stosowanymi technikami, tak aby wzrósł komfort Czytelnika przy interpretacji wyników badań klinicznych, tj. pewność, że właściwie rozumie założenia organizacji badania i przedstawione rezultaty, a jednocześnie pewniej umie wytyczyć granicę między tym, co łatwe do odczytania i właściwego zinterpretowania (mimo wielości technicznych terminów), a tym co faktycznie wymaga zrozumienia subtelności wnioskowania statystycznego.

W książce w kilku miejscach omówiono najczęściej spotykane przez autorów błędne wyobrażenia na temat znaczenia pojęć, interpretacji wyników, własności metod, itp. Oczywiście

zaproponowana przez nas lista błędów nie dotyczy wszystkich, więc z góry przepraszamy Czytelników, którzy mogliby poczuć się dotknięci sugerowaniem, że wpadliby w przedstawione pułapki.

Ponieważ język angielski jest obecnie *lingua franca* nauki, istnieje potrzeba zrozumienia pojęć technicznych w tym języku. Dlatego w opracowaniu, tam gdzie wydaje się to najbardziej zasadne, podajemy angielskie sformułowania – i tak na przykład pojęcia przedstawione w pierwszym akapicie Czytelnik najprawdopodobniej napotkał w następującej postaci: *randomization, blinding, Jadad scale, superiority vs non-inferiority trial, sample size calculation, allocation concealment, intention to treat analysis, type I error, 90% test power, Cox proportional hazard model, hazard ratio 95% confidence interval, multivariate analysis, logistic regression*.

Niniejszy podręcznik powstał na bazie materiałów wykorzystywanych przez autorów w czasie wykładów prowadzonych w ramach kursów organizowanych przez Polską Unię Onkologiczną. Przedstawione sposoby wyjaśniania i proporcje między formalizmami i intuicją zdawały się odpowiadać uczestnikom tych szkoleń i mamy nadzieję, że zostaną zaakceptowane także przez Czytelnika. W książce podano także przykłady z badań klinicznych z interpretacją. Celowo nie wykorzystano zbyt wielu takich przykładów, aby nie wyrobić nawyku uczenia się interpretacji na pamięć, a raczej zilustrować wprowadzane koncepcje jedną lub dwiema rzeczywistymi sytuacjami.

Istotne jest, że niniejsze opracowanie nie ma na celu przygotowania Czytelnika do samodzielnego prowadzenia analiz statystycznych (ani tym bardziej organizowania badań klinicznych, co odbywa się w dużych zespołach). Oczywiście niektóre omówione tematy mogą pomóc w samodzielnych pracach analitycznych i uniknięciu błędów, ale tylko jako uzupełnienie wiedzy statystycznej dostarczonej przez dedykowane podręczniki statystyki, często uwzględniające sposób pracy z konkretnym oprogramowaniem.

Czytelnikowi, który chciałby rozszerzyć wiedzę możemy polecić kilka pozycji. *Cochrane Handbook for Systematic Reviews of Interventions* jest chyba swoistą biblią oceny jakości badań klinicznych [3]. Znajduje się w niej wiele cennych wskazówek oceny zarówno sposobu przeprowadzenia, jak i raportowania wyników badania. W tym zakresie warto także poruszyć polskie podręczniki, przede wszystkim „Podstawy EBM czyli medycyny opartej na danych naukowych dla lekarzy i studentów medycyny" pod redakcją Brożka, Gajewskiego i Jaeschke [1]. Książka Peat, Barton i Elliott zawiera omówienie podstawowych technik ilościowych z reprodukcjami wielu opublikowanych artykułów i zestawami ćwiczeń – jest rekomendowaną pozycją, dla Czytelników, którzy chcieliby zobaczyć zastosowania statystyki na tle faktycznych opublikowanych badań [4]. Co do zakresu tematycznego pokrywa się z niniejszym opracowaniem, obejmuje dodatkowo omówienie metod stosowanych w epidemiologii oraz przy analizie testów diagnostycznych. Przy czytaniu anglojęzycznych publikacji pomocny może być słownik pojęć autorstwa Burzykowskiego i wsp. [2]. Przystępne omówienie kwestii metod ilościowych w badaniach medycznych zawierają także książki [5,6,7,8].

Aby ułatwić Czytelnikowi dalsze poszukiwania, przedstawiamy listę istotnych w kontekście interpretacji badań klinicznych pojęć, których nie udało się objąć niniejszym opracowaniem. Po pierwsze, jedynie zasygnalizowano kwestie ustalenia liczebności próby. Oczywiście zagadnienie to jest ważniejsze dla organizujących badanie kliniczne, ale umiejętność krytycznej oceny podejścia zastosowanego przez autorów może być przydatna także dla Czytelnika

badania. Po drugie, w świetle znacznej liczby badań klinicznych dotyczących tego samego problemu klinicznego, istotne staje się zagadnienie łączenia wyników kilku badań klinicznych. Wykorzystywane są tu takie metody, jak metaanaliza (ang. *meta-analysis*), metaregresja (ang. *meta-regression*), porównania pośrednie (ang. *indirect comparisons*), metody analizy sieciowej (ang. *network analysis*), itd. Spośród tych zasygnalizowano jedynie metaanalizę, pomijając pozostałe, jako wymagające zbyt rozległego omówienia. Istotne jest, że niniejszy podręcznik został przygotowany w duchu tzw. statystyki częstościowej (ang. *frequentist statistics*). W ilościowej ocenie technologii medycznych (zwłaszcza w kontekście analiz sieciowych) wykorzystywane jest tzw. podejście bayesowskie (ang. *Bayesian statistics*), często bardziej intuicyjne w odbiorze ale rzadziej spotykane w klasycznych analizach statystycznych dla badań klinicznych. Wreszcie w książce nie poruszono bardzo wielu konkretnych zagadnień statystycznych, rodzajów testów i technik, takich jak testy warstwowe (ang. *stratified tests*), metody badań prognostycznych oraz metod diagnostycznych i wiele, wiele innych.

Pierwsze rozdziały mają na celu zaznajomienie Czytelnika z badaniami klinicznymi, to jest wykorzystywanymi metodami oraz zasadami ich przeprowadzania i oceny. Szczególnie dużo miejsca poświęcono jakości i rzetelności badania klinicznego, tak aby Czytelnik po lekturze publikacji mógł ocenić, na ile wyniki tego badania istotnie wpływają lub mogą zmienić praktykę kliniczną. W rozdziale 2 omówiono typy badań klinicznych z uwagi na ich konstrukcję oraz miejsce i znaczenie w fazach rozwoju produktów leczniczych. W rozdziale 3 przedstawiono metodykę badań klinicznych, którą także pod kątem oceny jakości i rzetelności badań klinicznych omówiono w rozdziale 5, przywołując stosowane w tym celu skale. W rozdziale 4 wprowadzono z perspektywy klinicznej Czytelnika do ważnego i trudnego zagadnienia wpływu założeń statystycznych na plan badania klinicznego. Rozdział 6 rozpoczyna część podręcznika skoncentrowaną na metodach ilościowych. W rozdziale tym przedstawiono podstawowe pojęcia z obszaru statystyki – w szczególności statystyki opisowej i wnioskowania statystycznego. Pojęcia te są dalej wykorzystywane i rozszerzane w kolejnych rozdziałach. Rozdział 7 poświęcono kwestiom porównywania dwóch technologii z użyciem tzw. miar EBM, powszechnie spotykanych w badaniach klinicznych przy dychotomicznych punktach końcowych (np. wyleczenie lub brak wyleczenia). W części tej zasygnalizowano także problem metaanaliz. W rozdziale 8 przedstawiono zagadnienia porównywania technologii ze względu na zmienne ciągłe (np. zmiana poziomu hemoglobiny). Rozdział 9 poświęcono kwestiom analizy wieloczynnikowej, tj. jednoczesnej ocenie wpływu wielu zmiennych (np. cech demograficznych) na efekt leczenia. Wreszcie w ostatnim rozdziale omówiono kwestie analizy przeżycia.

Niektóre tematy pojawiają się w więcej niż jednym rozdziale. Dotyczy to na przykład idei testowania hipotez statystycznych, kwestii badań typu *non-inferiority*, wnioskowania w przypadku jednoczesnego testowania wielu hipotez, zaślepienia, randomizacji, oceny i analizy punktów końcowych, itp. Powtórzenia te mają na celu omówienie tych zagadnień z różnych perspektyw – w szczególności bardziej lekarskiej i bardziej statystycznej – a także w różnych kontekstach. Mamy nadzieję, że to raczej ułatwi odbiór tych – czasem niełatwych – tematów, niż znuży Czytelnika.

Bibliografia

1. Brożek, J,; Gajewski, P. Jaeschke, R.: Podstawy EBM czyli medycyny opartej na danych naukowych dla lekarzy i studentów medycyny. Medycyna Praktyczna, 2008.
2. Burzykowski, T.; Kawalec, E.; Kraszewska, E.; Kupść, W.: Angielsko-polski słownik terminów biostatystyki klinicznej. Warszawa, 2009.
3. Higgins, J.P.T.; Green, S. (red.): Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0, March 2011. [dostęp 4 października 2011]. Dostępna w Internecie: http://www.cochrane-handbook.org/
4. Peat, J.; Barton, B.; Elliott, E.: Statistics Workbook for Evidence-based Health Care. Wiley-Blackwell, 2008.
5. Petrie, A.; Sabin, C.; Moczko, J. (red. wyd. pol.): Statystyka medyczna w zarysie. Warszawa, 2006.
6. Stanisz, A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 1. Statystyki podstawowe. Kraków, 2006.
7. Stanisz, A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe. Kraków, 2007.
8. Stanisz, A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe. Kraków, 2007.

II. Eksperymentalne i nieeksperymentalne metody oceny interwencji medycznych

Wojciech MASEŁBAS

2.1. Prospektywne i retrospektywne metody oceny interwencji medycznych

Nieodłączną częścią *Evidence Based Medicine*[1] jest systematyczna ocena technologii medycznych (leków, wyrobów medycznych, testów diagnostycznych, urządzeń medycznych i sprzętu medycznego) [1]. Interwencje medyczne są szerszą kategorią i obejmują również metody leczenia operacyjnego, narzędzia stosowane przez psychologów i psychiatrów (kwestionariusze i skale ocen, metody psychoterapii), określone warunki środowiska (np. cechy klimatu), zachowania (dieta, aktywność fizyczna, stosowanie używek) czy rozwiązania organizacyjne (metody finansowania leku przez ubezpieczyciela, system organizacji kolejek na zabieg czy limity przyjęć). Wszystkie interwencje medyczne można analizować *ex post*, gdy ocena efektu ich wprowadzenia bazuje na skrupulatnej analizie obecnej sytuacji i cofaniu się w czasie w poszukiwaniu ewentualnego czynnika lub interwencji, które mogły doprowadzić do tego stanu w przeszłości. Przykładem może być poszukiwanie czynnika, który odpowiada za występowanie określonej choroby np. nowotworu płuc. Analizując uzyskane informacje możemy stwierdzić, że w przeszłości większość chorych np. paliła ponad 20 papierosów dziennie, nieliczni mieli kontakt z azbestem lub rozpuszczalnikami i farbami nitro. I nawet, jeśli w tej chwili żaden z pacjentów nie pali, to czynnik będący prawdopodobną przyczyną choroby działał na nich w przeszłości i doprowadził do rozwoju choroby.

Metody retrospektywne, to jest gromadzące informacje wstecznie, czyli po zaistnieniu zdarzeń, których analizy nie zaplanowano *a priori*, są bardziej narażone na błąd pominięcia istotnych informacji. Głównym narzędziem badawczym są kwestionariusze i ankiety, które zbierają informacje z przeszłości.

Metody prospektywne opierają się na obserwacji efektów interwencji, która została dokonana w teraźniejszości (w czasie rzeczywistym) zaś ocena jej skutków następuje wraz z upływem czasu (obecnie i w przyszłości). Przykładem może być sprawdzenie, który ze sposobów leczenia jest lepszy. U części chorych stosujemy jeden sposób terapii, u pozostałej inny, a następnie obserwujemy ich skutki i wyciągamy wnioski. W tym przypadku możemy ze znacznie większym przekonaniem stwierdzić, że w badanej populacji chorych z wybraną chorobą jeden z badanych sposobów terapii jest lepszy od drugiego. Kontrolujemy bowiem nie tylko

[1] Evidence Based Medicine (EMB) – czyli medycyna oparta na dowodach naukowych to postępowanie oparte na wiarygodnych i aktualnych publikacjach.

czynniki, których efekty działania oceniamy w badaniu (oba sposoby leczenia), ale również analizujemy te, które działają przypadkowo.

2.2. Badania pierwotne i wtórne

Wyniki badań biomedycznych powinny być publikowane w sposób umożliwiający weryfikację wyciągniętych wniosków. Na ich podstawie powstają rekomendacje dotyczące postępowania w danej jednostce chorobowej, podejmowane są decyzje dotyczące finansowania, refundacji lub wprowadzenia programów lekowych. Powyższe wnioski mogą być oparte na badaniach pierwotnych, w których analizowane są dane otrzymane bezpośrednio od chorych uczestniczących w testach, lub badania wtórne. Badania wtórne bazują na analizie danych uzyskanych we wcześniej prowadzonych próbach. Należą do nich:

- przegląd systematyczny,
- synteza jakościowa,
- metaanaliza,
- porównanie pośrednie.

Istotne znaczenie ma metaanaliza, czyli technika statystycznego łączenia wyników wielu pojedynczych badań klinicznych prowadzonych w tym samym wskazaniu oraz z użyciem tych samych metod oceny. Jej celem jest dostarczenie bardziej precyzyjnych informacji na temat rzeczywistej skuteczności technologii medycznej, która była przedmiotem analizowanych badań. Więcej informacji na ten temat znajduje się w rozdziale 7.

2.3. Metody nieeksperymentalne (obserwacyjne)

Metody oceny interwencji medycznych możemy podzielić na eksperymentalne i nieeksperymentalne (obserwacyjne). W pierwszym przypadku decyzja dotycząca wyboru zastosowanej interwencji, momentu jej wprowadzenia oraz długości stosowania zależy od badacza prowadzącego dany eksperyment medyczny. Badacz decyduje też o wyborze grupy kontrolnej oraz o interwencji, jakiej będzie ona poddana. Metody nieeksperymentalne zakładają bierną obserwację grupy lub grup osób wybranych z populacji, na którą działa dany czynnik, oraz analizę uzyskanych w ten sposób danych. Badacz nie ma przy tym wpływu, na które osoby w grupie działa badany czynnik, ani też nie ingeruje w warunki, w jakich przebiega badanie. Nieeksperymentalne metody oceny interwencji medycznych dzielimy na opisowe i analityczne [1].

2.4. Metody nieeksperymentalne o typie opisowym

Metody opisowe obejmują najprostsze sposoby oceny interwencji medycznych bazujące na obserwacji wystąpienia pewnego zjawiska (np. efektu zdrowotnego) i jego związku z zastosowaniem określonej interwencji. Zaliczamy do nich opis przypadku, opis serii przypadków, badania ekologiczne (epidemiologiczne) oraz rejestry. W większości przypadków są to badania retrospektywne, w których analizowany efekt zdrowotny występuje w czasie rzeczywistym, zaś czynnik sprawczy wystąpił w przeszłości.

2.3.1.1. Opis przypadku i opis serii przypadków

Opis przypadku (ang. *case report*) i opis serii przypadków (ang. *case report series*) dokumentują fakt wystąpienia pewnego efektu zdrowotnego (choroba, wyzdrowienie, pogorszenie, zgon itp.) w związku z zastosowaniem danej interwencji medycznej (leczenie lub jego brak) [4]. Opis serii przypadków zawiera dane dotyczące kilku podobnych zdarzeń. Ułomnością opisów jest brak informacji na temat częstości występowania opisywanych zdarzeń w całej populacji chorych z danym schorzeniem oraz wpływu interwencji na inne osoby chore. Przykładem opisu przypadku może być publikacja opisująca nieoczekiwaną remisję choroby nowotworowej po zastosowaniu leku przeciwbólowego. Ponieważ mechanizm przeciwnowotworowego działania leku oraz ciąg przyczynowo-skutkowy nie są znane, tego typu doniesienie ma bardzo ograniczony wpływ na kształtowanie się wytycznych dotyczących leczenia chorób nowotworowych.

2.3.1.2. Badania ekologiczne (epidemiologiczne)

W badaniach tego typu dokonuje się porównania wskaźników epidemiologicznych, dotyczących zachorowalności, chorobowości czy śmiertelności w odniesieniu do danych statystycznych w skali makro. Ocenie można np. poddać poziom zanieczyszczenia środowiska w poszczególnych regionach kraju (województwach) oraz częstość występowania astmy oskrzelowej czy przewlekłej obturacyjnej choroby płuc w populacji zamieszkującej te obszary. Wyciągane wnioski mogą odnosić się przy tym wyłącznie do miar statystycznych bez możliwości bezpośredniego potwierdzenia związku przyczynowo-skutkowego między występowaniem zanieczyszczeń powietrza za zachorowalnością na astmę.

2.3.1.3. Rejestry

Rejestry (ang. *registries*) to bazy danych odnoszące się do konkretnej jednostki chorobowej (np. nowotwór piersi), typu pacjenta (kobiety z mutacją genu BRCA1 i BRCA2) lub stosowanej interwencji (stosowanie herceptyny u kobiety z nowotworem piersi i nadekspresją antygenu HER2). Są one przydatne do określenia naturalnego przebiegu choroby, skuteczności i bezpieczeństwa prowadzonego leczenia, występujących efektów zdrowotnych. Na podstawie analizy danych zawartych w rejestrach można wyciągać wnioski na temat występowania określonych efektów zdrowotnych w populacji osób objętych rejestrem. Powyższe wnioski można próbować ekstrapolować na całą populację chorych z danym schorzeniem lub problemem zdrowotnym.

2.3.2. Metody nieeksperymentalne o typie analitycznym

Metody nieeksperymentalne o typie analitycznym obejmują: badania przesiewowe, przekrojowe, kliniczno-kontrolne, kohortowe oraz nieinterwencyjne badania kliniczne. Ich wspólną cechą jest brak interwencji podejmowanej przez badacza, a zatem nieeksperymentalny charakter prób. W większości przypadków są to badania prospektywne, w których wraz z upływem czasu analizujemy efekty zdrowotne wywołane pewnym działaniem lub jego brakiem. To, co je odróżnia od metod eksperymentalnych, to fakt, że oceniana interwencja zachodzi w sposób niezależny od badacza i miałaby miejsce nawet wtedy, gdyby badanie nie było prowadzone.

2.3.2.2. Badania przesiewowe

Badania przesiewowe (ang. *screening studies*) to analizy wykonywane przy okazji badań profilaktycznych. Pozwalają na określenie występowania pewnego zjawiska w grupie osób uczestniczących w programie badań profilaktycznych [4]. Przykładem badania przesiewowego może być analiza częstości występowania nowotworu piersi w oparciu o grupę kobiet uczestniczących w profilaktycznych badaniach mammograficznych. Otrzymane wyniki mogą być jak najbardziej zgodne z rzeczywistą częstością występowania choroby w populacji lub też znacznie od niej odbiegać. Na udział w programie badań profilaktycznych częściej mogą decydować się np. osoby zaniepokojone stanem swego zdrowia. Z powyższych powodów istnieją duże ograniczenia w przenoszeniu otrzymanych wyników na cała populację.

2.3.2.2. Badania przekrojowe

W badaniach przekrojowych (ang. *cross-sectional studies*) analizujemy występowanie pewnego zjawiska (np. efektu zdrowotnego) w próbie populacyjnej w określonym punkcie czasowym. Przykładem tego typu badania może być oznaczenie mutacji genu BRCA1 i BRCA2 w populacji kobiet między 18 a 35 rokiem życia, u których między styczniem a czerwcem 2011 wykonano w Instytucie Centrum Onkologii w Warszawie biopsję zmian zlokalizowanych w piersi. Wynikiem przeprowadzonego badania może być określenie występowania mutacji genu BRCA w populacji kobiet między 18 a 35 rokiem życia, u których występowały wskazania do biopsji. Na podstawie uzyskanych danych można również wyciągać wnioski wynikające z porównania wyników badania genetycznego i histopatologicznego. Błędem byłoby natomiast twierdzenie, że wyliczona w ten sposób częstość występowania mutacji genu BRCA1 i BRCA2 przekłada się na całą populację. Wskazania do biopsji stanowią bowiem czynnik predykcyjny, zwiększający prawdopodobieństwo stwierdzenia mutacji genów BRCA1 i BRCA2. Kobiety, które nie miały wskazań do wykonania biopsji będą zapewne miały znacznie mniejsze ryzyko występowania powyższych mutacji.

2.3.2.3. Badania kliniczno-kontrolne

Badanie kliniczno-kontrolne (ang. *case-control studies**) jest analizą retrospektyną i polega na wybraniu grupy osób z daną cechą (*case*) i bez tej cechy (*control*) oraz retrospektywnym zebraniu danych dotyczących ekspozycji badanych osób na pewien czynnik [3]. Są one użyteczne w identyfikacji czynników ryzyka dla rzadko występujących efektów zdrowotnych. Nie dostarczają jednak danych na temat rzeczywistego prawdopodobieństwa wystąpienia danego efektu zdrowotnego zarówno w grupie badanej, jak i kontrolnej. Przykładem może być analiza narażenia na kontakt z azbestem osób z nowotworem płuca i bez nowotworu. Może być ona jedynie tak dokładna, jak doskonała jest pamięć ludzka, a wiadomo, że po latach większość szczegółów umyka, zaś dokładne określenie ekspozycji może być po prostu niemożliwe.

2.3.2.4. Badania kohortowe

Badanie kohortowe (ang. *cohort studies*) jest zazwyczaj badaniem prospektywnym, w którym badacz określa grupę lub grupy osób poddane badaniu (kohorty) oraz działające na nie

* Termin „badanie kliniczno-kontrolne" jest nieco niefortunnym tłumaczeniem angielskiego case-control, ponieważ narzuca kontekst kliniczny, a nie oddaje porównawczego względem każdego przypadku charakteru badania.

czynniki. Wybrana grupa bądź grupy poddawane są następnie obserwacji a wszystkie zdarzenia zdrowotne odnotowywane. W przypadku jednej kohorty jest to badanie bez grupy kontrolnej (porównawczej), a każdy z uczestników badania stanowi kontrolę sam dla siebie. W tak prowadzonym badaniu porównujemy stan zdrowia określony przez pewne mierzalne parametry przed rozpoczęciem badania, w jego trakcie oraz po zakończeniu. W przypadku dwóch lub większej liczby kohort możemy również mówić o porównaniach między grupami [2]. Przykładem badania kohortowego byłaby analiza zachorowalności na nowotwór piersi w grupie 1000 kobiet między 18. a 35. rokiem życia z mutacją genu BRCA. Gdybyśmy, jako drugą grupę (kohortę) obserwowali grupę 1000 kobiet w tym samym przedziale wieku, ale bez powyższej mutacji to moglibyśmy z analizy wyciągnąć wnioski dotyczące wpływu mutacji genu BRCA na ryzyko zachorowania na nowotwór piersi.

2.3.2.5. Badania nieinterwencyjne

Badania nieinterwencyjne (ang. *non-interventional studies*) to według polskiego prawodawstwa odmiana badań klinicznych z użyciem produktu leczniczego, w których:

- produkt leczniczy jest stosowany w sposób określony w pozwoleniu na dopuszczenie do obrotu;
- przydzielenie chorego do grupy, w której stosowana jest określona metoda leczenia, nie następuje na podstawie protokołu badania, ale zależy od aktualnej praktyki, a decyzja o podaniu leku jest jednoznacznie oddzielona od decyzji o włączeniu pacjenta do badania;
- u pacjentów nie wykonuje się żadnych dodatkowych procedur diagnostycznych ani monitorowania stanu zdrowia, a do analizy zebranych danych stosuje się metody epidemiologiczne [5].

Można powiedzieć, że badanie nieinterwencyjne to typ prospektywnego badania kohortowego. To, co je wyróżnia, to fakt, iż analizowanym czynnikiem sprawczym jest lek. Ważne jest również to, że oceniana interwencja nie zależy od badacza i byłaby zastosowana w niezmieniony sposób niezależnie od tego, czy pacjent uczestniczyłby w badaniu klinicznym czy też nie. Przykładem badania nieinterwencyjnego może być analiza liczby erytrocytów u chorych z nowotworem piersi poddanych różnym metodom chemioterapii. Na podstawie rozmazów wykonywanych przed rozpoczęciem leczenia oraz typowo przed każdym cyklem chemioterapii można pokusić się o wyciągnięcie wniosków, który ze stosowanych leków ma najsilniejsze działanie hamujące wytwarzanie czerwonych krwinek.

2.4. Metody eksperymentalne

Badania, w których zastosowanie danej interwencji medycznej zależy od badacza, zaliczamy do metod eksperymentalnych. Badacz jest w stanie określać, w której grupie i kiedy zostanie zastosowany oceniany czynnik oraz kontrolować długość i siłę jego działania (dawkę leku). Eksperymentalne metody oceny interwencji medycznych z użyciem produktów leczniczych oraz wyrobów medycznych są w myśl ustawowych definicji określane mianem badań klinicznych [5,6].

2.4.1. Badania kliniczne produktów leczniczych i wyrobów medycznych

Badanie kliniczne produktu leczniczego to w myśl opisu zawartego w ustawie prawo farmaceutyczne [5] każde badanie prowadzone z udziałem ludzi w celu odkrycia lub potwierdzenia klinicznych, farmakologicznych w tym farmakodynamicznych skutków działania jednego lub wielu badanych produktów leczniczych, lub w celu zidentyfikowania działań niepożądanych jednego lub większej liczby badanych produktów leczniczych, lub śledzenia wchłaniania, dystrybucji, metabolizmu i wydalania jednego lub większej liczby badanych produktów leczniczych, mając na uwadze ich bezpieczeństwo i skuteczność.

Badanie kliniczne wyrobu medycznego to zgodnie z definicją zawartą w ustawie o wyrobach medycznych zaprojektowane i zaplanowane systematyczne badanie prowadzone z udziałem ludzi, podjęte w celu weryfikacji bezpieczeństwa lub działania określonego wyrobu medycznego, wyposażenia wyrobu medycznego albo aktywnego wyrobu medycznego do implantacji [6].

2.4.1.1. Badania kliniczne z grupą kontrolną

Badania kliniczne (ang. *clinical trial, CT*) można prowadzić, jako badania bez grupy kontrolnej (a tym samym bez zaślepienia[2] i randomizacji[3]), w którym każdy z uczestników jest kontrolą dla samego siebie a oceniane efekty zdrowotne analizowane na przestrzeni czasu, w jakiej trwa badanie [4] oraz jako badania kontrolowane (ang. *controlled trial*). Przykładem badania bez grupy kontrolnej może być próba zaplanowana w celu zbierania informacji na temat działań niepożądanych badanego produktu leczniczego, w której każdy uczestnik przyjmuje testowany lek. Skrupulatnie notujemy wszystkie niepożądane zdarzenia występujące po jego użyciu i obliczamy częstość ich występowania w odniesieniu do populacji eksponowanej na działanie tego produktu. Badanie z grupą kontrolną polegałoby na porównaniu częstości występowania danego efektu zdrowotnego (niepożądanych zdarzeń po użyciu badanego produktu leczniczego) w grupie eksponowanej na działanie danej interwencji medycznej (w tym przypadku leku) oraz grupie kontrolnej, która nie była poddana żadnej interwencji lub była eksponowana na działanie innego produktu leczniczego. Istotą badania z grupą kontrolną jest porównanie efektu zdrowotnego między grupami [7]. Badania kontrolowane mają o wiele większe znaczenie dla rzetelności wniosków wyciąganych na podstawie danych uzyskanych z badania i płynących z tego implikacji dla praktyki klinicznej.

2.4.1.2. Badania kliniczne z randomizacją (ang. *randomised clinical trial*, RCT)

W badaniach z grupą kontrolną badacz może wpływać na decyzję, u którego z uczestników badania zostanie zastosowana nowa metoda terapeutyczna, a który będzie leczony tradycyjnie. Stwarza to pokusę takiego wpływania na sposób leczenia by otrzymać wyniki, na których nam zależy. W celu eliminacji czynników zakłócających rzetelną i bezstronną ocenę

---

[2] Zaślepienie to jedna z metod usuwających wpływ badającego i badanego na uzyskane wyniki. Więcej informacji na ten temat znajduje się w rozdziale 5.

[3] Randomizacja to metoda użyta w celu przydzielenia uczestników do grupy badanej lub kontrolnej. Może odbywać się w sposób losowy (analogicznie do rzutu monetą) lub na podstawie skomplikowanych algorytmów obsługiwanych przez komputer. Jest to jedna z metod usuwających wpływ badacza oraz badanego na otrzymane wyniki.

wyników badania wprowadzono metody usuwające wpływ badanego i badającego na otrzymane wyniki [2].

Randomizacja to metoda alokacji uczestnika badania do grupy, w której będzie zastosowana oceniana interwencja medyczna (np. leczonej badanym produktem leczniczym), oraz grupy kontrolnej. Można w tym celu używać metod losowych, opartych na algorytmach czy programach komputerowych lub innych. Więcej informacji na ten temat znajduje się w rozdziale 5.

Każde badanie przebiegające z randomizacją jest jednocześnie badaniem kontrolowanym (muszą w nim być co najmniej dwie grupy – badana i kontrolna).

2.4.1.3. Fazy badań klinicznych produktów leczniczych

Podział badań klinicznych na fazy występuje wyłącznie w odniesieniu do badań klinicznych produktów leczniczych. Nie spotyka się go w przypadku badań klinicznych wyrobów medycznych ani oceny innych interwencji medycznych.

Faza 0 (zero) została wprowadzona stosunkowo niedawno i często jest łączona z fazą I. W badaniach fazy 0 nowa substancja zostaje po raz pierwszy podana człowiekowi w tzw. mikrodawce np. 1/500 dawki (w przeliczeniu na kilogram masy ciała) wywołującej efekt farmakologiczny u najbardziej wrażliwego zwierzęcia uczestniczącego w badaniach przedklinicznych. Jeśli to możliwe używa się drogi dożylnej. Najczęściej uczestnikami badania są zdrowi ochotnicy (gdy a priori wiadomo, że badany produkt może wywołać działania szkodliwe np. cytostatyk to badania fazy 0 i fazy I prowadzone są z udziałem pacjentów z zaawansowaną chorobą i brakiem alternatywy terapeutycznej). Celem fazy 0 jest ocena bezpieczeństwa badanego produktu leczniczego i stwierdzenie, czy można go w ogóle podawać człowiekowi. Celem fazy 0 może być również ustalenie wstępnych danych dotyczących farmakokinetyki: dystrybucji, metabolizmu i wydalania badanej substancji oraz farmakodynamiki: działanie na receptory lub inne punkty uchwytu. W fazie 0 uczestniczy niewielka liczba osób badanych – zwykle kilku ochotników. Często z grupy 4 uczestników tylko jeden otrzymuje testowaną substancję, zaś 3 pozostałych placebo. Negatywne wyniki fazy 0 prowadzą do zaprzestania dalszych badań z daną cząsteczką a tym samych niepotrzebnego narażania uczestników dalszych badań [2].

Faza I badań klinicznych – nowa substancja zostaje po raz pierwszy podana człowiekowi w dawce mającej wywołać działanie farmakologiczne. Pierwsze podanie badanego produktu ma na celu potwierdzenie/ustalenie wstępnych danych dotyczących tolerancji i bezpieczeństwa przyszłego leku. Poszukuje się minimalnej biologicznie aktywnej dawki (*Minimal Biologically Active Dose* MBAD) lub maksymalnej dawki tolerowanej (*Maximal Tolerated Dose* MTD). Projekty fazy I mają również za zadanie uzyskanie danych farmakokinetycznych: absorpcji, dystrybucji, metabolizmu i wydalania badanej substancji. W fazie I uczestniczy stosunkowo niewielka liczba osób badanych – zwykle kilkudziesięciu ochotników [2].

Faza II badań klinicznych – zazwyczaj po raz pierwszy przyszły lek podawany jest osobom chorym [4]. Głównym celem jest potwierdzenie danych zebranych w fazie I oraz uzyskanie podstawowych informacji dotyczących bezpieczeństwa i skuteczności przyszłego produktu leczniczego w populacji osób chorych (jego przyszłych konsumentów). Zasadnicze znaczenie ma także sprawdzenie kilku dawek leku tak, aby ustalić optymalne dawkowanie produktu leczniczego. Liczba pacjentów uczestniczących w tej fazie rozwoju leku sięga kilkuset osób.

W przypadku konieczności użycia placebo w badaniach onkologicznych stosuje się konstrukcje *placebo on the top of standard therapy* lub podaje badany lek/placebo chorym z brakiem alternatywy terapeutycznej. Badania fazy II dzieli się na dwie grupy:

- faza IIa:
  - poszukiwanie dawki terapeutycznej,
  - ocena skuteczności biologicznej,
  - dowód słuszności wyboru celu terapeutycznego (ang. *proof of principle*),
- faza IIb:
  - wstępne dane na temat skuteczności klinicznej – badania, terapeutyczne poznawcze,
  - dowód słuszności koncepcji (ang. *proof of concept*).

Faza III badań klinicznych – celem badań jest potwierdzenie skuteczności i bezpieczeństwa leku w większej populacji pacjentów. Z reguły badania fazy III obejmują chorych ze wskazaniem do farmakoterapii, które później będzie przedmiotem aplikacji o rejestrację produktu leczniczego. Typowo do celów rejestracji potrzebne są dwa podobne badania fazy III, łącznie obejmujące kilkuset do kilku tysięcy uczestników.

Zazwyczaj wyróżnia się wśród nich badania:

- fazy IIIa – prowadzone przed złożeniem wniosku o rejestracje nowego leku
- fazy IIIb – prowadzone po złożeniu wniosku o dopuszczenie produktu leczniczego do obrotu.

W niektórych przypadkach badanie może trwać na tyle długo, że przed jego zakończeniem wniosek zostanie pozytywnie rozpatrzony a lek dopuszczony do sprzedaży. W tym przypadku badanie zmienia swoją klasyfikację na fazę IV [3].

Faza IV badań klinicznych – obejmuje badania wykonywane po rejestracji (dopuszczeniu do obrotu) produktu leczniczego, gdy celem jest poszerzenie wiedzy na temat zastosowania leku w już zaaprobowanych wskazaniach, ocena częstości rzadkich działań niepożądanych, ocena działania leku w wybranych populacjach chorych, badania farmakoekonomiczne wspomagające wniosek o refundację leku. Tutaj klasyfikowane są także badania porównawcze prowadzone metodą *head to head* często prowadzone bez udziału przemysłu np. sponsorowane przez Europejską Organizację do Badania i Leczenia Nowotworów *European Organisation for Research and Treatment of Cancer* (EORTC).

2.5. Klasyfikacja znaczenia doniesień naukowych dla oceny technologii medycznych

Ze względu na wagę wniosków, które mogą być wyciągane w poszczególnych typach badań, ustalono klasyfikację znaczenia doniesień naukowych dla oceny badanych interwencji medycznych. Największe znaczenie przypisuje się w niej metaanalizie wykonanej na podstawie przeglądu badań z randomizacją i grupą kontrolną. Osobna klasyfikacja powstała dla doniesień naukowych dotyczących metod diagnostycznych.

Tabela 2.1. Klasyfikacja doniesień naukowych odnoszących się do diagnostyki [8].

Rodzaj badania	Opis
D I	Przegląd systematyczny badań poziomu D II.
D II	Badania kliniczne oceniające dokładność metody diagnostycznej, w których zastosowano metodę ślepej próby oraz porównano ocenianą metodę diagnostyczną z testem referencyjnym (złotym standardem) w grupie pacjentów z określonym stanem klinicznym kolejno włączanych do badania.
D III-1	Badania oceniające dokładność metody diagnostycznej, w których zastosowano metodę ślepej próby oraz porównano ocenianą metodę diagnostyczną z testem referencyjnym (złotym standardem) w grupie pacjentów z określonym stanem klinicznym włączanych do badania nie w sposób kolejny.
D III-2	Badania porównujące ocenianą metodę diagnostyczną z testem referencyjnym, które nie spełniają kryteriów poziomu D II i D III-1.
D III-3	Diagnostyczne badania kliniczno-kontrolne.
D IV	Badania opisujące wyniki diagnostyczne, bez zastosowania testu referencyjnego.

Tabela 2.2. Klasyfikacja doniesień naukowych odnoszących się do terapii [8].

Typ badania	Rodzaj badania	Opis podtypu
Przegląd systematyczny RCT	IA	Metaanaliza na podstawie wyników przeglądu systematycznego RCT.
	IB	Systematyczny przegląd RCT bez metaanalizy.
Badanie eksperymentalne	IIA	Poprawnie zaprojektowana kontrolowana próba kliniczna z randomizacją.
	IIB	Poprawnie zaprojektowana kontrolowana próba kliniczna z pseudorandomizacją.
	IIC	Poprawnie zaprojektowana kontrolowana próba kliniczna bez randomizacji.
Badanie obserwacyjne z grupą kontrolną	IIIA	Przegląd systematyczny badań obserwacyjnych.
	IIIB	Poprawnie zaprojektowane prospektywne badanie kohortowe z równoczasową grupą kontrolną.
	IIIC	Poprawnie zaprojektowane prospektywne badanie kohortowe z historyczną grupą kontrolną.
	IIID	Poprawnie zaprojektowane retrospektywne badanie kohortowe z równoczasową grupą kontrolną.
	IIIE	Poprawnie zaprojektowane badanie kliniczno-kontrolne (retrospektywne).
Badanie opisowe	IVA	Seria przypadków – badanie pretest/posttest.
	IVB	Seria przypadków – badanie posttest.
	IVC	Inne badanie grupy pacjentów.
	IVD	Opis przypadku.
Opinia ekspertów	V	Opinia ekspertów w oparciu o doświadczenie kliniczne, badania opisowe oraz raporty panelów ekspertów.

Bibliografia

1. Gajewski, P.; Jaeschke, R.; Brożek, J.: Podstawy EBM, Kraków, 2008.
2. Gryfin, J.P.; O'Grady, J.: The Textbook of Pharmaceutical Medicine. Oxford, 2006.
3. Hackshaw, A.: A Concise Guide to Clinical Trials. London, 2009.
4. Hutchinson, D.R.: Dictionary of Clinical Research. Richmond, 1998.
5. Ustawa z dnia 6 września 2001 r. Prawo farmaceutyczne (Dz.U. 2001, Nr 126, poz. 1381 z późn. zm.).
6. Ustawa z 20 maja 2010 r. o wyrobach medycznych (Dz.U. 2010, Nr 107, poz. 679)
7. Wulff, H.; Gotzsche, P.C.: Racjonalna diagnoza i leczenie. Łódź, 2005.
8. Wytyczne oceny technologii medycznych (HTA). Agencja Oceny Technologii Medycznych. Wersja 2.1. Warszawa, 2009.

III. Metodyka prowadzenia badania klinicznego

Wojciech MASEŁBAS

Prawidłowa metodyka prowadzenia badania klinicznego pozwala uniknąć błędów podważających wiarygodność uzyskanych wyników [2]. Pierwszym etapem badania powinien zawsze być przegląd literatury dotyczącej danego problemu. Ze względu na udział ludzi – uczestników badania klinicznego – prowadzenie badania, które już wcześniej zostało przeprowadzone i dostarczyło wiarygodnych i rzetelnych danych na temat ocenianej interwencji medycznej, byłoby bowiem zarówno postępowaniem nieetycznym, jak i nieuzasadnionym pod względem naukowym [1]. Metodyka badania klinicznego zakłada przestrzeganie przepisów prawa, wytycznych medycznych towarzystw naukowych oraz zaleceń etycznych odnoszących się do badanej interwencji, populacji osób mających uczestniczyć w badaniu czy schorzenia będącego przedmiotem badań.

Wymogiem wynikającym zarówno z metodyki, jak i przepisów prawa jest uprzednie przygotowanie protokołu badania, który musi uzyskać pozytywną opinię komisji bioetycznej, a w badaniach klinicznych produktów leczniczych i wyrobów medycznych również Prezesa Urzędu Rejestracji Produktów Leczniczych, Wyrobów Medycznych i Produktów Biobójczych [6]. Znajomość zasad metodyki prowadzenia badań pozwala na wybór takiego schematu badania i takich metod badawczych, by pogodzić interes nauki z zasadami etycznymi i wymogami prawa. Do metodyki prowadzenia badania klinicznego zaliczamy:

- ogólną charakterystykę projektu (ang. *study design*),
- wybór badanej populacji,
- badane interwencje medyczne (w grupie badanej i porównawczej),
- wybór narzędzi badawczych,
- określenie miary wyników,
- ramy czasowe badania,
- wybór metod usuwających wpływ badacza i badanego na otrzymane wyniki.

3.1. Ogólne założenia dotyczące badania klinicznego

Planując badanie kliniczne zakładamy, iż uzyskane wyniki będą nie tylko rzetelne i wiarygodne (wiarygodność wewnętrzna badania[1]) ale również reprezentatywne dla całej populacji osób z danym schorzeniem, które w przyszłości będą leczone przy użyciu testowanego produktu leczniczego (wiarygodność zewnętrzna[1]). Oznacza to, że niepodważalne wyniki prawidłowo prowadzonego badania z udziałem kilkudziesięciu czy kilkuset osób przekładają się na całą populację chorych z daną chorobą. Na ich podstawie agencje regulatorowe dopuszczają do obrotu nowe leki i wyroby medyczne a towarzystwa naukowe wydają rekomendacje postępowania klinicznego.

3.2. Cel badania

Cel badania powinien być precyzyjnie określony w protokole badania klinicznego. Niewystarczającym jest podanie, iż celem badania jest sprawdzenie, czy lek działa albo jak działa. Wybór celu badania w znacznym stopniu określa inne parametry badania wynikające z przyjętej metodyki. Przykładem celu badania klinicznego może być porównanie całkowitego przeżycia w okresie pięcioletnim w grupie chorych z przerzutami nowotworu piersi do kości poddanych co najmniej sześciu kursom chemioterapii przy użyciu paclitakselu (grupa badana) lub doksorubicyny (grupa kontrolna). Tak postawiony cel określa nam badaną populację, porównywane interwencje medyczne (paclitaksel i doksorubicyna), długość okresu obserwacji (5 lat od zakończenia leczenia), miarę oceny skuteczności badanej interwencji (przeżycie) oraz narzuca wybór metodyki badania (dwie równoległe grupy).

Jeszcze bardziej precyzyjnie określonym celem będzie ocena, czy łączne stosowanie paclitakselu i doksorubicyny wydłuża czas do niepowodzenia terapii (ang. *time to treatment failure*) w porównaniu z paclitakselem lub doksorubicyną stosowanymi w monoterapii w grupie chorych z przerzutową postacią nowotworu piersi. Mimo że na pierwszy rzut oka nie wszystko wydaje się tu jasne, to znajomość metodyki pozwala na stwierdzenie, iż badaną populacją będą chore z przerzutową postacią nowotworu piersi, w badaniu będziemy mieli 3 równoległe grupy i 3 testowane interwencje (paclitaksel z doksorubicyną, paclitaksel w monoterapii i doksorubicyna w monoterapii). Miarą skuteczności leczenia będzie czas do niepowodzenia terapii. Długość badania będzie uzależniona od przyjętych założeń statystycznych (opisywanych w rozdziale 4).

3.3. Opis badanej populacji

Badana populacja określana jest w protokole badania klinicznego poprzez kryteria włączenia i wykluczenia (ang. *inclusion and exclusion criteria*) [3]. Muszą one być na tyle precyzyjne, by nie powodować wątpliwości interpretacyjnych i nie prowokować błędów określanych jako odstępstwa od protokołu (ang. *protocol deviations*)[2]. W badaniu klinicznym zazwyczaj mogą

[1] Więcej informacji na ten temat znajduje się w rozdziale 5.

[2] Liczne odstępstwa od protokołu powodują, że otrzymane wyniki pochodzą od grupy chorych znacznie różniącej się od przyjętych założeń. Jeśli określimy, że w badaniu uczestniczą osoby powyżej 75. r.ż., zaś z powodu prob-

uczestniczyć zarówno chorzy do tej pory nieleczeni (ze świeżo postawioną diagnozą), jak i osoby, u których do tej pory stosowane leczenie było nieskuteczne. W celu uniknięcia wpływu uprzednio stosowanych leków na oceniane efekty zdrowotne[3] stosuje się u nich fazę wypłukania (ang. *wash out*) [3]. Polega ona na czasowym, choć znacznie dłuższym niż 5 okresów półtrwania, odstawieniu leków w celu usunięcia ich efektów klinicznych. Wypłukanie jest nieskuteczne w przypadku preparatów o bardzo długim czasie półtrwania, długo utrzymującym się efekcie farmakologicznym oraz przypadkach, w których efekty zdrowotne pojawiają się z dużym opóźnieniem od podania leku. Wypłukanie uprzednio stosowanego leku pozwala na sprowadzenie do tego samego mianownika i poddanie spójnej ocenie 2 grup chorych, które bez tej procedury znacznie by się między sobą różniły.

3.4. Wybór interwencji w grupie kontrolnej

Zdecydowana większość badań klinicznych prowadzona jest metodą porównawczą – badacz porównuje w nich średnią odpowiedź na lek uzyskaną w grupie badanej z grupą kontrolną[4]. Zarówno wybór grupy kontrolnej, jak i interwencji medycznej stosowanej w tej grupie zależą od metodyki badania klinicznego [6]. Największe znaczenie dla rekomendacji dotyczących przyszłego stosowania ocenianej interwencji medycznej mają wyniki badań, w których grupa badana i grupa kontrolna były niemal identyczne pod względem analizowanych parametrów badania (jak wiek, płeć, zaawansowanie choroby, wskaźniki biochemiczne, histopatologiczne i genetyczne, wcześniej stosowane metody terapii itp.) [1]. Możemy wtedy z dużym prawdopodobieństwem powiedzieć, że jedyna różnica, jaka w trakcie badania wystąpiła między grupami, wynikała z zastosowanych interwencji medycznych (badanej oraz tej jakiej użyliśmy do porównania). Porównanie obu interwencji jest zatem pozbawione wpływu innych czynników.

Zarówno z metodycznego, jak i etycznego punktu widzenia ważnym jest, by interwencja medyczna stosowana w grupie kontrolnej była odpowiednia do celów badania. Oznacza to, że w badaniach fazy I lub IIa, patrz rozdział II, zastosowanie placebo w grupie kontrolnej będzie miało o wiele większą akceptację niż w badaniach fazy IIb i III. Jednocześnie względy etyczne zazwyczaj wykluczają możliwość użycia placebo w schorzeniach onkologicznych, chorobach zakaźnych, zespołach metabolicznych (jak np. cukrzyca) [1]. W przypadku konieczności użycia placebo w badaniach onkologicznych stosuje się konstrukcje *placebo on the top of standard therapy,* gdy wszyscy chorzy otrzymują leczenie standardowe a ponadto grupie badanej podajemy testowany lek a grupie kontrolnej placebo [2]. Z punktu widzenia późniejszych rekomendacji terapeutycznych największe znaczenie mają badania, w których nowatorską interwencję medyczną (np.

---

lemów z rekrutacją pacjentów będziemy włączać do badania również pięćdziesięciolatków, to wyciągane wnioski nie mogą odnosić się do grupy osób powyżej 75. r.ż., gdyż będzie ona stanowiła jedynie podgrupę w całej populacji osób uczestniczących w badaniu.

[3] Wpływ uprzednio stosowanych leków na oceniane efekty zdrowotne nazywa się dryfem lub efektem z przeniesienia (ang. *carry over effect*). Nałożenie się efektu obu leków (stosowanego uprzednio oraz podawanego obecnie) powoduje duży problem w ocenie skuteczności działania każdego z nich.

[4] W większości przypadków operujemy średnią wraz z przedziałem ufności.

nowy lek) porównuje się ze standardem terapii lub lekiem z tej samej grupy terapeutycznej, który na dodatek jest uznawany za najlepszy w swojej klasie. Tego typu badania często określane mianem *head to head* najlepiej odpowiadają na pytanie, który z dwóch ocenianych leków jest lepszy. Należy jednak pamiętać, że wyciągnięte wnioski mogą odnosić się do populacji pacjentów, którzy uczestniczyli w badaniu klinicznym, oraz innych chorych, którzy odpowiadają ich charakterystyce. Jeśli badanie było prowadzone w grupie osób między 18. a 65. r.ż., to na podstawie jego wyników nie można z pełnym przekonaniem stwierdzić, że oceniany lek będzie działał tak samo skutecznie lub będzie tak samo bezpieczny jeśli podamy go osobom po 80. r.ż.

3.5. Hipotezy badania klinicznego

Niniejszy rozdział zawiera jedynie wprowadzenie do szerokiej dyskusji na temat problemu związanego z wyborem hipotez badania klinicznego. Więcej informacji na ten temat znajduje się w rozdziale 6.

Każdy projekt klinicznej oceny interwencji medycznej rozpoczyna się od sformułowania problemu badawczego oraz najbardziej prawdopodobnego ogólnego rozwiązania, czyli hipotezy badawczej. Poprawne sformułowanie hipotezy w dużej mierze przesądza o sukcesie badawczym. Hipoteza powinna być tak sformułowana, by łatwo można ją było przyjąć lub odrzucić. W badaniach biomedycznych powszechnie stosowana jest zasada falsyfikacji. Polega ona na postawieniu hipotezy, zwanej hipotezą zerową, która jest niejako zaprzeczeniem celu badania. Chcąc udowodnić, że badana interwencja jest lepsza niż dotychczas stosowane leczenie, w hipotezie zerowej zakładamy, że różnica między analizowanymi parametrami lub rozkładami wynosi zero. Dopiero po jej odrzuceniu możemy przyjąć hipotezę alternatywną będącą potwierdzeniem celu prowadzonego badania klinicznego (czyli, że nowy lek jest lepszy). Zasady falsyfikacji używa się w celu zmniejszenia ryzyka uzyskania zupełnie przypadkowych wyników oraz wyciągnięcia na ich podstawie wniosków dotyczących badanej interwencji w całej populacji chorych z danym schorzeniem. Należy ponadto pamiętać, że przyjęta hipoteza badawcza pozwala na wyciąganie wniosków wyłącznie w zakresie tych parametrów badania, które zostały w niej określone.

3.5.1. Hipoteza zerowa

Typowym przykładem hipotezy zerowej w badaniu testującym skuteczność nowej metody leczenia może być brak różnicy między efektami zdrowotnymi w grupie chorych poddanych badanej interwencji medycznej oraz grupie kontrolnej. Odrzucenie hipotezy zerowej, czyli występowanie różnicy między analizowanymi parametrami po użyciu dwóch różnych sposobów leczenia pozwala na rozważenie możliwości przyjęcia hipotezy alternatywnej. Należy też pamiętać, że test statystyczny nie jest dowodem prawdziwości czy fałszywości hipotezy. Za pomocą testu można jedynie albo odrzucić hipotezę zerową, albo też orzec, że wyniki doświadczenia jej nie przeczą.

W niektórych przypadkach (np. w badaniach typu badana interwencja jest równoważna – *equivalence study*[5]) hipoteza zerowa będzie brzmiała – między badanymi lekami jest różnica w zakresie parametrów będących przedmiotem badania (np. parametrów farmakokinetycz-

[5] Opis typów badania ze względu na przyjęte założenia statystyczne zawarty jest w rozdziale 4.

nych). Prawidłowe przeprowadzenie badania, zebranie danych oraz ich skrupulatna analiza pozwalająca na rzetelne stwierdzenie, że różnica jest na tyle niewielka, iż mieści się w pewnym akceptowalnym zakresie – pozwala na stwierdzenie, że badane leki są równoważne. Więcej informacji na temat typów badań znajduje się w rozdziale 4.

3.5.2. Hipoteza alternatywna

Hipoteza alternatywna jest zwykle potwierdzeniem celu badania klinicznego. Jeśli na podstawie analizy literatury oraz wcześniej przeprowadzonych badań przedklinicznych i klinicznych zakładamy, że łączne stosowanie paclitakselu i doksorubicyny będzie wiązało się ze zwiększeniem o 5 punktów procentowych liczby pacjentów zaliczanych do grupy „przeżycie bez choroby" (ang. *disease free survival*) i porównujemy je z paclitakselem lub doksorubicyną stosowanymi w monoterapii, a badanie prowadzone jest w grupie chorych po mastektomii z powodu nowotworu piersi, to możemy w ten sposób sformułować hipotezę alternatywną. Występowanie różnicy między 3 analizowanymi grupami (paclikaksel plus doksorubicyna *versus* paclitaksel w monoterapii oraz doksorubicyna w monoterapii) pozwala na odrzucenie hipotezy zerowej. Przyjęcie hipotezy alternatywnej wymaga ponadto potwierdzenia, że w grupie leczonej paclitakselem łącznie z doksorubicyną w trakcie trwania badania rzeczywiście zanotowano mniejszą o co najmniej 5 punktów procentowych liczbę nawrotów choroby nowotworowej. Warto zauważyć, że przyjęty punkt końcowy – przeżycie bez choroby – może być użyty wyłącznie w przypadku chorych z dobrym rokowaniem, zatem błędem metodycznym byłoby określenie kryteriów włączenia i wyłączenia pozwalających na udział w badaniu również chorych z przerzutową postacią choroby, dla których rokowanie byłoby znacznie gorsze. Przy nierównomiernym rozkładzie osób ze złym rokowaniem między grupą badaną i kontrolną nasz projekt byłby już na samym początku naznaczony błędem selekcji (ang. *selection bias*)[6].

3.6. Metoda grup równoległych

Metoda grup równoległych (ang. *parallel*) jest najczęściej stosowanym schematem badania klinicznego. Polega na takim podziale uczestników badania klinicznego na grupy badaną i kontrolną, by na etapie rozpoczęcia badania różnice między nimi były minimalne. Następnie w obu grupach równolegle stosujemy odpowiednie interwencje medyczne (np. badany produkt leczniczy w grupie badanej zaś w grupie kontrolnej lek standardowo stosowany w danym schorzeniu) i porównujemy efekty tych interwencji. Zaletą tego typu badania jest możliwość oceny efektu leczenia u każdego uczestnika oddzielnie (ocena typu *pretest* i *posttest*), jak również porównania średniego efektu leczenia między grupami [3].

Metoda grup równoległych może być użyta do porównania badanej interwencji z brakiem leczenia, z placebo czy ze standardowo stosowanym lekiem. Możliwe jest również porównanie farmakoterapii oraz inwazyjnej metody leczenia czy chemioterapii z radioterapią.

[6] Więcej informacji na temat błędu wyboru oraz innych typów błędów znajduje się w rozdziale 5.

Przy bardziej skomplikowanej konstrukcji badania możemy starać się porównać efekt chemioterapii lub radioterapii w grupie chorych z nowotworem płuca poddanych uprzednio leczeniu operacyjnemu lub nie. W takim przypadku będziemy mieli 4 grupy i odpowiednio dopasowywane do tej konstrukcji hipotezy badawcze.

3.7. Metoda grup naprzemiennych

W schemacie badania z grupami naprzemiennymi (ang. *cross over*) również mamy dwie grupy, jednak obie interwencje stosowane są w nich naprzemiennie. Kolejność i długość ich stosowania oraz częstość zmian stosowanej terapii określona jest w protokole badania. W zależności od użytych leków oraz schorzenia, które jest przedmiotem badania, pomiędzy poszczególnymi seriami leczenia może następować faza wypłukania (ang. *wash out*) lub też nie. Przykładem badania z użyciem metody grup naprzemiennych może być analiza skuteczności dwóch leków przeciwwymiotnych stosowanych w premedykacji leczenia przeciwnowotworowego. Protokół badania może zakładać zmianę sposobu leczenia po każdych trzech kursach chemioterapii, a ponieważ leki przeciwwymiotne podawane są jedynie przez kilka dni, to nie ma potrzeby umieszczania w protokole fazy wypłukania. W badaniu z użyciem metody grup naprzemiennych możliwe jest porównanie skuteczności jednego i drugiego leku u każdego uczestnika badania klinicznego. Zazwyczaj pozwala to na zmniejszenie liczby uczestników, gdyż eliminujemy zmienność pomiędzy pacjentami (ang. *between patients variation*). Ze względu na ryzyko efektu przeniesienia metody grup naprzemiennych nie stosuje się w przypadku preparatów o bardzo długim czasie półtrwania, długo utrzymującym się efekcie farmakologicznym oraz przypadkach, w których efekty zdrowotne pojawiają się z dużym opóźnieniem od podania leku. Metoda grup naprzemiennych nie powinna być również stosowana w przypadkach, w których objawy choroby (a tym samym mierzalne parametry skuteczności badanej terapii) są na tyle zmienne, że ich odczyt w różnych przedziałach czasu (ang. *between period variation*) może powodować u tego samego pacjenta na tyle duże różnice, co zmienność między pacjentami. W tych przypadkach preferowany jest model porównania grup równoległych [1].

3.8. Alternatywne modele badań klinicznych

W literaturze spotyka się alternatywne modele badań spełniających ustawową definicję badania klinicznego. Należą do nich m.in.:

- badanie z elementami metody grup naprzemiennych,
- badanie z pseudorandomizacją,
- model badania klinicznego „n of 1".

W badaniu z elementami metody grup naprzemiennych mamy dwie grupy chorych (badaną i kontrolną) poddane dwóm różnym interwencjom medycznym. Badana interwencja jest metodą o nieznanej skuteczności, zaś w grupie kontrolnej podawany jest lek o udowodnionym działaniu i dużej efektywności. Ze względów etycznych w planie badania zawieramy możliwość by w przypadku niepowodzenia terapii pacjenci leczeni przy użyciu eksperymentalnej metody mieli możliwość przejścia do grupy leczonej standardowo i dalszego udziału w badaniu. Zmiana w drugą stronę nie jest możliwa.

W badaniu z tzw. pseudorandomizacją pacjenci są jeszcze przed wyrażeniem zgody na udział w badaniu alokowani do grupy badanej lub kontrolnej. Chorzy z grupy badanej, którzy wyrażą zgodę na udział w próbie klinicznej otrzymują leczenie eksperymentalne. Grupie kontrolnej podaje się leczenie standardowe. Pacjenci uprzednio alokowani do grupy badanej, którzy nie wyrażą zgody na przyjmowanie eksperymentalnej terapii, mają możliwość dołączenia do grupy kontrolnej i otrzymać leczenie standardowe [5].

Badania w modelu „n of 1" to randomizowane, prowadzone metodą ślepej próby[7] wielokrotne porównania badanego produktu leczniczego zazwyczaj z placebo dokonywane u tego samego pacjenta (n razy u 1 uczestnika). W pewnym sensie przypominają one model badania z użyciem metody grup naprzemiennych. Badania tego typu są szczególnie przydatne w sytuacji, gdy lekarz ma do czynienia ze szczególnym przypadkiem klinicznym i nie jest pewien skuteczności danego leku. Wielokrotne porównanie ocenianego preparatu z placebo (lub innym lekiem) może dostarczyć niezmiernie cennych informacji na temat jego działania. Przykładem badania w modelu n of 1 może być próba kliniczna oceniająca skuteczność leku przeciwwymiotnego w porównaniu z placebo w leczeniu bólów migrenowych [4].

3.9. Ocena badanej interwencji i punkty końcowe badania

Skuteczność i bezpieczeństwo stosowania badanej interwencji medycznej winniśmy oceniać w sposób ustalony w protokole badania. Musimy w nim *a priori* określić, jakie zdarzenia zdrowotne będziemy traktować jako mierzalne efekty działania leku. Nazywamy je punktami końcowymi badania (ang. *end points*). Punkty końcowe dzielimy na pierwotne i wtórne.

Pierwotny punkt końcowy badania (ang. *primary end point*) to taki, który został zapisany w hipotezie badania [1]. W oparciu o prawidłowo przeprowadzone badanie twierdzenie dotyczące punktów końcowych zawartych w hipotezie badania może być bardziej kategoryczne. Dla przykładu, jeśli stosowanie leku A ma zmniejszać częstość występowania anemii (definiowanej, jako obniżenie poziomu Hb poniżej 12 g/dL) u chorych poddawanych chemioterapii, to pierwotnym punktem końcowym badania będzie poziom Hb mierzony laboratoryjnie. Jako wtórne punkty końcowe mogą być zdefiniowane: liczba przetoczeń krwi w grupie badanej oraz kontrolnej, wydolność organizmu definiowana, jako wynik sześciominutowego marszu określany w metrach czy ogólna jakość życia. Wykryty przy okazji tak zaprojektowanego badania wpływ testowanego leku na poziom glikemii jest zupełnie przypadkowym odkryciem i nie może być traktowany, jako dowiedziony z taką samą pewnością co pierwotny punkt końcowy. Więcej informacji na temat punktów końcowych badania znajduje się w rozdziale 5.

Punkty końcowe muszą odpowiadać celom, które określiliśmy na początku badania. Siła rekomendacji przygotowanych na podstawie badania klinicznego zależy od jakości punktów

[7] Metoda ślepej próby pozwala na ukrycie przed pacjentem (metoda pojedynczej ślepej próby) oraz pacjentem i lekarzem (metoda podwójnej ślepej próby) informacji, który lek jest stosowany. Pozwala to na usunięcie subiektywnego wpływu badanego i badacza na możliwość wypaczenia wyników badania. W przypadku leków zaślepienie osiąga się poprzez takie ich przygotowanie by wyglądały identycznie. Przy porównaniu różnych dróg podania lub różnych metod leczenia używamy metody maskowania. Więcej informacji na ten temat znajduje się w rozdziale 5.

końcowych. Największe znaczenie przypisuje się tzw. twardym punktom końcowym (ang. *hard end points*), które są najbardziej wiarygodne, gdyż subiektywność ich oceny jest ograniczona. W badaniu przeżywalności twardym punktem końcowym jest zgon z jakichkolwiek przyczyn. Inne możliwe określenia punktów końcowych, które będą zaliczane do twardych to:

- zgon związany z chorobą np. zgon z powodu choroby nowotworowej,
- odsetek wyleczeń w okresie 5 letnim (określany jako brak wznowy).

Mimo ewidentnej wiarygodności twardych punktów końcowych dla celów rekomendacji tworzonych na podstawie wyników badania nie zawsze udaje się zaplanować badanie z ich wykorzystaniem. W niektórych przypadkach byłoby to nieetyczne, gdyż np. we wczesnych fazach badań produktów leczniczych o testowanym leku wiemy zbyt mało, by planować badanie z twardym punktem końcowym jak zgon z jakichkolwiek przyczyn [1]. W tych przypadkach musimy zadowolić się zastępczymi/miękkimi punktami końcowymi (ang. *surrogate end points*), które mogą być dobrze skorelowane z twardym punktem końcowym.

Jako zastępcze punkty końcowe można wymienić:

- wielkość guza nowotworowego,
- częstość hospitalizacji,
- poziom markera skorelowanego z chorobą nowotworową.

W niektórych przypadkach zupełnie nie mają one związku z całkowitym przeżyciem, choć mogą pozwolić na wyciągnięcie wniosków dotyczących skuteczności testowanej terapii. Znane są również przypadki, w których entuzjastyczna ocena skuteczności nowej terapii dokonana w oparciu o zastępcze punkty końcowe rozpadała się w proch po weryfikacji z użyciem twardych punktów (patrz rozdział 5). Dla przykładu chemioterapia w nieoperacyjnym nowotworze płuca może prowadzić do zmniejszenia wielkości zmian nowotworowych, ale jednocześnie zwiększa ryzyko powikłań zatorowo-zakrzepowych, a tym samym może zwiększać całkowitą śmiertelność.

W ścisłym związku z wybranymi punktami końcowymi pozostaje metoda, jakiej będziemy używać do oceny badanej interwencji medycznej. W dużym uproszczeniu można ją sprowadzić do:

- metod opartych na liczeniu osób (dane binarne, nominalne):
  - przeżycie lub zgon (liczba osób żyjących i zmarłych),
  - hospitalizacja lub brak,
  - wyleczenie lub nie (w oparciu o obiektywne kryteria),
  - ORR (ang. *Objective Response Rate*) – odsetek pacjentów, u których wystąpiła pełna poprawa (ang. *complete response CR*) lub częściowa poprawa (ang. *partial response PR*);
- metod opartych na liczeniu efektu u badanych osób (dane ciągłe, przedziałowe):
  - wielkość guza nowotworowego,
  - poziom markera choroby nowotworowej np. PSA,
  - liczba dni hospitalizacji,
  - liczenie czasu do wystąpienia odpowiedzi (np. czas od randomizacji do wystąpienia progresji zmian guza nowotworowego – ang. *time to tumor progression* TTP).

By zobiektywizować ocenę punktów końcowych zależnych od subiektywnych opinii, często tworzy się jednolite kryteria np. często używane w onkologii *Response Evaluation Criteria In Solid Tumors (RECIST):*

- pełna poprawa (ang. *complete response CR*) – zanik wszystkich mierzalnych zmian nowotworowych,
- częściowa poprawa (ang. *partial response PR*) – 30% redukcja sumy najdłuższych wymiarów wszystkich mierzalnych zmian nowotworowych,
- progresja choroby (ang. *progressive disease PD*) – 20% wzrost sumy najdłuższych wymiarów wszystkich mierzalnych zmian nowotworowych,
- choroba stabilna (ang. *stable disease SD*) – niewielkie zmiany wielkości guzów nowotworowych nie pozwalające na ocenę w oparciu o powyższe kryteria.

Pozwalają one na dokonanie bardziej obiektywnej oceny, co ma szczególne znaczenie w badaniach wieloośrodkowych, w których dane do analizy mogą pochodzić od wielu różnych badaczy.

Tabela 3.1. Różne typy punktów końcowych najczęściej stosowane w onkologii – opracowanie własne.

Endpoint	**Punkt końcowy**	**Miara**	**Komentarz**
Overall survival	Przeżycie	Zgon z jakiejkolwiek przyczyny	Łatwy do określenia
Disease free survival	Przeżycie bez choroby	Nawrót choroby	U chorych z dobrą prognozą
Event free survival	Przeżycie bez zdarzenia	Wystąpienie specyficznego zdarzenia związanego z chorobą	Jak wyżej – prognoza nie musi być tak dobra
Progression free survivial	Przeżycie bez pogorszenia	Pierwszy objaw progresji choroby	Użyteczny w przypadkach zaawansowanej choroby
Disease specific survival	Przeżycie bez zgonu z powodu choroby	Zgon z powodu specyficznej przyczyny związanej z chorobą	Wymaga poznania dokładnej przyczyny zgonu
Time to treatment failure	Czas do niepowodzenia terapii	Pierwszy objaw progresji choroby	Użyteczny w przypadku zaawansowanej choroby

3.10. Czas trwania badania klinicznego

Długość, częstość oraz intensywność stosowania danej interwencji w grupie badanej i kontrolnej są ściśle określone w protokole badania klinicznego. Nie można rzetelnie porównać dwóch podobnie działających antybiotyków w terapii zapalenia płuc, jeśli jeden byłby podawany przez 5 dni a inny przez 10 dni. Dokonana po zakończeniu leczenia ocena zapewne wykazałaby znacznie mniejszą liczbę wyleczeń w grupie stosującej lek A przez 5 dni niż w przypadku 10-dniowej terapii lekiem B. Błędem byłaby również ocena stanu zdrowia wszystkich chorych w piątym dniu badania, bo część z nich jest dopiero w połowie terapii. Zatem oba leki w tym przypadku powinny być stosowane tak samo długo, zaś ocena wyleczeń dokonana w tym samym przedziale czasu od rozpoczęcia leczenia.

Zupełnie prawidłowe pod względem metodycznym byłoby natomiast porównanie jednej dawki leku o przedłużonym działaniu i podawanego podskórnie raz w miesiącu z tabletkami, które muszą być stosowane codziennie przez 30 dni. Oczywiście tak jak poprzednio musi ono nastąpić w odpowiednim przedziale czasowym od rozpoczęcia terapii, czyli po co najmniej 30 dniach lub wielokrotności tego okresu.

Są jednak przypadki, w których długość badania określona jest w zupełnie inny sposób i bez użycia skali czasowej. W protokole badania może np. pojawić się zapis, że jest ono prowadzone do momentu wystąpienia pewnej liczby zdarzeń określanych, jako punkty końcowe badania. W badaniu z udziałem chorych z przerzutami do kości mogą to być patologiczne złamania w miejscu występowania zmian nowotworowych. Jeśli w odpowiedni sposób zaplanujemy badanie pod względem metodycznym, dobierzemy wielkość badanych grup oraz określimy klinicznie istotną różnicę skuteczności między porównywanymi metodami terapii, to możliwe będzie również określenie, przy jakiej liczbie zdarzeń będzie ich wystarczająco dużo by stwierdzić, że badany lek działa[8].

W przypadku niektórych badań randomizowanych z podwójnie ślepą próbą (szczególnie sponsorowanych przez przemysł farmaceutyczny) w trakcie trwania projektu dokonuje się oceny etapowej (*interim analysis*). Polega ona na ocenie odkodowanych (i ujawniających przyjmowaną terapię) danych przez grono niezależnych ekspertów. W przypadku stwierdzenia, że jedno z ramion terapii jest zdecydowanie lepsze niż drugie mogą oni wydać rekomendację o przedwczesnym zakończeniu badania z powodów etycznych. Dalsze kontynuowanie próby narażałoby bowiem chorych leczonych lekiem mniej skutecznym na niepotrzebne ryzyko.

Bibliografia

1. Gryfin, J.P., O'Grady, J.: The Textbook of Pharmaceutical Medicine. Oxford, 2006.
2. Hackshaw, A., A Concise Guide to Clinical Trials. London, 2009.
3. Hutchinson, D.R., Dictionary of Clinical Research. Richmond, 1998.
4. Mahon, J., Lapaucis, A., Donner, A., Wood, T.: Randomised study of n of 1 trials versus standard practice. Br Med. J. 1996, 312. 1069-1074.
5. Nowakowska, M., Model badania klinicznego. W: Walter, M., (red.) Badania kliniczne. Organizacja, nadzór, monitorowanie. Warszawa, 2004.
6. Ustawa z dnia 6 września 2001 r. Prawo farmaceutyczne (Dz.U. 2001, Nr 126, poz. 1381 z późn. zm.).

[8] Więcej informacji na temat założeń statystycznych badania znajduje się w rozdziale 4.

IV. Wpływ założeń statystycznych na plan badania klinicznego

Wojciech MASEŁBAS

Wyniki każdego badania klinicznego analizowane są zarówno z punktu widzenia oceny statystycznej, jak i oceny klinicznej [5]. Te same punkty widzenia powinny być brane pod uwagę podczas planowania badania. Dyskusja – czy to względy metodyczne projektowanego badania wpływają na wybór założeń statystycznych, czy też założenia statystyczne narzucają wybór określonego modelu badania – jest niemal tak stara jak spór o to, co było pierwsze – jajko czy kura. Faktem jest, że w dobrze zaplanowanym badaniu zarówno jego plan, jak i założenia statystyczne muszą ze sobą współgrać. Ze względu na to, że badanie kliniczne jest pewnego rodzaju eksperymentem o nie do końca znanym przebiegu i zupełnie nieznanym wyniku, może się oczywiście okazać, że przyjęte założenia były błędne. Należy jednak dopełnić wszelkich starań, by podczas planowania uwzględnione zostały wszystkie możliwe czynniki a ryzyko błędu obniżone do minimum. Przyjęte założenia statystyczne wraz z uzasadnieniem oraz opis metod analizy danych uzyskanych w trakcie badania stanowią jeden z najważniejszych rozdziałów protokołu badania klinicznego.

Bardziej formalne wprowadzenie do kwestii statystycznych przedstawiono w rozdziale 6. Poniżej omówiono te zagadnienia bardziej z perspektywy medycznej, tak aby zacząć budować u Czytelnika odpowiednie rozumienie tych pojęć.

4.1. Determinanty wyboru założeń statystycznych

Jak wspomniano w rozdziale 3 pierwszym krokiem w przygotowaniu planu badania klinicznego jest zdefiniowanie celu [1]. Precyzyjnie określony cel projektu pozwala na dobór odpowiedniej metodyki badania. Znajomość literatury odnoszącej się do wcześniejszych badań w tym samym wskazaniu, wiedza z zakresu danej dziedziny terapeutycznej oraz umiejętność określenia rokowania znacznie zwiększają szanse na dokonanie właściwego wyboru założeń statystycznych. Znacznie upraszczając – chodzi o dobranie takich paramentów badania (li-

czebność badanych grup, czas obserwacji, miara oceny skuteczności i bezpieczeństwa stosowanych interwencji), by po jego przeprowadzeniu można było powiedzieć, że między badanymi grupami występuje różnica pozwalająca na odrzucenie hipotezy zerowej a tym samym przyjęcie hipotezy alternatywnej.

4.1.1. Wybór punktów końcowych

W przytoczonym w rozdziale 3 przykładzie badania klinicznego prowadzonego w celu oceny, czy łączne stosowanie paclitakselu i doksorubicyny wydłuża czas do niepowodzenia terapii (ang. *time to treatment failure*) w porównaniu z paclitakselem lub doksorubicynyną stosowanymi w monoterapii w grupie chorych z przerzutową postacią nowotworu piersi, miarą skuteczności leczenia będzie czas do niepowodzenia terapii. Jako punkt końcowy badania należy zatem zdefiniować każdy stan prowadzący do zaprzestania leczenia (łącznie ze zgonem, wystąpieniem działań niepożądanych uniemożliwiających dalsze leczenie, przerwaniem udziału w badaniu klinicznym połączonym z zaprzestaniem chemioterapii i niestawianiem się na wizyty kontrolne) lub jakikolwiek objaw świadczący o progresji choroby (pojawienie się nowych przerzutów lub wzrost średnicy już istniejących zmian). W przypadku tak określonego celu czas trwania badania będzie w dużej mierze uzależniony od znajomości danych statystycznych dotyczących częstości i czasu wystąpienia niepowodzenia terapii podczas leczenia paclitakselem lub doksorubicynyną stosowanymi w monoterapii, który można określić w oparciu o dotychczasowe publikacje.

4.1.2. Wybór interwencji stosowanej w grupie porównawczej

Planując badanie kliniczne zakładamy, że w grupie badanej stosowana będzie testowana interwencja medyczna [2]. Interwencję stosowaną w grupie porównawczej staramy się wybrać odpowiednio do celów badania. Pod uwagę bierzemy zarówno względy kliniczne, etyczne a nierzadko również ekonomiczne. Dla przykładu realny brak dostępności leku uważanego za obecnie najlepszy sposób terapii, wynikający z braku refundacji czy programu lekowego, powoduje, że jego użycie, jako produktu referencyjnego może dostarczyć danych mało istotnych pod względem klinicznym. Gdy zgodnie z przewidywaniami okaże się, że badana interwencja jest mniej skuteczna od leku uważanego za najlepszy w danym wskazaniu, to wciąż otwarte będzie pytanie, czy jest lepsza od leczenia, które jest realnie dostępne a tym samym oferowane wszystkim pacjentom. Oczywiście komisja bioetyczna może mieć tu inne zdanie i stanąć na stanowisku, że niedopuszczalnym jest proponowanie pacjentom udziału w badaniu klinicznym, w którym oceniane mają być dwa leki, o których widomo, że są mniej skuteczne od innej teoretycznie dostępnej i zdecydowanie lepszej terapii.

W przypadku programu badań i rozwoju nowego produktu leczniczego pod uwagę bierzemy również fazę badania klinicznego [3]. We wczesnych fazach badań (I i IIa) porównanie z placebo (oczywiście, jeśli jest ono akceptowalne pod względem etycznym) może być z punktu widzenia metodyki lepszym rozwiązaniem niż prowadzenie badania *head to head* ze standardem terapii [4]. Ze względów etycznych placebo nie byłoby akceptowalne w leczeniu z wyłączeniem chorób onkologicznych, zakażeń bakteryjnych czy wirusowych, schorzeń metabolicznych i wielu innych. Można natomiast myśleć o jego użyciu w schorzeniach alergicznych, łagodnej depresji, świądzie i innych chorobach, w których tzw. efekt placebo

(obserwowane po podaniu placebo korzystne efekty terapeutyczne) wynika z subiektywnego, indywidualnego nastawienia chorego i jego psychiki. [1]. Wykazanie wyższości nad placebo nie będzie oszałamiającym sukcesem pod względem klinicznym, ale może być osiągnięte przy narażeniu znacznie mniejszej liczby uczestników badania na ryzyko związane z udziałem w próbie. Ze względu na mały zakres wiedzy, jakim dysponujemy na temat badanego produktu, wydaje się to być o wiele bardziej sensownym rozwiązaniem niż prowadzenie porównania ze standardem terapii. Wybór interwencji stosowanej w grupie porównawczej implikuje również konieczność użycia specyficznych rozwiązań metodycznych. W przypadku chęci porównania w badaniu podwójnie zaślepionym (ang. *double blind*) różnych metod terapii (np. chemioterapii i radioterapii), czy choćby leków różniących się drogą podania i częstością dawkowania, musimy stosować metodę zwaną maskowaniem (ang. *double dummy*). Polega ona na poddaniu każdemu pacjentowi zarówno jednej, jak i drugiej terapii, ale w grupie badanej zastosowane będzie aktywne leczenie testowanym lekiem a jednocześnie placebo imitujące leczenie porównawcze. W grupie kontrolnej to leczenie porównawcze będzie aktywne, zaś testowany produkt będzie placebo. By badanie było uznane za podwójnie zaślepione, osoba oceniające efekty leczenia nie może wiedzieć, jakie leczenie otrzymał każdy z pacjentów. Gdy chcemy w badaniu podwójnie maskowanym porównać chemioterapię z radioterapią, musimy każdemu pacjentowi podać wlew dożylny (w grupie badanej aktywny, w grupie kontrolnej placebo) oraz wykonać naświetlania zmian nowotworowych (w grupie kontrolnej prawdziwe, w grupie badanej udawane – bez włączania lampy, ale ze wszystkimi innymi procedurami jak tatuaż ułatwiający pozycjonowanie, właściwe ułożenie pacjenta, itp.). W przypadku porównania chemioterapii z metodą leczenia operacyjnego, w badaniu podwójnie maskowanym konieczne będzie wykonanie u chorych z grupy kontrolnej tzw. udawanej operacji (ang. *sham operation*) polegającej na przecięciu powłok ciała ale bez usuwania guza nowotworowego. Tak jak i w poprzednio opisanym przykładzie osoba oceniająca efekty leczenia nie może wiedzieć, jakie leczenie otrzymał każdy z pacjentów. Ten wymóg nie musi być przestrzegany w badaniach z pojedynczym maskowaniem.

4.1.3. Określenie progu istotności klinicznej (oczekiwanej różnicy między grupami)

W przykładzie z rozdziału 3 na podstawie danych literaturowych oraz wcześniej przeprowadzonych badań przedklinicznych i klinicznych przyjęliśmy założenie, że łączne stosowanie obu leków będzie, o co najmniej 5 punktów procentowych (por. rozdział 6 przypis 18) bardziej skuteczne niż monoterapia. Jest to miara oczekiwanej różnicy w skuteczności terapii mierzonej liczbą zanotowanych punktów końcowych między grupą badaną (leczoną łącznie paclitakselem i doksorubicyną) oraz kontrolną (otrzymującej paclitaksel lub doksorubicynę). W celu rozwiania wątpliwości w założeniach statystycznych powinniśmy określić, że powodzeniem badania będzie uzyskanie wyniku o co najmniej 5 p.p. lepszego niż w przypadku leku z grupy kontrolnej.

4.2. Typy badań klinicznych określone charakterem hipotezy zerowej

Celem badania klinicznego jest sprawdzenie na podstawie przeprowadzonego eksperymentu pewnej tezy, którą stawiamy na podstawie dotychczas posiadanej wiedzy. Teza odnosi się do tej interwencji medycznej, którą chcemy ocenić. Z uwagi na zastosowaną metodykę badania

zazwyczaj oceniamy skuteczność i bezpieczeństwo testowanej interwencji w porównaniu do jednej lub nawet kilku innych metod terapii. Zarówno z punktu widzenia etyki, jak i zdrowego rozsądku zupełnie bezcelowym byłoby prowadzenie badania mającego udowodnić, że nowa interwencja (nowy lek, wyrób medyczny, technika operacyjna itp.) są gorsze niż dotychczas stosowane metody. Przyjętą praktyką jest zatem konstrukcja badań określanych jako:

- badania typu „badana interwencja jest lepsza" (ang. *superiority study*),
- badania typu „badana interwencja nie jest gorsza" (ang. *non-inferiority study*),
- badania typu „badana interwencja jest równoważna" (ang. *equivalence study*).

4.2.1. Badania typu badana interwencja jest lepsza

W tym typie badania zakładamy, że oceniana interwencja medyczna jest skuteczniejsza niż leczenie zastosowane w grupie kontrolnej. Celem badania klinicznego jest udowodnienie tego założenia. Ponieważ wyniki badania klinicznego analizowane są z punktu widzenia oceny statystycznej, koniecznym będzie określenie parametrów tej ocen. Zazwyczaj zadowala nas przyjęcie wartości p na poziomie niższym niż 0,05. W praktyce oznacza to, że przy prawidłowym przeprowadzonym badaniu ryzyko, iż otrzymane wyniki będą zupełnie przypadkowe a tym samym będą błędnie dowodziły prawdziwości postawionej tezy nie będzie większe niż 5%. Matematyczny zapis hipotez w przypadku badania typu „badana interwencja jest lepsza", przedstawia się następująco:

Hipoteza zerowa

H_0: T = S – skuteczność testowanego leczenia (T) jest równa skuteczności leczenia standardowego (S)

Hipoteza alternatywna

H_1: T > S – skuteczność testowanego leczenia (T) jest większa od skuteczności leczenia standardowego (S)

Badania w konstrukcji typu badana interwencja jest lepsza prowadzone są zwykle we wczesnych fazach badań klinicznych produktów leczniczych a szczególnie, gdy w grupie porównawczej stosowane jest placebo.

4.2.2. Badania typu badana interwencja nie jest gorsza

W badaniu opartym na konstrukcji typu badana interwencja nie jest gorsza zakładamy, że testowany lek nie będzie gorszy niż terapia wybrana do porównania. W tym przypadku w grupie kontrolnej stosujemy lek uznany za standard terapii, preparat uznany za najbardziej skuteczny, ewentualnie najlepszy w danej grupie terapeutycznej. W tego typu badaniu często nie mamy żadnych podstaw by oczekiwać, że nasza nowatorska metoda terapii będzie lepsza od leku użytego jako produkt referencyjny (porównawczy). Zakładamy zatem, że będziemy w pełni usatysfakcjonowani, gdy okaże się, że nie jest ona gorsza (nie jest mniej skuteczna) niż najbardziej skuteczny lek wybrany do porównania.

Jednocześnie musimy określić, jaka różnica w skuteczności między porównywanymi terapiami będzie dla nas akceptowalna pod względem klinicznym. Tę różnicę oznacza się zazwyczaj jako δ (delta) i określa jako próg istotności klinicznej. Oznacza ona maksymalną dopuszczalną różnicę między grupami na niekorzyść badanej interwencji, by można było uznać, że rzeczywiście nie jest ona gorsza z klinicznego punktu widzenia.

Wartość δ zależy od dziedziny terapeutycznej, konkretnego wskazania, skuteczności interwencji stosowanych w grupie kontrolnej, przyjętej miary skuteczności oraz wybranych punktów końcowych.

Jeśli na przykład wartość δ określimy na poziomie 5 p.p. zaś w badaniu klinicznym udowodnimy, że stary, dobrze znany i bezpieczny lek jest jedynie o 4 p.p. bardziej skuteczny w leczeniu świądu skóry niż innowacyjny, bardzo drogi i obarczony dużą liczbą działań niepożądanych produkt, to będzie to sukces naszego badania.

Matematyczny zapis hipotez w przypadku badań typu badana interwencja nie jest gorsza wygląda następująco.

Hipoteza zerowa

H_0: $T \leq S - \delta$ – skuteczność testowanego leczenia (T) jest gorsza niż skuteczność leczenia standardowego (S) o co najmniej δ.

Hipoteza alternatywna

H_1: $T > S - \delta$ – skuteczność testowanego leczenia (T) nie jest gorsza niż skuteczność leczenia standardowego (S) o wartość większą niż δ.

4.2.3. Badania typu badana interwencja jest równoważna

Tego typu konstrukcję spotyka się niemal wyłącznie w badaniach równoważności biologicznej. Celem badania jest wykazanie, że badany lek (zazwyczaj produkt generyczny) jest równoważny pod względem parametrów farmakokinetycznych z produktem referencyjnym (lek innowacyjny, dla którego wygasła ochrona patentowa). Głównym porównywanym parametrem jest całkowita wielkość wchłoniętego leku mierzona jako pole pod krzywą wykresu stężenia leku w surowicy krwi na linii czasu. Wytyczne wspólnotowe określają, że dla większości leków różnica całkowitej ilości wchłoniętego leku mieszcząca się w zakresie 80% do 125% pozwala uznać oba leki za równoważne. Różnica będzie w takim przypadku wynosiła od minus 20% do plus 25%. Hipoteza zerowa badania typu badana interwencja jest równoważna mówi, że porównywane leki różnią się między sobą bardziej niż ustalony zakres wartości. Jej odrzucenie pozwala na przyjęcie hipotezy alternatywnej określającej, że różnica między lekami zawiera się w określonym przedziale.

4.3. Określenie liczebności badanych grup

Określenie wielkości badanych grup jest doskonałym przykładem dokumentującym twierdzenie, że względy kliniczne projektowanego badania wpływają na założenia statystyczne i *vice versa*. Wielkość badanych grup zależy bowiem od:

- przyjętej dopuszczalnej wielkości błędu I rodzaju[1],
- przyjętej mocy testu[2],

[1] Błąd I rodzaju to błąd polegający na odrzuceniu hipotezy zerowej, mimo że jest ona prawdziwa. Określa się go poprzez wartość α zazwyczaj ustalaną na poziomie 5%. Więcej informacji na ten temat znajduje się w rozdziale 6.

[2] Moc testu to poziom akceptacji niewykrycia różnicy, która rzeczywiście istnieje, zazwyczaj określana na poziomie M= 1-β równym 80%. Oznacza to, że na 100 porównań badacz akceptuje ryzyko, że w 20 przypadkach (w 20% przypadków) nie stwierdzi istniejącej różnicy.

- zakładanej (często na podstawie wcześniejszych badań) różnicy między grupami, dla której chcemy ustalić moc testu,
- zmienności cechy uznanej za mierzalny efekt terapii w populacji (np. SD[3]).

Liczebność próby będzie tym większa:

- im mniejsze wartości przyjmą wartości α i β oznaczające akceptowane prawdopodobieństwo popełnienia błędów I i II rodzaju,
- im mniejsza będzie klinicznie istotna różnica między grupami, dla których ustalana jest moc testu – tę czasem oznacza się Δ (nie należy mylić z δ),
- im większa będzie zmienność wyników uzyskanych od różnych pacjentów uczestniczących w badaniu klinicznym (określana na podstawie danych literaturowych oraz wiedzy klinicznej).

Liczebność próby dla przykładu dwóch grup i porównywania zmiennej ciągłej określona jest wzorem[4]:

$$n = \frac{2\, SD^2 f(\alpha, \beta)}{\Delta^2}$$

gdzie f(α, β) to wartość funkcji zależnej od wartości α i β odczytana z tabel [1].

Dla przykładu załóżmy, że w badaniu oceniającym skuteczność nowego leku w terapii nadciśnienia tętniczego uznaliśmy, że przewidywana różnica między średnim ciśnieniem rozkurczowym między grupą badaną i kontrolną wynosi Δ=5 mm Hg, przyjęliśmy ryzyko popełnienia błędu I rodzaju na poziomie α = 5%, zaś błędu II rodzaju na poziomie β równym 20%. Załóżmy dalej, że rozrzut wyników otrzymanych w wyniku pomiaru ciśnienia tętniczego u różnych pacjentów mierzony odchyleniem standardowym będzie równy SD=10 mm Hg. Liczebność próby w tak określonym badaniu wyniesie wówczas: $n = 2 \times 10^2 / 5^2 \times 7{,}9 = 63{,}2$, czyli w zaokrągleniu 64 pacjentów w każdej z grup. Tak wyliczoną liczbę często się zwiększa, o czym mowa w dalszej części tego rozdziału.

4.4. Założenia statystyczne a przypadki złamania protokołu oraz wycofań z badania klinicznego

Mimo usilnych starań badanie kliniczne nigdy nie przebiega zgodnie z idealistycznymi założeniami. Zawsze może się okazać, że z analizy musimy wyłączyć pacjentów, którzy po weryfikacji kryteriów włączenia i wyłączenia okazują się nie spełniać kryteriów udziału w badaniu (ang. *protocol violations*) [4].

Z tego powodu zazwyczaj wyliczoną wielkość próby zwiększamy o co najmniej kilku uczestników, by nie okazało się, że przeprowadzone badanie nie pozwala na wyciągnięcie

[3] SD to odchylenie standardowe, szerzej opisywane w rozdziale 6.

[4] Jest to jeden z najprostszych przykładów na pokazanie skomplikowanego procesu ustalania liczebności próby. Osoby zainteresowane tym tematem mogą również skorzystać np. z możliwości oferowanych przez University of Iowa dotyczących kalkulacji niezbędnej wielkości próby przy pomocy kalkulatorów dostępnych on-line http://www.stat.uiowa.edu/~rlenth/Power/

wniosków ze względu na zbyt małą liczbę obserwacji. Więcej informacji na ten temat znajduje się w rozdziale 5.

Nieco innym problemem jest kwestia wycofań z badania klinicznego (ang. *study drop outs*). Z punktu widzenia etyki pacjent ma prawo w każdym momencie wycofać zgodę na udział w badaniu a tym samym nie uczestniczyć w dalszym leczeniu oraz zaplanowanych wizytach kontrolnych zbierających dane do analizy statystycznej. Przenosiny do innego miasta lub nawet kraju także nie są czymś niespotykanym i tworzą grupę określaną mianem *lost to follow up* [4]. W odniesieniu do niektórych uczestników badania lekarz może podjąć decyzję o ich przedwczesnym wyłączeniu z testów (ze względu na działania niepożądane, brak skuteczności terapii lub inne względy medyczne) i zaproponować leczenie standardowo stosowane w danym schorzeniu. Protokół badania musi w takim przypadku określać, jak traktować chorych, od których nie udało się zebrać danych niezbędnych do wykonania ostatecznej analizy. W zalecanej analizie w grupach wyodrębnionych zgodnie z protokołem badania (ang. *per protocol analysis*) ocenie poddane będą dane tych pacjentów, którzy uczestniczyli w wizycie kończącej badanie [3] (patrz rozdział 5).

Bibliografia

1. Gryfin, J.P.; O'Grady, J.: The Textbook of Pharmaceutical Medicine. Oxford, 2006.
2. Gajewski, P.; Jaeschke, R.; Brożek, J.: Podstawy EBM, Kraków, 2008.
3. Hackshaw, A.: A Concise Guide to Clinical Trials. London, 2009.
4. Hutchinson, D.R.: Dictionary of Clinical Research. Richmond, 1998.
5. Wulff, H.; Gotzsche, P.C.: Racjonalna diagnoza i leczenie. Łódź, 2005.

V. Ocena metodologicznej jakości badania klinicznego – wybrane aspekty

Maciej NIEWADA

5.1. Wprowadzenie

Na ocenę jakości (rzetelności) wyników badania klinicznego składa się naprawdę wiele elementów. Z tej przyczyny należy traktować jakość badania klinicznego jako „zmienną ciągłą". Trudno dzielić badania kliniczne jedynie na dobre, to jest rzetelne, i złe, których wyniki nie przynoszą żadnych cennych informacji. Nawet bardzo słabej jakości badanie może być źródłem ciekawych i przydatnych informacji lub przyczynkiem do pogłębionych studiów. Czasami z uwagi na uwarunkowania kliniczne jesteśmy ograniczeni w zakresie możliwych do zastosowania metod badawczych (np. podwójnie ślepa próba w przypadku zabiegu chirurgicznego). Z drugiej strony nie ma badań „idealnych" i nawet w tych opublikowanych w najbardziej prestiżowych periodykach można dopatrzeć się, lub autorzy sami o nich wzmiankują, ułomności i ograniczeń.

Z uwagi na dość złożony charakter oceny badania klinicznego, który jest przedmiotem osobnych, często bardzo obszernych, jak np. *Cochrane Handbook for Systematic Reviews of Interventions* [1], opracowań, w niniejszym rozdziale ograniczymy się do przedstawienia wybranych elementów oceny rzetelności badania klinicznego, zwracając szczególną uwagę na te, które odnoszą się do elementów bardziej statystyczno-ilościowych, a nie *stricte* klinicznych. Dlatego zrezygnujemy z formalnego podejścia do oceny dowodów klinicznych i ich przydatności w podejmowaniu decyzji w praktyce klinicznej (klasyczne podejście EBM), a skupimy na kwantum pierwotnej informacji, którą stanowi pojedyncze pierwotne badanie kliniczne (zrezygnujemy z prezentacji opracowań wtórnych, to jest przeglądów systematycznych i metaanaliz, które omówiono w rozdziale 7).

5.2. Jakie błędy można popełnić w badaniu klinicznym, których zagrożenia należy być świadomym?

Istotą oceny jakości badań klinicznych jest określenie błędów systematycznych to jest nieprzypadkowych[1], powtarzających się wpływów (zafałszowań; ang. *bias*), które powodują zniekształcenie wyników badania i podważają jego rzetelność. W zakresie błędów, ich rodzajów i klasyfikacji, można przytoczyć wiele opracowań; odnosząc się jedynie do najważniejszych typów błędów systematycznych należy rozróżnić:

- błąd selekcji (ang. *selection bias*) – błąd doboru chorych do porównywanych grup, który powoduje różnicę w ich charakterystyce,
- błąd przeprowadzenia lub nazywany także błędem związanym ze znajomością interwencji (ang. *performance bias*) – błąd wynikający z różnic w postępowaniu, opiece nad chorymi w porównywanych grupach lub z odmiennej ekspozycji na czynniki związane z opieką medyczną, które mogą wpływać na wynik terapeutyczny,
- błąd pomiaru, detekcji lub nazywany także błędem związanym z oceną punktów końcowych (ang. *detection bias*) – błąd wynikający z różnic w sposobie pomiaru i oceny punktów końcowych,
- błąd utraty – związany z wycofywaniem się lub wykluczeniem chorych z badania (ang. *attrition bias*) powodujący powstanie istotnych różnic w zakresie liczby i charakterystyki chorych, którzy ukończyli całe badanie.[2]

W celu zmniejszenia wpływu błędów na wyniki, badania często mają określony charakter, sposób przeprowadzania, odwołują się do specyficznych technik czy też wymagają szczególnego sposobu analizy statystycznej i prezentacji danych. Ich ocena pozwala na określenie rzetelności otrzymanych wyników.

Od rzetelności/jakości badania klinicznego należy odróżnić jego wiarygodność. W zakresie tego ostatniego pojęcia rozróżniamy wiarygodność wewnętrzną i zewnętrzną badania. Wiarygodność wewnętrzna dotyczy stopnia, w jakim otrzymane wyniki i sformułowane wnioski odpowiadają rzeczywistemu związkowi między analizowanym postępowaniem a jego wpływem na oceniane efekty terapeutyczne. Z kolei wiarygodność zewnętrzna jest pojęciem wykraczającym poza samo badanie kliniczne i odpowiada możliwości uogólniania wniosków badania na całą docelową populację chorych, u których dana technologia ma być zastosowana w warunkach rutynowej praktyki klinicznej. Wiarygodność wewnętrzna jest więc pojęciem bardziej związanym z jakością przeprowadzenia badania klinicznego, natomiast wiarygodność zewnętrzna dotyczy sposobu zaprojektowania badania z myślą o możliwości uogólniania jego wyników. To właśnie protokół badania (często istotnie odbiegający od rutynowej praktyki klinicznej w zakresie postępowania zarówno diagnostycznego, jak i terapeutycznego) oraz kryteria włączenia i wyłączenia (powodujące dobór do badania określonej grupy chorych) powodują, że badanie ma wpływ na rekomendacje i sposób postępowania oraz liczbę chorych nim objętych.

---

[1] Dla odróżnienia od błędu losowego, który jest przyczyną niedokładności pomiaru.

[2] Pomijajmy błąd publikacji (ang. *reporting bias*) wynikający w asymetrii publikacji badań z pozytywnymi i negatywnymi wynikami, który dotyczy przeglądów systematycznych i metaanaliz.

5.3. Randomizacja – dlaczego jest najważniejsza?

Najważniejszym mechanizmem ograniczającym błąd selekcji i mającym fundamentalne znaczenie dla jakości badania klinicznego jest randomizacja, czyli losowy dobór chorych do porównywanych grup (nie należy mylić randomizacji z losowym doborem chorych do badania!; procedurze randomizacji jest poddany chory spełniający kryteria włączenia i wyłączenia, który wyraził zgodę na udział w badaniu i jest następnie losowo przydzielany do jednej z porównywanych grup). Jest to jedyny sposób na uzyskanie na początku badania podobnych, nie tylko w zakresie znanych, ale także nieznanych cech mogących wpływać na wynik badania, grup chorych. Jak do tej pory nie udało się w tym zakresie znaleźć innego, lepszego, nierodzącego dylematów etycznych (losowość i jej nieprzewidywalność nie są przez wszystkich akceptowane w przypadku zdrowia i życia ludzkiego) sposobu prowadzenia badania eksperymentalnego. Chociaż w tym miejscu warto odnotować, iż coraz częściej pojawiają się badania wykorzystujące inne techniki statystyczne (ang. *propensity score matching*, metoda zmiennej instrumentalnej), które mogą odtwarzać warunki quasi randomizacyjne, jednak bazują one na znanych zmiennych i nie zapewniają możliwości pełnej kontroli wpływu nieznanych czynników i ich rozkładu w porównywanych grupach [2-6].

Randomizacja zależy przede wszystkim od docelowej liczby chorych w badaniu. W przypadku małych badań prawdopodobieństwo nierównego rozłożenia liczby i cech chorych w grupach jest duże i z tej przyczyny nie można polegać jedynie na prostej randomizacji (np. komputerowej liście liczb losowych lub tabeli liczb losowych). Konieczne staje się „wspomaganie" randomizacji (bez ograniczania jej losowego charakteru) i wykorzystanie technik określanych jako ograniczona randomizacja (ang. *restricted randomisation* [7]), jak na przykład:

- randomizację blokową (lub permutowanych bloków) – o przydziale do grupy decyduje losowanie nie dla każdego chorego z osobna, ale losowanie bloków, czyli równolicznych grup (bloków) o określonej sekwencji przydziału kolejnych chorych do każdej z badanych grup); ponieważ w blokach jest zawarty określony stosunek liczby chorych zakwalifikowanych do każdej z grup, ten typ randomizacji zapewnia możliwość kontrolowania ostatecznej liczby chorych w grupach,
- randomizację warstwową (ang. *stratified randomisation*) – randomizacja odbywa się niezależnie w każdej z góry określonej warstwie – grupie chorych, np. w każdym z ośrodków, w grupie mężczyzn i osobno w grupie kobiet, itp.
- randomizację adaptacyjną (ang. *adaptive randomisation*) – prawdopodobieństwo przydziału do danej grupy zmienia się w trakcie trwania badania tak, aby przez cały czas jego trwania kontrolować rozkład cech w porównywanych grupach. Najprostszym przykładem w przypadku badań z jedną grupą kontrolną jest randomizacja parami, w której z dwójki chorych o tej samej charakterystyce, pierwszy jest w pełni losowo przydzielany do jednej z grupy, a drugi chory trafia do przeciwnej (jego przydział do grupy jest więc ustalony losowo w przypadku randomizacji pierwszego chorego z pary). W przypadku randomizacji adaptacyjnej można jednak stosować dużo bardziej wyrafinowane, wspomagane komputerowo, metody losowej alokacji chorych do grup. Warto zwrócić uwagę, iż randomizacja adaptacyjna może dotyczyć nie tylko cech chorych określonych na początku badania, ale także może być podyktowana efektem terapeutycznym (ang. *response-adaptive randomization* lub *outcome-adaptive randomization*). W tym przypadku

prawdopodobieństwo przydzielenia chorego do grupy wzrasta, jeśli obserwowano w niej lepsze efekty terapeutyczne, co jest szczególnie cenne z perspektywy etyki. Ten rodzaj randomizacji nie jest jednak powszechnie stosowany z uwagi na inne ograniczenia.

Z perspektywy oceny rzetelności badania klinicznego znaczenie ma określenie skuteczności i poprawności randomizacji. Co do poprawności, to jest sposobu jej przeprowadzenia, w publikacjach znajdujemy dość szczątkowe informacje, które często nie ułatwiają oceny zaplanowania i przeprowadzenia randomizacji. Określenie „komputerowa" wydaje się być słowem kluczem często wymienianym, nie dającym jednocześnie żadnych szczegółów odnośnie wykorzystanego typu randomizacji. Z tej przyczyny istotna jest przynajmniej ocena skuteczności randomizacji, czyli zapewnienie wejściowej, na początku badania, bardzo zbliżonej, niemal identycznej charakterystyki chorych. W prawie każdej publikacji z badania klinicznego pierwsza tabela zestawia charakterystykę chorych w porównywanych grupach. Raportowane miary centralne (najczęściej średnia) i rozproszenia (najczęściej SD lub 95% przedział ufności) powinny być porównywalne, najlepiej niemal identyczne, w przypadku skutecznej randomizacji. Jeśli tak nie jest, to mógł wkraść się systematyczny błąd lub szczególnie w przypadku małego badania klinicznego, mimo wysiłków badaczy na etapie planowania badania, odnotowane różnice są przypadkowe. W przypadku odmienności w charakterystyce chorych na początku badania w porównywanych grupach zawsze należy poddać ocenie zasadność dalszego porównywania tych grup i konsekwencji dla stopnia zafałszowania wyników. Można także odwołać się do bardziej zaawansowanych metod regresji, które umożliwiają porównanie obu grup i korektę o wpływ wyjściowych różnic w charakterystyce chorych na uzyskane wyniki badania.

Z randomizacją, szczególnie jej przeprowadzeniem i implementacją, wiąże się także pojęcia utajnienia lub ukrycia kodu randomizacji bądź informacji o przydziale chorego do grup badanych (ang. *allocation concealment*). Jest to fundamentalna zasada, której niespełnienie łamie losowy charakter przydziału chorych do grup i polega na tym, że najczęściej badacz lub inne osoby świadomie kierują przydziałem chorych do grup. Można z praktycznego punktu *allocation concealment* zdefiniować jako procedurę, która ochrania proces randomizacji i uniemożliwia dostęp do informacji o grupie, do której został przydzielony pacjent przed włączeniem go do badania, a także w trakcie jego trwania. Procedurę tę zapewnia centralna randomizacja, niezależna od ośrodka, wykorzystanie numerów opakowań leków, na których nie znajdują się żadne dodatkowe, specyficzne informacje o zawartości, itp. Czasami pełna realizacja *allocation concealment* nie jest możliwa, na przykład z uwagi na specyficzne dla leku działania niepożądane.

Dość oczywistym jest stwierdzenie, iż w przypadku świadomego kierowania przydziałem chorych do grup nie sposób uniknąć zafałszowania i uzyskać rzetelne wyniki badania. Potwierdzają to analizy badań, w których *allocation concealment* nie był opisany lub był niewystarczający i w przypadku których częściej raportowano korzystne działanie w zakresie subiektywnych punktów końcowych [8].

5.4. Czego badanie może dowieść, a co jedynie zasugerować?

W odpowiedzi na to pytanie należy odnieść się do hipotezy badawczej sformułowanej w badaniu i wynikającej z niej kalkulacji wielkości próby. Po pierwsze badania mogą dowodzić wyższości (ang. *superiority*), równoważności (ang. *equivalence*) lub faktu, iż oceniana

technologia nie jest gorsza niż porównywana (ang. *non-inferiority*). Omówiono te zagadania w rozdziale 4 poświęconym wpływowi założeń statystycznych na plan badania klinicznego. W tym miejscu warto jedynie odnieść się do różnicy między porównywanymi grupami, która wyznacza margines porównywalności – próg istotności klinicznej (ang. *margin*), czyli maksymalną różnicę między grupami, powyżej której należy uznać badaną terapię za gorszą od kontroli – oznaczany jako delta (δ). W oszacowaniu delty, oczywiście w oparciu o przesłanki kliniczne, stosuje się różne metody statystyczne [9], jednak nie należy zapominać, o ogólnym fakcie, iż im mniejsza delta tym więcej chorych musi uczestniczyć w badaniu. Skutkuje to ryzykiem zastosowania zbyt dużej delty, która wymaga także interpretacji klinicznej. Z ostrożnością należy interpretować wyniki badań, w których dopuszczalna różnica efektu terapeutycznego jest zbyt wysoka i nie wydaje się klinicznie pomijalna.

Ponieważ o wielkości próby (liczebności badania) decyduje spodziewana liczba zdarzeń klinicznych, a dokładnie różnica w ryzyku tych zdarzeń między grupami, z tego powodu często w badaniach wykorzystywany jest złożony punkt końcowy. Złożony punkt kalkulowany jest jako łączna liczba chorych, u których wystąpił przynajmniej jeden ze składowych punktów końcowych. Umożliwia to zmniejszenie liczby chorych w badaniu, ale zawsze wymaga oceny wpływu terapii na poszczególne składowe punktu końcowego, jeśli ich znaczenie kliniczne jest różne. Chodzi o to aby wnioskowanie o wpływie na złożony punkt końcowy nie wynikało ze zmiany jednego, najmniej istotnego, efektu terapeutycznego składającego się na złożony punkt końcowy.

Charakter sformułowanej hipotezy badawczej pozwala na wnioskowanie jedynie w jej zakresie. Z tego też powodu badanie może dowieść wpływu na taki punkt końcowy, którego dotyczy hipoteza badawcza i który posłużył określeniu wielkości próby. Taki punkt końcowy jest nazywany pierwotnym (pierwszorzędowym, ang. *primary end-point*) w przeciwieństwie do wtórnego (drugorzędowego, ang. *secondary end-point*), o którym informacje są zbierane w trakcie badania, ale nie posłużył on ocenie niezbędnej liczebności próby.

Wszystkie wnioski wyciągane w zakresie wtórnych punktów końcowych, o ile liczba tych zdarzeń jest mniejsza niż pierwotnych punktów końcowych, wymagają bardzo ostrożnej interpretacji i weryfikacji w specyficznie pod ich kątem zaplanowanych badaniach klinicznych. Podobnie dzieje się w przypadku tzw. analiz *post-hoc* w podgrupach chorych określonych daną cechą, na przykład z całej grupy badanej wybieramy jedynie chorych w wieku powyżej 50 lat, chorych z cukrzycą lub chorych z przebytym zawałem serca w wywiadzie. Analizy takie są obciążone błędem pierwszego rodzaju i w przypadku sformułowania wystarczająco dużej liczby podgrup możemy niemalże z pewnością oczekiwać, że w kilku z nich otrzymane wyniki będą miały zupełnie nieprawdziwy, przypadkowy charakter (patrz rozdział 6 poświęcony także testowaniu hipotez wielokrotnych). Analizy *post-hoc*, w podgrupach chorych należy interpretować zbiorczo i ewentualnie wnioskować o heterogenicznym lub homogennicznym wyniku otrzymanym we wszystkich zdefiniowanych podgrupach. W przypadku różnic między grupami należy wnioskować o poszczególnych grupach z ostrożnością, mając na uwadze jej liczebność i potrzebę weryfikacji obserwacji w osobnym, dedykowanym badaniu klinicznym.

Analizy w podgrupach są czasami zdefiniowane na etapie planowania badania. W takiej sytuacji przewidywana liczba chorych w podgrupie jest podyktowana statystyczną kalkulacją wielkości próby lub wyniki w tej podgrupie są porównywane z całą badaną populacją pod

kątem weryfikacji odmienności obserwowanego wyniku (proszę zwrócić uwagę, że w tym drugim przypadku mamy do czynienia z analizą *post-hoc*, ale obejmuje ona nie tylko wybraną podgrupę, lecz całą badaną populację) [10].

5.5. Ocena punktów końcowych – efektu klinicznego

Zastosowane w badaniu punkty końcowe mają podstawowe znaczenie dla jego interpretacji, ponieważ szczególnie ważne dla praktyki klinicznej są badania z tzw. „twardymi punktami końcowymi" (ang. *hard end-points*), klinicznie istotnymi efektami zdrowotnymi, odgrywającymi ważną rolę w danej chorobie. Postuluje się także, aby szczególnie koncentrować się na badaniach, w których oceniany efekt zdrowotny jest maksymalnie zobiektywizowany. Oczywiście jedynie w pełni obiektywny efekt kliniczny to zgon, ponieważ pozostałe są określane zmieniającymi się definicjami lub są raportowane przez samych chorych. Zgodnie z wytycznymi Agencji Oceny Technologii Medycznych[11], proponuje się uznać za istotne klinicznie następujące efekty:

– zgony,
– zachorowania bądź wyleczenia,
– jakość życia,
– działania niepożądane (z podziałem na ciężkie i pozostałe) lub incydenty medyczne.

Zachorowania bądź wyleczenia (często także definiowane w sposób zmieniający się na przestrzeni lat, co utrudnia porównanie wyników badań) są szczególnie istotne w ocenie chorób, które nie wiążą się z ryzykiem zgonu. Jakość życia, lub szerzej zdefiniowane efekty terapeutyczne raportowane/oceniane przez chorych, są także zalecane w badaniach skuteczności leków przez odpowiadające za rejestracje leków agencje, jak np. FDA. W ocenie tych efektów zdrowotnych najczęściej posługujemy się specyficznymi kwestionariuszami, które powinny być trafne (mierzyć precyzyjnie efekt, do pomiaru którego zostały stworzone), niezawodne (gwarantujące stabilność (ang. *test-retest reliability*), porównywalność (ang. *inter-rater reliability*) i wewnętrzną spójność (ang. *internal consistency reliability*) wyników) i wrażliwe[3] (zdolne do wykrycia takiej zmiany mierzonego stanu, która jest istotna dla pacjenta, nawet jeśli różnica jest mała) [12].

Działania niepożądane lub incydenty medyczne, bezpośrednio związane z zastosowaniem i powodowane przez technologię medyczną, same *per se* trudno określić jako istotne punkty końcowe, o ile nie obejmują one zgonów, zachorowań lub istotnie wpływają na jakość życia.

W odróżnieniu od istotnych punktów końcowych, surogaty (zastępcze punkty końcowe), które są najczęściej wynikami badań dodatkowych, mogą nas informować o progresji choroby, jednak nie zawsze ich zmiana gwarantuje dobre przybliżenie lub oddaje faktyczną poprawę, lub pogorszenie związane z zastosowanym leczeniem (Tabela 1).

[3] Z wrażliwością łączy się pojęcie minimalnej istotnej różnicy (ang. *minimally important difference* – MID), to jest najmniejszej różnicy w wyniku danego pomiaru, którą pacjent – respondent uważa za istotną na tyle, że skłoniłaby go do rozważenia zmiany postępowania. MID ocenia się indywidualnie dla każdego kwestionariusza – narzędzia pomiarowego.

Tabela 5.1. Badania, w których nie obserwowano zależności między wpływem terapii na zastępczy i twardy punkt końcowy.

Lek	Zastępczy punkt końcowy (korzystne działanie)	Twardy punkt końcowy (niekorzystny wpływ)
Aprotinina [13]	Utrata krwi w trakcie zabiegu kardiochirurgicznego	Śmiertelność
Doksazosyna [14]	Nadciśnienie tętnicze	Zawał serca
Enkainid, flekainind [15]	Pobudzenia przedwczesne	Nagły zgon sercowy
Epoprostenol [16]	Kurczliwość serca – frakcja wyrzutowa	Śmiertelność
Erytropoetyna [17]	Hemoglobina w niewydolności nerek	Śmiertelność
Estrogeny, progestin [18]	Cholesterol	Udar, otępienie, rak piersi
Fluorek sodu [19]	Gęstość tkanki kostnej	Złamania pozakręgosłupowe
Ibopamina [20]	Kurczliwość serca – frakcja wyrzutowa	Śmiertelność
Intensywna leczenie w cukrzycy typu 2 [17]	HbA1c < 6%	Śmiertelność
Klofibrat [21]	Cholesterol	Śmiertelność
Metoprolol [22]	Niedokrwienie okołooperacyjne	Śmiertelność
Milrinon [23]	Kurczliwość serca – frakcja wyrzutowa	Śmiertelność
Nesiritid [24]	Ciśnienie zaklinowania (PCWP – *pulmonary capillary wedge pressure*) i duszność	Śmiertelność
Rozyglitazon [25]	HbA1c	Zawał serca
Tolbutamid, fenformina [26]	Glikemia	Śmiertelność
Torcetrapib [27]	Cholesterol	Śmiertelność

5.6. Metoda zaślepienia (maskowania) – kogo, jak i dlaczego?

W celu zmniejszenia ryzyka błędu przeprowadzenia lub detekcji stosuje się metodę zaślepienia, która może dotyczyć uczestników badania (chorych lub zdrowych), badaczy opiekujących się uczestnikami, oceniających efekty zdrowotne, analizujących dane i piszących raport, publikację z badania. Dość powszechnie znanymi pojęciami jest badanie otwarte (brak metody zaślepienia – zarówno pacjent uczestniczący w badaniu, jak i lekarz wiedzą, jaką interwencję otrzymuje konkretny uczestnik badania; badanie może być z lub bez randomizacji), przeprowadzone metodą pojedynczej (badacz wie jaką interwencję pacjent otrzymuje, natomiast pacjent nie wie, czy został przypisany do grupy badanej czy kontrolnej), podwójnej ślepej próby (zarówno badacz, jak i pacjent nie wiedzą, do której grupy został przypisany ten ostatni) lub potrójnej ślepej próby (zespół, który organizuje i analizuje dane z badania, jak również uczestnicy badania i badacze nie wiedzą, do której grupy terapeutycznej przydzieleni są uczestnicy). Co ciekawe, badania wskazują, że powyższe sformułowania i definicje są różnie rozumiane [28,29]. Z tej przyczyny obecne zalecenia (CONSORT 2010) wskazują na potrzebę precyzyjnej informacji, kto jest poddany metodzie zaślepienia, w jaki sposób

i dokładnie czego zaślepienie dotyczy [30,31]. Oczywiście zaślepienie dotyczy przede wszystkim stosowanej i ocenianej w badaniu metody, jednak należy pamiętać, iż brak zaślepienia w pozostałych elementach może wskazywać na fakt, czy chory jest w grupie kontrolnej lub nie. Tak na przykład stało się w badaniu PROOF, w którym duża liczba kobiet w grupie kontrolnej, znając wyniki okresowego badania gęstości tkanki kostnej, zrezygnowała z udziału w przypadku braku zmian w wynikach tego badania, przypuszczając, że są w grupie kontrolnej otrzymującej placebo [32,33]. Z tej przyczyny konstruując badanie i interpretując jego wyniki, należy mieć na względzie poprawność metody zaślepienia, która powinna zapewnić brak możliwości odczytania alokacji chorego do grup.

Z metodą zaślepienia wiąże się także pojęcie maskowania (ang. *dummy*), które zgodnie ze słownikiem EBM/HTA oznacza technikę zaślepienia próby w sytuacji kiedy porównywane interwencje różnią się postacią lub drogą podawania, np. porównanie produktu w postaci tabletek i komparatora w inhalatorze. Jedna z grup otrzymuje wtedy aktywny lek w tabletce i jednocześnie inhaluje placebo, natomiast druga inhaluje aktywny komparator i jednocześnie przyjmuje tabletkę z placebo.

Oczywiście w niektórych przypadkach nie sposób jest przeprowadzić badanie metodą zaślepienia – na przykład wtedy, kiedy badany jest wpływ wysiłku fizycznego i konieczne jest zaangażowanie chorego. Poza tym nie zawsze zaślepienie wszystkich badaczy i chorych jest niezbędne. W ostatnich latach coraz więcej publikowanych jest badań typu PROBE (ang. *prospective, randomized, open, blinded-endpoint evaluation*), w których metodzie zaślepienia poddani są badacze uczestniczący w badaniu w zakresie jedynie oceny punktów końcowych. Badacze ci oceniają wyniki, nie mają wiedzy, do której grupy należą poszczególni pacjenci. Im bardziej obiektywny i mniej raportowany przez samego chorego, oceniany punkt końcowy, tym konieczność zaślepienia mniejsza. Analizy wskazują, że im punkt końcowy i ocena efektu terapeutycznego jest bardziej subiektywna, tym większe niebezpieczeństwo zafałszowania wyników badania przeprowadzonego bez użycia metody zaślepienia na korzyść ocenianej metody, co na przykład potwierdzono w badaniach nad stwardnieniem rozsianym [8,34-36].

5.7. Analiza wyników badania

W celu ograniczenia błędu utraty istotne jest zapewnienie udziału przez cały okres badania możliwie wszystkich chorych włączonych do badania i losowo przydzielonych do grup. W praktyce jednak taka sytuacja zdarza się bardzo rzadko z uwagi na fakt, iż chorzy rezygnują z badania, są z niego wykluczani lub pojawią się problemy organizacyjno-techniczne, które powodują, iż udział w badaniu odbiega od zaplanowanego. W tym przypadku pojawia się bardzo ważna rozbieżność w liczbie i charakterystyce chorych włączonych do badania oraz tych, którzy je ukończyli. Mimo że intuicyjnie „na pierwszy rzut oka" analiza grupy chorych, która zakończyła badanie wydaje się być preferowanym postępowaniem z uwagi na komplet informacji o tych chorych, to jednak podejście takie (nazywane analizą w grupach wyodrębnionych zgodnie z protokołem badania – ang. *per protocol analysis*) nie jest „złotym standardem". Ma ono jedną bardzo poważną wadę, polegającą na wprowadzeniu błędu selekcji na skutek pominięcia efektu randomizacji. Jeśli bowiem ograniczymy się

do analizy chorych, którzy „wypełnili protokół badania", czyli uczestniczyli w nim zgodnie z planem i jest dla nich dostępny komplet informacji, to nie są to chorzy tożsami z całą grupą włączoną do badania i poddaną randomizacji, a więc mogą być wyselekcjonowani (za przerwaniem przez nich badania może stać systematyczna przyczyna). W celu zachowania efektu randomizacji preferowana jest analiza w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem (ang. *intention-to-treat analysis*), uwzględniająca całą włączoną do badania populację chorych i zgodnie z którą, częstości zdarzeń klinicznych obliczane są jako iloraz ich liczby w stosunku do liczebności całej grupy na początku badania. W praktyce mamy często do czynienia z różnego typu modyfikacjami analizy w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem, co podyktowane jest charakterem badanej metody (Tabela 2). Na przykład w szczepieniach ważne dla ich skuteczności i uzyskania pełnej odporności jest przeprowadzenie pełnego cyklu kilku podań szczepionki. Z tej przyczyny często analiza w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem *sensu stricto* nie jest możliwa i konieczne są jej modyfikacje.

Tabela 5.2. Opis protokołów obserwacji skuteczności szczepionki Silgard.

Protokół obserwacji	Opis pacjentów włączonych do obserwacji
Zmodyfikowana analiza w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem – typ-1	• otrzymali wszystkie 3 dawki szczepionki • seronegatywni w stosunku do istotnych typów HPV w dniu rozpoczęcia szczepień • brak DNA istotnych typów HPV w dniu rozpoczęcia szczepień i przez 7 kolejnych miesięcy od rozpoczęcia szczepienia (badania przy użyciu PCR) • obserwacja punktów końcowych rozpoczęta w 30 dni po zakończeniu całego cyklu szczepienia
Zmodyfikowana analiza w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem – typ-2	• otrzymali co najmniej 1 dawkę szczepionki • seronegatywni w stosunku do odpowiednich (tj. badanych) typów HPV w dniu rozpoczęcia szczepień • brak DNA odpowiednich (tj. badanych) typów HPV w dniu rozpoczęcia szczepień (badania przy użyciu PCR) • obserwacja punktów końcowych rozpoczęta w 30 dni po 1. dawce szczepienia • co najmniej 1 kontrolna wizyta po miesiącu od podania pierwszej dawki szczepienia
Zmodyfikowana analiza w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem – typ-3	• otrzymali co najmniej 1 dawkę szczepionki • dowolny status serologiczny w stosunku do HPV • dowolny status w badaniu DNA HPV (badania PCR) • obserwacja punktów końcowych rozpoczęta w 30 dni po 1. dawce szczepienia • co najmniej 1 kontrolna wizyta po miesiącu od podania pierwszej dawki szczepienia
Zmodyfikowana analiza w grupach wyodrębnionych zgodnie z zaplanowanym leczeniem – ograniczona typ-2	• seronegatywni w stosunku do wszystkich typów HPV w dniu rozpoczęcia szczepień • brak DNA wszystkich typów HPV w dniu rozpoczęcia szczepień (badania przy użyciu PCR) • brak zmian w badaniu cytologicznym w dniu rozpoczęcia szczepienia (test Papanicolau) • obserwacja punktów końcowych rozpoczęta w 30 dni po 1. dawce szczepienia

Analiza w grupach wyodrębnionych zgodnie z protokołem (*per-protocol*)	• otrzymali wszystkie 3 dawki szczepionki • seronegatywni w stosunku do istotnych typów HPV w dniu rozpoczęcia szczepień • brak DNA istotnych typów HPV w dniu rozpoczęcia szczepień i przez 7 kolejnych miesięcy od rozpoczęcia szczepienia (badania przy użyciu PCR) • nie naruszyli protokołu badania • obserwacja punktów końcowych rozpoczęta w 30 dni po zakończeniu całego cyklu szczepienia

Warto podkreślić, że w trakcie trwania badania mają często miejsce (nieplanowane, tak jak w badaniu naprzemiennym) zmiany sposobu leczenia chorych i stosowanie wymiennie porównywanych technologii medycznych. Oznacza to na przykład, że chory przydzielony losowo do grupy chorych otrzymujących lek A w trakcie trwania badania nie odnosi korzyści z jego zastosowania i zostaje u niego zmienione leczenie na lek B – alternatywny lek stosowany w grupie kontrolnej, z którym lek A jest porównywany. Dzieje się tak dość często w onkologii z uwagi na fakt kierowania się w praktyce klinicznej dobrem chorego, a nie protokołem badań klinicznych, i zmiany leczenia podyktowane są poszukiwaniem skutecznej metody terapeutycznej. W celu zachowania efektu randomizacji należy jednak analizować dane każdego chorego w grupie, do której został losowo przydzielony na początku badania, nawet mimo faktu, iż jego leczenie przebiegała po część tak jak w alternatywnych grupach.

W przypadku brakujących danych wynikających z niezakończenia obserwacji w przypadku wszystkich chorych stosowane, jednak nie polecane, są różne metody ich uzupełniania. Można w przypadku chorych, którzy nie ukończyli badania, przyjąć założenie o średnim efekcie obserwowanym w badaniu lub zastosować analizę scenariuszową i przyjąć alternatywnie, iż oceniany efekt kliniczny wystąpił u wszystkich lub żadnego z pacjentów, którzy nie ukończyli badania. Metody te nie są zalecane standardowo. Częściej stosowaną, jednak także nie pozbawioną ograniczeń, jest metoda ekstrapolacja ostatniej obserwacji (ang. *last observation carried forward* – LOCF), w której zakłada się, że odnotowane na ostatniej przeprowadzonej u chorego wizycie kontrolnej wyniki (obserwowane efekty kliniczne) nie ulegną zmianie do końca badania.

5.8. Klasyfikacja jakości badań – skale i narzędzia

Ponieważ rozdział ten jest poświęcony rzetelności i jakości badań klinicznych nie sposób pominąć skal, które są w tym celu wykorzystywane. Ocena rzetelności wyników badania klinicznego jest odmienna dla badań eksperymentalnych (skuteczności i bezpieczeństwa), obserwacyjnych czy diagnostycznych (te ostatnie nie są omówione z uwagi na specyfikę, której szczegóły wykraczają poza zakres tego opracowania). Wymieniono tylko najczęściej stosowane skale, które mogą i są często modyfikowane na doraźne potrzeby [37].

Chyba najbardziej popularną z nich jest skala Jadad do oceny jakości klinicznych badań eksperymentalnych.

5.8.1. Skala Jadad

Skala Jadad, zwana także oksfordzkim systemem oceny jakości badania klinicznego, została nazwana od nazwiska kolumbijskiego lekarza i naukowca Aleksandra Jadada-Bechara [38]. Według tej skali badanie może otrzymać od 0 (niska jakość) do 5 (najwyższa jakość) punktów. Oprócz oceny jakości pojedynczego badania skala Jadad często służy do ustalenia minimalnego poziomu badań włączonych do przeglądu systematycznego lub metaanalizy. Raportując w tej skali jakość kilku badań posługujemy się z reguły medianą, a nie średnią, jako miarą centralną, oraz zakresem.

Skala Jadad nie jest zbyt skomplikowana. Z tego powodu, a także z uwagi na fakt istotnego podkreślenia metody zaślepienia oraz różnice w ocenie dokonywane przez dwie różne osoby, była krytykowana [39-43]. Jest jednak najbardziej uniwersalną i właśnie prostą skalą, co stanowi o jej popularności [44]. Z drugiej strony *Cochrane Handbook for Systematic Reviews of Interventions* nie zaleca stosowania tej skali z uwagi na pominięcie w niej *allocation concealment* i zwrócenie uwagi jedynie na raportowanie, a nie faktyczne przeprowadzenie badania klinicznego.

Ramka 1. Skala Jadad.

Po przeczytaniu danego artykułu należy odpowiedzieć na następujące pytania:

- Pytanie 1: Czy praca opisywana jest jako randomizowana?
- Pytanie 2: Czy praca opisywana jest jako badanie z podwójnie ślepą próbą?
- Pytanie 3: Czy w pracy znajduje się opis pacjentów, którzy wycofali się lub zostali wycofani z badania?

Za odpowiedź „tak" na poszczególne pytanie dodać 1 punkt.
Za odpowiedź „nie" na poszczególne pytanie dodać 0 punktu.
Nie ma pośrednich ocen.

Dodać 1 punkt jeśli:

- Dla pytania 1: opisano metodę randomizacji i była ona właściwa i/lub
- Dla pytania 2: opisano metodę zaślepienia badania i była ona właściwa.

Odjąć 1 punkt jeśli:

- Dla pytania 1: opisano metodę randomizacji, lecz nie była ona właściwa i/lub
- Dla pytania 2: badanie było określone jako badanie z podwójnie ślepą próbą, lecz metoda zaślepienia badania była niewłaściwa.

5.8.2. Propozycja Cochrane Collaboration dotyczącą oceny ryzyka błędu systematycznego

Cochrane Collaboration zaproponowało własne narzędzie (nie skalę, ponieważ one z natury rzeczy agregują informacje w postaci liczby i tym samym nieprecyzyjnie odzwierciedlają poszczególne kryteria jakości badania klinicznego) w ocenie ryzyka błędu systematycznego (Tabela 5.3). Odpowiedź negatywna na sformułowane pytania odnośnie poszczególnych kryteriów oznacza wysokie ryzyko błędu.

Tabela 5.3. Propozycja Cochrane Collaboration dotycząca oceny ryzyka błędu systematycznego.

Kryterium - domena	Opis	Ocena
Określenie kodu randomizacyjnego (kolejności przydzielania chorych do grup).	Opisać możliwie najdokładniej metodę wykorzystaną do opracowania sekwencji przydziału chorych do grup tak, aby umożliwić ocenę jej skuteczności, to jest wygenerowania wyjściowo porównywalnych grup.	Czy prawidłowo określono sposób przydzielania chorych do grup?
Utajnienie kodu randomizacyjnego (kolejności przydzielania chorych do grup; *allocation concealment*).	Opisać możliwie najdokładniej metodę wykorzystaną do utajnienia informacji o przydziale chorych do grup tak, aby umożliwić ocenę możliwości przewidzenia przydziału chorych do grup przed lub w trakcie alokacji.	Czy kolejność przydzielania do grup była prawidłowo utajniona?
Metoda zaślepienia w zakresie każdego ważnego efektu terapeutycznego.	Opisać wszystkie metody wykorzystane do zaślepienia uczestników badania i badaczy co do badanego i stosowanego u chorych leczenia. Czy metoda zaślepienia była efektywna.	Czy informacja o przydziale do grup określonych różnym stosowanym leczeniem była tajna w trakcie badania?
Analiza danych niepełnych w zakresie każdego ważnego efektu terapeutycznego.	Opisać w jakim stopniu dane o efektach terapeutycznych były kompletne oraz jak przedstawiały się wyłączenia chorych z badania z poszczególnych przyczyn.	Czy publikacje z badania nie są oparte na selektywnym doborze chorych i tym samym danych o efektach terapeutycznych?
Selektywne raportowanie wyników o skuteczności klinicznej.	Opisać możliwość selektywnego raportowania wyników.	Czy publikacje z badania nie są oparte na selektywnym raportowaniu wyników o efektach zdrowotnych?
Inne źródła błędów.	Opisać inne aspekty badania mogące być źródłem błędów.	Czy badanie miało inne uchybienia metodologiczne mogące zwiększać ryzyko błędu?

5.8.3. Skala oceny jakości badań obserwacyjnych Newcastle-Ottawa Scale (NOS)

Jest narzędziem rekomendowanym do oceny wiarygodności badań obserwacyjnych przez *Cochrane Non-Randomized Studies Methods Working Group* oraz zalecanym przez wytyczne Agencji Oceny Technologii Medycznych z kwietnia 2009 roku. W ramach tej skali opracowano osobną wersję dla badań kohortowych[4] i badań kliniczno-kontrolnych[5] (ramki poniżej). Obie wersje zawierają 4 pytania dotyczące doboru pacjentów do badania i 1 pytanie o czynniki zakłócające. Wersja dla badań kohortowych zawiera dodatkowo 3 pytania o ocenę efektów terapeutycznych, natomiast wersja dla badań kliniczno-kontrolnych 3 pytania o ekspozycję. Wysoka jakość badania jest oznaczana gwiazdką dla każdego z pytań, za wyjątkiem pytań o czynniki zakłócające, za które można maksymalnie przyznać 2 gwiazdki.

[4] Def. – badanie obserwacyjne, w którym ocenia się prospektywnie wystąpienie określonego punktu końcowego w grupach (kohortach) osób narażonych i nienarażonych na dany czynnik lub interwencję, u których ten punkt końcowy na początku obserwacji nie występował. W badaniu kohorty historycznej grupy narażona i nienarażona są identyfikowane w przeszłości i „obserwowane" ku teraźniejszości pod względem występowania punktu końcowego (Polski Instytut EBM – www.ebm.org.pl).

[5] Def. – badanie obserwacyjne, w którym poszukuje się związku między daną ekspozycją a wystąpieniem określonego punktu końcowego, porównując retrospektywnie ekspozycję (odsetek narażonych) w grupie osób, u których punkt końcowy wystąpił, z ekspozycją w odpowiednio dobranej grupie osób kontrolnych, u których punkt końcowy nie wystąpił (Polski Instytut EBM – www.ebm.org.pl).

Ramka 2. Skala Newcastle-Ottawa Scale (NOS) dla badań kliniczno-kontrolnych

Badanie może otrzymać maksymalnie jedną gwiazdkę w przypadku każdego z pytań z części *Dobór pacjentów* oraz *Ocena efektów zdrowotnych*. Maksymalnie 2 gwiazdki mogą zostać przyznane w przypadku pytania w części *Czynniki zakłócające.*

Dobór pacjentów

1. Czy kryteria włączenia do grupy klinicznej zostały zdefiniowane we właściwy sposób?
   a. tak, niezależna walidacja kryteriów włączenia (np. > 1 osoba/zapis w dokumentacji medycznej/proces uzyskiwania informacji lub odwołanie do źródła danych pierwotnych takich jak wyniki prześwietleń RTG, odwołanie do dokumentacji medycznej/szpitalnej) ✵
   b. tak, np. łączenie rekordów (ang. *rekord linkage*)[6] lub sposób bazujący na zgłoszeniach spontanicznych przez pacjentów
   c. brak opisu
2. Reprezentatywność grupy klinicznej
   a. seria kolejnych przypadków/reprezentatywna (w sposób oczywisty) seria przypadków ✵
   b. możliwy błąd selekcji pacjentów do badania/nieokreślona
3. Dobór pacjentów do grupy kontrolnej
   a. pacjenci z grupy kontrolnej dobrani z tej samej społeczności, co pacjenci w grupie badanej ✵
   b. pacjenci z grupy kontrolnej dobrani z tego samego ośrodka, co pacjenci w grupie badanej
   c. brak opisu
4. Jak zdefiniowano kryterium włączenia do grupy kontrolnej?
   a. brak choroby w wywiadzie ✵
   b. brak opisu

Czynniki zakłócające

1. Czy grupa kontrolna była pod względem innych czynników determinujących stan zdrowia identyczna, jak grupa, w której występował potencjalny czynnik szkodliwy?
   a. grupy o zbliżonej charakterystyce pod względem _____ (wybierz najważniejszy czynnik zakłócający) ✵
   b. grupy o zbliżonej charakterystyce pod względem dodatkowych czynników zakłócających ✵ (to kryterium może być modyfikowane w celu kontroli czynników zakłócających o znaczeniu drugoplanowym)

Ekspozycja

1. Czy ekspozycję na badany czynnik oceniano w sposób obiektywny?
   a. wiarygodna dokumentacja (np. dokumentacja medyczna potwierdzająca wykonanie operacji chirurgicznych) ✵
   b. ustrukturyzowany wywiad, z zaślepieniem przynależności respondenta do grupy ✵
   c. ustrukturyzowany wywiad, bez zaślepienia
   d. spontaniczne raportowanie/tylko dokumentacja medyczna
   e. brak opisu
2. Czy zastosowano tę samą metodą oceny ekspozycji w grupie klinicznej i kontrolnej?
   a. tak ✵
   b. nie
3. Odsetek pacjentów z brakiem informacji o ekspozycji na czynnik chorobotwórczy
   a. ten sam odsetek pacjentów w obu grupach ✵
   b. opis pacjentów z brakiem odpowiedzi
   c. różne odsetki w obu grupach lub brak opisu

[6] łączenie danych zawartych w różnych zbiorach np. historii choroby czy dokumentach statystyki ludności, umożliwiające powiązanie informacji istotnych, ale odległych w czasie lub przestrzeni. Procedura łączenia rekordów pochodzących z różnych zbiorów jest możliwa tylko w oparciu o jednoznaczny system identyfikujący poszczególne osoby.

Ramka 3. Skala Newcastle-Ottawa Scale (NOS) dla badań kohortowych

Badanie może otrzymać maksymalnie jedną gwiazdkę w przypadku każdego z pytań z części *Dobór pacjentów* oraz *Ocena efektów zdrowotnych*. Maksymalnie 2 gwiazdki mogą zostać przyznane w przypadku pytania w części *Czynniki zakłócające.*

Dobór pacjentów

1. Reprezentatywność grupy poddanej ekspozycji na badany czynnik
   a. w sposób właściwy reprezentuje średni _____ (opisz) w populacji ✵
   b. w pewnym stopniu reprezentuje średni _____ w populacji ✵
   c. wyselekcjonowana grupa osób narażonych na badany czynnik, np. pielęgniarki, ochotnicy
   d. brak opisu
2. Dobór pacjentów do grupy nie poddanej ekspozycji na badany czynnik
   a. dobrani z tej samej populacji, co grupa poddana ekspozycji ✵
   b. dobrani w inny sposób
   c. brak opisu
3. W jaki sposób stwierdzano stopień narażenia pacjentów na badany czynnik?
   a. wiarygodna dokumentacja (np. dokumentacja medyczna potwierdzająca wykonanie operacji chirurgicznych) ✵
   b. ustrukturyzowany wywiad ✵
   c. spontaniczne raportowanie
   d. brak opisu
4. Wykazano, że badane efekty zdrowotne nie występowały na początku badania
   a. tak ✵
   b. nie

Czynniki zakłócające

1. Czy grupa kontrolna była pod względem innych czynników determinujących stan zdrowia identyczna, jak grupa, w której występował potencjalny czynnik szkodliwy?
   a. grupy o zbliżonej charakterystyce pod względem _____ (wybierz najważniejszy czynnik zakłócający) ✵
   b. grupy o zbliżonej charakterystyce pod względem dodatkowych czynników zakłócających ✵ (to kryterium może być modyfikowane w celu kontroli czynników zakłócających o znaczeniu drugoplanowym)

Ocena efektów zdrowotnych

1. Czy punkty końcowe oceniano w sposób obiektywny?
   a. tak, niezależna ocena, metodą ślepej próby ✵
   b. łączenie rekordów (ang. *rekord linkage*) ✵
   c. spontaniczne zgłoszenia pacjentów
   d. brak opisu
2. Czy okres obserwacji był wystarczająco długi, by mogły wystąpić efekty zdrowotne?
   a. tak (wybierz adekwatny czas obserwacji) ✵
   b. nie
3. Czy badany stan kliniczny oceniono, u wszystkich pacjentów, u których obserwację rozpoczęto?
   a. tak ✵
   b. niewielkie prawdopodobieństwo wprowadzenia błędu – wysoki odsetek pacjentów, którzy ukończyli badanie - > _____ % (wybierz adekwatny odsetek) lub opis pacjentów utraconych z badania ✵
   c. odsetek pacjentów, którzy ukończyli badanie < _____% (wybierz adekwatny odsetek) lub brak opisu pacjentów utraconych z badania
   d. nie podano

Biblografia

1. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1. The Cochrane Collaboration; 2008.
2. De Ridder A, De Graeve D. Can we account for selection bias? A comparison between bare metal and drug-eluting stents. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research 2011;14:3-14.
3. Zhehui L, Gardiner JC, Bradley CJ. Applying propensity score methods in medical research: pitfalls and prospects. Med Care Res Rev 2010;67:528-54.
4. Pirracchio R, Sprung C, Payen D, Chevret S. Benefits of ICU admission in critically ill patients: Whether instrumental variable methods or propensity scores should be used. BMC medical research methodology 2011;11:132.
5. Lalani T, Cabell CH, Benjamin DK, et al. Analysis of the impact of early surgery on in-hospital mortality of native valve endocarditis: use of propensity score and instrumental variable methods to adjust for treatment-selection bias. Circulation 2010;121:1005-13.
6. Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. J Clin Oncol 2001;19:1064-70.
7. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002;359:515-9.
8. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601-5.
9. Huitfeldt B, Hummel J. The draft FDA guideline on non-inferiority clinical trials: a critical review from European pharmaceutical industry statisticians. Pharm Stat 2011; 10:414-9.
10. Diener HC, Connolly SJ, Ezekowitz MD, et al. Dabigatran compared with warfarin in patients with atrial fibrillation and previous transient ischaemic attack or stroke: a subgroup analysis of the RE-LY trial. Lancet neurology 2010;9:1157-63.
11. Agencja Oceny Technologii Medycznych. Wytyczne oceny technologii medycznych (HTA). Warszawa; 2009.
12. Bryant D, Schunemann H, Brozek J, Jaeschke R, Guyatt G. [Patient reported outcomes: general principles of development and interpretability]. Polskie Archiwum Medycyny Wewnętrznej 2007;117:5-11.
13. Fergusson DA, Hebert PC, Mazer CD, et al. A comparison of aprotinin and lysine analogues in high-risk cardiac surgery. The New England journal of medicine 2008;358:2319-31.
14. Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). ALLHAT Collaborative Research Group. JAMA : the journal of the American Medical Association 2000;283:1967-75.
15. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. The New England journal of medicine 1991;324:781-8.
16. Califf RM, Adams KF, McKenna WJ, et al. A randomized controlled trial of epoprostenol therapy for severe congestive heart failure: The Flolan International Randomized Survival Trial (FIRST). American heart journal 1997;134:44-54.
17. Phrommintikul A, Haas SJ, Elsik M, Krum H. Mortality and target haemoglobin concentrations in anaemic patients with chronic kidney disease treated with erythropoietin: a meta-analysis. Lancet 2007;369:381-8.

18. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA : the journal of the American Medical Association 2002;288:321-33.
19. Haguenauer D, Welch V, Shea B, Tugwell P, Adachi JD, Wells G. Fluoride for the treatment of postmenopausal osteoporotic fractures: a meta-analysis. Osteoporos Int 2000;11:727-38.
20. Hampton JR, van Veldhuisen DJ, Kleber FX, et al. Randomised study of effect of ibopamine on survival in patients with advanced severe heart failure. Second Prospective Randomised Study of Ibopamine on Mortality and Efficacy (PRIME II) Investigators. Lancet 1997;349:971-7.
21. Oliver MF. Cholesterol, coronaries, clofibrate and death. The New England journal of medicine 1978;299:1360-2.
22. Devereaux PJ, Yang H, Yusuf S, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet 2008;371:1839-47.
23. Packer M, Carver JR, Rodeheffer RJ, et al. Effect of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group. The New England journal of medicine 1991;325:1468-75.
24. Sackner-Bernstein JD, Kowalski M, Fox M, Aaronson K. Short-term risk of death after treatment with nesiritide for decompensated heart failure: a pooled analysis of randomized controlled trials. JAMA : the journal of the American Medical Association 2005;293:1900-5.
25. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. The New England journal of medicine 2007;356:2457-71.
26. Meinert CL, Knatterud GL, Prout TE, Klimt CR. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II. Mortality results. Diabetes 1970;19: Suppl:789-830.
27. Barter PJ, Caulfield M, Eriksson M, et al. Effects of torcetrapib in patients at high risk for coronary events. The New England journal of medicine 2007;357:2109-22.
28. Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA : the journal of the American Medical Association 2001;285:2000-3.
29. Haahr MT, Hrobjartsson A. Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. Clinical trials 2006;3:360-5.
30. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.
31. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332.
32. Cummings SR, Chapurlat RD. What PROOF proves about calcitonin and clinical trials. The American journal of medicine 2000;109:330-1.
33. Chesnut CH, 3rd, Silverman S, Andriano K, et al. A randomized trial of nasal spray salmon calcitonin in postmenopausal women with established osteoporosis: the prevent recurrence of osteoporotic fractures study. PROOF Study Group. The American journal of medicine 2000;109:267-76.
34. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke; a journal of cerebral circulation 2008;39:929-34.
35. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. 1994 [classical article]. Neurology 2001;57:S31-5.
36. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology 1994;44:16-20.

37. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii-x, 1-173.
38. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled clinical trials 1996;17:1-12.
39. Berger VW. Is the Jadad score the proper evaluation of trials? J Rheumatol 2006;33:1710-1; author reply 1-2.
40. Bhogal SK, Teasell RW, Foley NC, Speechley MR. The PEDro scale provides a more comprehensive measure of methodological quality than the Jadad scale in stroke rehabilitation literature. Journal of clinical epidemiology 2005;58:668-73.
41. Bhandari M, Richards RR, Sprague S, Schemitsch EH. Quality in the reporting of randomized trials in surgery: is the Jadad scale reliable? Controlled clinical trials 2001;22:687-8.
42. Oremus M, Wolfson C, Perrault A, Demers L, Momoli F, Moride Y. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. Dementia and geriatric cognitive disorders 2001;12:232-6.
43. Clark HD, Wells GA, Huet C, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. Controlled clinical trials 1999;20:448-52.
44. Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: a systematic review. Physical therapy 2008;88:156-75.

VI. Statystyka opisowa i wnioskowanie statystyczne – podstawy

Michał JAKUBCZYK

Celem prowadzenia badań klinicznych jest określenie skuteczności badanych leków i zagrożeń związanych z ich stosowaniem (np. częstości działań niepożądanych). Oczywiście ten sam lek stosowany u różnych pacjentów może dać różny efekt ze względu na inny stopień zaawansowania choroby, ogólny stan zdrowia, styl życia i wiele innych czynników. Wiele z tych aspektów z perspektywy prowadzącego lub interpretującego badanie jest poza możliwością poznania, więc stanowi de facto element losowy.

W badaniach statystycznych rozróżnia się tzw. populację generalną (ang. *general population*) i próbę losową (ang. *random sample*). Przez populację generalną rozumiemy ogół pacjentów, wśród których chcemy ocenić skuteczność leku. Populacja ta w badaniach klinicznych jest na ogół zadana tylko teoretycznie, tj. obejmuje nie tylko pacjentów obecnie chorych na badaną chorobę, lecz wszystkich, którzy potencjalnie w przyszłości mogą zachorować. Oczywiście nie sposób objąć badaniem całej populacji generalnej, dlatego w praktyce bada się jej podzbiór – tj. próbę losową.

Nie podejmujemy kwestii sposobu doboru próby losowej – w szczególności jej wielkości i konstrukcji, tzn. procedury wyboru jednostek (elementy zasygnalizowano w rozdziale 4). Zagadnienia te są znacznie ważniejsze dla projektujących badanie. Zakładamy natomiast, że próba jest reprezentatywna, tj. ze względu na istotne dla badanego zjawiska charakterystyki przypomina populację generalną.

Tak więc o ile badaczy interesuje działanie leku w populacji generalnej, o tyle dostępne są wyniki w próbie, które z uwagi na wspomnianą losowość różniłyby się między różnymi próbami. Istnieje zatem konieczność oddzielenia losowości występującej na poziomie próby od prawidłowości, które można przypisać lekowi w populacji generalnej. Innymi słowy zachodzi potrzeba identyfikacji tych prawidłowości w wynikach obarczonych losowością. Wykorzystanie przy tym sformalizowanych narzędzi ilościowych po pierwsze pozwala na obiektywizację wnioskowania, a po drugie na analizę danych w mniej trywialnych sytuacjach, np. w przypadku jednoczesnej

analizy wielu zmiennych (por. rozdz. 9) czy analizy badań, w których część pacjentów odchodzi z badania w trakcie jego trwania (por. rozdz. 10). Metody statystyczne pozwalają na uniknięcie błędów popełnianych często przy „intuicyjnej" analizie prawdopodobieństw.[1]

W niniejszym rozdziale wprowadzono podstawowe słownictwo stosowane w analizach statystycznych. W pierwszej kolejności przedstawiono pojęcia stosowane w tzw. statystyce opisowej (ang. *descriptive statistics*), tj. przy charakteryzowaniu zbioru danych bez jego odnoszenia do populacji generalnej. W kolejnych podrozdziałach przedstawiono ideę tzw. wnioskowania statystycznego (ang. *statistical inference*), tj. procesu odnoszenia wyników próby losowej do populacji generalnej. Wnioskowanie statystyczne omówiono w rozbiciu na dwie podstawowe techniki – estymację i testowanie hipotez statystycznych. Rozdział ten ma ułatwić odbiór treści zawartych w dalszej części opracowania.

6.1. Statystyka opisowa

W badaniach klinicznych analizowanymi jednostkami są najczęściej poszczególni pacjenci. Są oni charakteryzowani ze względu na wiele cech obejmujących dane demograficzne, stan kliniczny w momencie włączenia, stosowane leczenie i jego skutki. Oczywiście nie sposób w publikacji zaprezentować szczegółowych informacji dotyczących całej próby losowej. Jednocześnie dla czytelnika interesujące jest zrozumienie, jacy pacjenci byli objęci badaniem, jakie były skutki leczenia, itd.

Statystyka opisowa to dział statystyki zajmujący się metodami syntetycznego charakteryzowania badanego zbioru danych. Definiuje się miary charakteryzujące wartości poszczególnych cech dla badanych pacjentów, aby czytelnik mógł wyrobić sobie wyobrażenie o badanej grupie. W zależności od charakteru cechy (zmiennej) dostępne są różne rodzaje miar, zacznijmy więc od zdefiniowania podstawowych typów zmiennych.

6.1.1. Skale zmiennych

W statystyce określa się przede wszystkim *skalę pomiarową* (ang. *measurement scale*), na której mierzona jest dana zmienna. Rodzaj skali[2] określa, jakiego typu porównania i operacje są dozwolone.

Najsłabszą skalą jest tzw. *skala nominalna* (ang. *nominal*), zmienne mierzone według tej skali określa się także jako zmienne kategoryczne (ang. *categorical variables*). Porównując wartości zmiennych nominalnych, można jedynie stwierdzić, czy dwóch pacjentów ma takie same czy też różne wartości tej cechy. Nie można np. stosować porównań typu „większe"/"mniejsze". Przykłady zmiennych mierzonych na skali nominalnej to: grupa krwi, genotyp wirusa, rasa pacjenta. Oczywiście w bazie danych można zakodować wartości zmiennych nominalnych z użyciem liczb, których porównywanie jest technicznie możliwe, ale nie zmienia to braku możliwości interpretacji wyników takich porównań.

---

[1] Często podaje się następujący przykład – co najmniej ile osób należy zebrać w pokoju, aby prawdopodobieństwo tego, że choć dwie obchodzą urodziny tego samego dnia, przekroczyło 50%? Często udzielane są odpowiedzi znacznie przewyższające prawdziwą, w pierwszym odruchu nawet ponad 366 osób! Prawdziwa odpowiedź to 23 osoby (czyli dla ułatwienia dwie drużyny piłkarskie i sędzia).

[2] Formalną definicję można znaleźć np. w opracowaniu [9].

Silniejszą skalą jest *skala porządkowa* (ang. *ordinal*). Ogólną regułą jest, że silniejsza skala umożliwia stosowanie wszystkich operacji dostępnych dla skal słabszych i pozwala także na dodatkowe typy porównań. W przypadku skali porządkowej można uszeregować wartości zmiennych w kolejności mającej interpretację. Tak więc porównując dwóch pacjentów ze względu na cechę mierzoną na skali porządkowej, można stwierdzić, że jeden pacjent ma większą wartość tej cechy. Skala porządkowa nie umożliwia jednak sensownego obliczania różnicy wartości cechy, więc nie można stwierdzić, ile wynosi różnica wartości cechy dla dwóch pacjentów. Przykłady zmiennych porządkowych to: wykształcenie (zakładając, że ograniczamy się np. do wartości: podstawowe, średnie, wyższe), klasa NYHA, klasyfikacja według skali ECOG. Zauważmy, że o ile np. dla klasy NYHA, technicznie biorąc, można obliczyć różnicę wartości zmiennej, o tyle ta różnica nie jest interpretowalna. Np. jeśli jeden pacjent ma klasę NYHA III, a drugi ma klasę NYHA II, to można by stwierdzić, że pierwszy ma klasę NYHA o jeden większą, ale takie zdanie nie ma sensu o tyle, że różnica między NYHA IV i NYHA III też wynosi jeden, a z klinicznego punktu widzenia, oznacza zupełnie inną zmianę sytuacji.

Trzecim, silniejszym typem skali jest *skala przedziałowa* (ang. *interval*). Dla skali tej można obliczać, o ile różnią się wartości zmiennej, ale nie można – ile razy się różnią. Tzn. uprawnione są stwierdzenia „wartość cechy wzrosła o 2", a nieuprawnione – „wartość cechy wzrosła dwukrotnie". Ten typ skali pojawia się, gdy wartość zerowa została ustalona arbitralnie przez definiującego skalę i rzadko występuje w naukach medycznych. Przykładem jest np. temperatura mierzona w stopniach Celsjusza. Oczywiście matematycznie można obliczyć iloraz dwóch temperatur, ale stwierdzenie „temperatura wzrosła dwukrotnie" jest mylące, gdyż dotyczy sytuacji zmiany z 1°C na 2°C, z 15°C na 30°C i z -10°C na -20°C! Innym przykładem jest skala pH.

Najsilniejszym typem jest *skala ilorazowa* (ang. *ratio*). Dla skali tej można dzielić wartości zmiennych dla dwóch pacjentów (albo dla tego samego pacjenta w różnych momentach). Przykłady cech mierzonych według skali ilorazowej to wiek, waga, BMI, obwód pasa, dawka leku, stężenie hemoglobiny, wartość ciśnienia, poziom triglicerydów, czas otrzymywania leku, czas do nawrotu objawów choroby, itd. W tabeli poniżej zestawiono informacje i przykłady dotyczące skal zmiennych.

Tabela 6.1. Skale zmiennych.

Typ skali	**Przykłady zmiennych i wartości**	**Dozwolone porównania**	**Przykładowe niedozwolone porównania**
nominalna	grupa krwi (0, A, B, AB) genotyp wirusa WZW B (A, B, …) rasa pacjenta (kaukaska, żółta, …)	równe, różne	większe/mniejsze
porządkowa	skala ECOG (0,1,…) klasa NYHA (I, II, III, IV)	równe, różne większe/mniejsze	o ile większe? ile razy większe?
przedziałowa	°C, pH	równe, różne większe/mniejsze o ile większe?	ile razy większe?
ilorazowa	wiek, waga	równe, różne większe/mniejsze o ile większe? ile razy większe	

Źródło: opracowanie własne

6.1.2. Miary położenia

Jak wspomniano, celem statystyki opisowej jest scharakteryzowanie wartości cechy w zbiorze jednostek. To syntetyczne przedstawienie może dotyczyć nieco różnych aspektów, upraszczając – czy wartości w próbie są duże czy małe? czy pacjenci są w miarę jednorodni czy zróżnicowani? czy odchylenia od wartości średniej następują raczej w górę czy w dół? Na powyższe pytania odpowiadają kolejno tzw. miary *położenia*, *zróżnicowania* i *asymetrii*.[3] W zależności od siły skali różna jest lista dostępnych miar. Poniżej zaprezentowano miary najczęściej wykorzystywane w publikacjach.

Miary położenia (ang. *location*) mają na celu określenie, czy wartości cechy w zbiorze są duże czy małe, np. czy pacjenci są młodzi czy starzy, czy są w zaawansowanym stadium choroby czy nie, czy są otyli czy nie, czy nastąpiła duża poprawa parametrów fizjologicznych czy nie. Do najczęściej stosowanych miar położenia należą: dominanta, mediana, kwartyle i percentyle (te trzy miary ogólnie są nazywane kwantylami) oraz średnia.

Dominanta (ang. *mode*) to wartość cechy, która występuje w zbiorze danych najczęściej. Jeśli np. w grupie pacjentów, 32% ma grupę krwi 0, 38% – grupę A, 18% – grupę B, zaś 12% – grupę AB, to dominantą jest wartość A – grupa krwi o największej częstości. Gdy dwie wartości cechy występują równie często, dominanta nie istnieje. Jak widać z przykładu, dominantę można obliczyć już dla cech o skali nominalnej. Można ja oczywiście obliczyć także dla skal silniejszych, choć ponieważ zmienne o skali przedziałowej/ilorazowej często przyjmują bardzo wiele różnych wartości i są dostępne inne miary, zazwyczaj nie oblicza się dominanty.

Mediana (ang. *median*) to przykład miary pozycyjnej, tj. obliczonej na podstawie uszeregowania obserwowanych jednostek według wartości cechy. Może być obliczona dla cech o skali porządkowej lub silniejszych. Mediana to wartość cechy znajdująca się w połowie szeregu uporządkowanych wartości albo inaczej mówiąc taka wartość, że co najmniej połowa jednostek ma wartości cechy nie większe i jednocześnie co najmniej połowa – nie mniejsze. Jeśli na przykład rozważymy grupę 7 pacjentów o wartościach wg klasy NYHA równych (po uszeregowaniu): I, I, II, **II**, III, IV, IV, to medianą jest wyróżniona wartość dla środkowego (tutaj czwartego) pacjenta w szeregu. Zauważmy, że faktycznie co najmniej połowa (dokładnie 4/7) pacjentów ma wartość cechy nie większą niż II (tzn. równą I lub II) i co najmniej połowa (dokładnie 5/7) – nie mniejszą niż II (tzn. równą II, III lub IV).[4] W przypadku publikacji z badań klinicznych dane częściej prezentowane są w postaci częstości występowania poszczególnych wartości, tak jak w tabeli 6.2. Wówczas należy zidentyfikować taką wartość cechy, dla której skumulowany procent pacjentów o wartości cechy mniejszej lub równej po raz pierwszy przekracza 50%. Na przykład w tabeli 6.2. jedynie 28% pacjentów ma wartość ECOG mniejszą lub równą 1, zaś 66% pacjentów – mniejszą lub równą 2. Tak więc mediana jest równa 2.

Analogicznie do mediany definiuje się kwartyle. O ile mediana znajduje się w połowie zbioru, o tyle 1. kwartylem (ang. *1st quartile*) jest wartość znajdująca się w jednej czwartej

[3] Często wyróżnia się także tzw. miary klasyczne i pozycyjne. To rozróżnienie nie wydaje się ważne z punktu widzenia celu niniejszego opracowania. Upraszczając – miary klasyczne obliczane są na podstawie wszystkich jednostek w zbiorze, miary pozycyjne – jedynie wybranych (najczęściej na podstawie miejsca w uporządkowanym zbiorze danych).

[4] Gdy liczba pacjentów jest parzysta, może wystąpić trudność z obliczeniem mediany, ale są to trudności techniczne, rzadko spotykane w praktyce, i pomijamy tutaj stosowane sposoby obliczania mediany.

zbioru, tzn. dzieląca zbiór na co najmniej 25% jednostek mających cechę na poziomie mniejszym lub równym i co najmniej 75% jednostek mających cechę na poziomie większym lub równym. W tabeli poniżej 1. kwartyl jest równy 1. Z kolei 3. kwartyl (ang. *3rd quartile*) znajduje się w trzech czwartych zbioru i w poniższych przykładzie przyjmuje wartość 3.

Dalej wprowadza się tzw. *percentyle* (ang. *percentile*) – i tak 10. percentyl znajduje się w 10% zbioru. W przypadku z tabeli poniżej 10. percentyl jest równy 1, 30. percentyl – 2, itp. Oczywiście 25. percentyl to 1. kwartyl, 50. percentyl to 2. kwartyl, czyli mediana.

Tabela 6.2. Obliczanie pozycyjnych miar położenia.

Wartość ECOG	% pacjentów	skumulowany % pacjentów	przykładowe miary położenia
0	1%	1%	
1	27%	28%	← 1. kwartyl
2	38%	66%	← mediana
3	25%	91%	← 3. kwartyl
4	9%	100%	

Źródło: opracowanie własne

Najczęściej używaną miarą położenia jest średnia, której wzór dla porządku podajemy poniżej:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

gdzie $\overline{x}$ to średnia, x_i to wartość cechy dla pacjenta i, n to łączna liczba pacjentów. Policzenie średniej wymaga, aby sensowna była operacja dodawania wartości cechy – dotyczy to zatem jedynie cech mierzonych na skali przedziałowej lub ilorazowej. O ile więc technicznie rzecz ujmując można policzyć średnią wartość ECOG w powyższym przykładzie, nie jest to poprawne z metodologicznego punktu widzenia.

Średnią można interpretować jako wartość, którą przyjęłaby cecha, gdyby sumaryczną jej wartość w całym zbiorze równomiernie rozłożyć między wszystkie jednostki. Średnia może przyjąć wartość, której nie ma żadna jednostka w badanym zbiorze (np. średnia z liczb 1, 2, 6 wynosi 3), a nawet wartość, która nie ma sensu dla danej cechy (np. średnia z liczby podań danego leku w próbie może być niecałkowita). Oczywiście tych wad nie mają mediana i dominanta.

Wartość średnia jest obliczana na podstawie wartości wszystkich jednostek w próbie, więc niejako wykorzystuje wszystkie dostępne informacje, z drugiej strony jest wrażliwa na pojedyncze obserwacje o skrajnych wartościach (na przykład średnia z liczb 1,1,2,2,3,3,100 wynosi 16, zaś mediana – 2).

Średnia często ma bardziej intuicyjne własności niż mediana. Jeśli w badaniu podany jest średni poziom stężenia hemoglobiny na początku badania i średnia zmiana tego poziomu u badanych pacjentów, to można obliczyć średni końcowy poziom hemoglobiny jako zwykłą sumę średnich. Takie proste obliczenia dla mediany nie muszą dawać poprawnych rezultatów. Jeśli na przykład wartości zmiennej na początku wynoszą: 10, 10, 11, 11, 11, 12, 12, zaś zmiany są równe (po uszeregowaniu rosnąco) 1, 1, 2, 2, 2, 3, 3, to mediana poziomu wyjściowego i zmiany wynosi odpowiednio 11 i 2. Tymczasem w zależności od tego, u którego pacjenta

wystąpiła jaka zmiana, końcowy poziom cechy może przyjąć wartości 11, 11, 13, 13, 13, 15, 15 lub 11, 11, 13, 14, 14, 14 albo też 12, 12, 12, 12, 13, 15, 15, tak więc mediana końcowej wartości cechy może być równa 13 (czyli sumie median), 14 lub 12.

6.1.3. Miary zróżnicowania

Oczywiście miary położenia tym więcej mówią o wartościach w próbie, im mniej wartości te są zróżnicowane. Miary zróżnicowania przeważnie oblicza się dla zmiennych przedziałowych/ilorazowych.[5] Rozstęp (ang. *range*) to różnica między maksymalną i minimalną wartością cechy w zbiorze.[6] Rozstęp kwartylowy (ang. *interquartile range*) to różnica między 3. i 1. kwartylem. Rozstęp kwartylowy jest bardziej odporny na wartości ekstremalne i mówi, w jakim zakresie danych znajduje się co najmniej połowa obserwacji.

Najczęściej wykorzystywaną miarą zróżnicowania jest wariancja, oznaczana często SD^2 i zdefiniowana wzorem:[7]

$$SD^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Tak więc wariancja danej cechy to średni kwadrat różnicy między wartościami tej cechy dla poszczególnych jednostek i średnią w całym zbiorze. Wariancja jest równa zero, jedynie gdy wszystkie jednostki mają tę samą wartość cechy, w przeciwnym wypadku jest dodatnia. Odchylenie standardowe (SD, ang. *standard deviation*) to pierwiastek kwadratowy z wariancji.

Podobnie jak dla mediany, w przypadku łączenia informacji o SD dla dwóch zmiennych lub dwóch grup, należy uważać na możliwe błędy we wnioskowaniu. Zacznijmy od łącznia dwóch zmiennych. Załóżmy, że średni wyjściowy poziom stężenia hemoglobiny w grupie n=100 pacjentów wynosi 10 g/dl, zaś SD tego poziomu 0,5. Przyjmijmy dalej, że średnia zmiana dla tych pacjentów jest równa 2 g/dl z SD równym 1. O ile możemy obliczyć średni końcowy poziom hemoglobiny (10+2 g/dl), o tyle nie ma możliwości dokładnego obliczenia odchylenia standardowego tego poziomu. Jeśli pacjenci o wyjściowo niskim poziomie Hb mieli większe wzrosty tego parametru, to zróżnicowanie końcowego poziomu może być mniejsze niż początkowe. Jeśli z kolei duże wzrosty następowały u pacjentów z wysokim początkowym poziomem – zróżnicowanie może wzrosnąć.

Należy także uważać przy analizie w przypadku łączenia wyników dla kilku grup. Załóżmy, że odchylenie standardowe masy w grupie 100 kobiet wynosi 10 kg oraz że odchylenie standardowe masy w grupie 100 mężczyzn również wynosi 10 kg. Powstaje pytanie, ile jest równe odchylenie standardowe w grupie ogółem tych 200 osób. Otóż to odchylenie będzie równe 10 kg, tylko jeśli także średnia masa w obu tych grupach była równa. W przeciwnym

[5] Istnieją miary, które można interpretować jako miary zróżnicowania już dla zmiennych nominalnych, np. entropia Shannona, ale nie są one wykorzystywane w badaniach klinicznych.

[6] Oczywiście same wartości minimalną i maksymalną można obliczyć już dla zmiennych porządkowych, nie można jedynie (sensownie) obliczyć ich różnicy.

[7] Z powodów wykraczających poza zakres niniejszego opracowania wykorzystuje się często wzór, w którym w mianowniku ułamka wykorzystuje się wartość (n-1) zamiast n. Istotne jest, że dla dużych n różnica jest minimalna.

razie odchylenie standardowe w połączonych grupach będzie większe, gdyż zróżnicowanie w połączonych grupach wynika ze zróżnicowania wewnątrz obu grup, a dodatkowo zróżnicowania między grupami.[8]

Aby móc porównywać zmienność dla różnych cech, wykorzystuje się tzw. współczynnik zmienności (ang. *variability coefficient*), zdefiniowany poniżej:

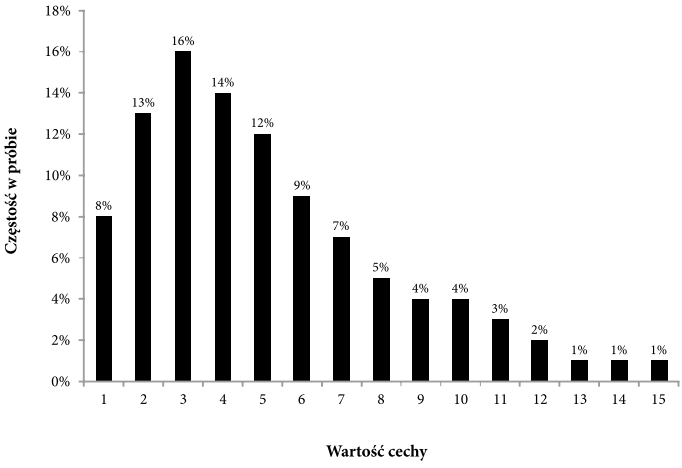
$$V = \frac{SD}{\bar{x}}$$

Warto zauważyć, że współczynnik ten można obliczyć jedynie dla zmiennych mierzonych na skali ilorazowej – o ile licznik nie zależy od umiejscowienia zera (które tylko dla ilorazowych nie jest arbitralne), o tyle mianownik jest na to wrażliwy.

6.1.4. Rozkłady skośne i symetryczne

Na wykresie poniżej przedstawiono przykładowy rozkład danej cechy w populacji, wskazując częstości poszczególnych wartości tej cechy. Parametry tego rozkładu są następujące: 1. kwartyl i dominanta są równe 3, mediana – 4, średnia – 5,16, 3. kwartyl – 6. Rozstęp jest równy 15-1 = 14, zaś rozstęp kwartylowy 6-3 = 3. Jak widać, średnia jest większa od mediany, a ta – od dominanty. Wynika to z faktu, że rozkład jest niesymetryczny – w próbie występuje wiele obserwacji niewiele mniejszych od średniej i nieliczne obserwacje o wiele od niej większe. Taki rozkład określa się jako prawostronnie skośny (ang. *right-skewed*) lub dodatnio skośny (ang. *positively skewed*). Stopień asymetrii mierzy się z wykorzystaniem różnych zdefiniowanych miar, rzadko wykorzystywanych w badaniach klinicznych. Prawostronna skośność pojawia się często wówczas, gdy wartości cechy są z definicji ograniczone z dołu przez jakąś wartość, nie są natomiast ograniczone z góry. Lewostronna skośność występuje zdecydowanie rzadziej.

Wykres 6.1. Przykładowy rozkład cechy w zbiorze danych.



Źródło: opracowanie własne

[8] Ta tzw. dekompozycja wariancji jest wykorzystywana np. w testach typu ANOVA, por. rozdz. 8.

Identyfikacja zmiennych o skośnych rozkładach jest istotna w kilku sytuacjach. Po pierwsze dla skośnych rozkładów zazwyczaj potrzebne są większe próby, aby założenia przyjmowane w procesie estymacji lub testowania hipotez dotyczących danej cechy (patrz niżej) były spełnione.[9] Po drugie przy interpretacji wyników intuicja może być zawodna dla skośnych rozkładów. Należy pamiętać, że w przypadku silnej skośności średnia wartość nie opisuje „typowego" pacjenta, w tym sensie, że relatywnie niewielu pacjentów może mieć wartości tej cechy blisko średniej. Typowi w tym sensie pacjenci znajdują się raczej w okolicy mediany, a tym bardziej dominanty rozkładu (o ile dominantę można wyznaczyć). W tym znaczeniu dla skośnych rozkładów stosowanie mediany zamiast średniej może być uzasadnione.

Z drugiej strony w niektórych sytuacjach to wciąż średnia mierzy to, co jest ważne dla odbiorcy. Załóżmy, że w badaniu mierzono liczbę dni hospitalizacji poszczególnych pacjentów. Ta zmienna bardzo często jest prawostronnie skośna (z definicji liczba dni nie może być mniejsza od zera, natomiast mogą zdarzyć się rzadko bardzo długie hospitalizacje). Jeśli chcemy na podstawie wyników takiej próby dokonać prognozy dotyczącej zużycia zasobów w przyszłości, tj. przewidzieć oczekiwaną całkowitą liczbę dni hospitalizacji dla pacjentów w przyszłości, to średnia jest lepszą miarą niż mediana, gdyż uwzględnia możliwość okazjonalnych bardzo długich hospitalizacji.[10]

6.1.5. Statystyka opisowa – uwagi końcowe

Oczywiście istnieją inne miary położenia i zróżnicowania, a także inne typy miar – np. miary koncentracji. Niemniej to te przedstawione powyżej spotyka się najczęściej w publikacjach wyników z badań klinicznych. Można także zdefiniować miary opisujące rozkłady wielowymiarowe, tj. związki między kilkoma cechami. Ta ostatnia kwestia jest poruszona w rozdziale 9. Należy pamiętać, że obliczanie statystyk opisowych, jak każda generalizacja, może prowadzić do mylnej interpretacji wyników. Wspomniano powyżej przykłady dotyczące oceny średniej czy mediany dla skośnych rozkładów. Ocena rozkładu na podstawie statystyk opisowych wymaga jednoczesnego spojrzenia na miary położenia, zróżnicowania i ocenę skośności, aby zrozumieć charakter danych. Pomocna jest także analiza wykresów prezentujących rozkład.

6.2. Wnioskowanie statystyczne – estymacja

W poprzednim podrozdziale zajmowaliśmy się kwestią analizy rozkładu wartości w próbie, przy czym traktowaliśmy tę próbę jako zbiór danych, który interesuje nas sam w sobie. W dalszej części rozdziału uwzględniamy fakt, że próba jest losowym wycinkiem populacji generalnej, zaś nas interesuje rozkład cechy (lub raczej niektóre jego parametry) w populacji generalnej.

6.2.1. Rozkład zmiennej losowej

Nie wchodząc w kwestie techniczne, warto intuicyjnie zrozumieć, czym jest rozkład. Dla próby losowej, czyli skończonego zbioru pacjentów, rozkład jakiejś cechy to po prostu opis,

[9] Dla porządku zauważmy, że istnieją także symetryczne rozkłady, dla których założenia te, nawet dla bardzo dużych prób, nie są spełnione.

[10] Abstrahujemy w tym miejscu od kwestii błędu estymacji średniej oraz błędu prognozy związanego z losowością danego zjawiska. Oczywiście kwestie te dotyczą także prognozowania na podstawie mediany.

jakie są wartości tej cechy u poszczególnych pacjentów. W przypadku populacji generalnej, która jest tworem teoretycznym, więc niejako obejmuje nieskończoną liczbę pacjentów, przez rozkład możemy rozumieć opis, jak te wartości mogą kształtować się u pacjentów wyłanianych z tej populacji, tj. jakie jest prawdopodobieństwo, że u danego pacjenta cecha przyjmie jakąś konkretną wartość lub wartość z jakiegoś przedziału. Rozważmy dwa przykłady.

Zacznijmy od rozważenia zmiennej przyjmującej jedynie dwie możliwe wartości, czyli zmiennej binarnej lub zerojedynkowej (ang. *binary variable*). Przykłady takich zmiennych w badaniach klinicznych to np. uzyskanie wyleczenia lub nie, zaobserwowanie określonego działania niepożądanego, zgon lub przeżycie (wszystko w określonym horyzoncie czasu). Za pełen opis rozkładu takiej zmiennej w populacji generalnej wystarczy określenie prawdopodobieństwa przyjęcia jednej z dwóch, wyróżnionej wartości (np. wyleczenie, wystąpienie działania niepożądanego, przeżycie) – prawdopodobieństwo przyjęcia drugiej wartości jest dopełnieniem do 100%. Mamy tu do czynienia z tzw. rozkładem dwupunktowym.

Rozważamy teraz przykłady tzw. zmiennych ciągłych (ang. *continuous variables*), np. zmianę stężenia hemoglobiny w czasie leczenia, zmianę BMI, itp. W populacji generalnej możemy teraz obserwować różne wartości tych zmiennych (tzw. różne realizacje) i do opisania rozkładu nie wystarczy jedna liczba. Przez rozkład rozumielibyśmy tutaj raczej pełen opis, jakie jest prawdopodobieństwo, że u pacjenta z populacji generalnej zaobserwujemy wartość cechy z jakiegoś przedziału. Decydenta zazwyczaj nie interesuje pełne opisanie rozkładu, a jedynie poznanie jego wybranych cech, najczęściej wartości oczekiwanej (ang. *expected value*), czasem także określanej pojęciem średniej.[11]

Dalsza część rozdziału dotyczy wnioskowania o parametrach rozkładu na podstawie statystyk opisowych z próby. Ten podrozdział (6.2) dotyczy kwestii estymacji, tj. szacowania, jakie te wartości parametrów są, na podstawie tego, jakie są wartości statystyk opisowych w próbie. Należy zauważyć, że losowość próby oznacza, że inna próba pacjentów dałaby najprawdopodobniej nieco inne wartości statystyk opisowych. Ta losowość jest uwzględniona w procesie estymacji.

6.2.2. Estymacja punktowa

Przedstawmy ideę procesu estymacji na przykładzie. Załóżmy, że rozważamy zmienną binarną – wyleczenie lub nie. Chcemy oszacować, jakie jest prawdopodobieństwo uzyskania wyleczenia w populacji generalnej. Dysponujemy danymi z próby losowej, obejmującej n=70 pacjentów, spośród których u k=42 uzyskano wyleczenie w interesującym nas horyzoncie czasu. Tak więc częstość wyleczeń w próbie wyniosła k/n=42/70=60%. Intuicyjnie, spodziewamy się, że ta liczba informuje nas także o prawdopodobieństwie uzyskania wyleczenia w ogóle (tj. w populacji generalnej). Estymacja punktowa polega na podaniu konkretnej wartości, na poziomie której szacujemy nasz parametr. Estymatorem nazwiemy formułę obliczaną na podstawie próby, którą wykorzystujemy do określenia oszacowania parametru

[11] Ze względu na fakt, że populacja generalna nie składa się ze skończonej liczby pacjentów, nie jest to po prostu średnia z próby policzona dla większej liczby pacjentów. Istnieją na przykład rozkłady, dla których średnia w ogóle nie istnieje. Na potrzeby niniejszego podręcznika można jednak intuicyjnie utożsamiać wartość oczekiwaną w populacji ze średnią z próby.

w populacji generalnej. W przypadku szacowania prawdopodobieństwa zajścia jakiegoś zdarzenia (np. uzyskania wyleczenia), zazwyczaj wykorzystywanym estymatorem jest po prostu częstość tego zdarzenia w próbie – jest to tak naturalne, że może umknąć uwadze fakt, że jest to przenoszenie jednej informacji na inny obszar. Tak więc dokonując estymacji punktowej, określilibyśmy prawdopodobieństwo uzyskania wyleczenia w populacji generalnej na poziomie 60% [12].

Przy estymacji musimy pamiętać o wpływie losowości. Pacjenci w próbie mogli mieć szczęście i uzyskano wyjątkowo dużo wyleczeń lub pecha i uzyskano mniej wyleczeń, niż należałoby oczekiwać. Aby opisać tę losowość bardziej precyzyjnie, musimy wprowadzić pojęcie rozkładu normalnego.

6.2.3. Rozkład normalny – podstawowe informacje

Istnieje cała rodzina rozkładów normalnych różniących się dwoma parametrami – średnią i odchyleniem standardowym. To znaczy stwierdzenie, że jakaś zmienna ma rozkład normalny i podanie średniej oraz odchylenia standardowego w pełni określa rozkład. Rozkład normalny jest rozkładem symetrycznym.

Na wykresie 6.2. zaprezentowano funkcję gęstości (ang. *density function*) rozkładu normalnego standardowego, tj. o średniej równej 0 i odchyleniu standardowym równym 1, często oznaczanego N(0,1). Funkcja gęstości pokazuje, jakie zakresy wartości są bardziej prawdopodobne. Prawdopodobieństwo wylosowania liczby z jakiegoś przedziału liczbowego, to pole powierzchni pod funkcją gęstości dla tego przedziału odłożonego na osi odciętych (osi poziomej). Pole powierzchni pod całą krzywą wynosi zawsze 1. Na wykresie zaznaczono na przykład pole powierzchni na odcinku od -1 do 1, które wynosi 0,6827, tzn. prawdopodobieństwo, że losowo wybrana liczba z rozkładu normalnego standardowego znajdzie się w przedziale [-1,1] wynosi ok. 68,27%. Dla nas ważna jest informacja, że w 95% przypadków wylosowana wartość z rozkładu normalnego standardowego będzie zawierała się w przedziale ok. (-1,96; 1,96). Jeśli więc jakaś zmienna ma rozkład normalny standardowy, to przez typowe wartości możemy rozumieć wartości między -1,96 i 1,96.

Popularność rozkładu normalnego wynika z dwóch powodów. Po pierwsze czasem przyjmuje się, że niektóre cechy (wyrażone bezpośrednio, a często logarytmicznie) są w populacji rozłożone w przybliżeniu zgodnie z rozkładem normalnym, np. poziom IQ, wyniki egzaminów, BMI, wzrost, waga.[13] Po drugie, jak wskazują twierdzenia matematyczne, niektóre miary z prób losowych mają w przybliżeniu rozkład normalny (także dla cech, które same nie mają w populacji rozkładu normalnego).[14]

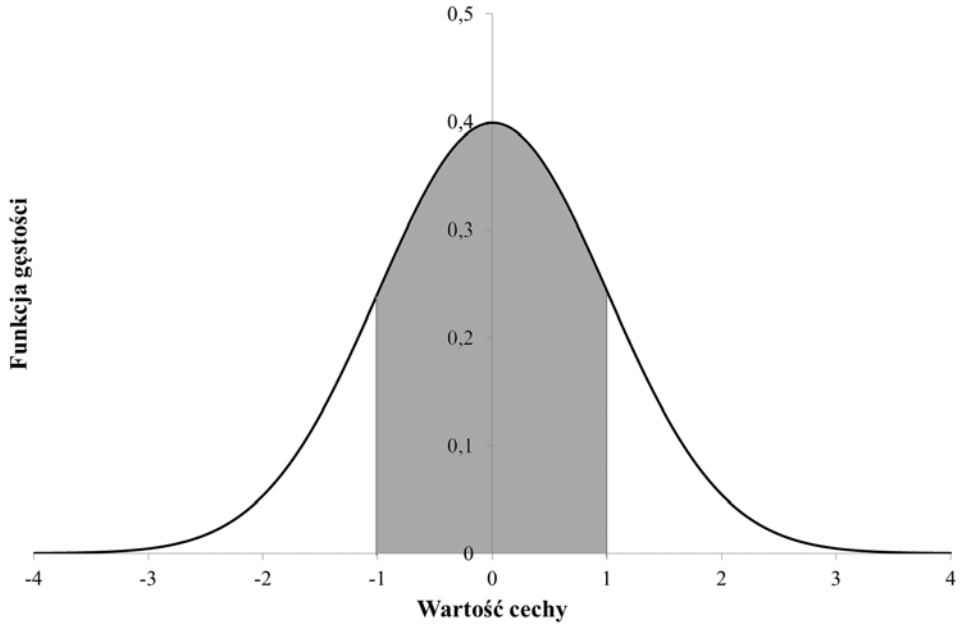
---

[12] Bardzo często estymatorem parametru w populacji generalnej jest analogiczny parametr policzony na podstawie próby. Jak zobaczymy za moment, estymatorem średniej w populacji generalnej jest średnia w próbie. Dodajmy jednak, że bywają sytuacje, gdy znalezienie odpowiedniego estymatora nie jest tylko zamianą słów. Bardzo często stosuje się na przykład nieco inne wzory do obliczenia wariancji w próbie i oszacowania wariancji w populacji na podstawie próby. W innych przypadkach wzory mogą różnić się znacznie bardziej.

[13] Należy podkreślić, że jest to jednie przybliżony rozkład normalny. Jak stwierdził Geary – „…normalność to mit; nigdy nie było i nie będzie rozkładu normalnego …", por. [5] za [8].

[14] Te dwa powody są ze sobą związane – otóż rozkład normalny często jest rozkładem powstającym, jeśli na daną wielkość wpływa addytywnie wiele czynników, z których żaden nie dominuje. Taka sytuacja ma np. miejsce w przypadku wyniku egzaminu – na końcowy rezultat wpływają w identycznym stopniu poszczególne pytania. Taka sytuacja ma także miejsce w przypadku prób losowych – na wartość statystyki opisowej wpływają wartości cechy dla poszczególnych obiektów w próbie.

Wykres 6.2. Funkcja gęstości rozkładu normalnego standardowego.



Źródło: opracowanie własne

6.2.4. Estymacja przedziałowa

Jak wspomniano, wynik konkretnej próby jest obarczony losowością. Załóżmy, że prawdziwe prawdopodobieństwo wyleczenia wynosi p. Przyjmijmy, że losujemy różne próby losowe, każda o rozmiarze n, tzn. prowadzimy dużo różnych badań klinicznych, z których każde obejmuje n pacjentów. Wtedy częstość uzyskania wyleczenia w tych badaniach nie będzie jednakowa. Gdybyśmy przeanalizowali rozkład uzyskanych częstości dla bardzo wielu prób (byłby to rozkład zbioru obejmującego tyle elementów, ile prób przeprowadziliśmy), to okazuje się, że rozkład ten (zakładając, że próby są duże, powiedzmy n≥30), byłby w przybliżeniu normalny o średniej równej p.

Zatem o ile konkretne próby mogą zawyżać lub zaniżać prawdopodobieństwo uzyskania wyleczenia, o tyle średnio dobrze je estymują.[15] Można oszacować odchylenie standardowe rozkładu estymatora, tj. oszacować rozmiar błędu estymacji. W naszym przypadku intuicyjnie zgodzimy się, że większej losowości wyniku spodziewamy się dla prób z mniejszą liczbą pacjentów i *vice versa*. Ocena tego błędu jest dana ogólnym wzorem (i po podstawieniu wartości z naszego przykładu):

$$\mathrm{SE} = \sqrt{\frac{k/n\left(1 - k/n\right)}{n}} = \sqrt{\frac{60\% \times 40\%}{70}} \approx 5{,}9\%.$$

[15] W statystyce mówi się, że taki estymator jest nieobciążony (ang. *unbiased*).

SE oznacza błąd standardowy (ang. *standard error*). Jeśli potrafimy ocenić rozmiar błędu[16], oprócz estymacji punktowej można wykonać estymację przedziałową, tj. podać zakres wartości, a nie tylko konkretną liczbę. Szerokość tego zakresu powinna uwzględniać ocenianą wielkość błędu, tj. informację, jak wartości estymatora w różnych próbach różniłyby się od prawdziwej, nieznanej i estymowanej wartości parametru w populacji. Najczęściej oblicza się tzw. 95% przedział ufności (ang. *95% confidence interval, 95%CI*). Ponieważ, jak już wspomniano, typowe (w znaczeniu – uzyskiwane z prawdopodobieństwem 95%) wartości dla rozkładu normalnego standardowego zawierają się w przedziale (-1,96; 1,96), to ogólny wzór na 95% przedział ufności jest postaci: 95%CI = (estymator punktowy – 1,96 * SE; estymator punktowy + 1,96 * SE). W naszym przypadku:

$$95\%CI = (60\% - 1{,}96 \times 5{,}9\%;\ 60\% + 1{,}96 \times 5{,}9\%) = (48{,}52\%;\ 71{,}48\%).$$

Niestety interpretacja przedziału ufności nie jest intuicyjna. Nie możemy[17] powiedzieć: *Z 95% prawdopodobieństwem przedział ufności zawiera prawdziwą wartość prawdopodobieństwa*. Wynika to z samej filozofii statystyki częstościowej, w której prawdziwa wartość prawdopodobieństwa nie jest zmienną losową, więc w ogóle nie możemy mówić o żadnym prawdopodobieństwie tego, że zawiera się lub nie w jakimś przedziale (albo jest w tym przedziale, albo nie, ale nie jest to coś podlegającego opisowi probabilistycznemu). Poprawna interpretacja brzmi: *Gdybyśmy generowali dużo prób losowych (o tym samym rozmiarze) i obliczali 95%CI według procedury powyżej, to 95% z tych przedziałów ufności zawierałoby prawdziwą wartość prawdopodobieństwa* albo *95%CI, który mam zamiar policzyć na podstawie jeszcze niezebranych danych, z 95% prawdopodobieństwem pokryje prawdziwą wartość*. Różnica w interpretacji polega na tym, że opis probabilistyczny stosujemy tutaj do przedziału ufności, który jest obliczony na podstawie próby losowej, a wartości próby podlegają opisowi probabilistycznemu, bo są losowe. Tak więc sam przedział ufności jest losowy. Podane w publikacjach badań przedziały ufności należy traktować jako wynikające z dostępnych dowodów oszacowania, w jakim zakresie znajduje się prawdziwa wartość parametru. To oszacowanie cechuje się większą lub mniejszą precyzją (tzn. długością) i większą lub mniej pewnością (tj. ilu procentowy jest to przedział).

Długość przedziału ufności dla prawdopodobieństwa zależy od liczebności próby (im większa próba, tym krótszy przedział), stopnia pewności (95%CI jest krótszy od 99%CI) i częstości szacowanego zjawiska (przedział ufności jest dłuższy, jeśli zjawisko ma ok. 50% prawdopodobieństwo, i krótszy, jeśli jest to zjawisko o mniejszym lub większym prawdopodobieństwie).

6.2.5. Estymacja średniej

Dla utrwalenia pojęć rozważmy teraz drugi przykład estymacji, tym razem dla zmiennej ciągłej. Załóżmy, że interesuje nas, o ile średnio w populacji generalnej zmienia się poziom

[16] Pamiętajmy, że mówimy o rozmiarze błędu w zbiorze różnych prób losowych, tj. dla sytuacji, w której losujemy różne próby losowe. Nie wiemy oczywiście, jaki jest konkretny poziom błędu dla naszej (w praktyce jedynej) próby. Nie wiemy także, czy próba zawyża, czy zaniża szacowaną wartość.

[17] Przynajmniej w ujęciu statystyki częstościowej.

hemoglobiny (dla określonej procedury pomiaru) w wyniku stosowania leczenia (w określonym horyzoncie czasu). W próbie 150 pacjentów średnia zmiana hemoglobiny wyniosła 2,4 g/dl, zaś odchylenie standardowe 1,1 g/dl.

W tym przypadku estymatorem punktowym średniej zmiany w populacji generalnej jest znów po prostu średnia zmiana w próbie. Tak więc średnią zmianę w populacji generalnej szacujemy na 2,4 g/dl. Oczywiście próba jest losowa, więc w innej próbie oczekiwalibyśmy nieco innej średniej zmiany poziomu hemoglobiny. Okazuje się, że także w tym przypadku rozkład średnich zmian jest w przybliżeniu rozkładem normalnym wokół prawdziwej oczekiwanej zmiany hemoglobiny. Odchylenie standardowe tego rozkładu mierzy błąd popełniany w pojedynczej próbie. Podobnie jak w poprzednim przykładzie oczekujemy, że wielkość tego błędu maleje z rozmiarem próby.

W obecnym przykładzie występuje dodatkowe zjawisko. Otóż zmiana hemoglobiny może być zmienną, która podlega większej lub mniejszej zmienności u indywidualnych pacjentów – tzn. terapia albo cechuje się stosunkowo dużą przewidywalnością, tj. uzyskuje się podobny efekt terapeutyczny u wszystkich pacjentów, lub też uzyskiwane wyniki silnie zależą od indywidualnych cech i są bardzo zróżnicowane. Tę zmienność na poziomie indywidualnych pacjentów przybliża odchylenie standardowe zmiany hemoglobiny w próbie, wynoszące 1,1 g/dl. Zgodzimy się, że jeśli badane zjawisko podlega dużej zmienności, to wynik pojedynczej próby losowej (tj. średnia wartość dla grupy n pacjentów) jest bardziej narażony na zaburzenie i potencjalnie gorzej przybliża rzeczywistą wartość parametru. Dlatego też we wzorze na wielkość błędu estymatora punktowego pojawia się informacja o odchyleniu standardowym, tj. zmienności zjawiska podlegającego estymacji:

$$SEM = \sqrt{\frac{SD}{n}} = \sqrt{\frac{1{,}1}{150}} \approx 0{,}086.$$

W kontekście estymacji średniej błąd oszacowania często oznacza się SEM, tj. błąd standardowy średniej (ang. *standard error of mean*). Mamy zatem bezpośredni związek między zmiennością wyników w próbie, a oszacowaniem błędu estymacji średniej na podstawie próby. Wzór na 95% przedział ufności jest postaci:

$$95\%CI = (2{,}4 - 1{,}96 \text{ x } 0{,}086;\ 2{,}4 + 1{,}96 \text{ x } 0{,}086) = (2{,}232;\ 2{,}568).$$

W przypadku estymacji średniej długość przedziału ufności wynika oczywiście z rozmiaru próby, pewności przedziału ufności oraz zmienności zjawiska w próbie.

6.2.6. Estymacja – uwagi końcowe

Powyżej przedstawiono jedynie dwa przykłady estymacji. Już nawet w tych sytuacjach można wykorzystywać inne techniki i wzory. I tak na przykład przy estymacji średniej, nieco inne wzory byłyby wykorzystane, gdyby próba była mniejsza, a za to można byłoby założyć, że zmiana hemoglobiny u indywidualnych pacjentów jest zmienną o rozkładzie normalnym. W takim przypadku rozkład błędu w próbie byłby dany tzw. rozkładem t-Studenta (różniącym się od

normalnego zwłaszcza dla małych prób). W praktyce badań klinicznych estymuje się nie tylko prawdopodobieństwa i średnie zmian parametrów. Szacować można np. różnice prawdopodobieństw (wyleczenia dla leku i placebo), różnice średnich zmian (dla leku i placebo), odchylenie standardowe zmiany, współczynnik korelacji dwóch cech, itd. Kwestie te poruszono na różnym poziomie dokładności w kolejnych rozdziałach. Bardziej niż szczegóły istotne jest zrozumienie ogólnej idei estymacji: i) na podstawie próby można obliczyć statystykę traktowaną jako estymator punktowy, ii) ze względu na losowość próby dopuszczamy, że prawdziwa wartość w populacji jest inna, iii) twierdzenia matematyczne pozwalają na oszacowanie rozkładu (typu i parametrów) błędu popełnianego w próbie, iv) uwzględniając ten błąd można dokonać oszacowania przedziałowego o różnym stopniu pewności (zatem i precyzji).

6.3. Wnioskowanie statystyczne – testowanie hipotez statystycznych

Poprzedni podrozdział dotyczył kwestii estymacji, tj. szacowania wielkości parametru w populacji generalnej. Drugim rodzajem wnioskowania statystycznego jest tzw. testowanie hipotez statystycznych (ang. *statistical hypothesis testing*), tj. rozstrzyganie o prawdziwości lub fałszywości stwierdzeń dotyczących populacji generalnej w świetle dostępnych danych.

6.3.1. Podstawowe terminy

W obszarze testowania hipotez statystycznych stosuje się specyficzne słownictwo. I tak, hipoteza, której prawdziwość jest rozstrzygana to tzw. hipoteza zerowa (ang. *null hypothesis*). Nawiązując do poprzedniego podrozdziału, przykładowe hipotezy zerowe mogłyby brzmień: „prawdopodobieństwo wyleczenia wynosi 50%" albo „średnia zmiana hemoglobiny wynosi 2 g/dl". W ocenie technologii medycznych i badaniach klinicznych częściej spotyka się jednak hipotezy zerowe dotyczące dwóch lub więcej leków jednocześnie (albo leku i placebo, różnych dawek leku, itp.), np.: „prawdopodobieństwo wyleczenia jest jednakowe dla leku i placebo", „średnia zmiana hemoglobiny jest identyczna dla obu leków". Jak widać, hipoteza zerowa często jest stwierdzeniem zrównującym jakieś parametry, niejako jest stwierdzeniem „na ostrzu noża". Hipotezę zerową zwykle skrótowo oznacza się H_0.

Hipotezę zerową rozważa się na tle hipotezy alternatywnej, którą przyjmiemy, jeśli odrzucimy hipotezę zerową, tj. uznamy, że hipoteza zerowa jest fałszywa. Hipoteza alternatywna (ang. *alternative hypothesis*) często jest dopełnieniem H_0, tj. stwierdzeniem: „prawdopodobieństwo wyleczenia nie jest równe 50%", „średnia zmiana hemoglobiny nie jest równa 2 g/dl", „prawdopodobieństwo wyleczenia różni się dla leku i dla placebo", „średnia zmiana hemoglobiny jest różna dla obu leków". Hipotezę alternatywną często oznacza się jako H_1.

Przy testowaniu hipotez rozstrzyga się o prawdziwości H_0 na podstawie danych. Oczywiście w praktyce uzyskane wyniki nie są logicznie sprzeczne ani z H_0, ani z H_1. Wyjątkowo mały odsetek wyleczeń w grupie leku może być spowodowany jego niską skutecznością lub pechem pacjentów w tej grupie. Na moment abstrahujmy od sposobu podjęcia decyzji o odrzuceniu lub nie H_0 (tj. uznaniu, że jest fałszywa lub nie). Wówczas możliwe sytuacje ilustruje tabela 6.3. Podejmiemy dobrą decyzję, jeśli nie odrzucimy prawdziwej H_0 lub odrzucimy fałszywą. Popełnimy tzw. błąd I rodzaju (ang. *type I error*), jeśli odrzucimy H_0, a jest ona prawdziwa. Popełnimy błąd II rodzaju (ang. *type II error*), jeśli nie odrzucimy H_0, a jest ona fałszywa. Oczy-

wiście badaczom zależy, aby procedura testowania niosła jak najniższe prawdopodobieństwo popełnienia błędu I rodzaju (dla prawdziwych H_0) i błędu II rodzaju (dla fałszywych H_0). Często wykorzystuje się pojęcie mocy testu (ang. *power*) – prawdopodobieństwa odrzucenia fałszywej H_0. Tak więc przy testowaniu zależy nam na wysokiej mocy testu i niskim prawdopodobieństwie błędu I rodzaju.

Tabela 6.3. Możliwe sytuacje przy testowaniu hipotez statystycznych.

Rzeczywistość:	**Decyzja:**	
	Nie odrzucam H_0	**Odrzucam H_0**
H_0 prawdziwa	poprawna decyzja	błędna decyzja, błąd I rodzaju
H_0 fałszywa	błędna decyzja, błąd II rodzaju	poprawna decyzja, moc testu

Źródło: opracowanie własne

Ponieważ, jak wspomniano, w praktyce często dostępne dane nie są logicznie sprzeczne ani z H_0, ani z H_1, nie ma możliwości skonstruowania testu o mocy 100% i zerowym ryzyku błędu I rodzaju. Jeśli przyjmiemy procedurę częściej odrzucającą H_0, to zazwyczaj rośnie moc testu (bo będąc w 2. wierszu powyższej tabeli, częściej znajdę się w 2. kolumnie), ale rośnie prawdopodobieństwo błędu I rodzaju (będąc w 1. wierszu, częściej znajdę się w 2. kolumnie).

6.3.2. Procedura testowania hipotez

Dalsze rozważania prowadźmy na przykładzie. Załóżmy, że chcemy przetestować H_0: „prawdopodobieństwo wyleczenia jest identyczne dla leku A i B", przy H_1 mówiącej o różnych prawdopodobieństwach. W badaniu podano oba leki 100 osobom i uzyskano 60 wyleczeń dla leku A, tj. 60%, i 50 – dla leku B, tj. 50%. Powstaje pytanie, czy uzyskana różnica[18] 10 p.p. jest wystarczająca do odrzucenia H_0. Zwróćmy dalej uwagę, że intuicyjnie – jeśli różnica 10 p.p. wystąpiła w małej próbie, jest to słabszy powód do odrzucenia H_0, niż gdyby to zdarzyło się w bardzo dużej próbie.

W testowaniu hipotez przyjmuje się następujące podejście. Załóżmy, że H_0 jest prawdziwa. Ponieważ jest to dobrze określone zdarzenie, można obliczyć, jakie jest prawdopodobieństwo uzyskania takiej lub większej różnicy, jaką dostaliśmy w próbie. Jeśli to prawdopodobieństwo jest małe, to znaczy, że obserwowane dane są niezgodne z H_0, tj. są mało prawdopodobne w jej świetle. W takim razie postuluje się odrzucenie H_0.

Badacz może sam określić, jak rozumie sformułowanie „mało prawdopodobne". W praktyce często przyjmuje się progowy poziom, tzw. poziom istotności (ang. *significance level*), równy 5% (por. [3]), często oznaczany symbolem α. Prawdopodobieństwo uzyskania takich (lub bardziej skrajnych) danych oznacza się symbolem p, *p-value*, p* lub nazywa empirycznym poziomem istotności. Jeśli $p<\alpha$, to odrzucamy H_0 i przyjmujemy H_1.

[18] W tym miejscu warto zasugerować stosowanie pojęcia „punkty procentowe", p.p. (ang. *percentage points*). Kiedy mówimy o różnicy – w znaczeniu wyniku odejmowania – dwóch wielkości, które są wyrażane w procentach (jak np. częstości zdarzeń), różnicę należy wyrażać w punktach procentowych, aby uniknąć nieporozumień. Jeśli np. mówimy, że skuteczność dla leku B wynosiła 50%, a dla leku A była o 10% większa, to może (i powinno) to być zrozumiane, że skuteczność dla leku A była równa 50%+10%*50%=55%.

Sposób obliczenia wartości p pozostawmy nieomówiony – wynika z twierdzeń statystycznych, które pozwalają przewidzieć, jakie są najbardziej prawdopodobne wyniki próby, przy założeniu prawdziwości H_0. Sposób obliczenia p zależy od rodzaju testowanych hipotez i rodzaju testu. Dla naszego przykładu wartość p wynosi 0,1552, co oznacza, że przyjmując standardowy poziom istotności α=0,05 uznalibyśmy taką różnicę w wynikach (przy tej liczebności próby) jako dość prawdopodobną ($p>\alpha$) i nie mamy podstaw do odrzucenia H_0.

Należy pamiętać, że wartość p=0,1552 nie oznacza, że prawdopodobieństwo tego, że H_0 jest prawdziwa, wynosi 15,52%! Interpretacja przebiega w odwrotnym kierunku – to pod warunkiem prawdziwości H_0 prawdopodobieństwo zaobserwowania tego, co zaobserwowaliśmy (lub większych odchyleń), wynosi 15,52%. Pamiętajmy o tym, że do prawdziwości H_0 nie możemy w ogóle odnosić stwierdzeń typu probabilistycznego.

W opublikowanych badaniach nie formułuje się zazwyczaj *explicite* hipotezy zerowej, ale jest ona jasna z kontekstu – często dotyczy porównania częstości albo porównania średnich parametrów (zmian parametrów) w dwóch lub więcej grupach.

6.3.3. Ustalania poziomu istotności i mocy testu

Warto uświadomić sobie, czym skutkuje przyjęcie α=5%. Oznacza to, że przez mało prawdopodobne zdarzenia będziemy uznawać wyniki zdarzające się w 5% prób. Konsekwencją tego jest, że testując prawdziwe hipotezy zerowe odrzucimy je w 5% przypadków, tzn. np. stwierdzimy występowanie różnic w skutecznościach tam, gdzie ich w rzeczywistości nie ma, w 5% sytuacji! Oczywiście możemy zredukować to prawdopodobieństwo popełnienia błędu I rodzaju, ale konsekwencją będzie rzadsze dostrzeganie różnic tam, gdzie one są, tj. mniejsza moc testu.

Minimalizację konsekwencji tego konfliktu przeprowadza się następująco (niekoniecznie w tej kolejności chronologicznej). Badacz określa akceptowany przez niego poziom istotności, tj. ryzyko błędu I rodzaju, często na poziomie 5%. Statystycy konstruują test (tj. z naszej perspektywy sposób obliczenia wartości p) tak, aby dla ustalonego α, moc testu była maksymalna możliwa.[19] Badacz tak dobiera rozmiar próby, aby ta moc testu przekroczyła akceptowalny przez niego poziom (często ok. 80%-90%). Warto tu zauważyć, że ponieważ H_1 jest zdaniem niebędącym na ostrzu noża, nie ma możliwości obliczenia w ogólności mocy testu – tę oblicza się dla jakiegoś reprezentanta H_1.

Rozważmy powyższe rozumowanie na przykładzie z badania [4], w którym porównywano natychmiastowe i opóźnione stosowanie paliatywnej radioterapii u pacjentów z nieoperacyjnym, miejscowo zaawansowanym niedrobnokomórkowym rakiem płuca z towarzyszącymi niewielkimi objawami płucnymi. Autorzy piszą (tłumaczenie własne): … *założono,*

[19] Może warto w tym miejscu uświadomić możliwość dobierania różnych testów. Zauważmy, że testem byłoby też wylosowanie z jednakowym prawdopodobieństwem liczby ze zbioru {1,…,20} i odrzucenie H_0, jeśli liczba ta jest równa 1. Taki test na mocy konstrukcji ma ryzyko błędu I rodzaju równe 5%, ale jest to bardzo słaby test – moc też wynosi tylko 5%. Ten test w istocie nie wykorzystuje **żadnych** informacji z próby! Statystycy mogą zaproponować różne metody uwzględnienia informacji z próby i niektóre sposoby mogą dawać mocniejsze testy niż inne. Stosowalność tych metod może jednak zależeć od tego, czy spełnione są jakieś założenia dotyczące badanego zjawiska (np. założenie, że cecha ma rozkład normalny). Z tego względu nie zawsze można zastosować najmocniejszy test – przykładem jest stosowanie tzw. testów nieparametrycznych (ang. *non-parametric testing).*

że odsetek porażek w osiągnięciu pierwszorzędowego punktu końcowego w grupie opóźnionego leczenia wyniesie 70%. Aby wykryć redukcję do 50% w grupie leczenia natychmiastowego przy poziomie istotności 5%, z mocą testu 90%, konieczne było randomizowanie po 150 pacjentów do każdej z obu grup … . W przykładzie tym ustalono poziom istotności na poziomie 5%. Jako reprezentanta H_1 wybrano sytuację, że różnica w poziomach skuteczności wynosi 20 p.p. (70%-50%). Celem badaczy jest uzyskanie takiej mocy testu, żeby tę różnicę (zakładając, że występuje) wykryć z prawdopodobieństwem 90% (tzn. z takim prawdopodobieństwem odrzucić H0 – która oczywiście przy takim założeniu jest fałszywa). Uzyskanie takiej mocy testu wymaga objęcia badaniem 150 pacjentów w każdej z grup.

Wróćmy do ustalania poziomu α. Standardowo przyjmuje się wartość 5%, ale oczywiście można przyjąć inny poziom, np. jeśli bardziej obawiamy się popełnienia błędu II rodzaju. Wyobrażalne jest, że porównując nowy lek do dotychczasowego standardu postępowania pod względem ryzyka wystąpienia działań niepożądanych, błąd II rodzaju (mylne uznanie, że nowy lek nie zwiększa tego ryzyka) martwi nas bardziej niż I rodzaju (błędne stwierdzenie, że lek zwiększa ryzyko). Przyjmowanie innych poziomów istotności często zdarza się przy testowaniu hipotez mających na celu weryfikację założeń pozwalających na stosowanie kolejnych technik (por. testy homogeniczności badań w rozdz. 7).

6.3.4. Pułapki w testowaniu hipotez

Wspomniano powyżej o niewłaściwej interpretacji wartości p. W testowaniu hipotez i interpretacji wyników tego testowania pojawiają się czasem innego typu błędy.

Po pierwsze nieuprawnione jest stwierdzenie „przyjmuję H_0", jeśli $p>\alpha$. Właściwe jest stwierdzenie – „brak jest podstaw do odrzucenia H_0". Hipoteza zerowa nie została udowodniona, jej nieodrzucenie może wynikać z małej liczebności próby – pamiętajmy, że moc testu jest określona dla wybranego „wariantu" H_1.

Po drugie w badaniach czasem spotyka się tzw. testy jednostronne (ang. *one-sided*). O takim teście mówimy, gdy np. dla naszego przykładu H_1 brzmiałaby „skuteczność leku A jest wyższa niż skuteczność leku B". Efektywnie konsekwencją takiego podejścia jest, że poziomy p są dwa razy mniejsze, tak więc częściej odrzucamy H_0 przy tym samym poziomie α. To podejście jest słuszne, jeśli z góry dopuszczalny jest tylko jeden kierunek, w którym H_0 może być nieprawdziwa (tj. albo A i B są równie dobre, albo A jest lepszy). W takim wypadku zastosowanie testu jednostronnego prowadzi przecież do wzrostu mocy testu. Problem powstaje, gdy H_0 może być naruszona w obu kierunkach, a badacz wybiera kierunek testu po zaobserwowaniu danych, zgodnie z tym, co wskazują dane. Nie wchodząc w szczegóły – takie podejście operacyjnie oznacza, że (być może nieświadomie) badacz podwaja rzeczywisty poziom istotności, tzn. zwiększa ryzyko popełnienia błędu I rodzaju do 2α.[20]

Najbardziej subtelnym problemem jest jednoczesne testowanie wielu hipotez zerowych. W badaniach klinicznych taka sytuacja występuje, jeśli na podstawie badania wyciągane są wnioski dotyczące: różnych subpopulacji, różnych punktów końcowych, różnych dawek leku, wyników

[20] Może pomóc następująca analogia. Gdyby badacz zawsze wybierał kierunek testu wbrew kierunkowi sugerowanemu przez dane, to **nigdy** nie odrzuciłby H_0, czyli efektywnie zredukowałby poziom istotności do 0.

w różnych horyzontach czasowych, wyników z użyciem różnych kwestionariuszy, itd. Skoro dla pojedynczej prawdziwej hipotezy zerowej ryzyko jej (błędnego) odrzucenia wynosi 5%, to przy testowaniu dwóch hipotez zerowych (zakładając, że są niezależne) ryzyko odrzucenia **choć jednej** wynosi 9,75%. Przy pięciu Ho – 22,6%, zaś przy 20 – 64%. Oznacza to, że jeśli interpretacja jakościowa wyników badania będzie oparta na choć jednej odrzuconej hipotezie zerowej (np. wykazaniu przewagi leku choć w jednym z momentów), to istnieje bardzo duże ryzyko wyciągnięcia mylnego wniosku. Np. testując lek zupełnie nieskuteczny (równy placebo) w dwudziestu różnych subpopulacjach, w 64% przypadków zidentyfikujemy subpopulację, w której ten lek, zgodnie z wynikami naszego testu, jest w sposób **statystycznie istotny** różny od placebo.[21]

Zapadającą w pamięć ilustrację tego faktu wskazano w badaniu [1], w którym autorzy (świadomi problemu testowania hipotez wielokrotnych, artykuł ma cel dydaktyczny) badają, czy ludzie urodzeni w poszczególnych znakach zodiaku częściej chorują na poszczególne choroby. Oznacza to, że de facto testują wiele hipotez zerowych postaci: „prawdopodobieństwo zachorowania na X jest jednakowe dla urodzonych w znaku Y i pozostałych". Ponieważ hipotez jest bardzo dużo (dla każdego znaku zodiaku tyle, ile testowanych chorób), udaje się dla każdego znaku wskazać choroby statystycznie istotnie częściej występujące. I tak na przykład dla Bliźniąt wyższe o 30% jest ryzyko uzależnienia od alkoholu, p=0,0154.

W pierwszej chwili widzącym ten przykład błędnie wydaje się często, że problemem może być np. rozmiar próby. Tymczasem badanie objęło wszystkich mieszkańców stanu Ontario w Kanadzie (ponad 10 mln osób), a przede wszystkim – rozmiar próby nie ma tu znaczenia o tyle, że ryzyko popełnienia błędu I rodzaju jest kontrolowane z uwzględnieniem rozmiaru próby. Problemem jest tylko i wyłącznie jednoczesne testowanie wielu hipotez.

Czasem w opublikowanych badaniach uwzględnia się ten problem. Redukuje się poziom α dla pojedynczych hipotez, tak aby ryzyko błędu dla grupy hipotez było kontrolowane. Bardzo często stosowane są tzw. korekty Bonferroniego lub Holma-Bonferroniego, por. [6]. Inne możliwości w kontekście analizy wariancji przedstawiono w rozdziale 8. W przypadku badania wpływu znaków zodiaku uwzględnienie korekty wskazało na brak istotnych statystycznie różnic między częstością występowania chorób.

6.3.5. Testowanie hipotez a przedziały ufności

Istotne jest, że między dwoma typami wnioskowania statystycznego – estymacją przedziałową i testowaniem hipotez zachodzi związek. Kontynuując nasz przykład z lekiem A i B na podstawie próby można oszacować przedziałowo o ile punktów procentowych większe jest prawdopodobieństwo wyleczenia dla leku A niż dla leku B. Uzyskany odpowiednimi wzorami[22] 95% CI byłby postaci (-3,7 p.p.; 23,7 p.p.). Zauważmy, że ten przedział ufności zawiera wartość neutralną, tj. świadczącą o braku różnicy, czyli 0. Analogia jest następująca – jeśli 95% przedział ufności parametru X (tutaj różnicy skuteczności) zawiera jakąś wartość (tutaj 0), to znaczy, że testując H_0, iż X jest równe tej wartości, przy poziomie istotności 5% (wynikającym z różnicy 100%-95%), nie będzie podstaw do jej odrzucenia.

---

[21] Oczywiście średnio w 10 subpopulacjach lek po prostu okaże się być lepszy w próbie, w 10 subpopulacjach gorszy – ale wyniki nie będą w większości istotne statystycznie, tj. p będzie >α.

[22] Do kwestii tych wrócimy w rozdziale 7 dotyczącym miar stosowanych w ocenie technologii medycznych.

Powyższa analogia jest bardzo wygodna w praktyce – często podaje się jedynie 95% przedziały ufności, z których wynika, czy H_0 zostałaby odrzucona, czy nie. Niestety w praktyce doprowadziło to też do traktowania estymacji jedynie jako kroku pomocniczego, służącego do odrzucenia lub nie odnośnej hipotezy. Należy pamiętać, że estymacja przedziałowa to alternatywna metoda wnioskowania statystycznego, o innej filozofii niż testowanie hipotez. Inną informację niesie przedział ufności postaci (-3,7 p.p.; 23,7 p.p.) niż np. (-23,7 p.p.; 3,7 p.p.), mimo że w obu przypadkach nie odrzucimy hipotezy zerowej o równoważności obu leków – wynik testowania hipotez często nie przekazuje wielu istotnych informacji, por. [7]. Większość Czytelników słusznie wolałaby jednak w pierwszym przypadku stosować lek A w miejsce B, zaś w drugim odwrotnie. Prowokacyjnie zauważmy, że np. przy 95%CI=(-3,7 p.p.; 23,7 p.p.) nie ma podstaw do odrzucenia hipotezy, że lek A jest lepszy od leku B o 10 p.p.

Reasumując, pamiętajmy o obu technikach wnioskowania – estymacja ma na celu określenie domniemanej wielkości efektu (z uwzględnieniem losowości), zaś testowanie hipotez jest (sproceduryzowaną) metodą rozstrzygania typu tak/nie.

6.3.6. Testy typu *non-inferiority*

Można wręcz znaleźć argumenty przeciwko testowaniu hipotez w niektórych sytuacjach. Co właściwie oznacza hipoteza zerowa, że skuteczność dwóch leków jest **dokładnie** taka sama? Czy z klinicznego punktu widzenia w ogóle dopuszczamy, żeby leki o różnej budowie chemicznej (albo lek i placebo) były **dokładnie** takie same? A jeśli nie, to znaczy, że cała procedura testowania sprawdza hipotezę, o której wiemy, że jest nieprawdziwa. Można przewrotnie stwierdzić, że testujemy jedynie, czy liczebność próby była wystarczająco duża, aby uzyskać odpowiednią moc i odrzucić H_0.

Powyższy problem jest rozwiązywany np. w podejściu typu *non-inferiority*. W takim ujęciu testowana hipoteza zerowa mówi, że badany lek jest gorszy od komparatora o więcej niż δ, gdzie δ jest ustalonym progiem tolerancji (por. rozdz. 4 i 5). Np. jeśli przez p_A, p_B oznaczymy prawdopodobieństwo wyleczenia dla leku A i B, to w podejściu klasycznym testowalibyśmy H_0: $p_A=p_B$, zaś w testach typu *non-inferiority* H_0: $p_A \leq p_B-\delta$. W drugim przypadku hipoteza zerowa jest logicznie możliwa, a jej odrzucenie oznacza, że przyjmujemy, że lek jest nie gorszy.

Dla zilustrowania powyższych kwestii przeanalizujmy następujący fragment badania [2]: *… in order to retain an experiment-wise two-sided type I error of 5%, the noninferiority test for [progression free survival] was studied with a two-sided significance level of 2.5% (the remaining 2.5% were used for the superiority testing). Noninferiority was concluded if the upper limit of the 97.5% CI of the hazard ratio (HR) was ≤ 1.23 …* – (tłumaczenie własne) „… aby zachować w całym badaniu ryzyko błędu I rodzaju przy testach dwustronnych na poziomie 5%, test typu *non-inferiority* dla przeżycia bez progresji analizowano przy użyciu testu dwustronnego i poziomu istotności 2,5% (pozostałe 2,5% wykorzystano do testów typu *superiority)*[23]. Wnioskowano o tym, że lek jest nie gorszy, jeśli górna granica 97,5% przedziału ufności dla ilorazu ryzyka[24] była ≤ 1,23 …". Po pierwsze widać uwzględnienie przez autorów kwestii testów wie-

[23] Test z klasyczną, równościową Ho.

[24] Por. rozdz. 4.

lokrotnych (dwie hipotezy zerowe), po drugie wykorzystanie analogii między przedziałami ufności i testowaniem hipotez, wreszcie podejście typu *non-inferiority* i ustalenie progu δ. Kwestię testowania wielokrotnych hipotez uwzględniono poprzez redukcję przyjętego prawdopodobieństwa popełnienia błędu I rodzaju na poziomie każdej pojedynczej hipotezy – ze standardowych 5% do 2,5%. Testowanie hipotez de facto odbywa się przez analizę górnej granicy przedziału ufności. Podejście typu *non-inferiority* powoduje, że sprawdzane jest nie to, czy w przedziale ufności zawiera się wartość 1 (oznaczająca dokładnie taką samą skuteczność leków), a czy przedział ten obejmuje wartość 1,23, oznaczającą akceptowaną klinicznie różnicę między lekami. Jak widać z powyższego przykładu, czasem w opublikowanych wynikach badań klinicznych występuje wiele aspektów statystycznych jednocześnie.

Bibliografia

1. Austin, P.C.; Mamdani, M.M.; Juurlink, D.N.; Hux, J.E.: Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. Journal of Clinical Epidemiology, 2006, 59, 964-969.
2. Cassidy, J.; Clarke, S.; Díaz-Rubio, E.; Scheithauer, W.; Figer, A.; Wong, R.; Koski, S.; Lichinitser, M.; Yang, T.-S.; Rivera, F.; Couture, F.; Sirzén, F.; Saltz, L.: Randomized Phase III Study of Capecitabine Plus Oxaliplatin Compared With Fluorouracil/Folinic Acid Plus Oxaliplatin As First-Line Therapy for Metastatic Colorectal Cancer. Journal of Clinical Oncology, 2008, 26 (12), 2006-2012.
3. European Medicines Agency, ICH Topic E 9: Statistical Principles for Clinical Trials [dostęp 30 września 2011]. Dostępne w Internecie: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf
4. Falk, S.J.; Girling, D.J.; White, R.J.; Hopwood, P.; Harvey, A.; Qian, W.; Stephens, R.J. i Medical Research Council Lung Cancer Working Party: Immediate versus delayed palliative thoracic radiotherapy in patients with unresectable locally advanced nonsmall cell lung cancer and minimal thoracic symptoms: randomised controlled trial. British Medical Journal, 2002, 325 (7362), 465-471.
5. Geary, R.C.: Testing for normality. Biometrika, 1947, 34, 209-242.
6. Holm, S.: A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 1979, 6 (2), 65-70.
7. Loftus, G.R.: A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. Behavior Research Methods, Instruments & Computers, 1993, 25 (2), s. 250-256.
8. Nester, M.R.: An Applied Statistician's Creed. Applied Statistics, 1996, 45 (4), s. 401-410.
9. Stevens, S.S.: On the Theory of Scales of Measurement. Science, 1946, 103 (2684), s. 677-680.

VII. Tabele 2x2 i miary EBM

Michał Jakubczyk

Celem niniejszego rozdziału jest przedstawienie miar ilościowych powszechnie stosowanych w ocenie technologii medycznych i w obszarze EBM (ang. *evidence based medicine*), takich jak np. iloraz szans, względna redukcja ryzyka czy NNT.

Poniżej rozważamy bodaj najprostszą, ale i często występującą w praktyce, sytuację – porównanie dwóch technologii (np. nowego leku z dotychczas stosowanym albo leku i placebo) ze względu na punkt końcowy, który jest wyrażony zmienną binarną (np. przeżycie w określonym horyzoncie). Dla takiego porównania wyniki badania można przedstawić w postaci tzw. tabeli wielodzielczej (ang. *contingency table*), tj. tabeli, która prezentuje liczbę pacjentów stosujących poszczególne technologie i uzyskujących lub nie punkt końcowy. W przypadku rozważanej uproszczonej sytuacji jest to tabela 2x2, gdyż obie rozważane cechy (stosowana terapia i punkt końcowy) przyjmują dwie możliwe wartości. Przykład zaprezentowano poniżej. Założono, że analizowana interwencja (np. nowy lek) porównywana jest z jednym komparatorem, zaś obserwowany punkt końcowy to przeżycie lub zgon. Oznaczenia symboliczne oraz konkretne wartości liczbowe z tego przykładu są wykorzystywane w całym rozdziale poniżej. W tabeli założono, że grupy pacjentów stosujących interwencję i komparator są jednakowej liczebności, ale nie jest to konieczne założenie.

Tabela 7.1. Przykładowa tabela wielodzielcza 2x2.

Grupa	**Zgon**	**Przeżycie**	**Suma**
Interwencja	A=84	B=16	A+B=100
Komparator	C=93	D=7	C+D=100
Suma	A+C=177	B+D=23	A+B+C+D=200

Źródło: opracowanie własne

Poniżej w pierwszej kolejności zdefiniowano miary pozwalające na ilościowe porównanie wyników dwóch technologii uzyskanych w próbie – podano formuły używane do ich wyliczenia, interpretacje oraz zalety i wady stosowania. W dalszej części wskazano, w jaki sposób wyniki uzyskane w próbie można odnosić do całej populacji generalnej z wykorzystaniem zasad wnioskowania statystycznego. Tak więc zastosowano rozważania przedstawione w rozdziale 6 do konkretnych parametrów EBM. Wreszcie w ostatniej części zasygnalizowano bardziej zaawansowane kwestie, powstające w sytuacji dostępności różnych badań porównujących dwa leki, tj. tzw. metaanalizy.

7.1. Definicja miar EBM

7.1.1. Miary oceny pojedynczej interwencji

Oczywiście najbardziej oczywistą miarą jest określenie częstości występowania punktu końcowego w analizowanej grupie – pacjentów stosujących badaną interwencję lub komparator. Już w tym momencie istnieją dwie możliwości, tj. jako wyróżniony punkt końcowy można potraktować zgon lub przeżycie. Wybór tego punktu końcowego ma następujące znaczenie. Po pierwsze, oczywiście ma konsekwencje dla pożądanego kierunku zmian – jeśli np. wyróżniono zgon, to oczywiście liczymy na to, że stosowanie aktywnego leczenia zmniejsza częstość zgonów w stosunku do placebo. Po drugie, okazuje się, że dla niektórych miar zdefiniowanych poniżej, wybór punktu końcowego może mieć konsekwencje dla uzyskiwanych wyników i możliwych wniosków. Wrócimy do tej kwestii w kolejnych podrozdziałach. Po trzecie, stosowane pojęcia mają konotacje wynikające z języka potocznego i mogą być mniej lub bardziej naturalne dla poszczególnych punktów końcowych. Na przykład sformułowanie „szansa" odpowiada intuicyjnie zdarzeniu pożądanemu, podczas gdy „ryzyko" – niepożądanemu. Wydaje się, że mniej niezręczności wystąpi, jeśli jako punkt końcowy w dalszej części rozdziału wybierzemy zgon, choć w kilku miejscach obliczenia wykonamy także dla przeżycia, jako wyróżnionego punktu końcowego, aby Czytelnik mógł zobaczyć różnice w wynikach.

Aby łatwiej było wprowadzić kolejne miary, zamiast o częstości wystąpienia punktu końcowego będziemy mówić o ryzyku. Ryzyko zgonu w gałęzi interwencji oznaczmy R_I, wówczas $R_I=A/(A+B)$ i $R_I=84\%$ w naszym przykładzie. W gałęzi komparatora odpowiednio mamy $R_C=C/(C+D)$ i $R_C=93\%$.[1]

W EBM oprócz częstości (ryzyka) stosuje się bardzo często pojęcie szansy (ang. *odds*). Popularność tej miary wynika po pierwsze z jej dobrych własności statystycznych w obszarze analiz 2x2, które przedstawiono w dalszych podrozdziałach, a po drugie, ze związków z tzw. modelami regresji logistycznej (por. rozdz. 9). Tak więc szanse (i miary pochodne, takie jak iloraz szans) są bardzo często stosowane i dobrze jest wypracować wyczucie w zakresie rozumienia tego parametru, mimo jego dość nieintuicyjnej definicji.

Otóż szansą dla danego zdarzenia jest iloraz prawdopodobieństwa wystąpienia tego zdarzenia i prawdopodobieństwa jego niewystąpienia, czyli:

$$\text{szansa} = \frac{\text{prawdopodobieństwo}}{1 - \text{prawdopodobieństwo}}.$$

W naszym przykładzie szansą zgonu[2] w próbie jest częstość zgonów podzielona przez częstość przeżycia. Oznaczmy szansę dla pacjentów stosujących badaną interwencję przez O_I, zatem

[1] W praktyce mówi się często o „prawdopodobieństwie" zgonu w próbie. Takie sformułowanie jest poprawne w tym sensie, że częstość jest estymatorem punktowym prawdopodobieństwa zdarzenia w populacji generalnej (por. rozdz. 6). „Częstość" byłoby najbardziej odpowiednim słowem w odniesieniu do próby.

[2] W tym miejscu pojawia się problem z wydźwiękiem sformułowań w języku polskim. Po angielsku słowo *odds* jest neutralne. Na potrzeby przykładu moglibyśmy odwrócić sytuację i mówić o częstości i szansie przeżycia, ale wtedy byłby problem z takimi miarami, jak np. redukcja ryzyka, odnoszonymi do przeżycia, gdyż sformułowanie „ryzyko" ma z kolei wydźwięk negatywny. Utrudnieniem wydawało się z kolei mówienie o różnych punktach końcowych w tych sytuacjach, tj. o przeżyciu przy definiowaniu szansy i zgonie przy definiowaniu np. redukcji ryzyka.

po odpowiednich uproszczeniach $O_I = A/B$, a u nas w przykładzie $O_I = 0{,}84/0{,}16 = 5{,}25$. Dla pacjentów stosujących komparator mamy $O_C = C/D$, czyli $O_C = 0{,}93/0{,}07 \approx 13{,}3$. Zmieniając analizowany punkt końcowy na przeżycie otrzymalibyśmy odpowiednio szanse równe $B/A \approx 0{,}19$ i $D/C \approx 0{,}075$. Z przykładu widać, że szanse zgonu i przeżycia (odrębnie dla leku i komparatora) nie sumują się do 100%. Ze wzoru na szanse łatwo natomiast widać, że iloczyn tych szans jest równy 1.

Widać dalej, że wartość szansy, w odróżnieniu od częstości lub prawdopodobieństwa, nie jest ograniczona do przedziału [0,1]. Wartość szansy jest zawsze nieujemna, natomiast może być dowolnie dużą liczbą dodatnią. Szansa jest równa zero dla zdarzenia o zerowym prawdopodobieństwie, natomiast nie jest zdefiniowana dla zdarzenia zachodzącego ze 100% pewnością (dąży do plus nieskończoności, kiedy prawdopodobieństwo dąży do 100%). Szansa nie zmienia się proporcjonalnie do prawdopodobieństwa, tj. zdarzenie o dwukrotnie większym prawdopodobieństwie będzie miało więcej niż dwukrotnie większą szansę, ale oczywiście większe prawdopodobieństwo oznacza większą szansę. Związek między prawdopodobieństwem a szansą zilustrowano na wykresie 7.1. Dla bardzo małych prawdopodobieństw wartości prawdopodobieństwa i szansy są do siebie zbliżone.

Czytelnikowi może być łatwiej przyswoić intuicyjnie pojęcie szansy przez analogię do miar wykorzystywanych przy zakładach bukmacherskich. W zakładach tych używa się np. sformułowania, że przy obstawieniu zwycięstwa drużyny X, wypłaty są 1:1, czyli stawiając 1 PLN na drużynę X, w przypadku jej zwycięstwa wygramy 1 PLN (czyli otrzymamy postawioną złotówkę z powrotem i jeszcze jedną). Zakładając brak zysku (i straty) bukmachera, tego typu wypłata byłaby oferowana, gdyby prawdopodobieństwo zwycięstwa drużyny X wynosiło 50%, wtedy z 50% prawdopodobieństwem wygrywamy 1 PLN i z takim samy prawdopodobieństwem – tracimy. Tak więc stosunek wypłat 1:1 odpowiada szansie zwycięstwa drużyny X, równej 1 (szansa porażki również jest równa 1).

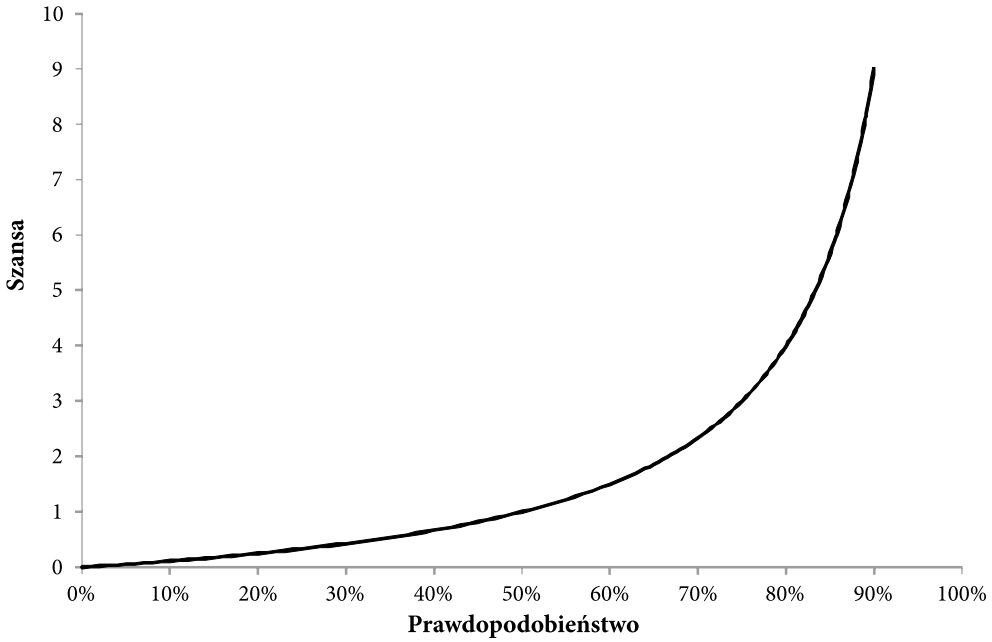
Stosunek wypłat 2:1 (rozumiany tak, że należy postawić 2 PLN, aby móc wygrać złotówkę) oznacza szanse zwycięstwa równe 2, co odpowiada (zakładając neutralność takiego zakładu) prawdopodobieństwu zwycięstwa równemu 2/3=66,(6)%.

Z kolei stosunek wypłat 2:3 (rozumiany tak, że musimy postawić 2 PLN, aby mieć możliwość wygrania 3 PLN) oznacza szanse zwycięstwa równe 2/3, co odpowiada (zakładając neutralność takiego zakładu) prawdopodobieństwu zwycięstwa równemu 2/5 = 40%.

Oczywiście na podstawie szansy można obliczyć prawdopodobieństwo:

$$\text{prawdopodobieństwo} = \frac{\text{szansa}}{\text{szansa} + 1}.$$

Wykres 7.1. Zależność szansy od prawdopodobieństwa



Źródło: opracowanie własne

7.1.2. Miary porównujące dwie interwencje

Bardziej interesujące jest oczywiście porównywanie dwóch technologii, tj. wyrażanie, czy i o ile badana interwencja jest lepsza od komparatora (potraktujmy komparator jako punkt odniesienia). Porównania te mogą być wykonane w oparciu o ryzyko (większość miar) lub o szanse (jedna miara). Ponadto istotne jest rozróżnienie między tzw. miarami bezwzględnymi (wykorzystującymi różnice między wartościami miar dla poszczególnych technologii) i względnymi (wykorzystującymi dzielenie).

Zacznijmy od miar bezwzględnych. Podstawową miarą jest różnica ryzyka (ang. *risk difference*) oznaczana często RD. Jest to po prostu różnica w poziomach ryzyka wyrażona w punktach procentowych. Ponieważ komparator jest punktem odniesienia (a zgon wyróżnionym punktem końcowym), mamy:

$$RD = R_1 - R_C = \frac{A}{A+B} - \frac{C}{C+D}$$

co w naszym przykładzie daje 84% - 93% = -9 p.p. Tak więc powiemy, że w stosowanie leku zamiast komparatora obniżyło w próbie ryzyko zgonu o 9 punktów procentowych.[3] Dodatnia wartość RD oznacza, że dla badanej interwencji częstość zdarzeń w próbie była

[3] Bardzo ważne dla uniknięcie niejednoznaczności jest stosowanie rozróżnienia między procentami a punktami procentowymi, wprowadzonego w rozdziale 6.

większa niż dla komparatora. Oczywiście neutralną wartością RD, świadczącą o braku różnic między technologiami zaobserwowanymi w próbie, jest 0.

Zmiana wyróżnionego punktu końcowego na przeżycie skutkowałaby zmianą wartości RD na przeciwną, tj. w naszym przykładzie usunięciem znaku minus.

Ponieważ często w analizach wyróżniony punkt końcowy jest zdarzeniem niepożądanym, stosuje się miarę, która od razu w sobie zawiera intencję minimalizowania ryzyka wystąpienia tego punktu końcowego. Bezwzględna redukcja ryzyka (ang. *absolute risk reduction*) jest oznaczana ARR i oznacza wartość redukcji ryzyka przy zmianie komparatora na interwencję, tj. jest dana wzorem:

$$ARR = R_C - R_I = \frac{C}{C+D} - \frac{A}{A+B} = -RD.$$

W naszym przypadku bezwzględna redukcja ryzyka wynosi 9 p.p. Jeśli interwencja zwiększyła ryzyko zdarzenia w próbie, to ARR jest ujemne – można także wówczas zastosować raczej pojęcie bezwzględnego wzrostu ryzyka (ang. *absolute risk increase*), ARI. Spotyka się także inne modyfikacje terminologiczne, np. tzw. bezwzględne zwiększenie korzyści (ang. *absolute benefit increase*), ABI, dla wyrażenia wzrostu częstości pożądanego punktu końcowego. W praktyce najczęściej spotyka się miary RD i ARR, istotne jest zaś przede wszystkim, że de facto wszystkie te miary wyrażają tę samą ideę – różnicę poziomów częstości zdarzeń, przy czym różni się jedynie kierunek interpretacji wartości ujemnych lub dodatnich.

Zarówno RD, jak i ARR (ARI, ABI) są miarami bezwzględnymi, tj. wykorzystują odejmowanie miar oceniających poszczególne technologie, mogą zatem przyjmować wartości dodatnie i ujemne. Jako bezwzględną miarę określa się także często (np. w [1]) tzw. miarę *number needed to treat*, NNT.[4] Intuicyjnie miara ta określa – ilu pacjentów należy leczyć z użyciem interwencji a nie komparatora, aby uniknąć jednego niekorzystnego punktu końcowego (uzyskać jeden dodatkowy korzystny punkt końcowy). Okazuje się, że zachodzi prosta zależność:

$$NNT = \frac{1}{ARR},$$

tak więc w naszym przykładzie NNT = 1 / 9 p.p. ≈ 11,1. Oznacza to, że lecząc nieco ponad jedenastu pacjentów lekiem zamiast komparatorem unikamy jednego zgonu. Jeśli analizowana technologia zwiększa ryzyko niekorzystnego punktu końcowego, to ARR jest mniejsze od zera. W takim wypadku najczęściej pomija się znak, a kierunek uwzględnia się w interpretacji, tj. wtedy NNT określa, przy zmianie leczenia u ilu pacjentów z komparatora na interwencję uzyskuje się jeden dodatkowy niekorzystny punkt końcowy. Aby podkreślić ten kierunek interpretacji, stosuje się czasem określenie *number needed to harm* i oznaczenie NNH.

Przejdźmy teraz do miar względnych, tj. odnoszących uzyskaną korzyść do skuteczności komparatora, traktowanego jako punkt odniesienia. Po pierwsze, definiuje się tzw. ryzyko względne (ang. *relative risk*), RR, tj. miarę wyrażającą, jak ma się ryzyko w grupie badanej technologii do ryzyka w grupie komparatora:

[4] Nie odnaleziono wygodnego polskiego odpowiednika.

$$RR = \frac{R_I}{R_C},$$

czyli dla naszego przykładu RR = 84%/93% ≈ 0,9. Wartość ta oznacza, że stosowanie interwencji w miejsce komparatora zmniejsza poziom ryzyka do 90% poziomu wyjściowego. Oczywiście wartością neutralną dla RR jest 1, tzn. dokładne zachowanie poziomu ryzyka. Wartości mniejsze od 1 oznaczają, że ryzyko punktu końcowego jest mniejsze dla interwencji niż dla komparatora, zaś większe od 1 – przeciwnie. Oczywiście RR zawsze przyjmuje wartości nieujemne.

Zamiast mówić o poziomie ryzyka w odniesieniu do ryzyka wyjściowego, można mierzyć, jaka część ryzyka została usunięta w wyniku zastosowania interwencji w miejsce komparatora. Wyraża to tzw. względna redukcja ryzyka (ang. *relative risk reduction*), RRR. Jest ona dana równoważnie następującymi wzorami:

$$RRR = \frac{R_C - R_I}{R_C} = \frac{ARR}{R_C} = 1 - RR.$$

Zatem dla naszego przykładu mamy RRR ≈ 0,1. Oznacza to, że oryginalne ryzyko, tj. ryzyko przy stosowaniu komparatora, zostaje zredukowane w 10% przy zastosowaniu interwencji. W tym przypadku widoczna jest korzyść ze stosowania punktów procentowych dla miar RD i ARR – przy konsekwentnym stosowaniu tego rozróżnienia, informacja, że ryzyko jest mniejsze o 10%, jest jednoznaczna (dotyczy miar względnych, tj. RRR). Wartością neutralną dla RRR jest 0 – w takim przypadku ryzyko nie zostało zmienione. Dodatnie wartości RRR oznaczają, że ryzyko się zmniejszyło, zaś wartości ujemne, że wzrosło.

Wreszcie ostatnia miara względna skuteczności interwencji bazuje nie na ryzyku, a na szansach. Otóż definiuje się tzw. iloraz szans (ang. *odds ratio*), OR, tj. stosunek szansy wystąpienia punktu końcowego dla pacjentów stosujących interwencję przez odpowiednią szansę dla pacjentów stosujących komparator.[5] Tak więc:

$$OR = \frac{O_I}{O_C},$$

czyli przy oznaczeniach z przykładu OR = AD/BC, co dla konkretnych użytych liczb daje OR ≈ 0,4. Oznacza to, że szansa zgonu maleje w wyniku stosowania interwencji. Jak wiemy, większa szansa oznacza większe prawdopodobieństwo, tak więc w naszym przykładzie spadek szansy oznacza zmniejszenie ryzyka zgonu. Wartością neutralną dla OR jest 1. Wartości mniejsze od 1 oznaczają spadek ryzyka punktu końcowego w grupie komparatora, zaś wartości większe – wzrost. OR jest zawsze wartością nieujemną.

[5] Uczyńmy tutaj uwagę terminologiczną. Wspomniano już, że w kolokwialnym użyciu stosuje się czasem określenie „szansa" w znaczeniu „prawdopodobieństwo". Zdarza się również wykorzystanie sformułowania „iloraz szans" dla określenia szansy (co być może wynika z tego, że szansa też jest liczona jako iloraz – choć prawdopodobieństw). Czytelnik powinien zatem uważnie podchodzić do interpretacji wyników.

7.1.3. Zalety i wady poszczególnych miar

Istnieje wiele miar pozwalających na ilościowe porównanie badanej interwencji i komparatora. Naturalne są zatem pytania – czy można wśród nich wyróżnić lepsze i gorsze? których należy używać? na które zwracać uwagę przy interpretacji?

Dobrą wiadomością jest, że przy analizie wyników z próby, wszystkie zdefiniowane powyżej miary zgadzają się jakościowo, tj. co do kierunku zmian. I tak poniższe zdania są zawsze równoważne (w szczególności wszystkie są prawdziwe dla naszego przykładu): częstość niekorzystnego punktu końcowego jest mniejsza dla grupy interwencji niż dla grupy komparatora; szansa niekorzystnego punktu końcowego jest mniejsza dla grupy interwencji niż dla grupy komparatora; RD<0 (ARR>0); NNT jest dodatnie; RR<1; OR<1.

Oczywiście wartości OR i RR są zbliżone, jeśli nie ma dużych różnic między interwencją i komparatorem (wtedy OR ≈ 1 ≈ RR). Ważniejsze jest, że są one także zbliżone do siebie (a już niekoniecznie do 100%), jeśli częstości punktu końcowego są niewielkie dla interwencji i komparatora (ale różne dla tych grup). Jeśli np. mielibyśmy R_I=1% i R_C=2%, to RR=0,5, zaś OR≈0,49. Zawsze wartości OR są dalej od wartości neutralnej (1) niż wartości RR.

Mimo że interpretacje wszystkich miar zgadzają się co do kierunku, mogą się one znacznie różnić co do wielkości i odbioru przez czytelnika. Badania sugerują, że silniej odbierane jest wyrażenie przewagi terapii w kategoriach redukcji ryzyka względnego (RRR) niż bezwzględnej redukcji ryzyka (ARR), por. [6]. Lekarze (decydenci) chętniej decydują się na stosowanie (finansowanie) danej technologii, jeśli jej skuteczność zaprezentowana jest z użyciem miary RRR (zakładając, że technologia generuje korzyści). Wynika to prawdopodobnie z faktu, że liczby są większe dla miary typu RRR. Jeśli np. technologia redukuje ryzyko zgonu z 5% do 4%, to o ile bezwzględna redukcja ryzyka wynosi „jedynie" 1 p.p., o tyle względna redukcja ryzyka „aż" 20%.

Oczywiście miary RD, NNT, RR i RRR są łatwiejsze w interpretacji niż OR (choć ta interpretacja może być inaczej odbierana, jak wynika z poprzedniego akapitu). Z drugiej strony iloraz szans ma lepsze własności matematyczne. Poniżej, w części nawiązującej do metaanaliz zostanie omówiona jedna z nich – tj. zgodność wyników uzyskiwanych w różnych badaniach klinicznych dla tego samego problemu klinicznego. Tutaj przedstawimy dwie inne.

Po pierwsze, odnosząc wyniki oszacowań ilorazu szans do innych subpopulacji, otrzymujemy dopuszczalne wartości, co nie zawsze musi mieć miejsce dla takich miar, jak RD (lub ARR) czy RR (lub RRR). Załóżmy, że w próbie zaobserwowaliśmy wartość RD=-10 p.p., tzn. interwencja redukowała ryzyko punktu końcowego o 10 p.p. Załóżmy teraz, że wiemy, że w innej subpopulacji ryzyko wynosi 8%. Mechaniczne odniesienie wyników próby dałoby ujemne (8%-10%=-2%) ryzyko wystąpienia punktu końcowego w tej subpopulacji leczonej badaną interwencją.

Rozważmy podobny przykład dla RR. Załóżmy, że w próbie zaobserwowaliśmy wartość RR=1,2, tzn. interwencja zwiększała ryzyko punktu końcowego o jedną piątą ryzyka dla komparatora. Załóżmy teraz, że wiemy, że w innej subpopulacji ryzyko wystąpienia punktu końcowego wynosi 85%. Mechaniczne odniesienie wyników próby dałoby teraz wartości przekraczające 100% (1,2 * 85% = 102%) ryzyko wystąpienia punktu końcowego w tej subpopulacji leczonej badaną interwencją.

Przypomnijmy, że szansa może być dowolną liczbą nieujemną. W takim razie odnoszenie dowolnego ilorazu szans (będącego zawsze liczbą nieujemną) do dowolnej szansy dla komparatora zawsze da legalną (czyli nieujemną) wartość szansy.

Należy tu oczywiście wspomnieć, że odnoszenie wyników z próby do (innej) subpopulacji nie jest dobrą praktyką, w związku z czym powyższe ograniczenie jest mniejsze niż się w pierwszym odruchu wydaje, por. [2]. Możliwość przenoszenia wyników z użyciem właśnie ilorazu szans jest natomiast uzasadniona w ramach założeń tzw. wieloczynnikowego modelu regresji logistycznej rozważanego w rozdziale 9.

Po drugie, stosowanie ilorazu szans zwiększa stabilność wyników ze względu na zmianę punktu końcowego. Przypomnijmy, że w naszym przykładzie OR≈0,4 zaś RR≈0,9 (jak widać wartości te znacznie się różnią ze względu na fakt, że punkt końcowy występuje często, oczywiście OR jest dalej od 1). Gdybyśmy teraz obliczyli te miary dla przeciwnego punktu końcowego, tj. przeżycia, to otrzymalibyśmy OR≈2,53 i RR≈2,29 (wartości te nie różnią się teraz tak znacznie, oczywiście OR jest dalej od 1). O ile ilorazy szans są po prostu swoimi odwrotnościami, o tyle nie ma prostego związku dla wartości ryzyka względnego. W dodatku zupełnie inny może być odbiór stwierdzenia, że ryzyko zgonu dla interwencji wynosi 90% ryzyka dla komparatora, niż stwierdzenia, że częstość przeżycia w grupie interwencji jest 2,29 razy większa niż częstość przeżycia w grupie komparatora.

Podsumowując, przy analizie wyników badań klinicznych nie warto ograniczać się do jednej miary, ale ocenić i zinterpretować kilka z nich. Warto wreszcie spojrzeć na dane źródłowe, tj. częstości zdarzeń w próbie, dla ułatwienia np. przeliczając je sobie na 1000 pacjentów.

7.2. Odnoszenie miar EBM do populacji generalnej

Powyżej obliczaliśmy wartości miar EBM dla wyników uzyskanych w próbie. Należy pamiętać, że w ocenie technologii medycznych próba interesuje nas o tyle, o ile dostarcza informacji o populacji generalnej. Jak wskazano w rozdziale 6, istnieją dwa typy wnioskowania statystycznego – estymacja przedziałowa i testowanie hipotez. Na podstawie wyników z próby w praktyce oblicza się nie tylko wartość np. ilorazu szans dla danej interwencji w porównaniu z komparatorem, ale przedział ufności dla tego parametru w populacji generalnej. Dzięki temu wiadomo, jaka jest skuteczność leczenia i na ile to oszacowanie skuteczności może wynikać z przypadkowości związanej z daną próbą. Nie jest naszym celem zaprezentowanie Czytelnikowi wszystkich odnośnych wzorów. Dostępne są darmowe programy (np. RevMan, por. [8]), które obliczają odpowiednie miary wraz z przedziałami ufności, zaś w dostarczonej dokumentacji zawarte są dla zainteresowanych odpowiednie formuły. Istotne jest, że sama idea jest zgodna z przedstawioną w rozdziale 6, tj. przedział powstaje poprzez uwzględnienie oszacowania punktowego pomniejszonego i powiększonego o błąd oszacowania z uwzględnieniem typu rozkładu tego błędu. Obliczane są także wartości p odpowiednich testów – hipoteza zerowa oznacza w nich, że badana miara EBM przyjmuje wartość neutralną, tj. interwencja nie różni się skutecznością od leku. Pamiętajmy, że między wynikami estymacji przedziałowej i testowania hipotez istnieje bliski związek.

W naszym przypadku otrzymalibyśmy następujące 95% przedziały ufności i wartości p (dla ułatwienia podajemy także wartość oszacowania punktowego i hipotezę zerową):

- RD=-9 p.p., 95%CI = (-17,8 p.p.; -0,2 p.p.), p=0,0439, H_0: RD=0;
- ARR=9 p.p., 95%CI = (0,2 p.p.; 17,8 p.p.), p=0,0439, H_0: ARR=0;
- RR=0,9, 95%CI = (0,82; 1), p=0,0483, H_0: RR=1;
- OR=0,4, 95%CI = (0,15; 1,01), p=0,0518, H_0: OR=1.

7.2.1. Przedziały ufności dla NNT

Oczywiście wyniki obliczeń dla ARR są zgodne z odpowiednimi wynikami dla RD. Ponieważ wartość parametru NNT bezpośrednio wynika z wartości ARR, podobnie granice 95% przedziału ufności dla NNT oblicza się na podstawie przedziału ufności dla ARR.[6] Sytuacja jest dość intuicyjna, jeśli przedział ten w całości jest dodatni, to znaczy w całości oznacza redukcję ryzyka dla interwencji w porównaniu z komparatorem. Wówczas oblicza się odwrotności granic 95%CI dla ARR, a dodatkowo zamienia się lewą i prawą granicę, tak aby były uporządkowane rosnąco. Tak więc dla naszego przypadku dostajemy NNT=11,1, 95%CI = (1/0,178; 1/0,002) = (5,6; 407).

Jeśli 95%CI dla ARR jest w całości ujemny, postępuje się analogicznie, z tym że uzyskany przedział ufności interpretuje się w kategoriach NNH. Jeśli np. 95%CI dla ARR byłby w postaci (-20 p.p.; -10 p.p.), to odpowiedni przedział dla NNT/NNH byłby równy (5; 10) i oznaczałby, że liczbę pacjentów, lecząc których uzyskujemy jeden dodatkowy niekorzystny punkt końcowy, szacujemy w zakresie między 5 a 10.

Problemy z interpretacją pojawiają się, gdy przedział ufności dla ARR zawiera 0, tj. granice przedziału ufności są różnych znaków (oznacza to także, że różnica między interwencją i komparatorem w sensie miary bezwzględnej nie jest istotna statystycznie). Przyjmijmy np. 95%CI dla ARR równy (-5 p.p.; 10 p.p.) zawierający oszacowanie punktowe równe 2,5 p.p. Oznacza to, że przewagę interwencji (bezwzględną redukcję ryzyka) szacujemy przedziałowo między -5 p.p. (czyli interwencja jest gorsza niż komparator) a 10 p.p. Mechaniczne zastosowanie powyższych reguł dałoby przedział ufności dla NNT postaci (-20; 10), niezawierający punktowego oszacowania NNT równego 40 (odwrotność 2,5 p.p.). Czasem postuluje się nieobliczanie przedziału ufności dla NNT w takim wypadku.

Metodę pokonania tej trudności pokazał np. Altman [3]. Taki przedział ufności można zapisać jako (NNH=20; NNT=10) i interpretować następująco – szacujemy, że wpływ zastąpienia komparatora przez interwencję zawiera się między następującymi granicami. Dolna (pesymistyczna) granica – zamiana komparatora na interwencję u 20 pacjentów powoduje uzyskanie jednego niekorzystnego punktu końcowego więcej. Górna (optymistyczna) granica – zamiana komparatora na interwencję u 10 pacjentów powoduje uzyskanie jednego niekorzystnego punktu końcowego mniej. Przedział ufności obejmuje przy tym wszystkie wyższe niż 20 wartości NNH i wszystkie wyższe niż 10 wartości NNT. Poza przedziałem są zbyt pesymistyczne oceny, np. – wystarczy stosowanie interwencji już u 15 pacjentów, aby uzyskać jeden niekorzystny punkt końcowy więcej – i zbyt optymistyczne, np. – wystarczy stosowanie interwencji już u 8 pacjentów, aby uniknąć jednego niekorzystnego punktu końcowego. Tak więc różnica polega na innej interpretacji uzyskanych granic przedziału ufności.

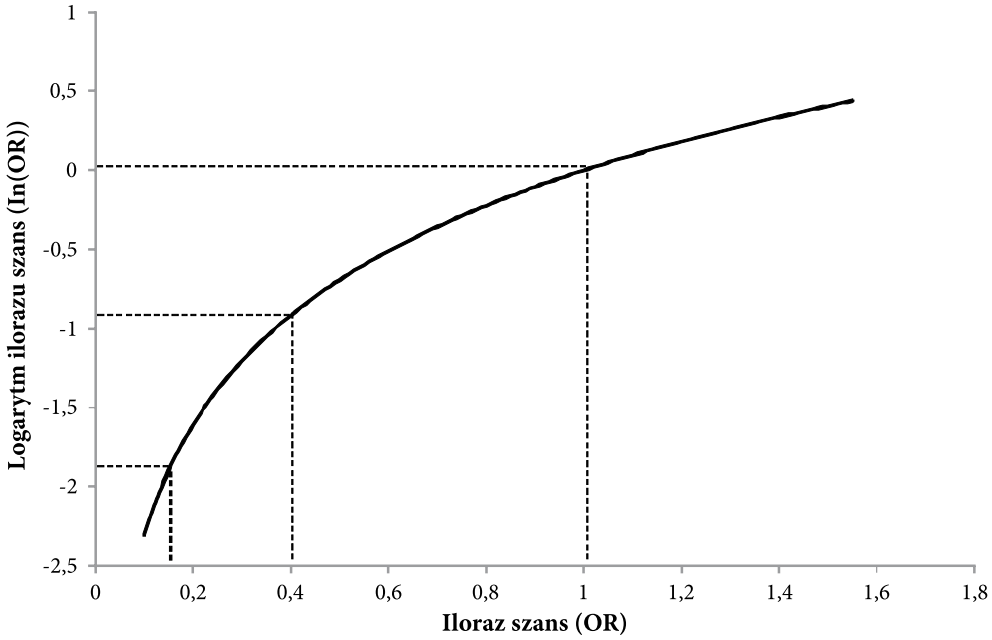
7.2.2. Symetryczne i asymetryczne przedziały ufności

Na przykładzie przedstawionym powyżej widać, że przedziały ufności dla takich miar jako RR, OR czy NNT nie są symetryczne, tzn. oszacowanie punktowe nie znajduje się pośrodku przedziału. Np. OR=0,4 zaś 95%CI = (0,15; 1,01) i jego środek jest równy ok. 0,58.

[6] Istnieją metody wykorzystujące przedziały ufności dla RR lub OR, ale są one rzadziej stosowane.

Wynika to z faktu, że miary te są dane jako ilorazy różnych wielkości i wzory pozwalające na ocenę błędu oszacowania i rozkładu tego błędu są dane nie np. dla samego ilorazu szans, a dla logarytmu tego ilorazu. Tzn. można ocenić błąd oszacowania logarytmu i błąd ten ma rozkład normalny wokół logarytmu oceny punktowej ilorazu szans. Tak więc przedział ufności dla logarytmów odpowiednich miar jest symetryczny wokół logarytmu oceny punktowej, a z kształtu funkcji logarytmicznej wynika, że przedział ufności dla samej miary jest asymetryczny. Kwestię tę ilustruje wykres 7.2. Tak więc asymetryczne przedziały ufności są wynikiem własności matematycznych. Przedziały te są symetryczne w tym sensie, że (np. dla 95%CI) zostawiają po 2,5% po obu stronach granic przedziału.

Wykres 7.2. Iloraz szans i logarytm ilorazu szans – symetryczne przedziały ufności dla logarytmu OR odpowiadają niesymetrycznym przedziałom ufności dla OR. Na wykresie oznaczono wartość oszacowania punktowego OR=0,4 i 95%CI=(0,15; 1,01).



Źródło: opracowanie własne

7.2.3. Testowanie hipotez dotyczących miar EBM

Jak wspomniano powyżej (i szerzej przedstawiono w rozdziale 6), na podstawie wyników estymacji przedziałowej można wnioskować o wynikach testowania odpowiednich hipotez statystycznych. W naszym przypadku – jeśli 95% przedział ufności dla danej miary nie zawiera wartości neutralnej, to znaczy, że odrzucimy hipotezę zerową, że komparator i badana interwencja nie różnią się ze względu na tę miarę.

Na przykład w naszym przykładzie 95%CI dla ARR jest postaci (0,2 p.p.; 17,8 p.p.) i nie zawiera wartości neutralnej, tj. 0. Oznacza to, że przy 5% poziomie istotności należy odrzucić

hipotezę zerową, że redukcja ryzyka jest równa zero. Obliczony odpowiedni poziom p wynosi 0,0439, co potwierdza wniosek.

W interpretacji mogą pojawić się trudności. W naszym przykładzie widać, że trudne może być wnioskowanie na temat statystycznej istotności przewagi interwencji. O ile dla ARR czy RR różnica jest statystycznie istotna (odpowiednio p=0,0439 i p=0,0483), o tyle dla OR – nie, p=0,0518. Powstaje problem, co to oznacza dla uznania interwencji za lepszą od komparatora.

Warto w tym miejscu przypomnieć, czemu służy procedura testowania hipotez statystycznych. Jest to algorytm rozstrzygania o prawdziwości zdań, przy czym nie jest on nieomylny, a jedynie tak skonstruowany, aby zapewnić utrzymanie w założonych graniach prawdopodobieństw popełnienia błędów I i II rodzaju. Tak więc wnioskowanie na podstawie samego tylko RR może mieć ryzyko błędu I rodzaju równe 5% (jeśli przyjmiemy taki poziom istotności) i określoną moc (zależną od rozmiaru próby), podobnie wnioskowanie tylko na podstawie OR. Jeśli chcemy wnioskować na podstawie obu tych miar jednocześnie, np. przyjąć, że odrzucamy hipotezę o równoważności, jeśli jednocześnie oba poziomy p<0,05 – to de facto definiujemy mechanizm decyzyjny, który ma swój poziom ryzyka błędu I rodzaju (mniejszy niż 5%, gdyż rzadziej odrzucimy prawdziwą H_0), ale też mniejszą moc (rzadziej odrzucimy fałszywą H_0). Symetryczna sytuacja powstanie, gdy uznamy, że wystarczy jeden poziom p <0,05.

Drugi problem wiąże się z dyskusją o wzajemnych przewagach RR i OR. Wspomniano powyżej, że OR ma wygodniejsze własności matematyczne. W przypadku testowania hipotez statystycznych ujawnia się kolejna z nich. Otóż dla OR testowanie hipotezy o równoważności komparatora i interwencji da taki sam wynik niezależnie od wyboru punktu końcowego (zgon lub przeżycie), tj. p=0,0518. Dla RR wykonanie analizy dla przeżycia daje 95% CI dla RR postaci: (0,98; 5,31), p=0,0548. Zatem przy poziomie istotności 5% podjęlibyśmy inną decyzje kierując się miarą RR w zależności od tego, czy wyróżnionym punktem końcowym byłyby zgony (statystycznie istotna przewaga interwencji) lub przeżycie (brak statystycznie istotnych różnic).

7.3. Wprowadzenie do metaanaliz

Często dostępne jest więcej niż jedno badanie porównujące dwie technologie ze względu na ten sam punkt końcowy. Intuicyjne jest, że warto uwzględnić wszystkie dostępne dowody, aby uzyskane oszacowania były jak najdokładniejsze. Łączenie wyników różnych badań, tj. praca na już wykonanych analizach, tak aby uzyskać syntetyczne ich wyniki, to tzw. metaanaliza (ang. *meta-analysis*).

Okazuje się, że mechaniczne połączenie wyników badań, poprzez zsumowanie liczby pacjentów w odpowiednich komórkach tabel 2x2 może prowadzić do błędnych wniosków. W tabeli 7.2 zaprezentowano przykład, w którym porównano interwencję i komparator ze względu na częstość zgonów. Badania są o tyle szczególne, że nie były zrównoważone liczebności ramion – w pierwszym więcej pacjentów otrzymywało komparator, w drugim – interwencję. Dodatkowo badania różniły się ogólnym poziomem ryzyka zgonu, które w II badaniu było większe. W obu badaniach indywidualnie interwencja redukuje ryzyko zgonu. Po zwykłym zsumowaniu liczb, w połączonych badaniach ocena wskazuje na pogorszenie rokowań przy stosowaniu interwencji. Sytuacja taka jest przykładem tzw. paradoksu Simpsona, który ogólnie ujmując dotyczy sytuacji, w której wyniki analizy podgrup nie są zgodne z wynikami analizy w całej grupie.

Tabela 7.2. Przykład paradoksu Simpsona w kontekście łączenia wyników badań klinicznych.

Badanie	Lek	Zgon	Przeżycie	Suma	Ocena interwencji
I	Interwencja	20 (20%)	80 (80%)	100	OR = 0,75
	Komparator	50 (25%)	150 (75%)	200	RR = 0,8
II	Interwencja	100 (50%)	100 (50%)	200	OR = 0,82
	Komparator	55 (55%)	45 (45%)	100	RR = 0,91
I + II	Interwencja	120 (40%)	180 (60%)	300	OR = 1,24
	Komparator	105 (35%)	195 (65%)	300	RR = 1,14

Źródło: opracowanie własne

Paradoks Simpsona pojawia się w rzeczywistych sytuacjach, często w kontekście łączenia wyników dla różnych grup pacjentów. W badaniu [4] porównywano dwie metody usuwania kamieni nerkowych – klasyczne leczenie chirurgiczne i PCNL. Dokonano porównania łącznie i w rozbiciu na grupy pacjentów o różnej średnicy kamienia (<2 cm i ≥2 cm). Dla obu grup pacjentów rozważanych odrębnie chirurgia była bardziej skuteczna niż PCNL: odsetek sukcesów 93% vs 87% dla pierwszej grupy i 73% vs 69% dla drugiej grupy. Po połączeniu obu grup chirurgia okazała się być mniej skuteczna: 78% vs 83%. Taki wynik był (matematycznie ujmując) skutkiem tego, że po pierwsze w pierwszej grupie dominowały przypadki PCNL, zaś w drugiej grupie na odwrót, a po drugie w pierwszej grupie odsetek sukcesów był w ogóle większy. Tak więc połączenie obu grup działa na korzyść PCNL, gdyż w obliczeniach uwzględniana jest znaczna liczba pacjentów z grupy, w których odsetek sukcesów jest wysoki. To odwrócenie proporcji liczby pacjentów w grupach o różnym ryzyku może wynikać np. z braku randomizacji i tendencji klinicystów do częstego stosowania danej technologii tylko w grupach o wysokim ryzyku (np. jeśli jest to technologia droga albo związana z dyskomfortem pacjenta i dopiero ciężki stan kliniczny uzasadnia jej użycie). To badanie, jak również inne przykłady omówione są np. w [5].

Podkreślmy, że właściwa interpretacja wyników powinna brzmieć tak, że to lek jest lepszy w przykładzie pokazanym w tabeli, zaś klasyczne leczenie chirurgiczne – w drugim przykładzie. Przewaga komparatora (w tabeli) i PCNL (w drugim przykładzie) jest sztucznym rezultatem, wynikającym z dysproporcji między ramionami w grupach w połączeniu z silniejszym efektem grupy niż samego leku.

Tak więc przy metaanalizach, aby uniknąć powyższych problemów, opracowano techniki statystyczne łączenia wyników różnych badań. Istotne jest, że techniki te nie łączą danych pierwotnych z poszczególnych badań (tj. nie sumują liczby pacjentów), a raczej uśredniają wartości wybranej miary EBM, przy czym wagi przyjęte dla poszczególnych badań wynikają z jakości danego badania – większe badania (oferujące dokładniejsze oszacowania) mają większą wagę. Wynikiem metaanalizy jest przede wszystkim łączne oszacowanie wybranych miar EBM na podstawie wszystkich badań, wraz z odpowiednim przedziałem ufności i poziomem p odpowiedniego testu.

Rozważmy przykład opracowany na podstawie metaanalizy przedstawionej w [7]. Metaanalizowano wyniki badań dotyczących skutków modyfikacji hipoksji tkanek w radioterapii raka płaskonabłonkowego głowy i szyi. Przeprowadzono przegląd systematyczny badań i wykonano m.in. metaanalizę 9 badań, w których stosowano terapię hiperbaryczną tlenem, ze względu na ryzyko zgonu z powodów związanych z chorobą.

Wykres 7.3. przedstawia wynik metaanalizy przeprowadzonej z wykorzystaniem oprogramowania RevMan [8]. Metaanalizę przeprowadzono dla ilorazu szans. W górnej części wykre-

su przedstawiono wyniki dla poszczególnych badań. Po lewej stronie w formie tabelarycznej – liczbę pacjentów stosujących porównywane technologie i liczbę pacjentów, u których wystąpił punkt końcowy. Po prawej stronie w postaci wykresu (tzw. *forest plot*), w którym zaznaczono linią pionową wartości neutralnej dla OR (1), niebieskimi punktami wartość oszacowania punktowego (wartości po lewej stronie wartości neutralnej świadczą o redukcji szansy zgonu przy stosowaniu interwencji), zaś liniami poziomymi – 95% przedziały ufności.

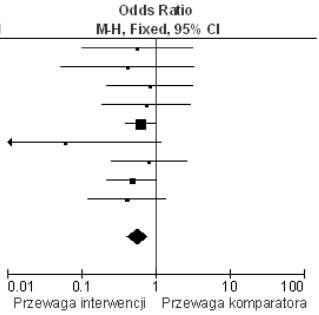
W dolnej części wykresu przedstawiono wyniki oszacowania łącznego. Po lewej stronie wartość OR. Po prawej, diamentem, graficznie wartość OR i zakres przedziału ufności. Jak widać, połączenie wyników pozwoliło na uzyskanie oszacowania o precyzji znacznie przewyższającej precyzję ocen z poszczególnych indywidualnych badań. Uzyskany wynik OR=0,56, 95%CI=(0,41; 0,78) oznacza, że stosowanie terapii hiperbarycznej tlenem redukuje ryzyko zgonu z powodów związanych z chorobą, co więcej różnica ta jest istotna statystycznie (p=0,0006).

Metaanalizę można przeprowadzić także dla innych miar EBM. Wykres 7.4. przedstawia wyniki dla RD, czyli miary bezwzględnej – jak widać, interpretacja jakościowa nie zmienia się – terapia hiperbaryczna tlenem redukuje ryzyko zgonu i redukcja ta jest statystycznie istotna (p=0,0005), tj. przypisywalibyśmy ją faktycznej przewadze technologii, a nie jedynie przypadkowości wyników w skończonej próbie losowej.

Wykres 7.3. Wyniki metaanalizy dla OR.

Study or Subgroup	T. hiperbaryczna tlenem Events	Total	Grupa kontrolna Events	Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Badanie 1	3	15	4	13	3.6%	0.56 [0.10, 3.17]
Badanie 2	5	9	6	8	3.0%	0.42 [0.05, 3.31]
Badanie 3	20	26	20	25	5.0%	0.83 [0.22, 3.18]
Badanie 4	10	18	10	16	5.0%	0.75 [0.19, 2.96]
Badanie 5	65	125	96	151	44.0%	0.62 [0.38, 1.01]
Badanie 6	0	9	7	15	5.8%	0.06 [0.00, 1.21]
Badanie 7	9	21	11	23	6.3%	0.82 [0.25, 2.69]
Badanie 8	21	53	29	50	19.0%	0.48 [0.22, 1.04]
Badanie 9	13	23	19	25	8.3%	0.41 [0.12, 1.41]
Total (95% CI)		**299**		**326**	**100.0%**	**0.56 [0.41, 0.78]**
Total events	146		202			

Heterogeneity: Chi² = 3.68, df = 8 (P = 0.88); I² = 0%
Test for overall effect: Z = 3.41 (P = 0.0006)

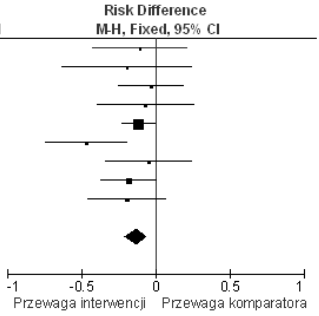


Źródło: opracowanie własne na podstawie wyników [7]

Wykres 7.4. Wyniki metaanalizy dla RD.

Study or Subgroup	T. hiperbaryczna tlenem Events	Total	Grupa kontrolna Events	Total	Weight	Risk Difference M-H, Fixed, 95% CI
Badanie 1	3	15	4	13	4.5%	-0.11 [-0.43, 0.21]
Badanie 2	5	9	6	8	2.7%	-0.19 [-0.64, 0.25]
Badanie 3	20	26	20	25	8.2%	-0.03 [-0.26, 0.19]
Badanie 4	10	18	10	16	5.5%	-0.07 [-0.40, 0.26]
Badanie 5	65	125	96	151	44.1%	-0.12 [-0.23, 0.00]
Badanie 6	0	9	7	15	3.6%	-0.47 [-0.75, -0.19]
Badanie 7	9	21	11	23	7.1%	-0.05 [-0.34, 0.24]
Badanie 8	21	53	29	50	16.6%	-0.18 [-0.37, 0.01]
Badanie 9	13	23	19	25	7.7%	-0.19 [-0.46, 0.07]
Total (95% CI)		**299**		**326**	**100.0%**	**-0.13 [-0.21, -0.06]**
Total events	146		202			

Heterogeneity: Chi² = 7.38, df = 8 (P = 0.50); I² = 0%
Test for overall effect: Z = 3.49 (P = 0.0005)



Źródło: opracowanie własne na podstawie wyników [7]

Metaanalizy można wykonać dla różnych miar – OR, RR, RD. Skoro wyniki dla pojedynczego badania mogą się różnić dla tych miar (w sensie statystycznej istotności), podobnie może być dla wielu badań. Cochrane Collaboration wskazuje, że przeprowadzone badania sugerują, że stosowanie miar względnych, tj. OR i RR, jest bezpieczniejsze o tyle, że mniej różnią się dla różnych badań i subpopulacji [2]. Dodatkowo przy wykonaniu metaanalizy istnieje możliwość zastosowania różnych technik ważenia pojedynczych badań (metody Mantela-Haenszela, Peto, odwrotności wariancji). Omówienie metod wykracza poza zakres podręcznika, poprzestańmy na komentarzu, że metody te są dedykowane i ujawniają swoje przewagi dla różnych sytuacji. Cochrane Collaboration w większości sytuacji sugeruje wykorzystanie metody Mantela-Haenszela.

Przy przeprowadzaniu i interpretacji wyników metaanalizy istotne jest zrozumienie zagadnienia heterogeniczności (ang. *heterogeneity*) włączonych badań klinicznych. Przez heterogeniczność rozumiemy tu zróżnicowanie wyników między pojedynczymi badaniami. Rozróżnijmy trzy sytuacje.

Po pierwsze może być tak, że wyniki są w miarę zbieżne dla wszystkich uzyskanych badań (tak jak w przykładzie powyżej). Bylibyśmy wówczas skłonni stwierdzić, że stosowanie interwencji generuje efekt kliniczny (np. w postaci parametru OR<1), który był taki sam w poszczególnych próbach, a jedynie w wyniku wpływu losowości na poziomie każdej próby konkretne oszacowanie OR z każdej próby dało trochę inny efekt. W takiej sytuacji chcielibyśmy estymować w populacji generalnej ten jeden, wspólny efekt leku. Wykorzystalibyśmy zatem tzw. model efektów stałych (ang. *fixed-effects model*). I odwrotnie – wykorzystanie takiego modelu oznacza, że wierzymy, że jest jeden poziom efektu, który szacujemy.

Po drugie może być tak, że wyniki badań różnią się między sobą i o ile wierzymy, że badania te dotyczą jednego problemu klinicznego i możemy je rozważać wspólnie, to widzimy, że z jakichś powodów efekt kliniczny leku był różny w różnych badaniach. Dodatkowo na ten różny poziom efektu nałożyła się losowość związana z przypadkowością w wynikach poszczególnych prób. W takiej sytuacji właściwe jest zastosowanie tzw. modelu efektów losowych (ang. *random-effects model*), tzn. zakładamy, że nie istnieje po prostu jedna wartość tego parametru, a raczej, że jest on inny w każdym badaniu i jest losowany z jakiegoś rozkładu. Wobec tego możemy szacować jedynie wybrany parametr tego rozkładu i zazwyczaj szacuje się średnią. Praktyczną konsekwencją wyboru modelu efektów losowych jest zastosowanie innych wag dla poszczególnych badań, większy błąd oszacowania i zazwyczaj nieco inny wynik estymacji punktowej. W naszym przypadku przy użyciu (tu nieuzasadnionym) modelu efektów losowych, otrzymalibyśmy OR=0,58, 95%CI=(0,42; 0,81).

Aby rozstrzygnąć między powyższymi dwiema sytuacjami, można wykorzystać podejście ilościowe. Istnieją testy statystyczne pozwalające na weryfikację hipotezy zerowej mówiącej o braku różnic parametru między badaniami. Na wykresach powyżej poziom p odpowiedniego testu wynosi kolejno p=0,88 i p=0,5. W praktyce często badacze bardziej obawiają się błędu II rodzaju, tj. niestwierdzenia heterogeniczności tam, gdzie ona występuje, więc stosują poziom istotności 0,1. Statystykę I^2 interpretuje się jako część zmienności wyników między badaniami wynikającą ze zmienności samego parametru EBM, a nie ze zwykłej losowości. W powyższych przykładach wynosi ona 0, tzn. oceniamy, że różnice w poszczególnych badaniach są po prostu skutkiem losowości procesu leczenia.

Z powyższych rozważań warto chyba wyciągnąć jeszcze ogólną informację dotyczącą znaczenia słowa model w analizach ilościowych, bo bywa ono niewłaściwie rozumiane. Jak widać, nie jest to tylko i wyłącznie pakiet technik pozwalający na ocenę jakiegoś interesującego parametru, a także przeważnie zestaw założeń przyjętych przez analityka dotyczących tego, jak według niego skonstruowany jest świat, generowane są dane, itd. Dopiero z tych założeń (często arbitralnych i trudno weryfikowalnych) wynika możliwość zastosowania jakiegoś wzoru itp.

Po trzecie wreszcie może zdarzyć się sytuacja, że heterogeniczność badań jest bardzo duża, gdyż np. dotyczą one różnych problemów klinicznych, np. rożnych subpopulacji, dawkowania, itp. Nie powinno się w takim wypadku przeprowadzać metaanalizy. Jej wynik byłby niemożliwy do zinterpretowania, gdyż uśrednialibyśmy wyniki uzyskane w różnych kontekstach klinicznych. Wynik takiego uśrednienia byłby zdominowany przez typy badań (np. subpopulacje), które dominowały w wynikach przeglądu, a odnosilibyśmy go także do innych subpopulacji, w których być może efekt kliniczny leku jest zupełnie inny. Uwzględnienie różnic ilościowych między badaniami może odbyć się np. z wykorzystaniem tzw. metaregresji (ang. *meta-regression*), zaś jakościowych – poprzez syntezę jakościową wyników, czy analizę w podgrupach.

7.4. Uwagi końcowe

W niniejszym rozdziale omówiono jedną z dwóch bardzo ważnych sytuacji w ocenie technologii medycznych – porównanie dwóch technologii z uwagi na binarny punkt końcowy (np. wyleczenie lub brak). Wbrew pozorom takie ograniczenie jest bardzo często zasadne. Nawet jeśli w badaniu analizowane są więcej niż dwie technologie i wiele punktów końcowych, porównania wykonuje się parami dla poszczególnych punktów końcowych.

Powyżej nie odniesiono się do miar ciągłych, ale w takim przypadku stosują się metody przedstawione w rozdziale 8. Dla miar ciągłych także możliwe jest przeprowadzanie metaanaliz, których wynikiem jest tzw. średnia ważona różnica (ang. *weighted mean difference*, WMD).

Poza samymi definicjami miar i interpretacjami, warto z powyższych rozważań zapamiętać, że na odbiór wyników wpływa także wybór metody (miary, punktu końcowego, metody metaanalizy). Przy interpretacji wyników badań warto zatem oceniać także metodykę, a w miarę wątpliwości i możliwości powtórzyć obliczenia na podstawie danych źródłowych dla innych miar EBM. Polskie wytyczne przeprowadzania analiz HTA wymagają zastosowania co najmniej jednej miary bezwzględnej i względnej, przedstawienia wyników testu homogeniczności i metody metaanalizy [1].

Bibliografia

1. Agencja Oceny Technologii Medycznych: Wytyczne oceny technologii medycznych (HTA). Wersja 2.1. Warszawa, 2009.
2. Alderson, P.; Green, S. (red.): Cochrane Collaboration open learning material for reviewers. Version 1.1, November 2002., [dostęp 21 września 2011]. Dostępny w Internecie: http://www.cochrane-net.org/openlearning/index.htm
3. Altman, D.G.: Confidence intervals for the number needed to treat. British Medical Journal, 1998, 317, 1309-1312.

4. Charig, C.R.; Webb, D.R.; Payne, S.R.; Wickham, J.E.A.: Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. British Medical Journal, 1986, 292 (6524), 879-882.
5. Julious, S.A.; Mullee, M.A.: Confounding and Simpsons's paradox. British Medical Journal, 1994, 309 (6967), 1480-1481.
6. McGettigan, P.; Sly, K.; O'Connell, D.; Hill, S.; Henry, D.: The Effects of Information Framing on the Practices of Physicians. Journal of General Internal Medicine, 1999, 14 (10), 633-642.
7. Overgaard, J.: Hypoxic modification of radiotherapy in squamous cell carcinoma of the head and neck – A systematic review and meta-analysis. Radiotherapy and Oncology, 2011, 100 (1), 22-32.
8. Review Manager, wersja 5.1, strona domowa [dostęp 21 września 2011]. http://ims.cochrane.org/revman

VIII. Najczęściej wykorzystywane testy statystyczne

Łukasz BOROWIEC

W rozdziale 6.3. przedstawiono najważniejsze pojęcia, założenia i przykłady zastosowania procedury testowania hipotez statystycznych. Celem niniejszego rozdziału jest pogłębienie tego tematu, a w szczególności bardziej szczegółowe omówienie najczęściej stosowanych w analizie danych medycznych testów dotyczących zmiennych ciągłych. W dalszej części rozdziału omówione zostaną czynniki, które decydują o użyciu danego rodzaju testu do analizy określonego rodzaju danych, a także przedstawiony będzie algorytm pomocny w wyborze właściwego testu statystycznego w zależności od danych.

8.1. Procedura testowania hipotez statystycznych

W rozdziale 6.3.2. procedurę testowania hipotez zaprezentowano na przykładzie problemu porównania częstości wyleczenia w grupach pacjentów leczonych lekami A i B. W tym miejscu przedstawimy procedurę testowania hipotez w nieco bardziej formalny sposób. Mianowicie, proces przeprowadzenia testu statystycznego w klasycznym ujęciu można podzielić na cztery etapy.

- Etap 1: sformułowanie hipotez adekwatnych do problemu badawczego. Jako hipotezę zerową przyjmuje się stwierdzenie, którego prawdziwość poddajemy w wątpliwość i którą chcemy odrzucić, jeśli tylko wyniki badań dadzą ku temu podstawę:
  H_0 : hipoteza zerowa,
  H_1 : hipoteza alternatywna (robocza).
- Etap 2: wybór właściwej procedury testowej i obliczenie wartości odpowiedniej statystyki testowej (ang. *test statistic*), na podstawie zgromadzonych danych.
- Etap 3: w oparciu o znajomość rozkładu wybranej statystyki testowej przy założeniu prawdziwości H_0, wyznacza się tzw. obszar krytyczny (lub obszar odrzuceń; ang. *rejection area*). W zależności od sformułowania hipotezy alternatywnej H_1, a także od

rodzaju statystyki testowej, obszar odrzuceń może być jedno- lub dwustronny (ang. *one-* oraz *two-sided*). Obszar ten jest zależny od przyjętego poziomu istotności (zaakceptowanego przez badacza prawdopodobieństwa popełnienia błędu I rodzaju - α).

- Etap 4: podjęcie decyzji na temat H_0: jeśli obliczona wartość statystyki testowej należy do obszaru odrzuceń, wówczas hipotezę zerową odrzuca się na korzyść hipotezy alternatywnej. W przeciwnym wypadku, brak jest podstaw do odrzucenia H_0. Pamiętajmy, że – zgodnie z tym co było sygnalizowane w rozdziale 6 – procedura testowania hipotez w klasycznym ujęciu częstościowym (wnioskowanie poprzez sprowadzenie do sprzeczności – *reductio ad absurdum*), nigdy nie może prowadzić do przyjęcia („udowodnienia") hipotezy zerowej – możliwym wynikiem rozumowania może być albo odrzucenie H_0 na korzyść H_1[1] albo stwierdzenie, że przy założonym poziomie istotności zgromadzone dane nie dają wystarczających podstaw do odrzucenia H_0 (por. [1]).

Współcześnie, obliczenia wykonywane są najczęściej przy pomocy wyspecjalizowanych pakietów statystycznych[2]. Zwalnia to badacza z konieczności przeprowadzania często żmudnych obliczeń numerycznych. Należy jednak pamiętać, że warunkiem koniecznym, aby wnioski płynące z przeprowadzonych testów statystycznych były uprawnione, jest zapewnienie doboru właściwej procedury testowej, a także sprawdzenie kluczowych założeń. Komputer bezbłędnie wykona obliczenia, ale to do badacza należy postawienie właściwego pytania (hipotezy badawczej), sprawdzenie założeń stosowanych metod oraz weryfikacja otrzymanych rezultatów z wiedzą kliniczną[3].

Przy zastosowaniu pakietów komputerowych procedura testowa prowadzona jest nieco inaczej: nie wyznacza się obszaru krytycznego dla ustalonego z góry poziomu istotności α, lecz na podstawie obliczonej wartości statystyki testowej wyznacza się graniczny poziom istotności, przy którym następuje zmiana decyzji nt. przyjęcia bądź odrzucenia H_0. Jest to wartość p (*p-value*) – prawdopodobieństwo, że przy założeniu prawdziwości H_0 otrzyma się wartość statystyki testowej taką, jak obliczona na podstawie zebranych danych empirycznych (lub wartość bardziej oddaloną od oczekiwanej przy założeniu prawdziwości H_0). Reguła wnioskowania o postawionych hipotezach jest zatem następująca:

- jeśli $p \leq \alpha$ – odrzucamy H_0 na korzyść H_1 („istotność statystyczna"),
- jeśli $p > \alpha$ – brak podstaw do odrzucenia H_0 („brak istotności statystycznej").

Jak podkreślono w rozdziale 6, należy pamiętać, że *p*-value nie jest prawdopodobieństwem prawdziwości hipotezy zerowej – wnioskowanie przebiega w przeciwnym kierunku! Innymi słowy, p jest miarą zgodności zaobserwowanych danych z założeniem o braku efektu

---

[1] Oczywiście taki wniosek jest warunkowy tj. zależny od przyjętego poziomu istotności – α.

[2] Omówienie możliwości dostępnych pakietów statystycznych wykracza poza zakres niniejszego opracowania. Warto jednakże wspomnieć, że do analizy danych z prób klinicznych na całym świecie powszechnie wykorzystywane jest oprogramowanie SAS (komercyjne), zaś w środowiskach naukowych popularność zdobywa pakiet R (bezpłatny). Referencje do polskich podręczników traktujących w przystępny sposób o analizie danych przy pomocy tych narzędzi znajdują się na końcu tego rozdziału (pozycje odpowiednio [6] oraz [3]).

[3] Por. Tadeusiewicz R.: Drogi i bezdroża statystyki w badaniach naukowych [dostęp 30 września 2011]. Dostępne w Internecie: www.statsoft.pl/czytelnia/badanianaukowe/d0ogol/nadrogi6.pdf

– jeśli to prawdopodobieństwo jest „odpowiednio małe" (tj. niższe od przyjętego poziomu istotności) wnioskujemy, że H_0 nie może być prawdziwa, a zatem istnieje efekt o którym mówi hipoteza alternatywna.

8.2. Test *t*-Studenta dla prób niezależnych

Test *t*-Studenta jest prawdopodobnie jednym z najczęściej raportowanych testów w zastosowaniach medycznych. W kontekście prób klinicznych, może być wykorzystany np. do porównania obserwowanego efektu mierzonego na skali przedziałowej pomiędzy równoległymi ramionami badania randomizowanego. Zakłada się, że analizowana cecha ma rozkład normalny z jednakową wariancją w obu populacjach. Nieznane wartości średnie μ_1 i μ_2 są estymowane średnimi $\bar{y}_1$ i $\bar{y}_2$ obliczonymi na podstawie niezależnie wylosowanych prób o liczności odpowiednio n_1 oraz n_2.

Statystyka testu (t) jest funkcją różnicy między średnimi w grupach, a także wielkości odchylenia standardowego w połączonej próbie. Hipoteza zerowa będzie odrzucana, przy otrzymaniu „dużych" wartości statystyki t. Przy założeniu prawdziwości H_0, statystyka testowa ma rozkład *t*-Studenta z parametrem $df = n_1 + n_2 - 2$ (tzw. liczba stopni swobody – ang. *Degrees of Freedom*). Formalnie procedura tego testu może być zapisana następująco:

- hipoteza zerowa $H_0 : \mu_1 = \mu_2$
- hipoteza alternatywna $H_1 : \mu_1 \neq \mu_2$ [4]
- statystyka testowa $t = \dfrac{\bar{y}_1 - \bar{y}_2}{s \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$,

gdzie $s = \sqrt{\dfrac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}$

- reguła decyzyjna: odrzucić H_0 jeśli tylko

$$|t| > t_{\alpha/2,\, n_1+n_2-2},$$

gdzie $t_{\alpha/2,\, n_1+n_2-2}$ to odpowiedni kwantyl rozkładu *t*-Studenta.

Obliczenia zilustrujemy przykładem dotyczącym badania satysfakcji z otrzymanego leczenia, u pacjentów leczonych z powodu nadciśnienia tętniczego w pewnym badaniu klinicznym. Do pomiaru satysfakcji z leczenia wykorzystano wizualną skalę analogową (VAS, ang. *Visual Analogue Scale*). Wynik wyrażony jest jako liczba z przedziału [0-100], przy czym wyższe wartości oznaczają większe zadowolenie pacjenta z leczenia. Tabela 8.1 przedstawia dane uzyskane na wizycie kończącej badanie, w grupach pacjentów otrzymujących odpowiednio placebo oraz leki A i B.

---

[4] Hipoteza alternatywna może być również sformułowana jako hipoteza jednostronna ($\mu_1 < \mu_2$ lub $\mu_1 > \mu_2$). W takim przypadku, odpowiedni obszar krytyczny zdefiniowany jest odpowiednio jako wyłącznie lewo- i prawostronny, zaś wartością krytyczną jest kwantyl rozkładu *t*-Studenta rzędu α.

Tabela 8.1. Wyniki badania satysfakcji z leczenia

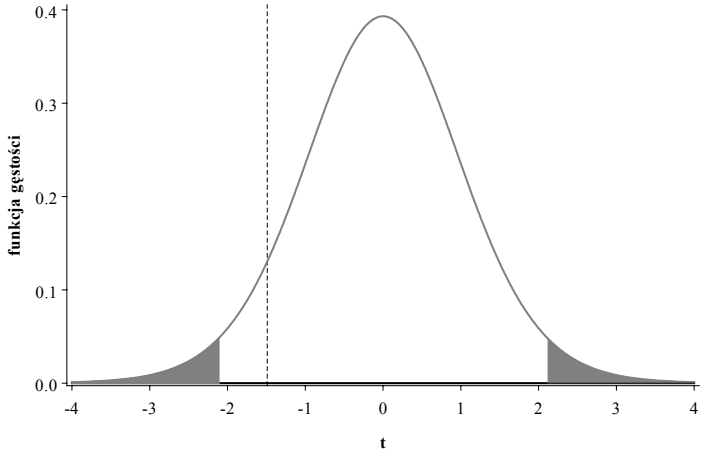
Grupa	VAS	Grupa	VAS	Grupa	VAS
Placebo	57	Lek A	69	Lek B	67
Placebo	65	Lek A	64	Lek B	72
Placebo	63	Lek A	57	Lek B	74
Placebo	60	Lek A	63	Lek B	69
Placebo	63	Lek A	59	Lek B	63
Placebo	51	Lek A	72	Lek B	62
Placebo	64	Lek A	65	Lek B	63
Placebo	56	Lek A	67	Lek B	76
Placebo	67	Lek A	58	Lek B	71
		Lek A	67	Lek B	65

Źródło: opracowanie własne

Porównamy wyniki leczenia pomiędzy grupami otrzymującymi placebo (n_1=9) oraz lek A (n_2=10). Średnie w grupach to odpowiednio $\bar{y}_1$=60,7 i $\bar{y}_2$=64,1 przy odchyleniach standardowych odpowiednio s_1=4,8 oraz s_2=4,7. Czy na podstawie tych wyników można stwierdzić, że lek A istotnie wpływa na wynik leczenia? Aby odpowiedzieć na to pytanie, przeprowadzimy test *t*-Studenta na poziomie istotności α=0,05.

W tym przypadku wartość statystyki *t* wynosi -1,49. Przy założeniu prawdziwości hipotezy zerowej, statystyka testowa ma rozkład *t*-Studenta z 17 stopniami swobody (kształt tego rozkładu przedstawiono na wykresie 1). Odpowiedni kwantyl tego rozkładu (rzędu 0,025) wynosi 2,11. Ponieważ 1,49<2,11, zatem zgodnie z podaną regułą nie ma podstaw do odrzucenia H_0. Wnioskujemy, że zgromadzone dane nie dają podstaw do twierdzenia o istnieniu efektu leczenia lekiem A (przy poziomie istotności 0,05).

Wykres 8.1. Krzywa gęstości rozkładu ***t***-Studenta z 17 stopniami swobody. Zaznaczono również wartość statystyki testowej (linia przerywana) oraz obszar odrzuceń (obszary zacienione)



Źródło: opracowanie własne

Przeprowadzając testowanie przy użyciu pakietu komputerowego uzyskamy wynik p-value=0,155. Uzyskana wartość p>α, zatem nie ma podstaw do odrzucenia hipotezy zerowej. Ściślej rzecz ujmując, przy założeniu prawdziwości H_0, prawdopodobieństwo uzyskania zaobserwowanych w tym badaniu wyników (lub wyników świadczących o jeszcze większej różnicy pomiędzy lekiem A i placebo) wynosi 15,5% i zazwyczaj nie zostałoby uznane za wystarczająco niskie, aby odrzucić hipotezę zerową.

8.3. Analiza wariancji

Analiza wariancji (ANOVA, ang. *ANalysis Of VAriance*) jest jedną z podstawowych technik wykorzystywanych do analizy danych medycznych. ANOVA jest uogólnieniem testu *t*-Studenta na przypadek więcej niż dwóch grup niezależnych. Test ten może być wykorzystany np. do porównania zmian obserwowanych w trzech ramionach badania klinicznego w schemacie grup równoległych.

Podobnie jak w przypadku testu Studenta, zakłada się, że analizowana cecha ma rozkład normalny, z jednakową wariancją w każdej z analizowanych k populacji. Koncepcja metody opiera się na dekompozycji wariancji całkowitej (por. rozdział 6) na dwa źródła: wariancja między średnimi w grupach i wariancja obserwacji indywidualnych wewnątrz grup. Przy założeniu prawdziwości H_0, zarówno wariancja międzygrupowa, jak i wewnątrzgrupowa powinny być niezależnymi oszacowaniami tej samej wariancji całkowitej, zaś ich iloraz zbliżony do 1. Statystyka testowa ma rozkład *F*-Snedecora o k-1 oraz N-k stopniach swobody. Obliczenia prowadzi się tradycyjnie w tzw. tabeli analizy wariancji.

Tabela 8.2. Ogólna postać tabeli analizy wariancji

Źródło zmienności	**Stopnie swobody (df)**	**Suma kwadratów odchyleń (SS, ang. *Sum of Squares*)**	**Średni kwadrat**
międzygrupowe (B, ang. *Between groups*)	dfB= $k-1$	SSB = $\sum_{i=1}^{k} n_i \cdot (\bar{y}_i - \bar{y})^2$	MSB=SSB/dfB
wewnątrzgrupowe (E, ang. *Within groups / "Error"*)	dfE= $N-k$	SSE = $\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	MSE=SSE/dfE
Ogółem (T, ang. *Total*)	dfT= $N-1$	SST = $\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	

Źródło: opracowanie własne

Wnioskowanie przebiega następująco:

- hipoteza zerowa $H_0 : \mu_1 = \mu_2 = ... = \mu_k$
- hipoteza alternatywna H_1 : istnieją takie i, j że $\mu_i \neq \mu_j$
- statystyka testowa $$F = \frac{MSB}{MSE},$$

  gdzie MSB i MSE zdefiniowano w tabeli 2
- reguła decyzyjna: odrzucić H_0 jeśli

  $$F > F_{\alpha,\, k-1,\, N-k},$$

  gdzie $F_{\alpha,\, k-1,\, N-k}$ to odpowiedni kwantyl rozkładu *F*-Snedecora.

Obliczenia zilustrujemy przykładem dotyczącym badania satysfakcji z leczenia. Tym razem porównujemy trzy grupy: pacjenci otrzymujący placebo (n_1=9), lek A (n_2=10) oraz lek B (n_3=10). Średnie w grupach to odpowiednio $\bar{y}_1$ =60,7, $\bar{y}_2$ =64,1 i $\bar{y}_3$ =68,2. Czy na podstawie tych wyników można stwierdzić, że średnie w grupach istotnie się różnią pod względem parametru VAS? Aby odpowiedzieć na to pytanie, przeprowadzimy test analizy wariancji, na poziomie istotności α=0,05. Wyniki obliczeń pomocniczych przestawia tabela 3.

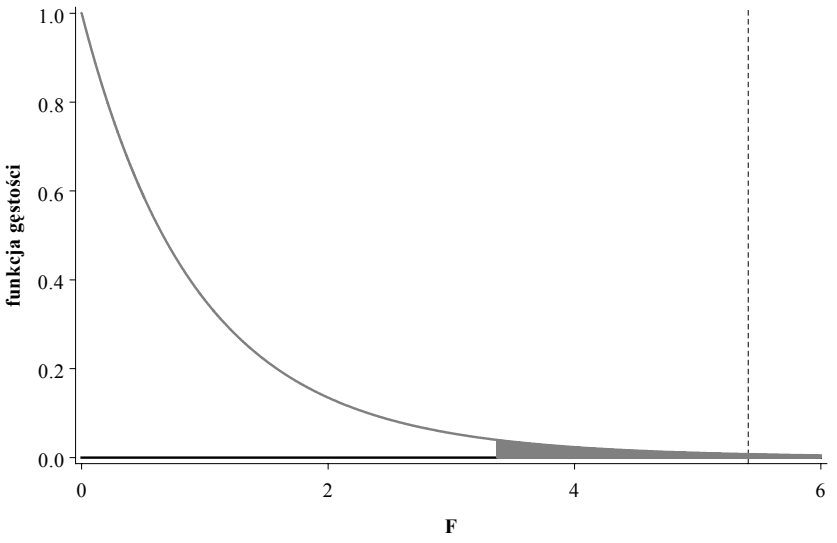
Tabela 8.3. Tabela ANOVA, na przykładzie analizy satysfakcji z leczenia w badaniu placebo vs lek A vs lek B

Źródło zmienności	Stopnie swobody	Suma kwadratów odchyleń	Średni kwadrat
międzygrupowe	dfB=2	SSB=270,67	MSB=135,34
Wewnątrzgrupowe	dfE=26	SSE=650,50	MSE=25,02
Ogółem	dfT=28	SST=921,17	

Źródło: opracowanie własne

W tym przypadku wartość statystyki $F = \frac{MSB}{MSE}$ = 5,41. Przy założeniu prawdziwości hipotezy zerowej, statystyka testowa ma rozkład *F* odpowiednio z 2 i 26 stopniami dla licznika i mianownika. Kształt tego rozkładu przedstawiono na wykresie 2. Odpowiedni kwantyl tego rozkładu rzędu 0,05 wynosi 3,37. Ponieważ 5,41>3,37, zatem zgodnie z podaną regułą wnioskowania należy odrzucić H_0. Na podstawie zaobserwowanych danych wnioskujemy, że istnieją istotne różnice pomiędzy badanymi grupami.

Wykres 8.2. Krzywa gęstości rozkładu F-Snedecora z 2 i 26 stopniami swobody. Zaznaczono również wartość statystyki testowej (linia przerywana) oraz obszar odrzuceń (obszar zacieniony)



Źródło: opracowanie własne

Przeprowadzając testowanie przy użyciu pakietu komputerowego uzyskamy wynik p-value=0,011. Oznacza to, że przy założeniu prawdziwości H_0 prawdopodobieństwo zaobserwowania danych z eksperymentu wynosi 1,1%. Zatem hipotezę zerową należy odrzucić, jeśli prowadzimy wnioskowanie przy poziomie istotności α=0,05. Jeśli akceptowane prawdopodobieństwo popełnienia błędu I rodzaju ustalimy na 1%, wówczas decyzja będzie odmienna (dane niewystarczające do odrzucenia hipotezy o równości średnich).

8.4. Procedury porównań wielokrotnych

Po odrzuceniu hipotezy zerowej w teście analizy wariancji, powstaje naturalne pytanie, które grupy różnią się między sobą. Jak wspomniano w rozdziale 6.3.4, łączny poziom istotności dla prowadzonego wnioskowania czyli prawdopodobieństwo błędnego odrzucenia co najmniej jednej hipotezy zerowej (ang. *experiment-wise error rate*), bardzo szybko wzrasta wraz z liczbą prowadzonych porównań. Jeśli k oznacza liczbę analizowanych grup, wówczas liczba wszystkich możliwych porównań parami wyraża się wzorem $K = \frac{k \cdot (k-1)}{2}$, zaś łączny poziom istotności – przy nominalnym poziomie istotności dla pojedynczego porównania (ang. *comparison-wise error rate*) ustalonym na 0,05 – dany jest wzorem $\alpha = 1-(1-0{,}05)^K$. Zależności przedstawia tabela.

Tabela 8.4. Zależność łącznego poziomu istotności od liczby grup i porównań w problemie porównań wielokrotnych

Liczba analizowanych grup (k)	Liczba porównań (K)	Łączny poziom istotności (α)
2	1	0,050
3	3	0,143
4	6	0,265
5	10	0,401
6	15	0,537
7	21	0,659
8	28	0,762
9	36	0,842
10	45	0,901

Źródło: opracowanie własne

Przykładowo, dla k=5 grup mamy K=10 możliwych porównań między grupami, zaś prawdopodobieństwo błędnego odrzucenia co najmniej jednej hipotezy zerowej przekracza 40%. Dla k=10 grup, łączny poziom istotności przekracza 90%.

Istnieje zatem potrzeba stosowania procedur wnioskowania, umożliwiających kontrolowanie łącznego poziomu istotności na założonym poziomie[5]. Jedną z procedur testów wielokrotnych

[5] Warto wspomnieć, że stosunkowo często w doniesieniach z badań medycznych raportowane jest użycie procedury najmniejszej istotnej różnicy (LSD – ang. *Least Significant Difference*), zaproponowanej przez R.A. Fishera w 1949 r.

(sygnalizowaną już w rozdziale 6) jest tzw. korekta Bonferroniego (ang. *Bonferroni correction*), polegająca na prowadzeniu poszczególnych porównań przy odpowiednio zmniejszonym poziomie istotności. Wadą procedury Bonferroniego jest jej zbytnia konserwatywność – rzeczywisty łączny poziom istotności jest niższy od założonego, zaś rzeczywisty poziom ufności dla łącznej analizy przedziałów ufności wyznaczonych tą metodą jest wyższy niż założony poziom ufności 1-α.

Istnieje wiele innych metod pozwalających na kontrolę łącznego poziomu istotności dla eksperymentu. Do najczęściej raportowanych w czasopismach medycznych należą procedury zaproponowane przez Tukeya i Scheffégo. Procedura Tukeya jest szczególnie użyteczna w przypadku wielokrotnych porównań par wartości średnich w grupach o jednakowej liczności. Metoda Scheffégo jest zalecana zwłaszcza wtedy, gdy analizuje się nie tylko różnice średnich w parach, ale także tzw. kontrasty między więcej niż dwiema średnimi (por. [4], s. 336-337]).

8.5. Weryfikacja założeń testu *t*-Studenta oraz analizy wariancji

Jak wspomniano w rozdz. 8.1, warunkiem koniecznym do tego, aby wnioski płynące z przeprowadzonej analizy danych były uprawnione, jest zastosowanie metody, która opiera się na założeniach dobrze opisujących dane, które mają być poddane analizie.

W niniejszym podrozdziale opisane zostaną możliwe podejścia do sprawdzenia założeń testu *t*-Studenta oraz analizy wariancji. Zastosowanie obu wspomnianych metod wymaga spełnienia założeń dotyczących 1) normalności rozkładu badanej cechy w poszczególnych grupach, 2) jednorodności wariancji. Istnieją dwa podejścia do weryfikacji tych założeń.

- Metody graficzne: przykładowe wykresy przydatne do badania normalności rozkładu to histogram i wykres kwantylowy (ang. *quantile plot*). W celu zbadania jednorodności wariancji można posłużyć się np. wykresem pudełkowym (ang. *box plot*).
- Metody formalne, polegające na weryfikacji stosownych hipotez na podstawie zebranych danych. Najczęściej stosowane testy służące do badania normalności rozkładu to testy Shapiro-Wilka i Kołmogorowa-Smirnowa. Przykładowe testy służące do oceny jednorodności wariancji to testy Levene'a oraz Bartletta.

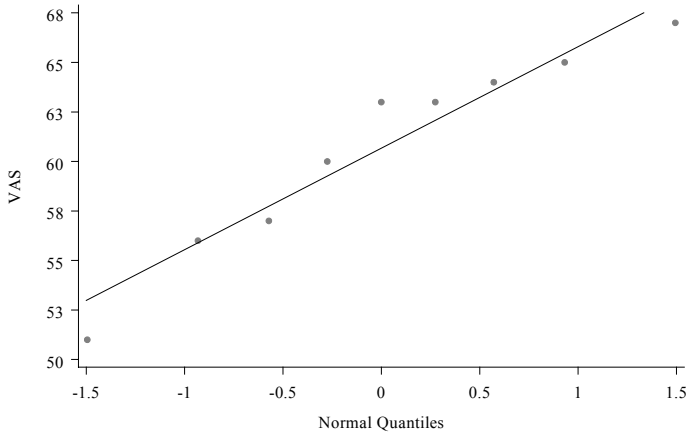
Niestety, oba podejścia nie są pozbawione wad. Metody graficzne są ze swej istoty podatne na subiektywizm osoby oceniającej wykres i wymagają pewnego doświadczenia (czy zaobserwowane odstępstwa są na tyle istotne, aby uznać, że jakieś założenie nie jest spełnione?). Z drugiej strony, testy dotyczące kształtu rozkładu czy też jednorodności wariancji, nie zawsze charakteryzują się wystarczającą mocą pozwalającą na wykrycie istotnych odstępstw od założeń. Dodatkowym mankamentem jest fakt, że w ten sposób dochodzi do jednoczesnego testowania kilku hipotez na tych samych danych, zatem w efekcie prawdopodobieństwo popełnienia błędu I rodzaju w całej procedurze wnioskowania jest wyższe niż nominalny poziom istotności (por. rozdział 6.3.4. oraz 8.4.).

Na wykresach 8.3. i 8.4. przedstawiono przykładowe wykresy kwantylowe oraz pudełkowe[6], dla danych dotyczących badania satysfakcji z leczenia. Wskazują one na brak istotnych odstępstw od założeń.

Metoda ta jest równoważna z przeprowadzeniem porównań między parami średnich przy użyciu testu *t*-Studenta omawianego w rozdziale 8.2., przy nominalnym poziomie istotności. Procedura nie zapewnia kontroli prawdopodobieństwa popełnienia błędu I rodzaju na poziomie eksperymentu, dlatego nie jest zalecana.

[6] Dokładny opis konstrukcji wykresów kwantylowych oraz wykresów pudełkowych można znaleźć np. w pracy [4].

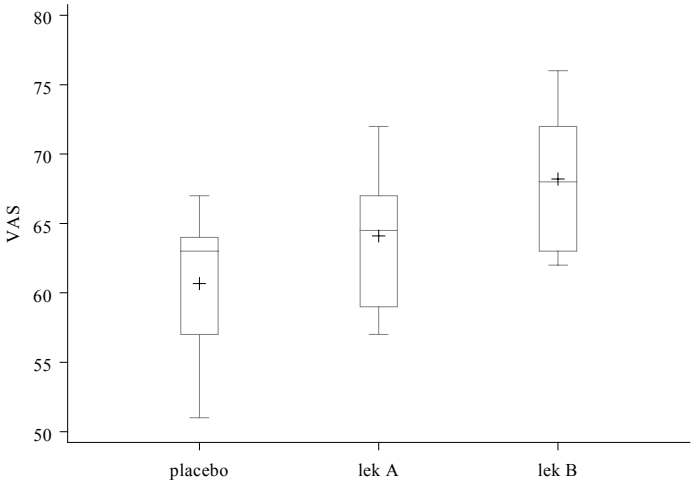
Wykres 8.3. Wykres kwantylowy dla danych dotyczących badania satysfakcji z leczenia (grupa placebo)



Źródło: opracowanie własne

Jeśli próba losowa pochodzi z rozkładu normalnego, punkty odpowiadające wartościom badanej cechy powinny układać się na linii prostej. Wydaje się, że dane nie wykazują istotnego odstępstwa od założenia o normalności.

Wykres 8.4. Wykres pudełkowy dla danych dotyczących badania satysfakcji z leczenia



11

Źródło: opracowanie własne

Wysokość „pudełek" odpowiadających poszczególnym grupom jest pozycyjną miarą zróżnicowania (rozstęp międzykwantylowy – ang. *interquartile range*) oznaczającą obszar w jakim znajduje się środkowe 50% obserwacji. Na podstawie wykresu można stwierdzić, że założenie o jednakowym rozrzucie zmiennej VAS w trzech grupach jest spełnione.

8.6. Metody nieparametryczne

Powyżej omówione testy statystyczne należą do procedur parametrycznych, tzn. do swojej stosowalności wymagają założenia, że dane pochodzą z określonego typu rozkładu (w tym przypadku normalnego). Jeśli nie można przyjąć tego założenia, można posłużyć się odpowiednią procedurą nieparametryczną[7] (ang. *non-parametric*). Nazwa tej grupy procedur wywodzi się z braku założeń dotyczących postaci rozkładu badanych cech, zaś obliczenia oparte są na rangach (kolejności) obserwacji. Przykładowo, nieparametrycznym odpowiednikiem testu *t*-Studenta jest test Manna-Whitneya[8], zaś odpowiednikiem parametrycznej analizy wariancji jest nieparametryczny test Kruskala-Wallisa.

Często stosowaną praktyką jest też prowadzenie wnioskowania przy użyciu testu nieparametrycznego również w sytuacji, gdy założenia testu parametrycznego są spełnione – wówczas przeprowadzenie takiego testu stanowi tzw. analizę wrażliwości (ang. *robustness analysis*). Przykładowo, dla danych dotyczących badania satysfakcji z leczenia, wyniki testów nieparametrycznych są następujące:

- test Manna-Whitneya dla porównania placebo vs lek A: p=0,160;
- test Kruskala-Wallisa dla porównania efektu pomiędzy trzema grupami: p=0,032.

Oczywiście podejście nieparametryczne prowadzi do identycznych wniosków jak poprzednio zaprezentowane procedury parametryczne.

8.7. Od czego zależy wybór właściwego testu?

Istnieje bardzo wiele testów statystycznych. W niniejszej części opracowania przedstawiony zostanie schemat pozwalający na wybór właściwego testu statystycznego, w sytuacjach z którymi najczęściej można się spotkać w kontekście analizy danych medycznych.

Wybór właściwej procedury testowej zależy od wielu czynników, wśród których jako najważniejsze należy wymienić:

- rodzaj badania[9],
- liczbę porównywanych grup,
- charakter zebranych danych (np. dane niezależne lub powiązane),
- liczność próby[10],
- rodzaj użytej skali pomiarowej[11],
- spełnienie założeń dotyczących rozkładu analizowanej cechy.

---

[7] Innym podejściem może być zastosowanie takiego przekształcenia zebranych danych, aby możliwe było zastosowanie procedur parametrycznych dla danych transformowanych. Warto również wspomnieć, że niektóre parametry analizowane są zwyczajowo po dokonaniu pewnej transformacji. Przykładowo, w badaniach biorównoważności leków generycznych (por. rozdz. 4), wartości stężeń badanego leku we krwi analizowane są po dokonaniu transformacji logarytmicznej. W tym przypadku, przekształcenie to wynika z teorii dotyczącej badanego zjawiska, a nie z badania rozkładu stężeń zaobserwowanych w konkretnej próbie.

[8] Lub równoważny test Wilcoxona.

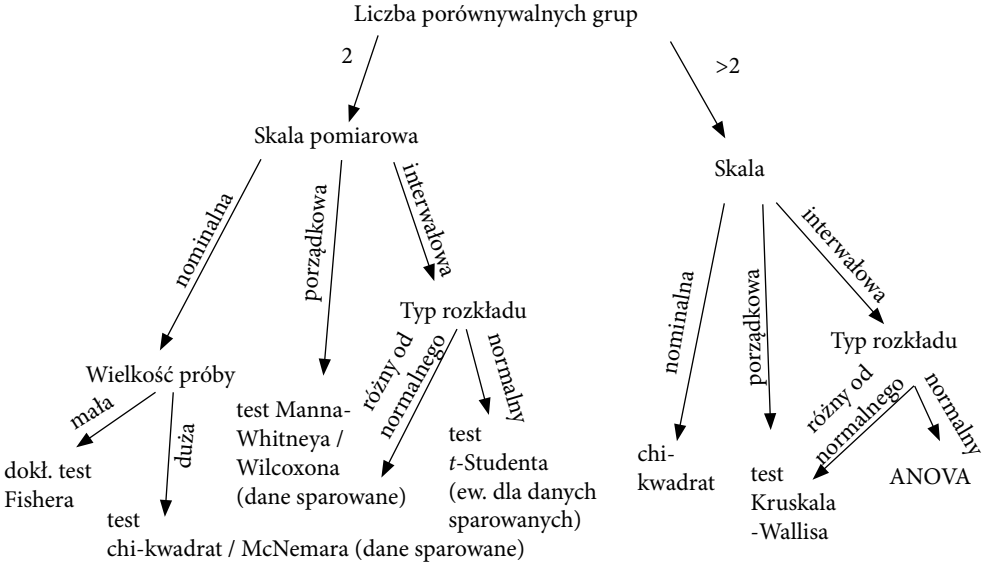
[9] Najważniejsze schematy badań eksperymentalnych i obserwacyjnych omówiono w rozdziale 2

[10] Dla „odpowiednio dużych" prób dokładny rozkład statystyk testowych może być przybliżany odpowiednimi rozkładami asymptotycznymi (np. rozkładem normalnym). Można przyjąć, że „duża próba" powinna liczyć co najmniej 30 obserwacji – por. rozdział 6.

[11] Typologię skal pomiarowych omówiono w rozdz. 6.1.1

Na wykresie 8.5. przedstawiono sposób wyboru procedury testowej, w zależności od wymienionych czynników. Warto przypomnieć, że jak to zasygnalizowano w rozdziale 6.1.1. między skalami pomiarowymi zachodzi hierarchiczna zależność – zawsze jest możliwe przejście od skali mocniejszej do słabszej i wykorzystanie technik właściwych dla tej słabszej skali. Przykładowo, poprzez kategoryzację wartości mierzonych na skali ciągłej można przejść do skali porządkowej i np. BMI mierzone w [kg/m^2] można analizować jako cechę porządkową: ≤25, (25-30], (30-35], >35.

Wykres 8.5. Sposób doboru testu statystycznego w zależności od wybranych czynników



Źródło: opracowanie własne

Bibliografia

1. Altman, D.G.; Bland, J.M.: Absence of evidence is not evidence of absence. British Medical Journal, 1995, 311 (7003), 485.
2. Armitage, P.; Berry, G.; Matthews, J.N.S.: Statistical Methods in Medical Research. John Wiley and Sons, 2008.
3. Biecek, P.: Przewodnik po pakiecie R. Wrocław, 2008.
4. Koronacki, J.; Mielniczuk, J.: Statystyka dla studentów kierunków technicznych i przyrodniczych. Warszawa, 2009.
5. Moczko, J.; Bręborowicz, G.; Tadeusiewicz, R.: Statystyka w badaniach medycznych. Warszawa, 1998.
6. Roterman-Konieczna, I.: Statystyka na receptę. Wprowadzenie do statystyki medycznej. Kraków, 2010.
7. Zar, J.H.: Biostatistical Analysis. Prentice Hall, 2010.

IX. Analizy korelacji i analizy wieloczynnikowe

Łukasz BOROWIEC

W rozdziale 8 omówiono najczęściej stosowane testy statystyczne, służące to badania różnic pomiędzy dwiema lub więcej grupami, pod względem pewnej wyróżnionej cechy. W procesie prowadzenia medycznych badań naukowych, zazwyczaj zbierane są dane nie tylko nt. cechy, która jest głównym kryterium dalszej analizy (Y), ale też szereg informacji towarzyszących (X), które mogą mieć znaczenie w analizie interesującej nas wielkości (np. dane demograficzne i kliniczne pacjenta, informacje o czynnikach ryzyka). W przypadku randomizowanych prób klinicznych (por. rozdz. 2) uwzględnienie tych dodatkowych informacji pozwala na zwiększenie efektywności wnioskowania nt. cechy Y, a także wyjaśnienie charakteru zależności Y od zmiennych towarzyszących. W przypadku analizy eksploracyjnej danych rejestrowych czy obserwacyjnych, gdzie rozkład zmiennych towarzyszących nie jest kontrolowany, uwzględnienie wszystkich istotnych informacji towarzyszących jest warunkiem prawidłowego wnioskowania o zjawisku Y.

Przykładowo, jeśli w grupie leczonej lekiem A u większości pacjentów występuje dodatkowo niewydolność nerek, zaś w grupie placebo odsetek pacjentów z tym czynnikiem ryzyka jest niższy, wówczas proste porównanie między grupami może dać mylny wynik, zafałszowany wpływem niewydolności nerek na obserwowany wynik leczenia (skuteczność leku A będzie systematycznie zaniżana poprzez chorobę współistniejącą). Z kolei uwzględnienie w analizie tej dodatkowej informacji pozwala na oszacowanie różnicy efektu leczenia, skorygowanej o wpływ wyjściowego stanu klinicznego pacjenta. W ogólnym przypadku, analiza wieloczynnikowa może pomóc w oszacowaniu badanego efektu, z uwzględnieniem braku równowagi pomiędzy porównywanymi grupami ze względu na pewne istotne cechy, mogące modyfikować badany punkt końcowy. Istnieją też schematy eksperymentalne, tzw. plany czynnikowe (ang. *factorial design*), gdzie jednocześnie bada się wpływ więcej niż jednej interwencji. W najprostszym układzie 2x2, pacjenci są randomizowani do czterech grup:

- aktywny lek A + aktywny lek B,
- aktywny lek A + placebo B,

- placebo A + aktywny lek B,
- placebo A + placebo B.

Taki schemat naturalnie narzuca zastosowanie analizy wieloczynnikowej. Podejście wieloczynnikowe dodatkowo pozwala na zbadanie tzw. interakcji pomiędzy interwencjami A i B (jednoczesne zastosowanie leku A i B może dać łącznie większy lub odpowiednio mniejszy efekt niż każda z terapii osobno, efekt leku A może też być niezależny od działania leku B). Niniejszy rozdział poświęcono omówieniu najważniejszych metod analizy wieloczynnikowej, pozwalającej na wyjaśnienie bardziej złożonych zależności pomiędzy zmiennymi.

9.1. Analiza korelacji

W rozdziale poświęconym statystyce opisowej omówiono miary położenia i rozproszenia dla pojedynczej zmiennej. Podstawową techniką służącą do analizy łącznego rozkładu zależności między dwiema cechami ilościowymi jest wykres rozproszenia (ang. *scatterplot*), oraz analiza współczynnika korelacji. Wykres rozproszenia pozwala na wstępną ocenę czy występuje współzależność pomiędzy cechami X i Y, a także czy zależność ta jest monotoniczna (wzrostowi X przeciętnie odpowiada wzrost lub spadek Y, w całym zakresie wartości) i w przybliżeniu liniowa (tj. wzrostowi X o jednostkę przeciętnie odpowiada taka sama zmiana Y). Współczynnik korelacji liniowej Pearsona (ang. *Pearson linear correlation coefficient*) w zbiorze obserwacji (x_i, y_i) wyraża się wzorem

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

gdzie $\bar{x}$, $\bar{y}$ oznacza wartość średnią.

Miara ta ma następujące własności:

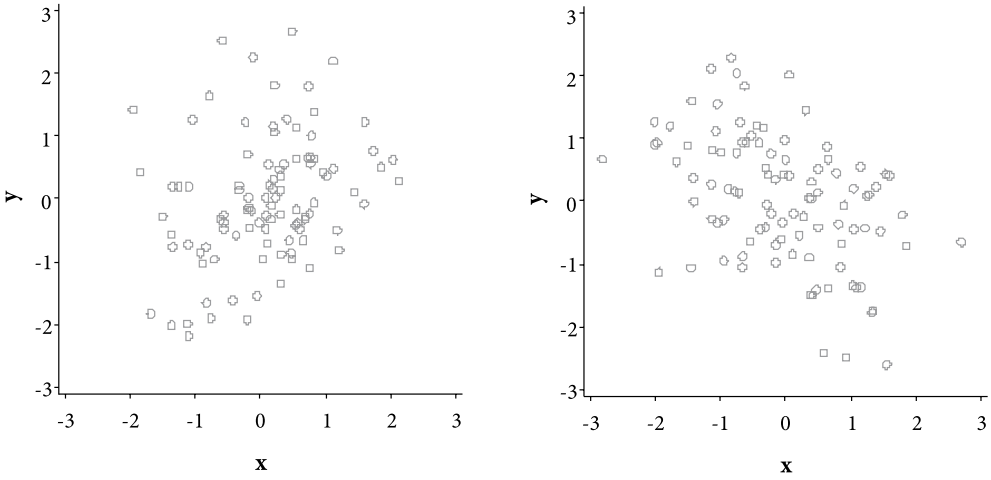
- współczynnik korelacji jest ograniczony wartościami $-1 \le r \le 1$,
- miara jest niewrażliwa na transformacje liniowe[1] X i Y (np. współczynnik korelacji pewnej cechy ze stężeniem cholesterolu LDL nie zależy od tego czy wartość wyrażona jest w [mg/dl] czy [mmol/l]; podobnie tę samą wartość będzie miał współczynnik korelacji z temperaturą, niezależnie od tego czy temperatura jest wyrażona w °C czy °K),
- wartości bliskie -1 lub 1 wskazują, że wykres rozproszenia jest skupiony wokół linii prostej, wartościom bliskim 0 odpowiada „chmura punktów", zatem współczynnik korelacji mierzy siłę zależności między X i Y,
- znak współczynnika informuje o kierunku zależności, tzn. czy wzrostowi X odpowiada wzrost (zależność dodatnia) czy spadek Y (zależność ujemna).

Wykres 1 przedstawia przykładowe wykresy rozproszenia, dla różnych wartości współczynnika korelacji. W pierwszym przypadku dane układają się w dość nieregularną chmurę

[1] Oczywiście przekształcenia nieliniowe (np. logarytm) zmieniają wartość współczynnika korelacji liniowej, tzn. korelacja pomiędzy X a Y nie jest na ogół równa korelacji X z ln(Y).

punktów – niewielki związek o kierunku dodatnim (r=0,1). W drugim przypadku, skupienie punktów jest wyraźniej zaznaczone, zaś zależność ma kierunek ujemny (r=-0,6).

Wykres 9.1. Wykresy rozproszenia dla różnych wartości współczynnika r

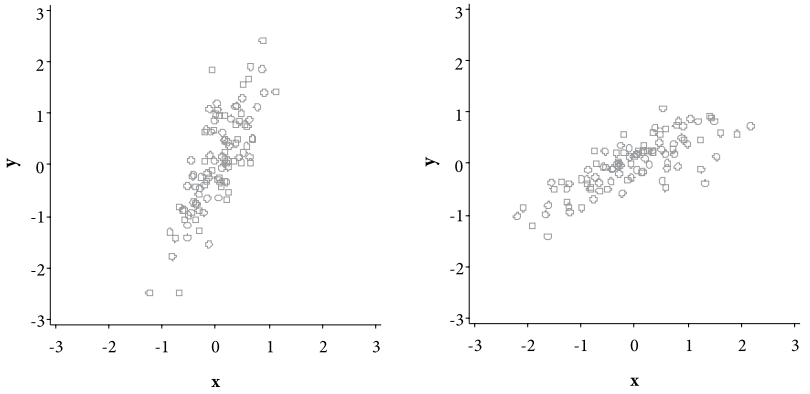


Źródło: opracowanie własne

Warto podkreślić, że współczynnik korelacji nie mierzy wielkości wpływu X na Y (tzn. tego, jaka jest przeciętna zmiana Y przypadająca na jednostkę X), lecz jakość dopasowania (tzn. w jakim stopniu zmiany Y są tłumaczone przez X, a na ile są losowe z perspektywy badacza, czyli tłumaczone przez inne czynniki). Ilustrację zawiera wykres 2. W obu przypadkach współczynnik korelacji liniowej wynosi 0,8. Skupienie układu danych wokół linii prostej jest jednakowe na obu wykresach, mimo ewidentnie różnej „stromizny" wykresu.

Goldsmith *et al.* [4, s. 7] zbadali związek indeksu jakości życia EQ-5D z parametrami demograficznymi i klinicznymi u pacjentów z chorobą wieńcową. Przykładowo, dla wieku, czasu próby wysiłkowej oraz skali percepcji choroby DPS zaraportowano współczynniki korelacji równe odpowiednio 0,05, 0,42 oraz 0,57. Tak więc indeks jakości życia korelował dodatnio z wiekiem pacjenta (w tym przypadku związek należałoby jednak uznać za nieistotny klinicznie), stwierdzono także korelację dodatnią o dość dużej sile z czasem próby wysiłkowej oraz skalą percepcji choroby. Co ciekawe, ocena jakości życia okazała się być mocniej skorelowana z subiektywnym postrzeganiem choroby, niż z wynikiem obiektywnego testu wydolnościowego.

Wykres 9.2. Wykresy rozproszenia dla tej samej wartości współczynnika r



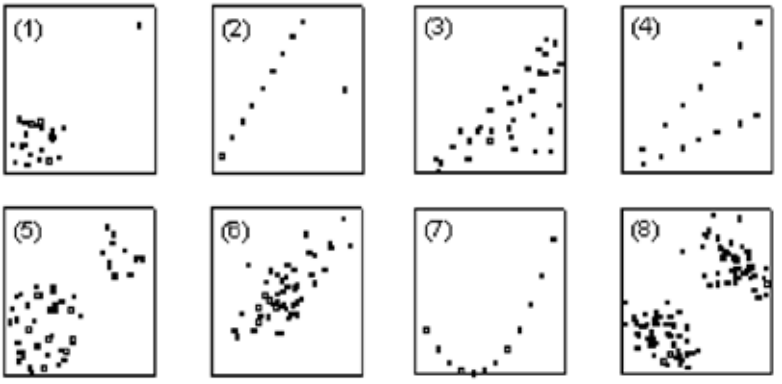
Źródło: opracowanie własne

Należy pamiętać, że współczynnik korelacji liniowej zawiera syntetyczną ocenę liczbową całej informacji nt. łącznego rozkładu cech X i Y. Jak każda miara syntetyczna, musi być interpretowana z uwzględnieniem pozostałych informacji o łącznym rozkładzie (np. z użyciem wykresów rozproszenia). Doskonałą ilustracją potencjalnych błędnych interpretacji jest wykres 1. W każdej z przedstawionych sytuacji, współczynnik korelacji liniowej wynosi 0,7, tymczasem jakościowa interpretacja jest bardzo różna w poszczególnych sytuacjach!

W 1. przypadku właściwy zbiór danych w lewej dolnej ćwiartce wykresu nie wykazuje żadnej korelacji. Współczynnik korelacji dla całego zbioru obserwacji jest sztucznie zawyżony przez tzw. obserwację odstającą (ang. *outlier*), np. pacjenta z nadciśnieniem tętniczym omyłkowo włączonego do badania prowadzonego w grupie pacjentów z ciśnieniem kontrolowanym.

W 2. przypadku dane również zawierają obserwację odstającą. Tym razem, współczynnik korelacji dla zbioru po jej usunięciu wynosi 1. Warto zwrócić uwagę, że w przeciwieństwie do poprzedniej sytuacji, wartość odstająca nie zostałaby wykryta przy analizie jednowymiarowej X oraz Y (np. przy analizie histogramów dla każdej z tych cech osobno).

Wykres 9.3. Pułapki w interpretacji współczynnika korelacji liniowej



Źródło: opracowanie własne na podstawie [2], za zgodą

Przykład 3. to typowy kształt wykresu rozproszenia wynikający z faktu, że jedna z analizowanych zmiennych jest składnikiem drugiej (przykładowo: całkowity cholesterol vs LDL). Współczynnik korelacji liniowej nie jest tu właściwym narzędziem analizy.

Dla przypadku 4. wykres sugeruje że dane są mieszaniną dwóch różnych populacji. Np. zależność badanych cech jest inna w populacji kobiet i mężczyzn. W tym przypadku należałoby osobno przeanalizować obie podgrupy – zaraportowanie współczynnika korelacji dla połączonej próby jest błędne.

W 5. przypadku zbiór zawiera dwie rozłączne grupy danych, w każdej z nich korelacja jest zerowa. Taka sytuacja może wynikać np. z niewłączenia do badania jednostek ze średnimi wartościami X lub Y.

W sytuacji przedstawionej w przykładzie 6. współczynnik korelacji liniowej dobrze opisuje współzależność.

Dla 7 przypadku współczynnik korelacji wynosi 0,7 pomimo zależności funkcyjnej Y od X. Oczywiście spowodowane to jest nieliniowym charakterem zależności. W takiej sytuacji należałoby poddać analizie odpowiednie przekształcenie Y (w tym przypadku pierwiastek kwadratowy).

Wreszcie przykład 8. przedstawia zbiór, który zawiera rozłączne grupy danych. Tym razem po rozdzieleniu podgrup, korelacja wewnątrz każdej z nich jest ujemna.

Na koniec omówimy modyfikacje i warianty współczynnika korelacji liniowej. Współczynnik korelacji rang Spearmana (ang. *Spearman rank correlation coefficient*) może być stosowany w sytuacjach, gdy jedna albo obie z cech X, Y są mierzone na skali porządkowej. Przykładowo, przy pomocy współczynnika korelacji rang można badać zależność pomiędzy klasą NYHA a spoczynkową częstością akcji serca u pacjentów z niewydolnością serca. Użycie w tym miejscu współczynnika Pearsona jest nieuprawnione, bowiem opiera się na założeniu, że odległości pomiędzy wartościami obu cech są dobrze zdefiniowane (por. rozdział 6.1.1.). Współczynnik korelacji wielorakiej (ang. *multiple correlation coefficient*) ocenia łączną zależność zmiennej Y z wieloma zmiennymi ilościowymi – jest związany z modelem regresji liniowej, omawianym w kolejnym podrozdziale.

Współczynnik korelacji obliczony na podstawie próby można traktować jako estymator korelacji w populacji generalnej. Można do tej miary zastosować techniki wnioskowania omówione w rozdziale 6, a więc wyznaczać błąd oszacowania i przedziały ufności lub przeprowadzać testy statystyczne hipotez dotyczących współczynnika korelacji w populacji. Do tego wnioskowania konieczne jest jednak dodatkowo przyjęcie założenia o normalności rozkładu.

Należy podkreślić, że korelacja nie oznacza związku przyczynowo-skutkowego! Obserwowana korelacja między dwoma cechami może być wynikiem działania trzeciej zmiennej, nieuwzględnionej w analizie. Dlatego też istnieje potrzeba stosowania bardziej zaawansowanych analiz, uwzględniających wpływ wielu czynników jednocześnie.

9.2. Model regresji liniowej

Analiza regresji jest techniką służącą do badania zależności wybranej cechy Y (zmienna objaśniana/zależna) od najczęściej wielu zmiennych $X_1, X_2, \ldots, X_k$ (zmienne objaśniające/niezależne). Modele regresji pozwalają na realizację dwóch zasadniczych celów:

Cel poznawczy – wyjaśnienie, w jaki sposób zmienna objaśniana jest modyfikowana przez wartości zmiennych niezależnych. Przykładowo, analiza regresji może być użyta w badaniu klinicznym w celu modelowania zależności odpowiedzi na leczenie hipotensyjne (Y: ciśnienie skurczowe krwi mierzone po 12 tygodniach leczenia) od zastosowanej dawki leku (X_1), wyjściowej wartości ciśnienia skurczowego (X_2), płci pacjenta (X_3), poziomu cholesterolu (X_4), współistniejącej cukrzycy (X_5).

Cel prognostyczny – opracowanie równania, pozwalającego na prognozę wartości Y w zależności od wartości zmiennych niezależnych, dla pacjentów spoza próby objętej badaniem. Przykładowo, model regresji może posłużyć do zbudowania narzędzia pozwalającego na prognozowanie dziesięcioletniego ryzyka zgonu sercowo-naczyniowego w zależności od wieku i płci pacjenta, ciśnienia skurczowego, poziomu cholesterolu, informacji dotyczących chorób współistniejących, stylu życia pacjenta, etc.

Ogólna postać modelu regresji liniowej to:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k + \varepsilon ,$$

gdzie Y to modelowana zmienna ilościowa. Zmienna ε nazywana jest błędem losowym w modelu regresji. Zakładamy zatem, że cecha Y jest zależna liniowo od zmiennych niezależnych X (przy czym mogą to być zmienne ilościowe lub jakościowe[2]), z dokładnością do pewnego błędu poczynionego podczas obserwacji. Błąd ten może obejmować np. wpływ innych zmiennych niezależnych nieuwzględnionych w modelu regresji, niedokładności pomiaru, etc. Przyjmuje się założenie, że ma on wartość średnią równą 0 (tzn. nie popełniamy błędu systematycznego) i nieznaną wariancję. Bardzo ważne jest również założenie o wzajemnej niezależności zmiennych objaśniających.

Współczynniki występujące przy zmiennych z prawej strony równania noszą nazwę parametrów strukturalnych modelu regresji. Parametry strukturalne w modelu regresji liniowej szacuje się metodą najmniejszych kwadratów (ang. *least squares*). W modelu regresji prostej (k=1), oszacowania parametrów wyrażają się wzorami („daszek" w poniższych wzorach oznacza estymator, tj. wartość oszacowaną na podstawie próby, w odróżnieniu od nieznanej wartości w populacji):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Z postaci modelu regresji łatwo wynika interpretacja oszacowań parametrów. Mianowicie, parametr β_i interpretuje się jako przeciętną zmianę cechy Y, spowodowaną jednostkowym wzrostem X_i (przy pozostałych warunkach niezmienionych – *ceteris paribus*). Oczywiście wzrost X_i o p jednostek wiąże się ze średnią zmianą Y o $p \cdot \beta_i$. W przypadku binarnej

[2] W tym przypadku zmienne jakościowe (nominalne) są na potrzeby obliczeń kodowane liczbowo – przykładowo, informacja o współistniejącej cukrzycy może być zakodowana jako 0=brak choroby, 1=choroba współistniejąca.

zmiennej objaśniającej, oszacowanie β_i oznacza przeciętną różnicę Y pomiędzy wartościami X_i kodowanymi jako 0 i 1. Przykładowo, jeśli w modelu zależności ciśnienia skurczowego od wielu zmiennych, wartość oszacowania dla cechy „współistniejąca cukrzyca" wynosi 5, oznacza to, że przy ustalonych wartościach pozostałych zmiennych niezależnych, pacjenci z cukrzycą mają przeciętnie wyższe ciśnienie o 5 mm Hg.

Dla każdego parametru $\beta_0, \beta_1, \ldots, \beta_k$, oprócz oszacowania punktowego można wyznaczyć jego błąd standardowy (pamiętajmy, że estymatory są obliczane na podstawie próby, w której wartości zmiennej Y są także zaburzone przez ε), a także przedział ufności odzwierciedlający precyzję oszacowania poszczególnych parametrów. Przeprowadza się również testy statystyczne służące weryfikacji hipotez postaci:

- $H_0 : \beta_i = 0$ (weryfikacja istotności statystycznej wybranego parametru β_i),
- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ (weryfikacja istotności statystycznej wszystkich parametrów jednocześnie, a zatem całego modelu).

W analizie regresji liniowej wyznacza się też sumaryczny wskaźnik jakości dopasowania modelu do danych empirycznych, jakim jest współczynnik determinacji R^2 (ang. *determination coefficient*). Ideę tej miary przedstawia poniższy wzór:

$$R^2 = \frac{\text{zmienność wyjaśniona przez model}}{\text{zmienność całkowita}} .$$

Współczynnik determinacji interpretować można jako część zmienności Y wyjaśnioną przez oszacowany model regresji liniowej. Wartość współczynnika determinacji jest unormowana w przedziale [0; 1], oczywiście model jest tym lepiej dopasowany do danych im współczynnik determinacji znajduje się bliżej wartości 1 (czyli 100%). Trudno jest podać granicę wartości współczynnika determinacji, przy którym model liniowy można uznać za wystarczająco dopasowany do danych – kwestia ta jest mocno związana ze specyfiką i pochodzeniem danych.

W tabeli 9.1. przedstawiono przykładowe wyniki analizy regresji liniowej. Prezentowany model opisuje zależność indeksu masy tkanki tłuszczowej (FMI, ang. *Fat Mass Index*) od obwodu talii (WC, ang. *Waist Circumference*), wskaźnika masy ciała (BMI, ang. *Body Mass Index*) oraz wieku. Parametry modelu oszacowano na podstawie próby 165 mężczyzn.

Tabela 9.1. Wyniki analizy regresji liniowej

Zmienna objaśniająca	Oszacowanie parametru	Błąd standardowy szacunku	*p*-value	Współczynnik determinacji
WC [cm]	0,136	0,032	<0,001	0,94
BMI [kg/m^2]	0,282	0,066		
Wiek [lata]	-0,010	0,011		

Źródło: opracowanie własne na podstawie [7]

Zgodnie z oszacowanym modelem, wzrost obwodu talii o 1 cm wiąże się ze zwiększeniem indeksu masy tkanki tłuszczowej średnio o 0,136 pkt. Pacjent o BMI wyższym o 1 kg/m^2 będzie miał wyższy indeks tkanki tłuszczowej niemal o 3 punkty. Indeks FMI spada przeciętnie

o 0,01 pkt wraz ze zmianą wieku o każdy rok. Wynik jednoczesnego testu istotności dla wszystkich parametrów modelu wskazuje że hipotezę o nieistotności parametrów należy zdecydowanie odrzucić. Jednocześnie warto zauważyć, że błąd względny oszacowania parametru dla wieku jest bardzo wysoki (110%). Współczynnik determinacji R^2 wskazuje na bardzo dobre dopasowanie modelu do danych (model wyjaśnia 94% całkowitej zmienności indeksu FMI).

Na koniec warto zauważyć, że model regresji liniowej może być użyteczny w wielu przypadkach, gdzie zależności między obserwowanymi zmiennymi nie mają charakteru liniowego. Często możliwe jest zastosowanie odpowiedniej transformacji dla części danych (np. logarytmowanie, pierwiastkowanie, potęgowanie, odwrotność), i szacowanie modelu liniowego na zmiennych przekształconych. Mogą one być predyktorami w regresji liniowej, o ile dane przekształcone spełniają założenie o liniowej współzależności.

9.2.1. Problem doboru zmiennych objaśniających

Zagadnienie optymalnego doboru zmiennych objaśniających do modeli regresyjnych wykracza poza zakres niniejszego opracowania. Warto jednak zasygnalizować najważniejsze kwestie związane z tym bardzo istotnym w praktyce aspektem.

Z jednej strony pożądane jest, aby model uwzględniał jak najwięcej informacji zawartej w zgromadzonych danych, mającej istotny wpływ na modelowane zjawisko. Z drugiej strony, dodawanie kolejnych zmiennych objaśniających do modelu może prowadzić do pogorszenia jego własności numerycznych (np. współliniowość), a także utrudniać interpretację uzyskanych rezultatów. Istnieją metody pozwalające na poszukiwanie optymalnych zbiorów zmiennych objaśniających, biorąc pod uwagę wymienione cele.

Metody raportowane w czasopismach medycznych najczęściej opierają się na algorytmach krokowych (sekwencyjnych). Polegają one na dodawaniu lub usuwaniu w kolejnych etapach zmiennych objaśniających, kierując się pewnym kryterium opartym o testy istotności. Istnieją trzy warianty tej metody. Dołączanie zmiennych (selekcja postępująca ang. *forward selection*) – metoda startuje od modelu z jedną zmienną objaśniającą, następnie dodając kolejne w zależności od ich istotności statystycznej w modelu. Eliminacja zmiennych (selekcja wsteczna ang. *backward selection*) – punktem wyjścia jest model zawierający wszystkie potencjalne predyktory, zmienne są następnie usuwane aż do uzyskania istotności wszystkich zmiennych w modelu na poziomie pewnego ustalonego progu. Metoda krokowa (ang. *stepwise selection*) – w kolejnych etapach możliwe jest zarówno włączenie nowej zmiennej, jak i usunięcie z modelu.

Warto zauważyć że w ten sposób nie rozpatruje się wszystkich możliwych podzbiorów zmiennych objaśniających. Zastosowanie metod dołączania i eliminacji nie musi też doprowadzić do uzyskania tej samej postaci modelu. Budowa modeli regresyjnych jest pewnego rodzaju sztuką. Należy pamiętać że za ostateczny kształt modelu odpowiedzialny jest zawsze badacz posiadający wiedzę z danej dziedziny i będący w stanie krytycznie zweryfikować związki obserwowane w zgromadzonych danych z wiedzą teoretyczną[3]. Metody służące do „mechanicznej" konstrukcji modelu optymalizującego pewne jego własności matematyczne,

[3] Por. przypis 3 w rozdziale 8.

nie zawsze muszą doprowadzić do konstrukcji modelu optymalnego ze względu na jego wartość naukową. Warto pamiętać, że postać modelu może też zależeć od celu analizy: czy model ma służyć do wyjaśnienia mechanizmów badanego zjawiska (pomóc w weryfikacji teorii naukowej), czy też przede wszystkim do prognozowania.

9.3. Model regresji logistycznej

W poprzedniej części przedstawiono model regresji liniowej, służący do modelowania zależności zmiennej ilościowej Y od (najczęściej wielu) zmiennych objaśniających. W badaniach medycznych, bardzo często rozpatrywaną miarą wyniku jest zmienna binarna (np. odpowiedź na leczenie, przeżycie określonego odcinka czasu, etc). Zwróćmy uwagę, że model regresji liniowej nie może być w tej sytuacji użyty, ze względu na ograniczenie zakresu zmiennej objaśnianej do dwóch wartości. Dlatego w tym rozdziale przedstawiony zostanie model regresji, pozwalający na analizę wyników o charakterze dychotomicznym.

Regresja logistyczna jest techniką analizy danych dychotomicznych (zero-jedynkowych), pozwalającą na jednoczesne uwzględnienie wpływu wielu zmiennych objaśniających. Innymi słowy, metoda ta jest odpowiednikiem modelu regresji liniowej, przydatnym do analizy binarnych punktów końcowych.

Ogólna postać modelu regresji logistycznej to

$$\text{logit}\,(P) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k,$$

gdzie $\text{logit}\,(P) = \ln\left(\frac{P}{1-P}\right)$ jest przekształceniem logitowym, przekształcającym prawdopodobieństwa modelowanego zdarzenia (P) na zbiór liczb rzeczywistych[4]. W rozdziale 6 zdefiniowane zostało pojęcie szansy (O, ang. *odds*). Wykorzystując związek między prawdopodobieństwem i szansą, postać modelu logistycznego można przedstawić w równoważnej postaci

$$\ln\left(\frac{P}{1-P}\right) = \ln(O) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k,$$

zatem model logistyczny objaśnia logarytm z szansy analizowanego zdarzenia.

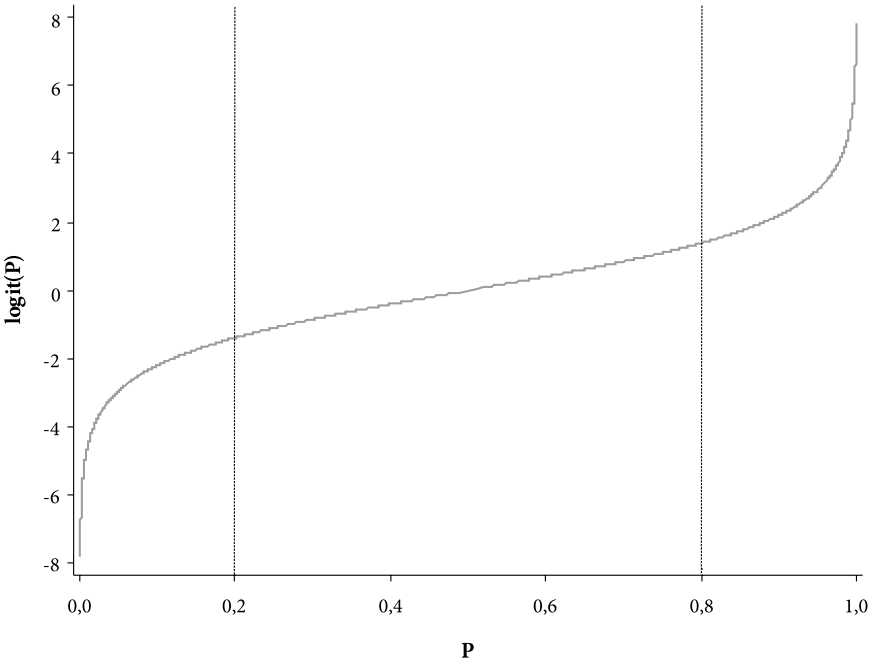
Kształt funkcji logit zaprezentowany został na wykresie 9.4.

---

[4] Istnieją inne przekształcenia stosowane do modelowania częstości (np. funkcja probit); popularność regresji logistycznej w zastosowaniach medycznych wynika m.in. z prostej interpretacji oszacowań parametrów w terminach ilorazów szans.

Warto zauważyć, że model logistyczny jest szczególnym przypadkiem tzw. uogólnionych modeli liniowych (ang. *generalized linear models*). Idea tych modeli polega na zastąpieniu lewej strony równania regresyjnego pewną funkcją modelowanego zjawiska. Innym przykładem takiego modelu użytecznym w medycynie jest model regresji Poissona, którego można użyć np. do analizy liczby pewnych zdarzeń zaobserwowanych u pacjentów w zdefiniowanym przedziale czasowym.

Wykres 9.4. Przekształcenie logitowe



Źródło: opracowanie własne

Warto zwrócić uwagę, że funkcja logit ma przebieg zbliżony do liniowego dla dość szerokiego zakresu prawdopodobieństw (między 0,2 a 0,8). Natomiast dla bardzo małych lub bardzo dużych wartości P, przebieg funkcji logit jest zdecydowanie bardziej „stromy". Własność ta jest zgodna z intuicją, zgodnie z którą np. zmiana o 5 p.p. w okolicy P=0,5 może nie mieć takiego samego znaczenia klinicznego jak zmiana z 0,9 na 0,95.

Parametry strukturalne w modelu regresji logistycznej szacuje się metodą największej wiarogodności (ang. *maximum likelihood*). Jest to metoda numeryczna tzn. nie jest możliwe podanie bezpośrednich formuł na obliczenie oszacowań parametrów.

Po przekształceniu wzoru definiującego kształt modelu logistycznego otrzymujemy następującą zależność szansy zdarzenia od zmiennych objaśniających:

$$O = e^{\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_k \cdot X_k}.$$

Z wzoru wynika kierunek wpływu poszczególnych zmiennych na modelowaną szansę (wzrost wartości X_i będzie powodował wzrost szansy, o ile oszacowanie β_1 jest dodatnie). W publikacjach medycznych, wyniki analizy logistycznej są najczęściej raportowane jako odpowiednie ilorazy szans (OR, ang. *odds ratio*). Parametr $e^{\beta i}$ interpretuje się jako iloraz szans, odpowiadający jednostkowej zmianie X_i, przy pozostałych wartościach niezmienionych. W przypadku binarnej zmiennej objaśniającej, oszacowanie $e^{\beta i}$ oznacza iloraz szans dla wartości X_i kodowanych jako 1 i 0 (przykładowo, iloraz szans dla pacjenta z chorobą współistniejącą vs brak choroby).

Podobnie jak w modelu regresji liniowej, można prowadzić wnioskowanie statystyczne dotyczące parametrów β_0, β_1, …, β_k, tzn. wyznaczyć błędy standardowe oszacowań, konstruować przedziały ufności[5], przeprowadzać testy hipotez dotyczących poszczególnych parametrów lub całego modelu regresji.

Tabela 9.2. przedstawia przykładowe wyniki analizy regresji logistycznej. Celem badania było wyjaśnienie związku pomiędzy zapadalnością na nowotwory jamy ustnej i gardła, w zależności od wybranych zachowań seksualnych pacjentów. Dane pochodzą z badania typu case-control (typu *case-contro)l*[6]; model oszacowano na podstawie danych dotyczących 100 pacjentów z diagnozą oraz 200-osobowej grupy kontrolnej. W celu wyeliminowania wpływu potencjalnych zmiennych zakłócających na wyniki wnioskowania, do szacowanego modelu regresji włączono dodatkowo następujące zmienne: wiek, płeć, palenie papierosów, picie alkoholu, higiena jamy ustnej, wywiad rodzinny w kierunku raka głowy i szyi.

Tabela 9.2. Wyniki analizy regresji logistycznej

Zmienna objaśniająca	**Skorygowany iloraz szans (OR)**	**Przedział ufności (95% CI)**
Liczba partnerów oralnych w ciągu życia		
0	1,0	
1-5	1,9	(0,8 - 4,5)
≥6	3,4	(1,3 - 8,8)*
Seks analny		
nie	1,0	
tak	1,3	(0,8 - 2,2)
Wiek inicjacji seksualnej		
≥18	1,0	
≤17	1,3	(0,7 - 2,3)
Użycie prezerwatywy		
zazwyczaj lub zawsze	1,0	
nigdy lub rzadko	2,2	(1,2 - 3,8)
Seks homoseksualny		
nie	1,0	
tak	1,0	(0,4 - 2,6)

** p-value dla trendu 0,009*

Źródło: opracowanie własne na podstawie [3]

Liczba partnerów oralnych w ciągu życia została uwzględniona w modelu jako zmienna porządkowa[7]. W porównaniu do wartości „0" (poziom referencyjny), zwiększenie tej liczby do 1-5 powoduje 90% wzrost szansy na zachorowanie, zaś przejście do kategorii ≥6 oznacza zwiększenie tej szansy 3,5 krotnie. Test istotności dla trendu potwierdza istotny wzrost prawdopodo-

[5] Przy czym w raportach z badań medycznych, przedziały ufności są zwyczajowo wyrażone w terminach odpowiedniego ilorazu szans – por. rozdz. 7.2.2.

[6] Por. rozdz. 2.

[7] Por. rozdz. 8.7.

bieństwa zachorowania wraz ze wzrostem liczby partnerów (hipotezę o jednakowym ryzyku zachorowania we wszystkich trzech grupach odrzucamy przy poziomie istotności 1%). Pozostałe zmienne objaśniające mają charakter dychotomiczny. Uprawianie seksu analnego oraz inicjacja w wieku ≤17 lat powoduje wzrost szansy zachorowania o 30%, jednak przedział ufności wokół oszacowania punktowego zawiera wartość 1 – zatem zależność ta nie jest istotna statystycznie. Rzadkie lub brak stosowania prezerwatywy zwiększa szansę zachorowania ponad dwukrotnie (parametr istotny statystycznie). Ostatnia rozpatrywana zmienna objaśniająca nie ma wpływu na zachorowalność – szansa jest jednakowa w obu grupach „tak" i „nie".

Bibliografia

1. Armitage P., Berry G., Matthews J.N.S.: Statistical Methods in Medical Research. John Wiley and Sons, 2008.
2. Dallal G.E.: Correlation Coefficients. W: The Little Handbook of Statistical Practice [dostęp 30 września 2011]. Dostępne w Internecie: http://www.jerrydallal.com/LHSP/corr.htm
3. D'Souza G. *et al*: Case–Control Study of Human Papillomavirus and Oropharyngeal Cancer. N Engl J Med 2007;356:1944-56.
4. Goldsmith *et al*: Mapping of the EQ-5D index from clinical outcome measures and demographic variables in patients with coronary heart disease. Health and Quality of Life Outcomes 2010, 8:54.
5. Hosmer D.W., Lemeshow S.: Applied logistic regression. Wiley-Interscience, 2000.
6. Koronacki J., Mielniczuk J.: Statystyka dla studentów kierunków technicznych i przyrodniczych. Warszawa, 2009.
7. Peltz G., Aguirre M.T., Sanderson M., Fadden M.K.: The role of fat mass index in determining obesity. Am J Hum Biol. 2010 ; 22(5): 639–647.
8. Zar J.H.: Biostatistical Analysis. Prentice Hall, 2010.

X. Analiza przeżycia

Michał JAKUBCZYK

Przedmiotem niniejszego rozdziału jest bodaj najbardziej naturalne spojrzenie na ocenę skuteczności terapii, tj. analiza czasu przeżycia pacjenta w wyniku stosowania leczenia. Tak więc o ile w rozdziale 7 zajmowaliśmy się przeżyciem w ujęciu tak/nie, o tyle poniżej analizowany jest nie sam fakt, a czas przeżycia.

Przy analizie przeżycia można by na pozór ograniczyć się do metod porównywania średnich, omówionych w rozdziale 8, gdyby nie następujące zagadnienia. Po pierwsze w analizie przeżycia pacjentów objętych badaniem klinicznym rzadko występuje komfortowa sytuacja, że grupa pacjentów jest jednocześnie włączona do badania i obserwowana zgodnie z protokołem aż do zakończenia. Występuje tzw. cenzorowanie (ang. *censoring*), tj. pacjenci włączeni do badania są z niego wyłączani z różnych powodów i nie ma możliwości ich obserwowania w całym założonym horyzoncie. Pacjenci mogą odejść z badania, może wystąpić konieczność zmiany terapii z powodów medycznych, itp.

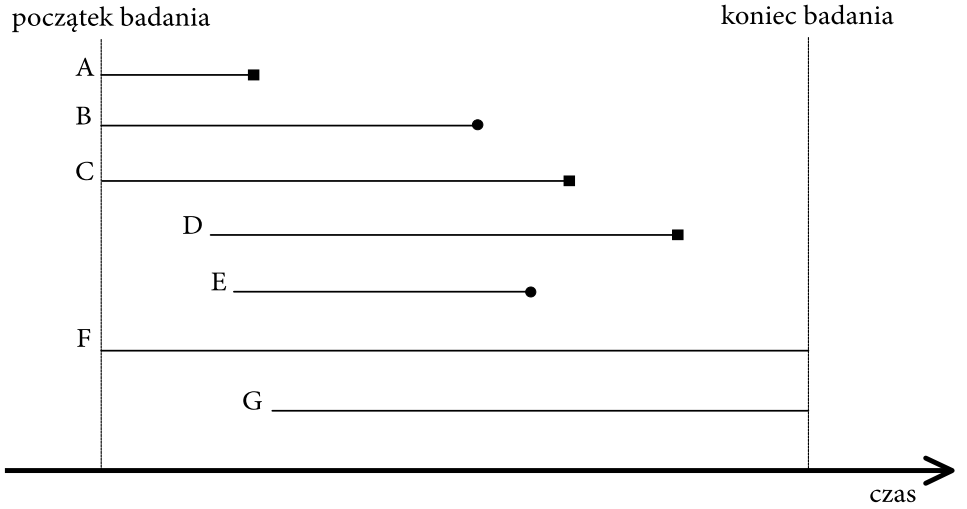
Dodatkowo oczywiście analiza prowadzona jest nie w perspektywie kilkudziesięciu lat, kiedy będzie dostępna informacja o czasie przeżycia dla wszystkich (niecenzorowanych) pacjentów, lecz w kilka miesięcy lub lat po rozpoczęciu badanego leczenia. Tak więc powstaje pytanie, jak uwzględnić informacje o żyjących pacjentach, zwłaszcza że przy organizacji badania ma raczej miejsce sytuacja, że pacjenci nie są włączani do niego w jednym momencie, a są dołączani sukcesywnie.

W związku z tym może okazać się, że w momencie analizy mamy informacje o pacjencie, który żył 6 miesięcy (np. od momentu zabiegu, rozpoczęcia terapii, itp.), lecz nie ma dostępnych informacji o jego dalszych losach. Może być także sytuacja, że mamy aktualne informacje o żyjących pacjentach, z których jeden rozpoczął leczenie 4 miesiące temu, a drugi 9 miesięcy. Każdy z tych pacjentów stanowi informację do wykorzystania, ale trudno jest uwzględnić ją w zwykłej analizie średniego czasu.

Wykres 10.1. przedstawia możliwe sytuacje związane z cenzorowaniem i niejednoczesnym włączaniem pacjentów do badania.

Po drugie w analizie przeżycia z klinicznego punktu widzenia interesujące jest poznanie nie tylko średniego czasu przeżycia, a raczej zrozumienie całego rozkładu, tj. tego czy zgony następują od razu po rozpoczęciu leczenia, czy też dopiero w kilka lat po.

Wykres 10.1. Przykładowe możliwe sytuacje w analizie przeżycia. Zgony oznaczono kwadratami, obserwacje cenzorowane – kółkami. Pacjenci A, C, D zmarli w trakcie badania. Pacjenci B i E odeszli z badania (przedwcześnie zakończono ich obserwację). Pacjenci F i G żyli w momencie zakończenia. Pacjentów D, E, G dołączono w trakcie badania.



Źródło: opracowanie własne

Analizę przeżycia można podzielić na kilka obszarów tematycznych, które wyznaczają plan podrozdziałów poniżej. Po pierwsze omówiono metody charakteryzowania czasu przeżycia w danej grupie (populacji generalnej) pacjentów i sposoby interpretowania tzw. krzywej przeżycia (ang. *survival curve*). W dalszej części przedstawiono metody porównywania przeżycia w dwóch grupach, w szczególności z użyciem testowania statystycznego. Wreszcie w ostatniej części omówiono metody identyfikowania wpływu (najczęściej wielu różnych) charakterystyk pacjenta na przeżycie z użyciem tzw. modelu ryzyka proporcjonalnego Coksa (ang. *Cox proportional hazard model*). Poniżej wykorzystano ogólne informacje dotyczące zagadnień estymacji i testowania hipotez statystycznych z rozdziału 6, a także dotyczące analizy wieloczynnikowej z rozdziału 9.

Dodajmy wreszcie, że o ile mówimy o analizie przeżycia, o tyle przedstawione techniki można odnosić także do innych zdarzeń niż zgon. Często w analizach onkologicznych bada się tzw. przeżycie bez progresji choroby (ang. *progression-free survival, PFS*), czyli wyróżnionym zdarzeniem jest zgon lub progresja choroby.[1]

[1] Podkreślmy tu, że obserwowanym zdarzeniem jest alternatywa – progresja lub zgon. Obserwowanie samej progresji choroby mogłoby dać myląće wyniki. Mogłoby się zdarzyć np., że progresja w jednej grupie byłaby mniejsza niż w drugiej z tego powodu, że w pierwszej grupie pacjenci w gorszym stanie zdrowia mają wyższe ryzyko zgonu i sama progresja jest nieobserwowana.

10.1. Opisywanie czasu przeżycia – krzywa Kaplana-Meiera

Zacznijmy od metod opisywania czasu przeżycia, tj. przedstawiania, jak kształtuje się ryzyko zgonu w kolejnych momentach. Sposobem prezentacji jest tzw. krzywa przeżycia (ang. *survival curve*), tj. krzywa pokazująca prawdopodobieństwo dożycia kolejnych momentów. Standardowo do wykreślenia tej krzywej jest stosowany tzw. estymator Kaplana-Meiera (ang. *Kaplan-Meier estimator*). Zanim przejdziemy do interpretacji tej krzywej uzyskanej metodą Kaplana-Meiera, warto przyjrzeć się idei jej konstrukcji.

Standardowo dane do analizy przeżycia są w postaci przedstawionej w tabeli 10.1. poniżej. Mamy zatem dla każdego pacjenta najbardziej aktualną informację. Informacja ta de facto może albo oznaczać, że pacjent zmarł po kilku miesiącach, albo po prostu, że według najbardziej aktualnych informacji żyje (te najbardziej aktualne informacje mogą pochodzić z momentu analizy albo z wcześniejszego w przypadku obserwacji cenzorowanej).

Tabela 10.1. Przykładowe dane do analizy przeżycia zebrane 7 miesięcy po rozpoczęciu badania.

ID pacjenta	Czas (liczba miesięcy)	Informacja
1	1	zgon
2	2	ostatni kontakt (obs. cenzorowana)
3	3	zgon
4	4	zgon
5	4	zgon
6	4	pacjent żyje (włączony w 3. miesiącu)
7	5	zgon
8	5	zgon
9	6	zgon
10	7	pacjent żyje

Źródło: opracowanie własne

Naiwne podejście do obliczenia krzywej przeżycie mogłoby mieć np. jedną z trzech postaci. Po pierwsze można byłoby przyjąć pesymistycznie, że brak informacji o pacjencie oznacza, że pacjent nie żyje. I tak dla pacjenta 2. założylibyśmy wówczas, że zmarł w 2. miesiącu po włączeniu. Oczywiście takie podejście zaniża przeżycie pacjentów. Po drugie można byłoby przyjąć pesymistycznie, że brak informacji interpretujemy na korzyść leku, tj. np. dla pacjenta 2. zakładamy, że dożył 7. miesiąca. Takie podejście zawyża przeżycie pacjentów. Można byłoby także rozważyć pomijanie w analizach pacjentów, co do których nie mamy pełnej informacji, u nas pacjentów 2. i 6. Zauważmy, że wtedy nie wykorzystujemy ogółu dostępnych informacji, a dodatkowo częściej będziemy pomijać pacjentów, którzy przeżyli (co do pacjentów, którzy umarli, mamy pełne informacje i ich nie odrzucimy) i to przeżyli raczej dłuższy czas (bo jest większe prawdopodobieństwo cenzorowania).

Estymator Kaplana-Meiera wykorzystuje pełną informację z próby, a dodatkowo nie przyjmuje żadnych – ani pesymistycznych, ani optymistycznych założeń. Wymaga oczywiście, aby cenzorowanie było niezależne od ryzyka zgonu.[2] Obliczając estymator Kaplana-Me-

[2] Zauważmy dla przykładu, że gdyby wszyscy pacjenci tuż przed zgonem byli cenzorowani, dostalibyśmy zafałszowane oszacowania.

iera analizuje się kolejne momenty i dla każdego z nich bada się: i) jak liczna była grupa ryzyka (tj. grupa dostępnych do analizy pacjentów); ii) jak wiele zgonów nastąpiło. Wystarczy przy tym badać momenty, w których nastąpił zgon. Wówczas jesteśmy w stanie obliczyć warunkowe (ang. *conditional*) prawdopodobieństwo zgonu pod warunkiem dożycia danego momentu, np. prawdopodobieństwo zgonu w 2 miesiące po rozpoczęciu leczenia (dla uproszczenia przyjmijmy comiesięczną analizę). Wtedy bezwarunkowe (ang. *unconditional*) prawdopodobieństwo przeżycia poza dany moment to iloczyn prawdopodobieństw warunkowych przeżycia kolejnych momentów, np. przeżycie ponad 2 miesiące wymaga przeżycia 1. miesiąca i przeżycia 2. miesiąca pod warunkiem dożycia 2. miesiąca. Ten schemat obliczeń dla naszego przykładu przedstawia tabela 10.2.

Tabela 10.2. Obliczanie krzywej przeżycia metodą Kaplana-Meiera.

miesiąc	Liczba		Prawdopodobieństwo		
	pacjentów	zgonów	zgonu (warunkowe)	przeżycia (warunkowe)	przeżycia
1	10	1	10%	90%	90%
3	8	1	12,5%	87,5%	78,8% = 90% * 87,5%
4	7	2	28,57%	71,4%	56,3%
5	4	2	50%	50%	28,1%
6	2	1	50%	50%	14,1%

Źródło: opracowanie własne

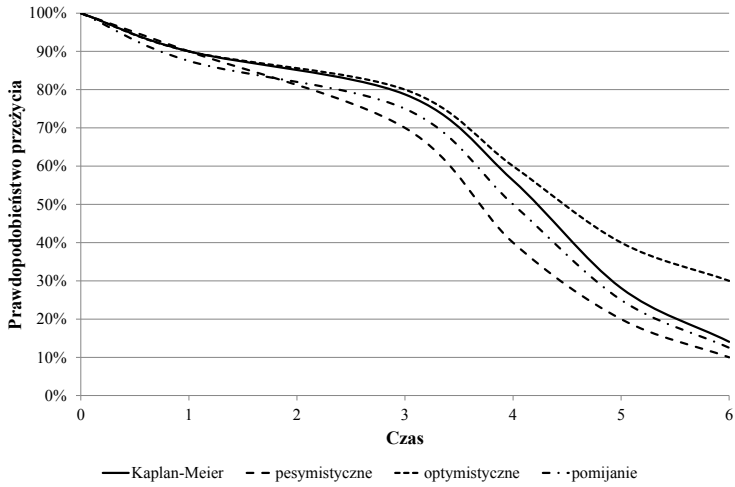
Wyniki estymacji, a także wyniki podejść naiwnych opisanych powyżej, przedstawiono na wykresie 2.

Często pojawia się pytanie, co właściwie opisuje krzywa Kaplana-Meiera, tj. czy opisuje sytuację, która miała miejsce w próbie, czy raczej jakie jest przeżycie w populacji generalnej. Jeśli chodzi o odniesienie krzywej przeżycia do próby, to należy pamiętać, że z uwagi na cenzorowanie, de facto nie wiadomo, co się stało w próbie. Tak więc dosłownie rzecz ujmując krzywa Kaplana-Meiera być może nie opisuje faktycznego przeżycia w próbie. Natomiast krzywa ta uwzględnia tylko informacje z próby i wszystkie dostępne informacje z próby. Krzywą Kaplana-Meiera przede wszystkim należy traktować jako estymator krzywej przeżycia w populacji generalnej (z której została wylosowana próba). W tym sensie opisuje zarówno próbę, jak i – szerzej – populację generalną. Oczywiście należy tu pamiętać o błędzie estymacji, tj. krzywa jest tylko estymatorem punktowym i prawie na pewno różni się od prawdziwej, nieznanej krzywej przeżycia.

Pamiętając o powyższych zastrzeżeniach, zobaczmy, co można wywnioskować z przebiegu krzywej przeżycia. Po pierwsze krzywa ta wskazuje na prawdopodobieństwo dożycia poszczególnych momentów czasu. Wartości te można odczytać bezpośrednio z osi rzędnych (pionowej) dla danych momentów czasu na osi odciętych. Po drugie na krzywej tej widać np. medianę czasu przeżycia, tj. taki czas, że prawdopodobieństwo dożycia tego momentu wynosi 50%. Wartość tę odczytujemy z osi odciętych jako taki moment, dla którego krzywa przeżycia spada do poziomu 50%. Jeśli przeżycie jest bardzo duże (albo okres obserwacji bardzo krótki), może zdarzyć się, że krzywa nie spada do tego poziomu w analizowanym horyzoncie. Wów-

czas nie ma możliwości obliczenia tej mediany na podstawie danych. Ewentualne przybliżenia mogłyby wykorzystać ekstrapolację dostępnych danych np. z wykorzystaniem tzw. krzywych ryzyka opisanych w kolejnym podrozdziale.

Wykres 10.2. Przykładowe krzywe przeżycia – uzyskane metodą Kaplana-Meiera oraz z użyciem naiwnych podejść.



Źródło: opracowanie własne

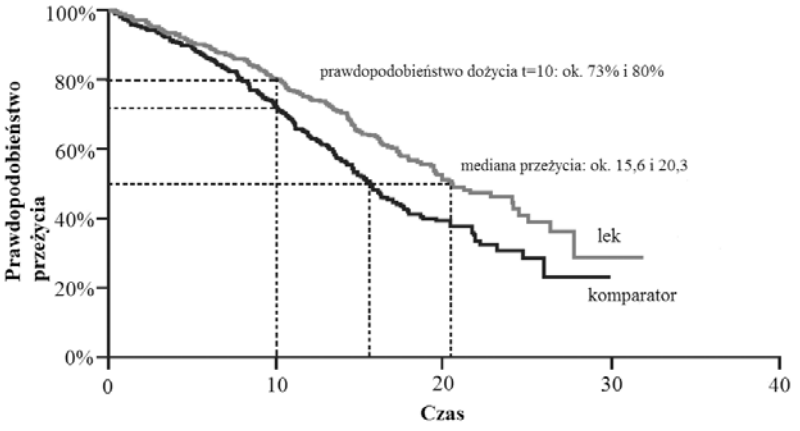
Interpretację ma także pole powierzchni pod krzywą – jest ono równe oczekiwanej długości życia, ale oczywiście nie ma możliwości wizualnego obliczenia tego pola. Dodatkowo bardzo często w badaniach klinicznych krzywa przeżycia nie osiąga zera, więc nie ma możliwości obliczenia tego czasu – nie wiadomo, co dzieje się z pacjentami poza dostępnym horyzontem.

Na wykresie 10.3. przedstawiono przykładową krzywą przeżycia z badania klinicznego. Czasem w opublikowanych badaniach pod wykresem przedstawia się dane liczbowe dotyczące liczby pacjentów obserwowanych (narażonych, ang. *at risk*) w danym momencie. Warto zwrócić uwagę, że nie jest błędem, jeśli liczba pacjentów obserwowanych w danym momencie jest równa 0, a krzywa przeżycia przyjmuje wartość większą od zera, tj. krzywa wskazuje, że prawdopodobieństwo dożycia jest dodatnie (przykład przedstawiono na wykresie 10.3.). Krzywa Kaplana-Meiera osiągnie zero, jeśli w danych zdarzy się moment, w którym wszyscy obserwowani pacjenci umrą. Tymczasem może okazać się, że żaden pacjent po prostu jeszcze nie osiągnął określonego horyzontu czasowego obserwacji z uwagi na moment analizy, włączenie pacjentów do badania w czasie jego trwania, cenzorowanie obserwacji.

Należy pamiętać, że krzywa Kaplana-Meiera jest estymatorem prawdopodobieństwa przeżycia, a zatem dotyczą jej kwestie błędów estymacji. Z tego powodu w publikacjach czasem przedstawia się krzywą wraz z przedziałami ufności dla wartości krzywej w określonych momentach. Przykład przedstawiono na wykresie 4. Może zdarzyć się, tak jak na przedstawionym przykładzie, że górna granica przedziału ufności dla późniejszego momentu przekracza dolną granicę przedziału ufności dla wcześniejszego momentu. Nie oznacza to, że krzywa przeżycia

może rosnąć! Przedziały ufności są obliczone niezależnie dla poszczególnych momentów i należy je odrębnie rozpatrywać. Przy łącznej interpretacji oczywiście należałoby uwzględnić związki między nimi, co bardzo utrudniłoby wizualizację możliwych przebiegów krzywej. Można obliczać także przedziały ufności dla takich parametrów jak mediana czasu przeżycia, czy oczekiwany czas przeżycia. Czytelnikowi zainteresowanemu tymi kwestiami możemy polecić np. [4].

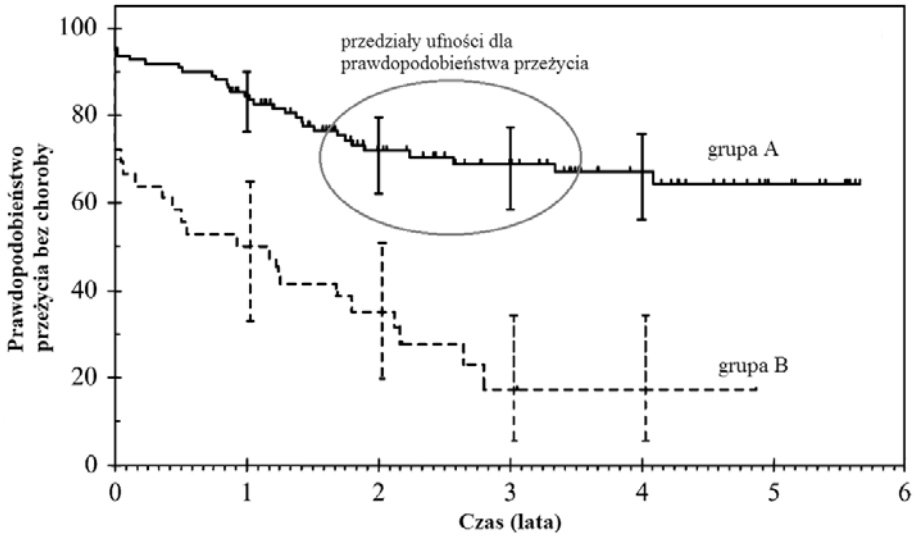
Wykres 10.3. Krzywe przeżycia dla leku i komparatora. Na wykresie przedstawiono sposób odczytania mediany czasu przeżycia oraz prawdopodobieństwa dożycia określonego momentu.



Liczba pacjentów								
lek	402	362	320	178	73	20	1	0
komparator	411	363	292	139	51	12	0	0

Źródło: zmodyfikowany wykres na podstawie [2]

Wykres 10.4. Estymator Kaplana-Meiera wraz z przedziałami ufności.



Źródło: zmodyfikowany wykres na podstawie [8]

10.2. Porównywanie przeżycia w dwóch grupach

Przy ocenie technologii istotne jest odpowiedzenie na pytanie, czy dany lek przynosi korzyści w sensie przeżycia w porównaniu do komparatora. W tym podrozdziale poruszamy tę kwestię w dwóch ujęciach – porównania samych krzywych przeżycia i porównywania tzw. krzywych ryzyka.

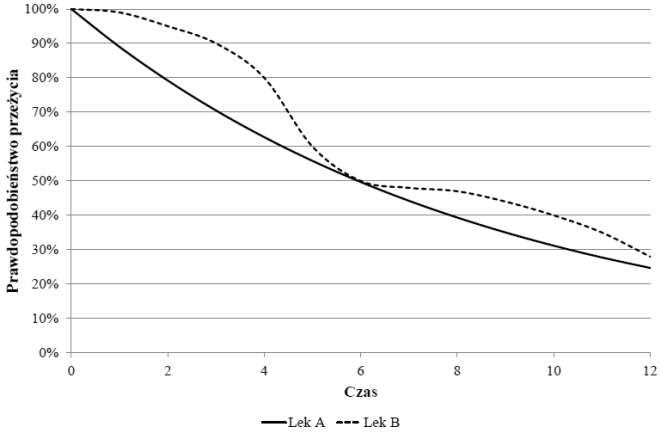
10.2.1. Porównanie krzywych i testowanie hipotez

Bardzo często w badaniach klinicznych, w których prowadzi się analizę przeżycia, wynik jest zaprezentowany w postaci jednego rysunku zawierającego krzywe dla dwóch lub więcej grup pacjentów. Umiejąc interpretować pojedyncze krzywe przeżycia, tj. odczytywać prawdopodobieństwa dożycia poszczególnych momentów, medianę i inne percentyle, można porównywać krzywe dla kilku leków.

Oczywiście o przewadze analizowanego leku świadczy krzywa leżąca wyżej niż krzywa dla komparatora. Taka sytuacja oznacza, że prawdopodobieństwo dożycia każdego konkretnego momentu jest większe dla analizowanego leku, a co za tym idzie większa jest także mediana czasu przeżycia (krzywa leżąca wyżej leży także bardziej na prawo) i oczekiwany czas przeżycia (większe jest pole powierzchni pod krzywą). Np. dla przykładu przedstawionego powyżej (Wykres 3) bardziej korzystna jest krzywa dla leku niż dla komparatora.

Uważać należy formułując interpretacje na podstawie porównań. Można np. porównać średnie czasy przeżycia w obu grupach (zakładając, że są możliwe do obliczenia, tj. krzywe przeżycia osiągnęły zero) i powiedzieć, że średni czas przeżycia rośnie o jakąś wartość, albo inaczej – ile wynosi średni wzrost czasu przeżycia. Nieco inaczej jest z medianą – można oczywiście obliczyć różnicę median i powiedzieć, że mediana czasu przeżycia rośnie o tyle i tyle. Nie oznacza to jednak, że ta liczba oznacza medianę wzrostu czasu przeżycia. Nie obserwowaliśmy pacjentów w dwóch wersjach – po przyjęciu leku i komparatora, w związku z czym nie wiadomo, u poszczególnych pacjentów, ile wyniósłby hipotetyczny wzrost. Mamy tu problem analogiczny do przedstawionego w rozdziale 6.1.2. Aby zilustrować tę kwestię możemy spojrzeć na wykres 10.5.

Wykres 10.5. Przykład w trudności interpretacji różnicy median czasów przeżycia.



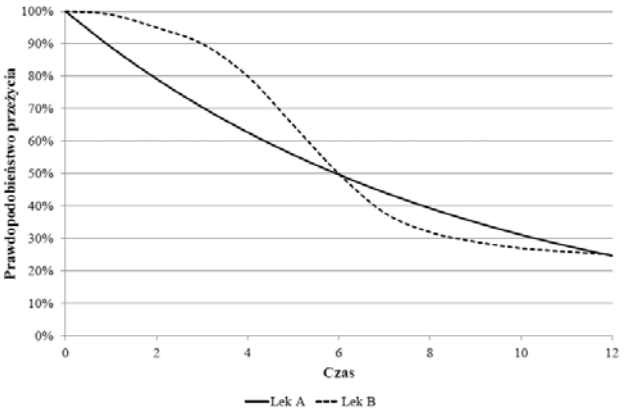
Źródło: opracowanie własne

Przywołany przykład ilustruje jednocześnie inną jeszcze kwestię. Pojawia się czasem następujące pytanie – czy skoro dwie krzywe przeżycia łączą się dla jakiegoś momentu w czasie (w szczególności dla bardzo długiego horyzontu obie dążą do zera), to znaczy, że nie ma różnic między porównywanymi lekami? Oczywiście tak nie jest. Rzeczywiste krzywe przeżycia dla bardzo długiego horyzontu osiągają zero, co odzwierciedla fakt, że wszyscy pacjenci umrą – istotne jest natomiast, że wyżej położone krzywe opóźniają ten moment zgonu. Podobnie dla krzywych spotykających się w danym momencie. Np. dla przykładu poniżej krzywe łączą się w t=6 na poziomie 50%. Oznacza to, że w obu grupach połowa pacjentów umrze w ciągu 6 miesięcy, a jednak w jednej grupie pacjenci ci średnio umierają później.

Krzywe przeżycia mogą się krzyżować, co oznacza, że profil ryzyka jest bardzo różny dla różnych podokresów. Przykład przedstawiono na wykresie 6. I tak lek B jest bezpieczniejszy w pierwszych miesiącach (do ok. 4. miesiąca), lecz później jego początkowe korzyści są systematycznie niwelowane przez wyższe ryzyko zgonu i w 6. miesiącu odsetek dożywających zrównuje się (aczkolwiek średni czas przeżycia jest ciągle wyższy dla grupy stosującej B). Ryzyko pozostaje wyższe dla leku B i prawdopodobieństwo dożycia np. 8. miesiąca jest już niższe niż dla leku A. W kolejnych miesiącach (od 9.) ryzyko jest większe dla leku A. Zakładając, że powyżej 12. miesiąca przebieg krzywych byłby identyczny, oczekiwana długość życia byłaby wyższa dla leku B, co widać porównując wizualnie pola powierzchni pod krzywymi.

Powyżej zajmowaliśmy się porównaniami przeżycia abstrahując od faktu, że krzywa przeżycia jest estymatorem. Fakt ten uwzględnia się, stosując metody wnioskowania statystycznego przedstawione w rozdziale 6. Do testowania statystycznego wykorzystuje się najczęściej tzw. test log-rank. Hipoteza zerowa mówi o tym, że krzywe przeżycia są jednakowe dla obu grup. Wynikiem testu jest wartość parametru p, na podstawie której H_0 odrzuca się lub nie. W badaniach (np. [16]) spotyka się także dość często test log-rank uwzględniający potencjalne zróżnicowanie ryzyka ze względu na inny czynnik lub czynniki (ang. *stratified log-rank test*), np. inną przeżywalność wśród grup o innym stanie klinicznym w momencie włączenia do badania. Nieuwzględnienie tego wpływu innych czynników, jeśli byłyby one nierównomiernie rozłożone między porównywanymi grupami, mógłby prowadzić do zaburzeń à la paradoks Simpsona (por. rozdz. 7).

Wykres 10.6. Przykład krzyżujących się krzywych przeżycia.



Źródło: opracowanie własne

10.2.2. Krzywe ryzyka

Przykład, który ilustruje Wykres 6, sugeruje następujący problem. Analiza krzywych przeżycia może utrudniać natychmiastowe dostrzeżenie, jak kształtuje się ryzyko w różnych analizowanych podokresach. Stwierdziliśmy powyżej, że dla tego przykładu już od 4. miesiąca ryzyko dla leku B jest wyższe. Wynika to z tego, że krzywa przeżycia zaczyna od tego momentu szybko maleć, mimo że wciąż jest wyżej niż dla krzywej dla leku A. Tak więc, zmiany ryzyka mogą być mało widoczne na wykresie krzywych przeżycia.

W analizie przeżycia poza krzywymi przeżycia analizuje się także często tzw. krzywe ryzyka (ang. *hazard curve*).[3] Przez ryzyko nie rozumiemy tutaj prawdopodobieństwa zgonu w danym momencie. Nie wchodząc w szczegóły matematyczne, możemy tu jedynie stwierdzić, że jest to pewna miara natężenia tego prawdopodobieństwa – tj. krzywa ryzyka o większych wartościach oznacza wyższe prawdopodobieństwo zgonu w danym okresie (pod warunkiem dożycia tego okresu). Ryzyko może przyjmować dowolne wartości dodatnie, także przekraczające 1 (gdyż nie jest to prawdopodobieństwo). Istotne jest, że istnieją formuły matematyczne pozwalające na swobodne przechodzenie między krzywymi przeżycia i krzywymi ryzyka.

Praca z krzywymi ryzyka jest o tyle wygodna, że krzywe te wyraźniej pokazują zmiany zagrożenia życia pacjenta w czasie. I tak np. intuicyjnie rozumiane pojęcie jednakowego ryzyka (opisujące np. życie człowieka w średniowieczu, który w młodości mógł umrzeć ze względu na warunki sanitarne, w wieku dojrzałym – na wojnie, a potem – „ze starości") odpowiada stałej krzywej ryzyka. Jeśli to stałe w czasie ryzyko jest różne dla dwóch badanych grup pacjentów, to krzywe ryzyka utrzymają to zróżnicowanie, które zanika dla krzywych przeżycia, bo obie dążą do zera.

W analizie często pracuje się na krzywych ryzyka zadanych matematycznymi wzorami, tj. należącymi do konkretnych rodzin krzywych ryzyka. Jeśli krzywą ryzyka określimy funkcją h(t), gdzie t oznacza czas, to możemy wyróżnić np. następujące popularne w analizach HTA typy krzywych:

- stałe ryzyko, $h(t) = \lambda$, $\lambda > 0$,;
- rozkład Gompertza, $h(t) = \lambda e^{\gamma t}$, $\lambda > 0$,;
- rozkład Weibulla, $h(t) = \lambda\gamma t^{\gamma-1}$, $\lambda > 0$, $\gamma > 0$,;
- rozkład Silera, $h(t) = \alpha e^{-\beta t} + \gamma + \delta e^{\zeta t}$, $\alpha, \beta, \gamma, \delta, \zeta \geq 0$,.

Rozkłady Gompertza i Weibulla pozwalają na uwzględnienie wzrostu lub spadku ryzyka w czasie. Rozkład Silera pozwala na uwzględnienie zwiększonego ryzyka na początku (np. tuż po operacji w wyniku powikłań), spadku i ponownego wzrostu (np. w wyniku starzenia się lub spadku skuteczności leczenia).

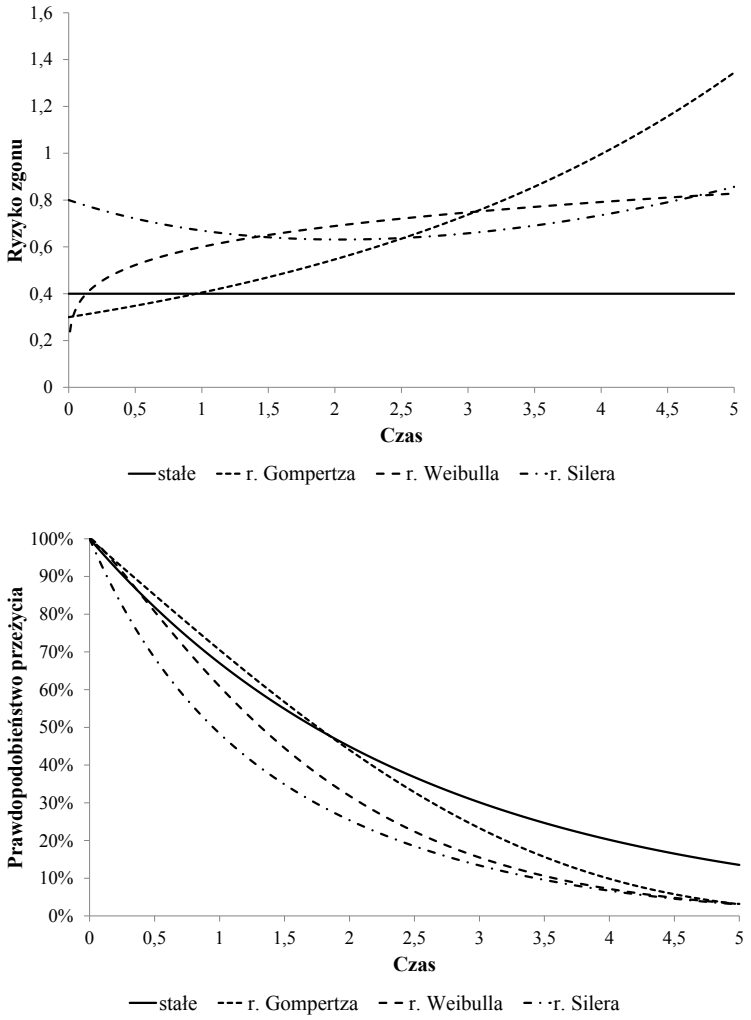
Przykłady krzywych ryzyka dla powyższych rozkładów przedstawiono na wykresie 10.7. Naszkicowano również odpowiadające im krzywe przeżycia. Jak widać, istotne jakościowe różnice w profilach ryzyka są widoczne dla krzywych ryzyka i są trudne do identyfikacji dla krzywych przeżycia.

---

[3] W polskiej praktyce spotkać także można określenie „krzywe hazardu", które właściwie nie jest tłumaczeniem, tylko utożsamieniem jednakowo zapisanych, lecz znaczących co innego słów. Proponujemy stosowanie sformułowania „krzywe ryzyka".

W analizach często przyjmuje się, że przeżycie w dwóch grupach jest opisane przez krzywe przeżycia z tej samej rodziny, różniące się jedynie współczynnikiem proporcjonalności (np. ryzyko w jednej grupie może być dwukrotnie większe niż w drugiej). W takim wypadku ten współczynnik proporcjonalności (załóżmy, że obliczony dla analizowanego leku względem grupy przyjmującej komparator) określa się jako iloraz ryzyka (ang. *hazard ratio, HR*). Jeśli HR=1, to oczywiście ryzyko jest jednakowe w obu grupach, a krzywe przeżycia tożsame. Jeśli HR>1, to ryzyko jest wyższe w analizowanej grupie niż w grupie komparatora, a zatem krzywa przeżycia będzie malała szybciej i znajdzie się niżej. Rokowania w grupie stosującej analizowany lek są gorsze. Jeśli HR<1, to rokowania w grupie stosującej analizowany lek są lepsze.

Wykres 10.7. Krzywe ryzyka i odpowiadające im krzywe przeżycia.



Źródło: opracowanie własne

Wynikiem analiz w badaniach klinicznych jest często właśnie oszacowana wartość HR dla danej grupy względem grupy referencyjnej wraz z 95% przedziałem ufności i wartością p odpowiedniego testu (H_0 mówi o tym, że HR=1). Wartości te mogą być podane na wykresie krzywych przeżycia, np. w badaniu [5]. Podano tam wartość ilorazu ryzyka dla całkowitego przeżycia w grupie stosującej erlotynib względem grupy stosującej placebo, HR=0,82, 95%CI=(0,69; 0,99). Oznacza to, że w grupie stosującej erlotynib ryzyko jest mniejsze i stanowi ok. 82% ryzyka wyjściowego. 95% przedział ufności jest w całości poniżej 1, to znaczy, że uznamy korzyści ze stosowania erlotynibu dla przeżycia za istotne statystycznie przy poziomie istotności 5% (wartość p odpowiedniego testu wynosi dokładnie p=0,038). Należy pamiętać, że ten iloraz ryzyka nie oznacza, że prawdopodobieństwo zgonu w grupie erlotynibu stanowi 82% prawdopodobieństwa zgonu w grupie placebo! Ryzyko (ang. *hazard*) w analizie przeżycia nie jest prawdopodobieństwem. Stosunek prawdopodobieństw zależy od rozważanego horyzontu czasu – np. w bardzo długim horyzoncie iloraz prawdopodobieństw jest równy 1, gdyż prawdopodobieństwo zgonu jest dla obu grup równe 100%.

Oczywiście założenie o proporcjonalnych różnicach ryzyka między populacjami dla każdego momentu jest uproszczeniem, które jest spełnione jedynie w przybliżeniu.

Zakończmy uwagą terminologiczną, że w publikacjach zdarza się, że zamiast oznaczenia HR spotyka się RR, co może powodować skojarzenia z analizą tabel 2x2 omówioną w rozdziale 7. Należy z kontekstu odczytać, czy chodzi o porównanie częstości, czy ryzyka dla analizy przeżycia. Zdarzają się także sytuacje odwrotne, tj. wykorzystanie oznaczania HR na ryzyko względne.

10.3. Identyfikacja czynników wpływających na przeżycie

W poprzednim podrozdziale wprowadziliśmy pojęcie ilorazu ryzyka, HR. Można powiedzieć, że szacowanie tego parametru jest identyfikowaniem, czy wybór leku (o ile to stosowany lek definiował porównywane podgrupy) wpływa na przeżycie. Np. wartość HR<1 oznacza, że wpływał w sposób korzystny (dodatkowo można oceniać np. przedział ufności, aby sprawdzić, czy ten korzystny wpływ był statystycznie istotny).

W rozdziale 9 wskazano, że często istnieje potrzeba jednoczesnego uwzględnienia potencjalnego wpływu kilku czynników w ramach tzw. analizy wieloczynnikowej (ang. *multivariate analysis*). Oczywiście taka konieczność często pojawia się także w kontekście analizy przeżycia. Możemy chcieć ocenić wpływ leku na przeżycie, dopuszczając jednocześnie, że to przeżycie jest także determinowane przez płeć, wiek i stan początkowy pacjenta. Jeśli np. obawiamy się, że randomizacja mogła nie w pełni zrównoważyć porównywane ramiona ze względu na te czynniki, to warto wykonać analizę wieloczynnikową.

Najczęściej stosowanym narzędziem jest tzw. model ryzyka proporcjonalnego Coksa (ang. *Cox proportional hazard model*). W ramach tego modelu zakłada się, że krzywe ryzyka między pacjentami różnią się jedynie proporcjonalnie, przy czym ten współczynnik proporcjonalności może zależeć od wszystkich badanych czynników. Z każdym czynnikiem możemy utożsamić mnożnik, który wpływa na tę krzywą ryzyka. Wynikiem obliczeń jest oszacowanie tych współczynników (i odpowiednich błędów, zatem przedziałów ufności i wartości p) dla

wszystkich czynników. Wartości równe 1 oznaczają, że dany czynnik nie wpływa na przeżycie. Wartości <1 oznaczają, że dany czynnik redukuje ryzyko, itd.

Dla cech binarnych, np. dla płci, należy podać, która grupa jest traktowana jako referencyjna. Np. sama informacja, że HR=0,5, nie określa czy ryzyko jest zredukowane dla mężczyzn (względem kobiet), czy odwrotnie. Grupę referencyjną określa się w publikacjach na różne sposoby. Np. zapis „epoetyna beta v placebo", jak w badaniu [6], oznacza, że podana wartość HR odpowiada grupie stosującej epoetynę beta w porównaniu do placebo. Wartość HR=0,555 oznacza, że stosowanie epoetyny zmniejsza ryzyko niemal dwukrotnie, o 44,5% wyjściowego poziomu. Wartość 95% przedziału ufności (0,396; 0,776) oznacza, że redukcję tę umiejscowilibyśmy w tym przedziale i (z punktu widzenia testowania hipotez), że redukcja jest istotna statystycznie, tj. przypisywalibyśmy ją faktycznej korzyści klinicznej w populacji generalnej, a nie jedynie przypadkowości w próbie.

Czasem wartość referencyjną podaje się w nawiasie. Np. w badaniu [1] zapis „Primary tumor status (stable)" i wartość HR=1,62 oznacza, że dla pacjentów o wyjściowo niestabilnym stanie choroby nowotworowej (czyli spoza grupy referencyjnej) ryzyko jest większe o ponad połowę. Wartość 95% przedziału ufności jest równa (1,11; 2,36), co oznacza statystycznie istotny wpływ stanu guza na rokowania.

Dla cech ciągłych, np. wieku, istnieje możliwość oszacowania wpływu każdego wzrostu wartości zmiennej o 1. Często wynikiem analizy jest jednak porównanie dwóch podgrup o wartości czynnika mniejszej i większej niż ustalony próg. Np. w badaniu [1] zapis w postaci „Wiek (<65. r.ż.)" i wartość HR=1,48 oznacza, że wśród pacjentów starszych (65 lat i więcej) ryzyko rośnie w porównaniu do pacjentów młodszych niż 65 lat. Tak więc zmienną ciągłą analizowano po jej zredukowaniu do zmiennej dychotomicznej.

Czasem sposób interpretacji nie jest oczywisty tylko na podstawie tabeli. W badaniu [1] podano wartość HR=1,36 dla czynnika oznaczonego jako „Liczba przerzutów do mózgu (1)". Nie jest oczywiste, czy podana wartość ilorazu ryzyka odnosi się do każdego dodatkowego przerzutu. Analiza tekstu badania wskazuje, że tę zmienną też zdychotomizowano, dzieląc pacjentów na grupy z jednym i 2-4 przerzutami.

Należy uważać, aby nie wziąć mylnie za analizę wieloczynnikową analizy pojedynczych czynników przeprowadzonej w podziale na podgrupy. W tabeli 10.3. przedstawiono przykładowe wyniki na podstawie badania [3]. W tym przypadku czynniki (np. płeć lub wiek) i ich wartości wyznaczają podgrupy, w ramach których porównywane są leki A i B. I tak na przykład jeden z wierszy tabeli oznacza, że wśród kobiet stosowanie leku B zmniejsza ryzyko (HR=0,83), ale wynik nie jest statystycznie istotny przy poziomie istotności 5% (95% przedział ufności zawiera neutralną wartość 1). Przy okazji interpretacji wyników tego typu analiz warto, aby Czytelnik uwzględnił kwestie testowania wielu hipotez omówione w rozdziale 6. Analizując wiele podgrup rośnie prawdopodobieństwo znalezienia choć jednej takiej, w której analizowany lek generuje istotne statystycznie korzyści kliniczne, nawet jeśli w rzeczywistości w żadnej z podgrup leki się nie różnią.

Tabela 10.3. Przykładowe dane do analizy przeżycia zebrane 7 miesięcy po rozpoczęciu badania.

Czynnik	L. pacjentów	Lek A		Lek B		HR (95%CI)
		L. pacjentów	Mediana przeżycia (miesiące)	L. pacjentów	Mediana przeżycia (miesiące)	
wszyscy	209	105	12,94	104	16,56	0,8 (0,58; 1,11)
płeć:						
kobiety	97	51	12,45	46	16,56	0,83 (0,51; 1,34)
mężczyźni	112	54	13,5	58	17,58	0,76 (0,48; 1,19)
wiek:						
<65	41	25	17,51	16	18,43	0,56 (0,23; 1,38)
≥65	168	80	10,35	88	15,26	0,81 (0,57; 1,17)

Źródło: opracowanie własne na podstawie [3]

Bibliografia

1. Aoyama, H.; Shirato, H.; Tago, M.; Nakagawa, K.; Toyoda, T.; Hatano, K.; Kenjyo, M.; Oya, N.; Hirota, S.; Shioura, H.; Kunieda, E.; Inomata, T.; Hayakawa, K.; Katoh, N.; Kobashi, G.: Stereotactic Radiosurgery Plus Whole-Brain Radiation Therapy vs Stereotactic Radiosurgery Alone for Treatment of Brain Metastases. A Randomized Controlled Trial. Journal of American Medical Association, 2006, 295 (21), 2483-2491.
2. Hurwitz et al. Bevacizumab plus Irinotecan, Fluorouracil, and Leucovorin for Metastatic Colorectal Cancer, NEJM, 2004, 350 (23), 2335-2342.
3. Kabbinavar, F.F.; Schulz, J.; McCleod, M.; Patel, T.; Hamm, J.T.; Hecht, R.; Mass, R.; Perrou, B.; Nelson, B.; Novotny, W.F.: Addition of Bevacizumab to Bolus Fluorouracil and Leucovorin in First-Line Metastatic Colorectal Cancer: Results of a Randomized Phase II Trial. Journal of Clinical Oncology, 2005, 23 (16), 3697-3705.
4. Machin, D.; Cheung, Y.B.; Parmar, M.K.B.: Survival Analysis. A Practical Approach. Second Edition. John Wiley & Sons Ltd, 2006.
5. Moore, M.J.; Goldstein, D.; Hamm, J.; Figer, A.; Hecht, J.R.; Gallinger, S.; Au, H.J.; Murawa, P.; Walde, D.; Wolff, R.A.; Campos, D.; Lim, R.; Ding, K.; Clark, G.; Voskoglou-Nomikos, T.; Ptasynski, M.; Parulekar, W.: Erlotinib Plus Gemcitabine Compared With Gemcitabine Alone in Patients With Advanced Pancreatic Cancer: A Phase III Trial of the National Cancer Institute of Canada Clinical Trials Group. Journal of Clinical Oncology, 2007, 25 (15), 1960-1966.
6. Österborg, A.; Brandberg, Y.; Molostova, V.; Iosava, G.; Abdulkadyrov, K.; Hedenus, M.; Messinger, D. for the Epoetin Beta Hematology Study Group: Randomized, Double-Blind, Placebo-Controlled Trial of Recombinant Human Erythropoietin, Epoetin Beta, in Hematologic Malignancies. Journal of Clinical Oncology, 2002, 20 (10), 2486-2494.
7. Shepherd, F.A.; Pereira, J.R.; Ciuleanu, T.; Tan, E.H.; Hirsh, V.; Thongprasert, S.; Campos, D.; Maoleekoonpiroj, S.; Smylie, M.; Martins, R.; van Kooten, M.; Dediu, M.; Findlay, B.; Tu, D.; Johnston, D.; Bezjak, A.; Clark, G.; Santabárbara, P.; Seymour, L.; for the National Cancer Institute of Canada Clinical Trials Group: Erlotinib in Previously Treated Non-Small-Cell Lung Cancer. The New England Journal of Medicine, 2005, 353 (2), 123-132.
8. Smith, N.; Norman, A.; Swift, I.; Brown, G.: MRI detects extra-mural vascular invasion and predicts outcome in colorectal cancer. Annals of Oncology, 2006, 17 (Supp. 6), vi19-vi27.

Projekt współfinansowany przez Unię Europejską z Europejskiego Funduszu Społecznego
„Kształcenie w ramach procesu specjalizacji lekarzy deficytowych specjalności
tj. onkologów, kardiologów i lekarzy medycyny pracy”

ISBN 978-83-62110-29-2

EGZEMPLARZ BEZPŁATNY

KAPITAŁ LUDZKI
CZŁOWIEK – NAJLEPSZA INWESTYCJA!