VIA MEDICA

PRACA PRZEGLĄDOWA

Marcin Kaczor[1, 2], Rafał Wójcik[2], Joanna Połowinczak-Przybyłek[3], Piotr Potemski[3]

[1]Uniwersytet Jagielloński Collegium Medicum w Krakowie
[2]Aestimo s.c., Kraków
[3]Klinika Chemioterapii Nowotworów Uniwersytetu Medycznego w Łodzi; WWCOiT im. M. Kopernika w Łodzi

Krytyczna ocena badań klinicznych w onkologii — część I

Critical appraisal of clinical trials in oncology — part I

Artykuł jest tłumaczeniem pracy:
Kaczor M, Wójcik R, Połowinczak-Przybyłek J, Potemski P. Critical appraisal of clinical trials in oncology — part I. Oncol Clin Pract 2019; 15. DOI: 10.5603/OCP.2018.0057.
Należy cytować wersję pierwotną.

Adres do korespondencji:
Dr n. med. Marcin Kaczor
II Katedra Chorób Wewnętrznych
im. prof. Andrzeja Szczeklika
Uniwersytet Jagielloński
Collegium Medicum w Krakowie
e-mail: marcin.kaczor@uj.edu.pl

STRESZCZENIE

Zgodnie z zasadami medycyny opartej na dowodach (EBM, *evidence based medicine*) decyzje terapeutyczne powinny być podejmowane na podstawie analizy wyników wiarygodnych badań klinicznych. Dostępność publikacji prezentującej w poprawny sposób metodykę próby klinicznej i jej wyniki o wysokiej jakości wydaje się być sytuacją idealną z punktu widzenia praktykującego lekarza. Jednak czasami przeczytanie samego streszczenia, a zwłaszcza jedynie kilkuzdaniowych wniosków z badania klinicznego może okazać się niewystarczające. Ostatecznie, aby podjąć w pełni świadomą i racjonalną decyzję terapeutyczną, należy szczegółowo ocenić jakość i metodykę badania klinicznego, wiarygodność zewnętrzną i wewnętrzną, istotność statystyczną i kliniczną wyników oraz wewnętrzną i zewnętrzną spójność prezentowanych wniosków. Takie umiejętności oceny i interpretacji wyników badań klinicznych mogą być przydatne w onkologii, zwłaszcza w przypadku innowacyjnych leków lub leczenia pacjentów w szczególnym stanie klinicznym i konieczności rozważania dostępu do leczenia ratunkowego. Warto też podkreślić specyfikę oceny do leczenia punktów końcowych w badaniach onkologicznych, w których wyjątkową rolę odgrywa dość trudna w interpretacji analiza przeżycia. W artykule przedstawione zostały w sposób przystępny podstawy teoretyczne oraz kolejne kroki krytycznej oceny metodyki i wyników badań klinicznych prowadzonych w obszarze onkologii.

Słowa kluczowe: onkologia, badania kliniczne z randomizacją, krytyczna ocena, analiza statystyczna, analiza przeżycia

ABSTRACT

The main concept of the Evidence Based Medicine is that all therapeutic decisions should be always based on results from relevant, credible and up-to-date clinical trials. Availability of a publication presenting description of a clinical trial conducted with reliable methods and its high-quality results seems to be an ideal situation from the practitioners point of view. However, reading only the abstract or just the authors conclusions may not always be sufficient to make the right clinical decision. For this purpose, several aspects of the clinical trial should be put under assessment, namely the methodology, its quality, internal and external credibility, clinical and statistical significance, as well as consistency of the results. Ability of performing the proper assessment of clinical trials may prove to be very helpful for practicing oncologists, especially in case of new, emerging therapies, specific clinical situations or when salvage treatment is necessary. It is also worth emphasizing that the outcomes assessment in oncology trials is specific, mainly due to the role of the survival analysis which is relatively difficult for interpretation. In this paper we tried to present in a clear and intelligible way the theoretical basis and subsequent steps in the critical appraisal of methods and results of clinical trials in oncology.

Key words: oncology, randomized clinical trial, critical appraisal, statistical analysis, survival analysis

ISSN 2450–1646

Oncol Prakt Klin Edu 2019; 5: 116–131

Wstęp

Podstawą podejmowania decyzji o wyborze leczenia są poprawnie przeprowadzone i wiarygodne badania kliniczne. Na bazie ich wyników, zgodnie z zasadami EBM, tworzy się obowiązujące wytyczne praktyki klinicznej. Aby ocenić, czy wnioski z badania są poprawne, przede wszystkim należy dokonać jego krytycznej analizy pod kątem **wiarygodności wewnętrznej**. W tym celu trzeba ocenić, czy badanie zostało poprawnie przeprowadzone (odpowiednia metodyka badania zapewniająca wiarygodne i niezaburzone wnioskowanie oraz prawidłowa analiza statystyczna), a także czy występuje **spójność wewnętrzna** wniosków w zakresie poszczególnych punktów końcowych. Tutaj pomocna może być również ocena **spójności zewnętrznej** — stwierdzenie, czy podobny efekt został zaobserwowany w innych próbach klinicznych. Następnie ocenie należy poddać **wiarygodność zewnętrzną**, odpowiadając na pytanie, czy ocenione jako wiarygodne wewnętrznie wyniki próby klinicznej mogą być ekstrapolowane na populację, która będzie poddana leczeniu w warunkach rzeczywistej praktyki klinicznej, i czy będzie można oczekiwać podobnych efektów klinicznych (charakterystyka pacjentów, medyczne postępowanie towarzyszące, odpowiedni komparator, stosowanie się uczestników do zaleceń — *compliance*). W końcowym etapie należy jeszcze ocenić **istotność kliniczną** wyników, tzn. odpowiedzieć na pytanie, czy wielkość obserwowanego efektu wskazuje na zauważalną korzyść kliniczną (przy uwzględnieniu rokowania w danej grupie pacjentów) i czy rzeczywiście powinna nastąpić zmiana praktyki klinicznej [1].

Poniżej omówiono kolejno poszczególne podstawowe elementy krytycznej oceny badania klinicznego. Dodatkowo, biorąc pod uwagę specyfikę prób klinicznych w onkologii, bardziej szczegółowo zaprezentowano analizę punktów końcowych typu „czas do zdarzenia".

Metodyka badania klinicznego

Poprawnie zaprojektowane i przeprowadzone badania kliniczne z randomizacją (RCT, *randomized clinical trial*) i zaślepieniem (*blinding*) dostarczają dowodów cechujących się najwyższym poziomem wiarygodności [2]. Są to próby eksperymentalne, w których ocenia się przynajmniej dwie interwencje terapeutyczne, a ich stosowanie u chorych podlega ścisłej kontroli, zgodnie z opracowanym wcześniej protokołem badania. W onkologii badania te zazwyczaj mają charakter prób z grupami równoległymi (*parallel*). Zdarza się czasem, że w niektórych populacjach pacjentów onkologicznych trudno jest przeprowadzić badanie z randomizacją, co może być spowodowane niskim rozpowszechnieniem niektórych nowotworów bądź też niewielką liczbą chorych w określonych stadiach zaawansowania lub liniach leczenia. Wówczas koniecznością staje się prowadzenie badań bez grupy kontrolnej (*single-arm*). Ich jakość metodologiczna jest jednak wyjściowo niższa w odniesieniu do prób z randomizacją. Podobnie jest w przypadku badań kohortowych (*cohort study*), w których mamy wprawdzie grupę kontrolną, ale ze względu na brak randomizacji obciążenie stanowi nielosowy rozkład czynników zakłócających i wnioskowanie na temat obserwowanych różnic w skuteczności terapii jest ograniczone [1].

Randomizacja

Randomizację, czyli losowy przydział chorych do grup, stosuje się po to, aby wyjściowe charakterystyki kliniczne i demograficzne pacjentów były możliwie zbliżone lub niemal identyczne, dzięki czemu, przy odpowiednio licznej populacji, zapewniony zostaje równomierny rozkład wszystkich potencjalnych, również nieznanych, czynników zakłócających. Procedura randomizacji nie może więc być prowadzona na podstawie prostych założeń, jak np. numer historii choroby czy data urodzenia, gdyż pozwala to w dalszym ciągu przewidzieć, do jakiej grupy trafi dany chory (takie działania nazywa się pseudorandomizacją). Do metod randomizacji zapewniających całkowitą losowość, a więc nieprzewidywalność, należą te, w których komputerowo lub przy wykorzystaniu specjalnych tabel tworzy się listy liczb losowych (taki sposób nazywa się randomizacją prostą). W przypadku małej docelowej liczby chorych w badaniu większe jest z kolei prawdopodobieństwo nierównego rozłożenia liczby i cech pacjentów w poszczególnych grupach — w takim wypadku można zastosować bardziej złożone metody randomizacji. Ich przykładami są: randomizacja blokowa (chorzy są rozdzielani do poszczególnych interwencji w blokach, czyli grupach o określonej sekwencji przydziału kolejnych pacjentów), randomizacja warstwowa — stratyfikacja (odbywają się niezależnie w każdej zdefiniowanej wcześniej warstwie, jak np. kraj pochodzenia, płeć czy rodzaj wcześniej stosowanego leczenia, zwłaszcza gdy w podgrupach tych spodziewane są różnice w zakresie skuteczności ocenianej interwencji) lub randomizacja adaptacyjna (w której prawdopodobieństwo przydziału do danej grupy zmienia się w trakcie trwania badania, tak by możliwa była kontrola rozkładu poszczególnych cech w wyodrębnionych grupach) [3]. W części badań obserwuje się nierównomierny **rozkład do grup badanych**, np. w stosunku 2:1, co może zwiększać liczbę informacji na temat nowej terapii, szczególnie w kontekście bezpieczeństwa, jak również możliwości rekrutacyjne (pacjenci wykazują większą chęć uczestnictwa w badaniu w związku z większą szansą otrzymania terapii eksperymentalnej), ale wpływa na spadek mocy statystycznej i wymaga zwiększenia łącznej liczebności próby w porównaniu z zastosowaniem przydziału 1:1 [3].

Ukrycie kodu alokacji i zaślepienie

Przy losowym doborze chorych do grup ważny jest proces jego utajenia (*allocation concealment*), a więc uniemożliwienia dostępu do informacji o grupie, do której został przydzielony dany pacjent — co jest możliwe przy zastosowaniu centralnej randomizacji, wykonanej niezależnie od poszczególnych ośrodków badawczych biorących udział w próbie. Zastosowanie ukrytej alokacji pozwala wyeliminować wpływ badacza na przypisanie pacjentów do grup, redukując tym samym ryzyko błędu systematycznego doboru próby (*selection bias*). Drugim krokiem zapewniającym większą wiarygodność jest wprowadzenie **zaślepienia**, a więc działań, które powodują, że pacjent (zaślepienie pojedyncze) albo pacjent i badacz (zaślepienie podwójne; *double blinded*) bądź pacjent, badacz oraz zespół analizujący wyniki (zaślepienie potrójne) nie mają świadomości, którą interwencję otrzymuje poszczególny chory. Zapewnia to większą wiarygodność badania, gdyż może eliminować niektóre czynniki zakłócające — terminalnie chory, który wie, że został przypisany do grupy placebo zamiast do aktywnej interwencji, może prezentować znacznie gorsze wyniki niż pacjent tego nieświadomy [4, 5]. W przypadku leków zaślepienie zapewnia się poprzez ich przygotowanie w identycznej postaci (np. w tak samo wyglądających fiolkach), a w przypadku różnych dróg podania lub zestawienia różnych metod leczenia istotną rolę odgrywa dodatkowo właściwe maskowanie (*dummy*) interwencji, czyli np. jednoczesne podawanie dwóch interwencji różniących się drogą podania, z tym że w każdym z ramion badania inna spośród nich zostaje zastąpiona placebo. W niektórych przypadkach, np. zróżnicowanych procedur medycznych, trudno jest zapewnić zaślepienie, bądź wiąże się to z dużym obciążeniem chorych. Należy pamiętać, że brak zaślepienia wpływa istotnie przede wszystkim na ocenę subiektywnych punktów końcowych, samodzielnie ocenianych przez pacjentów (PRO, *patient-reported outcome*; np. ocena nasilenia objawów, jakość życia) czy też ocenę bezpieczeństwa, nie zaburza zaś jednoznacznie obiektywnych punktów końcowych, takich jak zgon (a co za tym idzie — analizy przeżycia całkowitego) [3]. W przypadku oceny punktów końcowych przy użyciu badań patologicznych lub obrazowych, albo za pomocą wystandaryzowanych kryteriów (np. RECIST, *Response Evaluation Criteria in Solid Tumors*) ryzyko błędu systematycznego jest niejednoznaczne. Natomiast w badaniach onkologicznych często spotykamy się z sytuacją, że pomimo braku zaślepienia badania ocena obrazowa progresji (odpowiedzi na leczenie) jest potwierdzana centralnie przez niezależną i zaślepioną komisję. Istnieje jednak ryzyko, zwłaszcza w badaniach z placebo w grupie kontrolnej, dopuszczających zmianę ramienia po stwierdzeniu progresji, że brak zaślepienia spowoduje w przypadku wystąpienia nawet niewielkich objawów u chorych otrzymujących placebo szybsze niż planowane wykonanie badań obrazowych. Zdarzają się również próby kliniczne, w których nie jest wymagane zaślepienie wszystkich badaczy czy pacjentów, a analityków oceniających punkty końcowe — badania te określa się mianem PROBE (*prospective, randomized, open, blinded-endpoint evaluation*).

Ocena jakości badania

Najprostsza **ocena wiarygodności** badania może się odbyć za pomocą 5-punktowej skali Jadada [6]. W jej ramach ocenia się, czy badanie opisano jako randomizowane, czy zastosowano podwójne zaślepienie oraz podano informacje o tym, ilu pacjentów zakończyło udział w badaniu i z jakich powodów. Dodatkowe punkty można przyznać lub odjąć w zależności od tego, czy randomizację i zaślepienie przeprowadzono poprawnie bądź niepoprawnie. Skala ta pozwala jednak tylko na bardzo ogólną ocenę jakości badania i nie uwzględnia wielu innych czynników mogących prowadzić do błędu systematycznego (*bias*). Bardziej kompleksowym narzędziem są zalecenia opracowane przez *Cochrane Collaboration* [7], zgodnie z którymi ocenia się:

— dobór próby (*selection bias*) — czy zastosowano prawidłową metodę randomizacji oraz jej utajenia;
— zaślepienie pacjentów i personelu medycznego (*performance bias*);
— zaślepienie oceny wyników (*detection bias*) — czy wprowadzono zaślepienie badacza dokonującego oceny (*assessor*) lub czy autorzy publikacji uzasadnili, że brak takiego zaślepienia nie ma wpływu na ocenę danego punktu końcowego; w przypadku oceny punktów końcowych o różnej podatności na ryzyko zafałszowania wynikające z braku zaślepienia konieczne jest przeprowadzenie oceny dla każdego z nich osobno;
— niekompletność wyników i utrata pacjentów z badania (*attrition bias*) — niskie ryzyko błędu stwierdza się, jeśli dane utracone nie zaburzają oceny punktów końcowych, badacze zastosowali właściwą metodę imputacji brakujących danych [np. LOCF (*last observation carried forward*), w której w przypadku utraty pacjenta z obserwacji wartości poszczególnych punktów końcowych ocenione podczas ostatniej wizyty kontrolnej z jego udziałem są przepisywane dla każdego kolejnego punktu czasowego po jego utracie aż do zakończenia próby], a odsetek chorych wykluczonych z próby nie różni się między wyodrębnionymi grupami; w praktyce przyjmuje się, że jeśli z badania utracono więcej niż 10% pacjentów, to ryzyko błędu systematycznego wynikającego z niekompletności wyników jest wysokie, chyba że częstość poszczególnych przyczyn wykluczeń chorych jest podobna, a odsetek pacjentów utraconych z obserwacji (*lost to follow-up*) niewielki;

— selektywną prezentację wyników (*reporting bias*) — czy dostępny jest protokół badania i czy w publikacji przedstawiono wyniki dla wszystkich zdefiniowanych w nim punktów końcowych;
— inne czynniki (*other bias*) — czy nie stwierdzono żadnych innych potencjalnych źródeł obniżenia wiarygodności prezentowanych wyników (jak np. nieprawidłowy projekt badania, czy też zarzut nieuczciwości).

Warto zaznaczyć, że obecnie preferowane są badania wieloośrodkowe (*multicenter trial*) z odpowiednią reprezentacją geograficzną [8], choć wiążą się one z ryzykiem obniżenia standaryzacji zarówno stosowanych interwencji, jak i ocenianych wyników [1].

Określenie badanej populacji

Populacja docelowa badania powinna być szczegółowo zdefiniowana kryteriami włączenia. Ich analiza służy przeprowadzeniu oceny wiarygodności zewnętrznej, czyli określeniu charakterystyki pacjentów, na którą mogą być uogólniane wnioski z badania. Zbyt wąskie i szczegółowe kryteria włączenia mogą ograniczać możliwości rekrutacji chorych do badania i możliwość uogólniania wniosków, natomiast zbyt ogólne — powodować rozproszenie ocenianego efektu w podgrupach o zróżnicowanych charakterystykach, utrudniać losowy rozkład czynników zakłócających oraz uniemożliwiać analizę w obrębie podgrup [3].

Określenie komparatora

Kolejnym kluczowym elementem jest zastosowanie w badaniu właściwego **komparatora (grupy kontrolnej)**, od czego zależą też możliwość dalszej ekstrapolacji jego wyników na populację docelową i jego wiarygodność zewnętrzna. Optymalnym i pożądanym komparatorem jest aktualna praktyka kliniczna, zgodna z powszechnie przyjętymi zaleceniami i wytycznymi [9]. Często spotyka się jednak zastosowanie w grupie kontrolnej placebo. Jest to uzasadnione, kiedy nowa terapia jest leczeniem dodanym (*add-on*) do obecnego standardu (wówczas placebo stosowane jest tylko w celu zaślepienia, a komparator stanowi *de facto* obecna praktyka) lub gdy w warunkach rzeczywistej praktyki nie występuje inna opcja terapeutyczna poza leczeniem objawowym, np. gdy oceniana interwencja stanowi ostatnią linię leczenia. Porównanie z placebo zazwyczaj ma na celu wykazanie wyższości nowego sposobu leczenia (*superiority*). Wybór aktywnej interwencji jako komparatora zawsze niesie ze sobą dodatkowe wyzwania, także w kontekście wielkości próby badanej, często jednak zastosowanie placebo byłoby po prostu nieetyczne. W przypadku porównania z aktywnym postępowaniem można rozważyć testowanie hipotezy *non-inferiority* [3]. Zasadność doboru aktywnej interwencji jako komparatora należy oceniać także w kontekście zmieniających się zaleceń klinicznych, zwłaszcza w przypadku prób klinicznych planowanych kilka lat wcześniej. W onkologii, wobec zróżnicowanych schematów chemioterapii i sposobów postępowania, jako komparator przyjmuje się często terapię z wyboru lekarza. W takiej sytuacji należy ocenić, jakie interwencje i w jakich udziałach zastosowano w grupie komparatora, zwłaszcza w przypadku gdy możliwy jest wybór leczenia objawowego jako opcji postępowania, oraz czy odzwierciedlają one praktykę kliniczną i możliwość ekstrapolacji wniosków (wiarygodność zewnętrzna).

Określenie punktów końcowych

Punkty końcowe (*endpoints*, *outcomes*) powinny być dokładnie zdefiniowane w protokole badania, z określeniem punktów głównych/pierwszorzędowych (dla których szacuje się liczebność próby i moc statystyczną; *primary endpoints*) oraz dodatkowych/drugorzędowych (*secondary endpoints*). Pożądana jest ocena istotnych klinicznie punktów końcowych, takich jak czas przeżycia całkowitego (OS, *overall survival*) oraz jakość życia (ocena całościowa oraz ukierunkowana na ocenę objawów związanych z danym typem nowotworu). Możliwości oceny wpływu interwencji na przeżycie całkowite będą zależały od rodzaju nowotworu i stadium jego zaawansowania. Analiza taka na pewno będzie utrudniona w przypadku oceny wczesnych stadiów na etapie terapii z intencją wyleczenia (np. leczenie neo- i adiuwantowe), gdy oczekiwane dalsze przeżycie może wynosić dziesiątki lat i dodatkowo wpływ na obserwowane różnice w przeżyciu będą miały liczne kolejne linie leczenia po wystąpieniu późniejszych nawrotów lub progresji. W takich wypadkach zastępczymi punktami końcowymi mogą być czas przeżycia wolnego od choroby/nawrotu/wznowy (DFS — *disease-free survival*, EFS — *event-free survival*, RFS — *relapse-free survival*, czyli czas liczony od daty włączenia do badania do daty wystąpienia pierwszego udokumentowanego zdarzenia klinicznego lub zgonu, cokolwiek wystąpi wcześniej) w przypadku terapii stosowanych we wczesnych stadiach nowotworów, lub czas przeżycia wolnego od progresji (PFS, *progression-free survival*, czyli czas od daty randomizacji do daty wystąpienia progresji lub zgonu) w stadiach zaawansowanych. Zdarzenia kliniczne (*events*) określone w definicji PFS/DFS z reguły są obserwowane wcześniej niż zgon, zatem okres obserwacji konieczny do wykazania istotnej statystycznie różnicy między porównywanymi interwencjami jest krótszy niż w przypadku OS. Stąd ocena PFS jest preferowana np. w sytuacji dużej, niezaspokojonej potrzeby klinicznej (*unmet need*), czyli braku innego skutecznego leczenia, gdyż rejestracja leku w danym wskazaniu może zostać uzyskana znacznie szybciej (nawet o kilka lat), niż gdyby konieczne było oczekiwanie na wyniki OS. Do-

datkowo, obserwowane różnice w PFS nie są zaburzane przez kolejne linie leczenia i ewentualną zmianę ramienia badania (*cross-over*), ponieważ wprowadzenie nowego leczenia standardowo nie następuje przed wystąpieniem progresji choroby. Czas przeżycia wolnego od progresji jest oceniany w większości badań dotyczących leczenia w zaawansowanych stadiach nowotworów, niemniej traktuje się go jako surogatowy punkt końcowy. W literaturze można znaleźć wiele opracowań oceniających korelację PFS i OS w kontekście użyteczności PFS jako predyktora OS, choć jak dotąd wnioski prezentowane przez licznych autorów nie są jednoznaczne [10]. Do dodatkowych punktów końcowych można zaliczyć również odpowiedź obiektywną na podstawie badań obrazowych (ORR, *overall response rate*) w przypadku guzów narządowych (litych) lub remisję hematologiczną (*hematologic remission*) w hematoonkologii oraz czas utrzymywania się tych stanów (DoR, *duration of response*). Alternatywę względem przeżycia wolnego od progresji (PFS) o mniejszej wartości klinicznej stanowi czas do progresji choroby (TTP, *time to progression*), różniący się od PFS tym, że uwzględnia jako zdarzenie wyłącznie przypadki progresji choroby, natomiast obserwacje chorych, którzy zmarli przed jej wystąpieniem, są ucinane w momencie zgonu (obserwacje cenzorowane). Pokrewnymi punktami końcowymi, choć znacznie rzadziej stosowanymi w ocenie skuteczności leczenia paliatywnego, są także: czas do niepowodzenia leczenia (TTF, *time to treatment failure*) oraz czas do rozpoczęcia kolejnej linii leczenia (TTNT, *time to next treatment*).

Mając na uwadze różnorodność ocenianych punktów końcowych, należy poszukiwać spójności wewnętrznej prezentowanych wyników, tzn. starać się wykazać znamienny wpływ ocenianej interwencji na ORR, PFS, a następnie na OS. Trzeba jednak zawsze pamiętać o zróżnicowaniu ocenianych populacji pod względem rodzaju nowotworu, stadium zaawansowania, rokowania i czasu oczekiwanego dalszego przeżycia, a nawet rodzaju zastosowanej interwencji, np. wykazanie wpływu immunoterapii na OS, przy braku wpływu na PFS m.in. w związku ze zjawiskiem pseudoprogresji [11]. W przypadku badań w onkologii szczególną uwagę powinno się poświęcić ocenie bezpieczeństwa, m.in. działań niepożądanych w 3. i 4. stopniu nasilenia lub prowadzących do zgonu, w tym specyficznych toksyczności charakterystycznych dla danej interwencji. Ostatecznie należy ocenić stosunek korzyści do ryzyka, biorąc pod uwagę rokowanie w konkretnej populacji chorych [12, 13].

Informacje o zaplanowanej analizie statystycznej

Zakres i rodzaj **analizy statystycznej** w poprawnie przeprowadzonym badaniu klinicznym powinno się predefiniować w ramach wcześniej zarejestrowanego protokołu, w tym również powinno się predefiniować czynniki dopasowania i analizę w podgrupach.

Wyjściowe oszacowanie **wielkości badanej próby** (mocy statystycznej badania) jest jednym z kluczowych elementów oceny statystycznej. Pozwala ocenić, czy próba jest wystarczająco liczna, by potwierdzić lub wykluczyć różnice między interwencjami. Ocena wielkości próby dotyczy głównego (pierwszorzędowego) punktu końcowego (lub punktów końcowych). Wymaga ona określenia oczekiwanej częstości zdarzeń w grupie kontrolnej, wielkości efektu interwencji, który badanie ma wykryć (hipoteza alternatywna), założenia zdolności do wykrywania prawdziwego efektu (moc statystyczna testu) i wyboru poziomu istotności statystycznej. W badaniach onkologicznych przy długim założonym okresie obserwacji należy uwzględnić także oczekiwany stopień utraty pacjentów z badania (*discontinuation rate*). Ponieważ moc statystyczna zależy od liczby chorych doświadczających danego zdarzenia w trakcie obserwacji, w przypadku badań w onkologii często zakłada się obserwację pacjentów do wystąpienia założonej liczby zdarzeń (np. zgonów lub zgonów i przypadków progresji w ocenie OS lub PFS) [3].

W ocenie statystycznej badacze mogą przyjąć różne podejścia analityczne — najczęściej testuje się hipotezę o wyższości ocenianej interwencji (*superiority*), szczególnie we wczesnych fazach badań klinicznych i wówczas, gdy grupę kontrolną stanowi grupa przyjmująca placebo. Drugie podejście polega na ocenie, czy dana interwencja nie jest gorsza w zakresie skuteczności klinicznej od obecnie stosowanej (*non-inferiority*), zwłaszcza przy lepszym profilu bezpieczeństwa. W tym wypadku istnieje konieczność założenia akceptowalnej klinicznie zmienności skuteczności w zakresie głównego punktu końcowego i jeśli odpowiednia granica przedziału ufności (CI, *confidence interval*) dla różnicy między interwencjami nie przekracza ustalonego poziomu, interwencję uznaje się za nie gorszą od kontroli. Zastosowanie podejścia *non-inferiority* pozwala zmniejszyć wymaganą wielkość próby [14, 15]. Spotykane są również badania oceniające równoważność (*equivalence*) interwencji, gdzie zakłada się dopuszczalną zmienność w obu kierunkach, ale rzadko są one stosowane do oceny klinicznych punktów końcowych, raczej parametrów laboratoryjnych lub farmakokinetyki. Znaczenie ma też **populacja, jaką badacze uwzględniają w przeprowadzanych analizach** — może ona być różna w zależności od ocenianego punktu końcowego. Populacja ITT (*intention-to-treat*) uwzględnia w analizie wyników wszystkich chorych poddanych randomizacji, bez względu na to, czy otrzymali przypisaną interwencję i jak długo pozostawali w obserwacji (zazwyczaj dotyczy to oceny OS lub PFS). Niekiedy uwzględnia się też zmodyfikowaną populację ITT (mITT), czyli chorych poddanych randomizacji, którzy otrzymali co najmniej jedną dawkę leku — w niej zazwyczaj przeprowadzana jest analiza bezpieczeństwa. Populacja PP (*per-protocol*) oznacza pacjentów, którzy dodatkowo nie przerwali leczenia, nie złamali protokołu

i jest dla nich dostępny komplet informacji — często stosuje się ją do porównywania skuteczności interwencji w próbach typu *non-inferiority* [14]. Analiza ITT ma charakter bardziej konserwatywny, gdyż wykazuje tendencję do zaniżania korzystnego efektu klinicznego, podczas gdy analiza PP pozwala na porównanie opcji terapeutycznych w warunkach przeprowadzenia pełnej obserwacji. Jeżeli wyniki uzyskane w analizach ITT i PP wyraźnie się różnią, może to świadczyć o obniżonej wiarygodności badania. Ocenę obiektywnej odpowiedzi na leczenie często przeprowadza się w populacji chorych, dla których dodatkowo dostępne są wyniki badań obrazowych, czyli istnieje możliwość oceny progresji.

Ocena wyników

Analizę wyników badania klinicznego rozpoczynamy od szczegółowej oceny opisu włączonej populacji oraz tabel z **charakterystykami wyjściowymi**, które powinny obejmować podstawowe dane demograficzne, zaawansowanie choroby, przebieg wcześniejszego leczenia, a także inne czynniki mogące mieć wpływ na skuteczność ocenianej terapii — ich zakres i rodzaj są zależne od rodzaju nowotworu i powinny być dostosowane również do stopnia zaawansowania choroby. Analiza tych parametrów może służyć ocenie poprawności randomizacji i zniesienia wpływu czynników zakłócających (oczywiście analizując taką tabelę, możemy się odnieść tylko do tych znanych, jednak równocześnie zakładamy, że sam dobór losowy do grup, przy ich odpowiedniej liczebności, zapewnia nam także równomierny rozkład pozostałych, nieznanych czynników prognostycznych). Powinny być tu też wyróżnione podgrupy określone przy randomizacji ze stratyfikacją oraz podgrupy, w obrębie których zostaną wykonane predefiniowane analizy czy ewentualne wcześniej niezaplanowane analizy *post-hoc*. Te informacje są również pomocne do rozstrzygnięcia o wiarygodności zewnętrznej wyników próby klinicznej. Dzięki nim możemy ocenić, czy analizowana populacja jest zbliżona do tej, w której będziemy chcieli stosować daną interwencję [8].

Wyniki oceny w badaniu mające postać zmiennych kategorycznych (nominalne) zazwyczaj są przedstawiane jako liczebności i odsetki, natomiast zmienne ciągłe — za pomocą miary centralnej i miary rozrzutu — zazwyczaj wartości średniej i odchylenia standardowego (SD, *standard deviation*), a w przypadku zmiennych nieprzyjmujących rozkładu normalnego — mediany i zakresu, ewentualnie rozstępu międzykwartylowego (IQR, *interquartile range*) (por. poniżej). Ponadto niektóre zmienne ciągłe mogą być przekształcane w zmienne porządkowe (np. odsetek chorych powyżej danego wieku). Nawiązując do wspomnianej wcześniej oceny poprawności randomizacji, sprawdzamy, czy nie występują znamienne różnice w charakterystykach wyjściowych pomiędzy wyróżnionymi grupami — autorzy powinni podać wartości p w tabeli lub zadeklarować brak istotnych różnic w tekście publikacji [8].

W badaniu powinny być również szczegółowo przedstawione (zazwyczaj na odpowiednim diagramie) informacje na temat **przepływu pacjentów** od okresu skriningu do badania (czyli od wyrażenia zgody na udział w badaniu aż do włączenia do niego) po ewentualny dodatkowy okres obserwacji po zakończeniu badania. Jak wspomniano, jest to istotny element oceny wiarygodności badania — należy ocenić wielkość utraty chorych z obserwacji, to, jak może to wpływać na wiarygodną analizę wyników oraz występowanie różnic między grupami.

Kolejny etap obejmuje ocenę ilościową różnic między interwencjami, wyrażenie niepewności tych oszacowań za pomocą przedziałów ufności oraz ocenę siły dowodów, czyli potwierdzenie za pomocą wartości p (test istotności statystycznej), że obserwowana różnica jest prawdziwa, a nie jest dziełem przypadku [8, 16–18].

Ponieważ z oczywistych przyczyn nie można przebadać wszystkich chorych w rozważanym stanie klinicznym, należy wybrać pewną próbkę (grupę objętą badaniem klinicznym) i na podstawie obserwowanych w niej efektów wnioskować z pewnym przybliżeniem o rzeczywistej, ogólnej efektywności leczenia. W języku statystyki mówimy, że na podstawie wybranej miary efektu określanej w losowej próbie z tej populacji szacujemy (estymujemy) pewien parametr w populacji ogólnej, która w tym ujęciu jest jakby zbiorem wszystkich możliwych wyników naszego eksperymentu.

Na początek trzeba precyzyjnie sformułować **hipotezę badawczą**, testowaną następnie metodami statystycznymi w celu jej przyjęcia lub odrzucenia. Standardowo zakłada się, że oceniane interwencje w podobny sposób wpływają na efekt zdrowotny (tzw. hipoteza zerowa, *null hypothesis*), a obserwowane różnice są wyrazem zmienności losowej i wynikają jedynie z ograniczeń eksperymentu (np. zbyt mała grupa chorych). Hipotezą alternatywną jest stwierdzenie, że obserwowane różnice są prawdziwe i nie są jedynie losową obserwacją. Zadanie statystyki polega na wskazaniu, która z tych hipotez jest bardziej prawdopodobna.

Kiedy jednak będziemy wiedzieć, że nasze grupy nie różnią się od siebie? Intuicyjnie możemy stwierdzić, że o braku różnic między tymi grupami na pewno będzie świadczyć uzyskanie takiego samego odsetka chorych z odpowiedzią w grupie interwencji co w grupie kontrolnej. Jeśli jednak jakieś różnice występują, szacuje się prawdopodobieństwo — oznaczane jako **wartość p** (*p-value*) — otrzymania różnicy w leczeniu co najmniej tak dużej, jak obserwowana (w obu kierunkach, czyli na korzyść lub niekorzyść analizowanej interwencji) w sytuacji, gdyby hipoteza zerowa była prawdziwa. Jeżeli to prawdopodobieństwo braku różnic między grupami znajduje się poniżej przyjętego progu **istotności statystycznej**, wynoszącego w naukach biomedycznych 5% ($p < 0,05$),

to przyjmuje się, że różnice te rzeczywiście istnieją, a nie są wynikiem przypadku, i odrzucamy hipotezę zerową, wnioskując o znamiennych różnicach między grupami. Innymi słowy, oznacza to, że prawdopodobieństwo uzyskania co najmniej takiej różnicy jak wykazana wynosi mniej niż 0,05. Zatem im niższa wartość p dla danego oszacowania, tym mocniejsze dowody przeciw hipotezie o braku różnic i większe przekonanie o rzeczywistej skuteczności interwencji. Oczywiście, stwierdzenie istotności statystycznej wyniku wskazuje jedynie, że obserwowana zależność jest bardziej prawdopodobna, niż wynikałoby to ze zwykłego losowego przypadku, ale nie oznacza, iż obserwujemy prawdziwy efekt. W krytycznej ocenie należy uwzględnić także wiarygodność wewnętrzną i wpływ czynników zakłócających związany z metodyką badania i jego przeprowadzeniem (m.in. randomizacja, zaślepienie, utrata pacjentów). Ponadto trzeba pamiętać o różnicy między istotnością statystyczną a kliniczną i w dalszej kolejności ocenić wielkość obserwowanego efektu w kontekście rokowania w konkretnej populacji.

Niepewność oszacowań można ocenić analizując **95-procentowy przedział ufności** (CI, *confidence interval*), przyjmując, że istnieje 2,5-procentowe prawdopodobieństwo, że prawdziwy efekt jest poniżej, i 2,5-procentowe prawdopodobieństwo, że znajduje się powyżej tego przedziału (taki przedział ufności wynika z założenia wartości $p < 0,05$). Precyzja estymacji zwiększa się wraz z liczebnością badanej próby: im większe badanie, tym dokładniejsze oszacowanie i węższy przedział ufności dla ocenianego parametru. Przy wielu powtarzalnych badaniach dla pomiaru efektu możemy wyliczyć CI w każdym z tych badań, a 95% z nich powinno zawierać prawdziwy wynik. Wyznaczony CI może służyć także do oceny znamienności statystycznej wyniku, jeżeli cały przedział wskazuje na spójny efekt, tzn. nie zawiera wartości oznaczającej brak różnic (0 w przypadku różnicy zmiennych ciągłych lub prawdopodobieństw oraz 1, gdy rozpatrujemy stosunek hazardów lub prawdopodobieństw w obu grupach). Przedziały wykluczające te wartości wskazują na istotne różnice między porównywanymi grupami.

Zazwyczaj w badaniach spotka się ocenę statystyczną trzech rodzajów zmiennych: dychotomicznych (binarnych), np. odpowiedź na leczenie/brak odpowiedzi, ciągłych, np. średnia masa ciała, oraz zmiennych typu czas do zdarzenia, stosowanych w analizie przeżycia (np. OS).

Ocena zmiennych dychotomicznych

Jednym z najczęściej spotykanych rodzajów zmiennych jest liczba przypadków chorych, u których oceniane zdarzenie wystąpiło lub nie. Zazwyczaj jest ona wyrażana w postaci liczebności i odsetków. Zawsze należy pamiętać, że jest to liczba przypadków z pierwszym ocenianym zdarzeniem, co nie stanowi problemu, jeżeli są one unikatowe (np. zgon) albo rzadkie. Natomiast jeżeli mogą one być wielokrotnie powtarzalne, jak np. gorączka neutropeniczna, bardziej informatywna może być ocena łącznej liczby zdarzeń, najlepiej w przeliczeniu na czas obserwacji (*incidence rate per patient-year*). Liczbę zdarzeń możemy zatem analizować jako zmienną ciągłą lub dychotomiczną.

Załóżmy, że w badaniu przydzielono pacjentów w sposób losowy do dwóch grup liczących po 100 chorych, z których jedna otrzymała lek, a druga placebo. Po roku obserwacji stwierdzamy, że w grupie otrzymującej lek zdefiniowaną przez nas odpowiedź na leczenie uzyskało 80 pacjentów (80%), natomiast w kontrolnej — jedynie 40 (40%) chorych. Autorzy naszej przykładowej publikacji podali cztery parametry, odzwierciedlające różnice między analizowanymi grupami w częstości uzyskiwania odpowiedzi na leczenie (pamiętamy, że prawdopodobieństwo zdarzenia wynosi 0,8 w interwencji oraz 0,4 w kontroli):

— RB = 2,00 (95% CI: 1,54–2,59); $p < 0,0001$;
— OR = 6,00 (95% CI: 3,19–11,29); $p < 00001$;
— RD = 0,40 (95% CI: 0,28–0,52); $p < 0,0001$;
— NNT = 3 (95% CI: 2–4).

Możemy zadać sobie pytanie, ile razy prawdopodobieństwo (ryzyko) danego zdarzenia jest wyższe w grupie interwencji w porównaniu z kontrolą — odpowiedzią jest wtedy parametr względny, który jeśli oceniane zdarzenie ma charakter negatywny, nazywamy ryzykiem względnym (RR, *relative risk*), natomiast gdy zdarzenie ma charakter pozytywny — korzyścią względną (RB, *relative benefit*). Jeżeli częstości zdarzeń (prawdopodobieństwa ich wystąpienia) są takie same w obu grupach, ich stosunek wyniesie 1 — jest to więc wartość neutralna, wskazująca na brak różnic między interwencjami. Wartości większe od 1 wskazują na zwiększenie prawdopodobieństwa w grupie interwencji, natomiast mniejsze od 1 — na jego zmniejszenie. W naszym przykładzie RB wynosi 2, a więc prawdopodobieństwo uzyskania odpowiedzi jest 2-krotnie wyższe po zastosowaniu leku badanego przez autorów publikacji niż po placebo. Widzimy też, że skonstruowany dla tej wartości przedział ufności wynosi od 1,54 do 2,59 i nie zawiera wartości 1, więc możemy stwierdzić, że różnice są istotne statystycznie, co potwierdza również przytoczona wartość p ($p < 0,0001$).

Zamiast prawdopodobieństwa możemy obliczyć tzw. szansę wystąpienia danego zdarzenia w każdej z grup. Szansa (*odds*) jest zdefiniowana jako stosunek liczby pacjentów, u których zaobserwowano dane zdarzenie, do liczby chorych bez takiego zdarzenia — określa więc, ile razy częściej obserwujemy dane zdarzenie, niż go nie obserwujemy. W naszym przykładzie w grupie interwencji u 80 chorych wystąpiła odpowiedź, a u 20 nie wystąpiła, szansa wynosi więc 80/20 = 4 (można powiedzieć, że szansa wystąpienia odpowiedzi w tej grupie jest jak 4 do 1). Z kolei w grupie kontrolnej szansa uzyskania

odpowiedzi była dużo niższa i wyniosła 40/60 = 0,67. Możemy teraz, analogicznie do korzyści względnej, obliczyć stosunek tych dwóch szans w analizowanych grupach. Taki parametr nazywamy ilorazem szans (OR, *odds ratio*) i w naszym przypadku widzimy, że szansa uzyskania odpowiedzi na leczenie była 6-krotnie wyższa w grupie interwencji niż w grupie kontroli, co było różnicą znamienną statystycznie, o czym świadczą wartość p oraz przedział ufności, niezawierający neutralnej wartości 1. Chociaż wartości RR/RB są bardziej intuicyjne w interpretacji, w publikacjach często podawane są właśnie obliczenia w postaci OR, które stanowią naturalny wynik użycia metod statystycznych stosowanych powszechnie w ocenie dychotomicznych punktów końcowych (regresji logistycznej, często także z uwzględnieniem czynników dopasowania). Warto dodać, że w przypadku gdy w obu porównywanych ramionach badania uzyskiwane są bardzo wysokie częstości jakiegoś zdarzenia, obliczenie OR może mieć przewagę nad RR, które w takich sytuacjach będzie bliskie jedności i nie będzie dobrze obrazować faktycznej różnicy między grupami.

Oprócz przedstawionych powyżej parametrów względnych można oszacować także parametry bezwzględne, które uznaje się za bardziej informatywne, bo dodatkowo pokazują rzeczywistą częstość zdarzeń — czy były one skrajnie rzadkie czy też występują u znaczącego odsetka populacji. Różnica ryzyka (RD, *risk difference*) lub bezwzględne zmniejszenie ryzyka (RRR, *relative risk reduction*) — w naszym konkretnym przypadku możemy mówić o bezwzględnym zwiększeniu korzyści (RBI, *relative benefit increase*) — jest prostą różnicą prawdopodobieństw między grupami. W naszym przykładzie różnica ta wynosi 0,40 i możemy stwierdzić, że prawdopodobieństwo wzrasta o 40 punktów procentowych w grupie interwencji w stosunku do kontroli. W praktyce oznacza to, że na każdych 100 chorych otrzymujących interwencję u dodatkowych 40 w stosunku do leczenia kontrolnego zaobserwujemy odpowiedź na leczenie. W naszym przykładzie podano również przedział ufności skonstruowany dla obliczonej różnicy ryzyka oraz wartość p. Widzimy, że nasz przedział (0,28–0,52) nie zawiera wartości 0, możemy więc przyjąć, że obserwowane różnice między grupami są istotne statystycznie.

Możemy też odwrócić tę zależność i zadać pytanie, przy jakiej liczbie leczonych chorych uzyskamy 1 przypadek zdarzenia więcej. Taki parametr, jeśli obserwujemy korzystny efekt leczenia, nazywamy NNT (NNT, *number needed to treat*), a gdy zdarzenie jest niekorzystne — NNH (*number needed to harm*). Z proporcji wynika, że liczba ta wyniesie 1/0,40 (NNT = 1/RD), a więc 2,5, ponieważ jednak mówimy o liczbie pacjentów, wynik należy zaokrąglić do liczby całkowitej w górę — możemy wówczas powiedzieć, że lecząc 3 chorych interwencją zamiast kontroli przez dany czas (rok), oczekujemy 1 dodatkowego przypadku odpowiedzi na leczenie [16–18]. Interpretując wyniki dla zmiennych dychotomicznych, w szczególności wartości parametrów bezwzględnych, należy mieć na uwadze okres obserwacji dla danego punktu końcowego. Przykładowo, NNT uzyskane w badaniach o różnym czasie trwania mogą nie być bezpośrednio porównywalne, gdyż wartość NNT może się zmieniać wraz z okresem obserwacji.

Ocena zmiennych ciągłych

W badaniach klinicznych często oceniane są parametry określające zmiany nasilenia objawów choroby, wyników badań laboratoryjnych lub jakości życia na pewnej skali. W każdej z tych sytuacji uzyskane wyniki mają charakter ciągły — czyli przyjmują dowolną wartość wyrażoną liczbą rzeczywistą z danego zakresu. Na przykład, chorzy mogą być proszeni o wskazanie swojego samopoczucia na skali od 0 do 100, od najgorszego do najlepszego. Z wynikami ciągłymi mamy również do czynienia przy pomiarze ciężaru ciała, wzrostu, ciśnienia tętniczego, średniej liczby białych komórek krwi, stężenia hemoglobiny itd.

Takie wyniki dla analizowanych grup chorych najczęściej podsumowuje się, przedstawiając miarę centralną — wartość średnią lub medianę — dla przypomnienia, średnia (arytmetyczna) jest sumą wyników uzyskanych dla każdego pacjenta z grupy, podzieloną przez liczbę chorych w tej grupie, natomiast **mediana** jest wartością środkową, czyli taką, która dzieli grupę pacjentów na połowę (tzn. połowa pacjentów ma wynik poniżej wartości mediany, a druga połowa — powyżej). Zbiór wyników danego efektu wyrażonego zmienną ciągłą charakteryzuje też pewna zmienność, którą obrazowo można sobie przedstawić jako rozrzut obserwowanych wartości wokół średniej. Parametrem wskazującym na wielkość tej zmienności jest **odchylenie standardowe** (SD) — niższe wartości wskazują na mały rozrzut wyników wokół średniej, natomiast wyższe — na dużą zmienność i duże różnice między uzyskanymi wynikami a średnią.

Podobnie jak to było w przypadku analizy częstości zdarzeń, również w analizie danych ciągłych oceniamy ogólny efekt na podstawie próby z populacji ogólnej. Ponieważ takie próbkowanie daje nam w efekcie różne średnie, rozrzucone wokół wartości średniej w tej ogólnej populacji, możemy wprowadzić dodatkową miarę rozrzutu, określającą właśnie rozrzut średnich z prób wokół średniej w populacji ogólnej — taki parametr nazywamy **błędem standardowym** (SE, *standard error*) i jest on równy odchyleniu standardowemu w próbie podzielonemu przez pierwiastek z liczebności tej próby. Błąd standardowy maleje wraz ze wzrostem liczebności próby, a jego mniejsza wartość oznacza, że nasza próba lepiej przybliża wartość w populacji ogólnej.

W przypadku, gdy wyniki badania są prezentowane w postaci median, najczęściej podaje się również zakres, w jakim znajdują się obserwowane wyniki. Jak już wspomniano, mediana to wartość środkowa — taka, która dzieli nam uszeregowany zbiór danych na 2 równe części. Jednak takich „podzielników" zbioru możemy wyznaczyć wiele, w zależności od przyjętych kryteriów — ogółem nazywamy je kwantylami, a mediana stanowi szczególny przypadek takiego kwantylu. Możemy również podzielić zbiór na 4 części — wtedy nasze „podzielniki" nazywamy kwartylami (warto zauważyć, że mediana jest też drugim kwartylem zbioru), a w przypadku podziału zbioru na 100 równych części — percentylami (mediana jest wówczas 50. percentylem zbioru). Autorzy badań czasem przedstawiają wartości mediany wraz z tzw. **rozstępem międzykwartylowym** (IQR, *interquartile range*), czyli odległością pomiędzy pierwszym a trzecim kwartylem.

Rozumowanie przy ocenie statystycznej istotności różnic między grupami dla zmiennych ciągłych jest analogiczne do tego, jakie przeprowadziliśmy przy okazji opisu różnicy częstości zdarzeń w dwóch grupach — ogółem należy wyznaczyć wartości średnie danego parametru w analizowanych grupach, a potem policzyć ich różnicę i wyznaczone dla niej przedziały ufności lub obliczyć wartość p. W przypadku zmiennych o rozkładzie normalnym (także po odpowiednim przekształceniu danych) do oceny różnic używa się zazwyczaj testu t Studenta lub ANOVA, a w innych przypadkach któregoś z testów nieparametrycznych — np. U Manna–Whitneya. Ponieważ możemy oczekiwać, że u pacjenta z większymi wartościami wyjściowymi danego parametru mogą wystąpić większe jego zmiany w trakcie badania, stosowana jest również analiza kowariancji (ANCOVA), która porównuje średnie skorygowane o wartości wyjściowe. Podobnie jak w przypadku opisywanej wcześniej różnicy ryzyka, istotność statystyczną możemy oceniać na podstawie przedziałów ufności, a „0" jest wartością wskazującą na brak różnic między grupami.

Parametrem najczęściej przedstawianym w badaniu jest **różnica średnich** (MD, *mean difference*) między analizowanymi grupami. Przy ocenie wyników trzeba jednak zachować ostrożność, bo autorzy badania mogą je prezentować na kilka sposobów. Na przykład, oceniając jakość życia, można wyznaczyć średni wynik na końcu okresu obserwacji w obu grupach i policzyć różnicę średnich między nimi (warto wtedy zawsze się upewnić, że wyjściowo nie obserwowano istotnych różnic w pomiarach). Można również ocenić średnią zmianę tego wyniku w trakcie leczenia w odniesieniu do wartości wyjściowej i obliczyć różnice średnich dla takich zmian — takie podejście stosuje się częściej, bo pozwala ocenić wpływ leczenia z dopasowaniem względem już istniejącego początkowo efektu.

W badaniach klinicznych można również często napotkać średnią liczoną metodą **najmniejszych kwadratów**, określaną jako *least square mean* (LSM) — jest to po prostu średnia dopasowana względem dodatkowych czynników. Dla przykładu, w danej grupie osób można policzyć średni wiek, sumując liczby lat życia każdej z osób i dzieląc tę sumę przez liczbę analizowanych osób, dzięki czemu uzyskujemy zwykłą średnią. Wiemy jednak, że akurat w tej grupie znajduje się wiele kobiet w podeszłym wieku, co może zawyżać tak wyznaczoną średnią — w takim wypadku można obliczyć średni wiek najpierw wśród kobiet, potem wśród mężczyzn i dopiero uśredniając ten wiek dla obu grup, uzyskamy wartość średnią w całej grupie, dopasowaną pod względem płci [16–18].

Ocena zmiennych typu „czas do zdarzenia" — analiza przeżycia

Analiza danych typu czas do zdarzenia (np. zgonu bez względu na przyczynę w analizie OS, zgonu lub progresji nowotworu w ocenie PFS) wiąże się z kilkoma problemami. Ogólnie analizę przeżycia wykonujemy, gdyż w odpowiednio długim okresie obserwacji zdarzenia kliniczne (progresja/zgon) wystąpią u wszystkich lub niemal wszystkich chorych. W takim przypadku policzenie prostego parametru, jak RR, będzie bezużyteczne (prawdopodobieństwa zdarzeń będą bliskie 100% w obu grupach). Dodatkowo, w długookresowej obserwacji oprócz przypadków oznaczonych jako „ze zdarzeniem" lub „bez zdarzenia" pojawią się również pacjenci, których utracimy z obserwacji w trakcie badania, a których dalszych losów nie znamy, bądź też w momencie wykonywania analizy statystycznej są oni nadal w obserwacji, ale nie wiemy, jakie będą ich dalsze losy. Wreszcie, biorąc pod uwagę zróżnicowanie okresów rekrutacji i dat włączenia pacjentów do badania, w obserwacji będziemy mieli chorych cechujących się różnymi okresami przebywania w badaniu [18–22].

Odpowiedzią na te ograniczenia jest właśnie analiza przeżycia, uwzględniająca nie tylko fakt wystąpienia zdarzenia, ale także czas do jego wystąpienia, oraz umożliwiająca ucinanie (cenzorowanie) przypadków pacjentów traconych z obserwacji lub o nieznanych dalszych losach w momencie analizy danych. Należy zaznaczyć, że termin „analiza przeżycia" nie jest zarezerwowany wyłącznie dla oceny przeżycia całkowitego (tj. czasu do zgonu), lecz odnosi się do wszystkich punktów końcowych typu *time-to-event* (np. czas do uzyskania odpowiedzi, czas przeżycia wolnego od progresji choroby).

Najprostszą formą analizy przeżycia jest wykreślenie **krzywych Kaplana–Meiera**, z których możemy odczytać prawdopodobieństwa przeżycia do danego punktu czasowego oraz mediany przeżycia. Następnie za pomocą odpowiedniego testu statystycznego (najczęściej testu log-rank) dokonuje się porównawczej oceny różnic w czasie do wystąpienia zdarzenia między grupami. Bardziej zaawansowaną analizę, pozwalającą

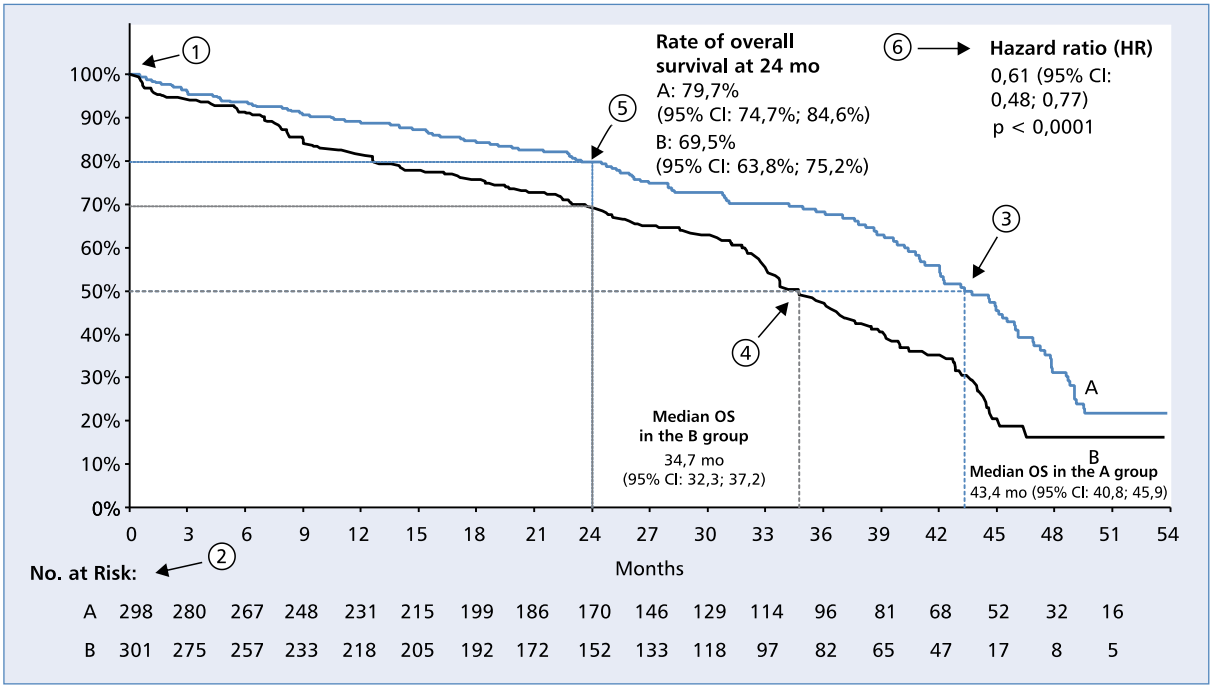
uwzględnić także czynniki dopasowania (zmienne objaśniające, niezależne), mające wpływ na czas przeżycia, przeprowadza się przy użyciu modeli regresji, najczęściej modelu proporcjonalnych hazardów Coxa (należy pamiętać, że przebieg krzywych Kaplana–Meiera, jak również wartości median czasu do zdarzenia w takiej sytuacji nie podlegają nadal dopasowaniu). Ogólnie analiza wielkości i kierunku różnic w przeżyciu polega na ocenie wartości **hazardu względnego** (HR, *hazard ratio*), **median czasu przeżycia** (do zdarzenia) oraz **prawdopodobieństwa przeżycia w określonym punkcie czasowym** (np. przeżycie 12-miesięczne). Wartość HR podsumowuje względne różnice w przeżyciu między grupami w całym okresie obserwacji. Ocena różnic w przeżyciu, polegająca wyłącznie na prostym porównaniu wartości median czasu przeżycia w grupach, nie jest wystarczająca dla zobrazowania różnic w pełnym horyzoncie badania i może być wręcz myląca, zwłaszcza w przypadku braku proporcjonalności hazardów (zagadnienie to omówiono w dalszej części opracowania).

Wykreślenie krzywej Kaplana–Meiera odbywa się na podstawie wyników badania dla poszczególnych pacjentów, np. czy i kiedy wystąpił zgon. Należy jednak pamiętać, że niektórzy chorzy „wypadają" z badania z innych przyczyn i informacja o tym, jak długo żyją, zostaje utracona. Zatem w momencie wykonywania analizy statystycznej uwzględnia się pacjentów, którzy żyją, pacjentów, którzy zmarli, oraz takich, którzy opuścili badanie jakiś czas temu i nie wiadomo, czy nadal żyją, co nazywamy obserwacją cenzorowaną.

Ryzyko (hazard) określa prawdopodobieństwo zajścia zdarzenia w określonym czasie, przy założeniu, że zdarzenie to do tej pory nie wystąpiło. Stosunek wartości hazardów oszacowanych dla grupy z interwencją i kontrolnej w danym czasie nazywamy hazardem względnym (HR). Koncepcyjnie, w uproszczonej interpretacji HR jest zbliżony do RR, ale trzeba pamiętać, że HR uwzględnia dane pochodzące z całego okresu obserwacji przeżycia w badaniu oraz przypadki cenzorowane, podczas gdy ocenę RR przeprowadza się w ustalonym punkcie czasowym (np. odsetek zgonów po 12 miesiącach leczenia). Interpretując przedstawioną w badaniu wartość HR, zakłada się, że stosunek ten jest w przybliżeniu stały w każdym czasie w trakcie obserwacji (założenie proporcjonalności hazardów), czyli jeżeli np. wartość HR wynosi 0,61, to zakładamy, że w chorzy w grupie z interwencją cechują się w przybliżeniu o 39% mniejszym ryzykiem zgonu niż ci z grupy kontrolnej w każdym punkcie czasowym w trakcie obserwacji. Tę zależność można też przedstawić jako przeciętne wydłużenie czasu przeżycia o 64% (1/0,61 = 1,64) w grupie z interwencją w porównaniu z kontrolą [18–22]. Należy przy tym zaznaczyć, że wartości HR nie można jednak w prosty sposób przełożyć na bezwzględne różnice w czasie przeżycia — przykładowo, w populacji o niskiej śmiertelności zmniejszeniu ryzyka zgonu o 30% (tj. HR = 0,70) może towarzyszyć wydłużenie średniego czasu przeżycia o 12 miesięcy, podczas gdy taka sama względna redukcja ryzyka zgonu (HR = 0,70) w populacji o wysokiej śmiertelności może się wiązać ze znacznie niższym efektem bezwzględnym (np. 3 miesiące).

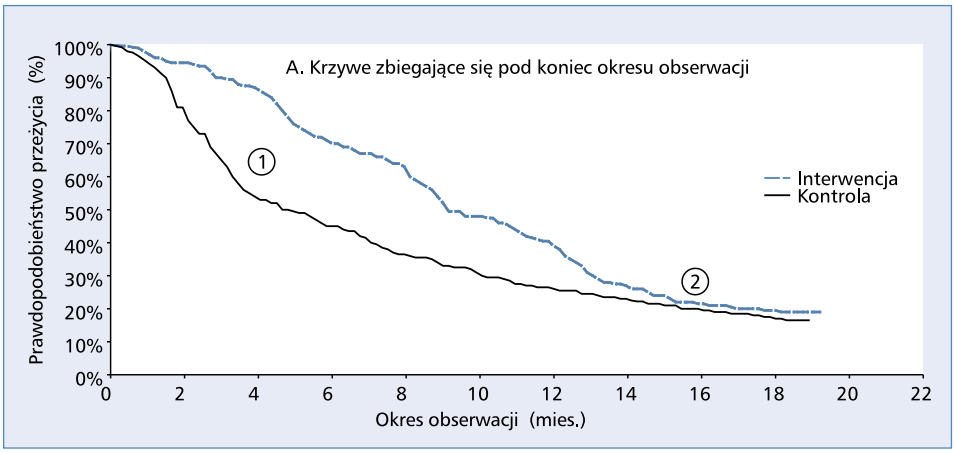
Najprościej założenie o **proporcjonalności hazardów** można potwierdzić, analizując wizualnie przebieg krzywych na wykresie Kaplana–Meiera, oceniając, czy różnica między nimi jest w przybliżeniu stała i utrzymuje się w czasie. Dopuszczalne są niewielkie odstępstwa (zmniejszenie się lub zwiększenie w czasie różnic w przebiegu krzywych) (ryc. 1). W sytuacji wystarczająco długiej obserwacji, zwłaszcza w końcowych liniach leczenia nowotworów w stadiach zaawansowanych, gdy dojdzie do wystąpienia zdarzeń u wszystkich pacjentów (przy bardzo dojrzałych danych, gdy niewielu chorych pozostaje w obserwacji), po początkowym okresie utrzymywania się różnic w przebiegu krzywych możemy obserwować ich zejście się (ryc. 2A). Przeciwstawnym przypadkiem jest sytuacja w populacji o niewielkim ryzyku zgonów i oczekiwanym długoletnim przeżyciu, kiedy to krzywe mogą osiągać płaski przebieg (*plateau*), ponieważ przypadki zgonów są nieliczne (ryc. 2D). Czasami też na samym początku obserwacji następuje przecięcie się krzywych, co może się zdarzyć, gdyż w grupie z interwencją w początkowym okresie może wzrastać ryzyko powikłań (zwłaszcza jeżeli występuje znacząca różnica między obserwowanymi procedurami — np. zabieg operacyjny z chemioterapią *vs*. leczenie zachowawcze i chemioterapia), a oczekiwany zysk kliniczny obserwujemy dopiero w dalszym okresie, gdy krzywe się rozchodzą (ryc. 2B). Ogólnie, jeżeli zmienność przebiegu krzywych w czasie obserwacji dotyczy wielkości efektu, a nie jego kierunku, to możemy uznać, że odstępstwo od założenia o proporcjonalności hazardów jest nieznaczne, i pozostajemy przy przedstawionej interpretacji HR. Jeżeli jednak następuje znacząca zmiana kierunku działania (ryc. 2C), to nie możemy interpretować obliczonej wartości HR, bo zmienia się ona istotnie w czasie. Pewnym rozwiązaniem jest próba analizy podgrup, po to by wykryć przyczynę różnic w skuteczności w czasie przy zastrzeżeniach związanych z taką analizą (por. poniżej) [18–22].

Należy pamiętać, że HR jest wartością względną, pozwalającą ocenić istotność statystyczną obserwowanych różnic w przeżyciu, ale jak wspomniano, przy podejmowaniu decyzji terapeutycznej konieczna jest również ocena istotności klinicznej, także w odniesieniu do rokowania w danej populacji. Bezwzględny wpływ interwencji w porównaniu z kontrolą można ocenić, analizując różnice w medianach albo porównując prawdopodobieństwo przeżycia w określonym czasie (np. przeżycie roczne lub 2-letnie).

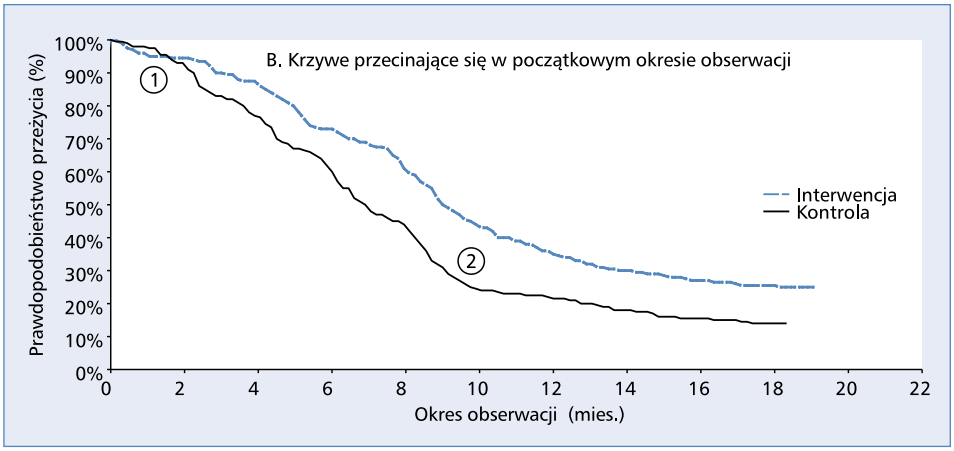


Rycina 1. Wykres Kaplana–Meiera znajduje się na płaszczyźnie ograniczonej osią Y opisującą prawdopodobieństwo przeżycia całkowitego i osią X, na której prezentowany jest czas od rozpoczęcia leczenia (obserwacji) w badaniu. Konstruując krzywą Kaplana–Meiera, w kolejnych przedziałach czasowych bierze się pod uwagę liczbę pacjentów z możliwością wykonania pomiaru na początku danego przedziału czasu (*at-risk*), liczbę pacjentów ze zdarzeniem oraz liczbę pacjentów utraconych z obserwacji. Wykres prezentuje skumulowane prawdopodobieństwo w danym czasie. W momencie rozpoczęcia obserwacji (1) wszyscy pacjenci żyją (OS = 100%). W rzeczywistości nie jest to ten sam moment (data randomizacji) dla każdego z badanych, gdyż są oni włączani do badania w różnych ośrodkach w różnym czasie. Następnie obserwujemy zmniejszanie się prawdopodobieństwa przeżycia całkowitego w obu ramionach wraz z upływem czasu, jest ono jednak zawsze wyższe w grupie A (pomimo pewnej zmienności w wielkości różnic między grupami kierunek efektu jest spójny i możemy uznać, że odstępstwo od założenia o proporcjonalności hazardów jest nieznaczne — por. tekst). Pod wykresem w regularnych interwałach czasowych powinny być podane liczby pacjentów w obserwacji (*at-risk*) na początku danego przedziału czasu (2). Jeżeli na końcu wykresu są one niewielkie (< 10% wartości wyjściowej), wnioskowanie z krzywych jest na tym odcinku ograniczone. Patrząc na wykres Kaplana–Meiera, wartości z obu krzywych można porównywać horyzontalnie, poszukując różnicy w czasie, kiedy skumulowane prawdopodobieństwo przeżycia osiągnie wartość 50% — (3) i (4) stanowią odpowiednio wartości median czasu przeżycia dla grup A i B (*median OS in the A group, median OS in the B group*), a różnica median wynosi 8,7 miesiąca. Różnice w przeżyciu możemy również oceniać wertykalnie, porównując wartości przeżycia w danym punkcie czasowym. W naszym przykładzie 2-letnie przeżycie (*rate of overall survival at 24 mo*) wynosi 79,7% w grupie z interwencją oraz 69,5% w kontroli (5). Jest to prawdopodobieństwo skumulowane dla tego okresu, często jego wartość jest podawana wraz z przedziałem ufności (co pozwala nam ocenić dokładność oszacowania), a w tekście publikacji odnajdziemy zazwyczaj ocenę statystyczną wyniku (wartość p dla różnic w przeżyciu skumulowanym w tym punkcie czasowym). Wnioskowanie na temat różnic w przeżyciu w całym okresie obserwacji umożliwia nam podana wartość hazardu względnego (*hazard ratio*) (6); widzimy, że w danym punkcie czasowym ryzyko zdarzenia (zgonu) jest niższe w grupie z interwencją, a wynik jest znamienny statystycznie (patrząc zarówno na przedział ufności, jak i podaną wartość p)

W trakcie trwania RCT, zwłaszcza przy długim okresie obserwacji, konieczne są **analizy wstępne** (*interim*) przeprowadzone przez niezależnych badaczy w sposób odślepiony, ale poufny. Niezależna, niezwiązana z badaniem komisja może podjąć decyzję o przedwczesnym przerwaniu badania, np. ze względu na zastrzeżenia dotyczące bezpieczeństwa albo na spektakularny efekt nowej interwencji (w takiej sytuacji decyzja o przerwaniu powinna być szczegółowo oceniona, ponieważ wyraźny efekt często bywa przeszacowany w krótkim okresie obserwacji, a różnice między interwencjami stają się mniej wyraźne w dłuższym okresie obserwacji) [14]. W badaniach w onkologii w protokole statystycznym często też predefiniowane są kolejne analizy wstępne (gdy ocena mocy statystycznej jest zależna od wystąpienia danej liczby zdarzeń). Ponieważ mamy tu sytuację wielokrotnego testowania hipotezy i wzrasta ryzyko przypadkowego zaobserwowania wyników „znamiennych statystycznie", wprowadzana jest korekta poziomu istotności (tym większa, im więcej jest planowanych analiz *interim*,



Rycina 2 A. Krzywe Kaplana–Meiera zbiegające się pod koniec okresu obserwacji. Od samego początku obserwacji pojawiają się różnice w przeżyciu (① krzywe wyraźnie się „rozchodzą"), ale pod koniec okresu obserwacji (②) skumulowane prawdopodobieństwo zgonu jest praktycznie takie samo w obu ramionach badania. Sytuacja taka może wystąpić np. w krańcowych stadiach nowotworów, gdzie ostatecznie bez względu na zastosowane leczenie zdarzenie (zgon) wystąpi u prawie wszystkich chorych. Jest to dopuszczalne odstępstwo od założenia o proporcjonalności hazardów. Zbieganie się krzywych w końcowym okresie obserwacji może być również spowodowane wysokim odsetkiem obserwacji cenzorowanych (tj. małą liczbą pacjentów *at-risk*), w konsekwencji czego estymacja Kaplana–Meiera w „ogonie" krzywej jest obciążona ograniczeniami



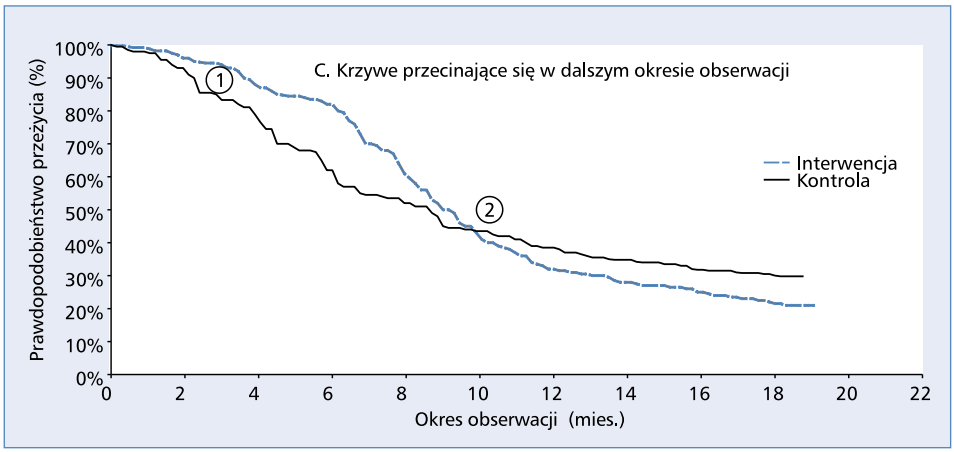
Rycina 2 B. Krzywe Kaplana–Meiera przecinające się w początkowym okresie obserwacji. W początkowym okresie obserwacji krzywe się przecinają i przez krótki okres przeżycie jest niższe w grupie z interwencją w porównaniu z kontrolą — np. w przypadku zastosowania interwencji obciążonej początkowo większym ryzykiem powikłań (①), następnie jednak pojawiają się wyraźne różnice w przeżyciu w dalszej obserwacji (②). W takiej sytuacji jest to również dopuszczalne odstępstwo od założenia o proporcjonalności hazardów

np. zgodnie z kryteriami O'Briena–Fleminga); warto wówczas zwrócić uwagę, że o wynikach znamiennych nie będziemy mówić, odnosząc się do wartości $p < 0{,}05$, lecz przy założeniu znacznie niższego progu, np. $< 0{,}001$.

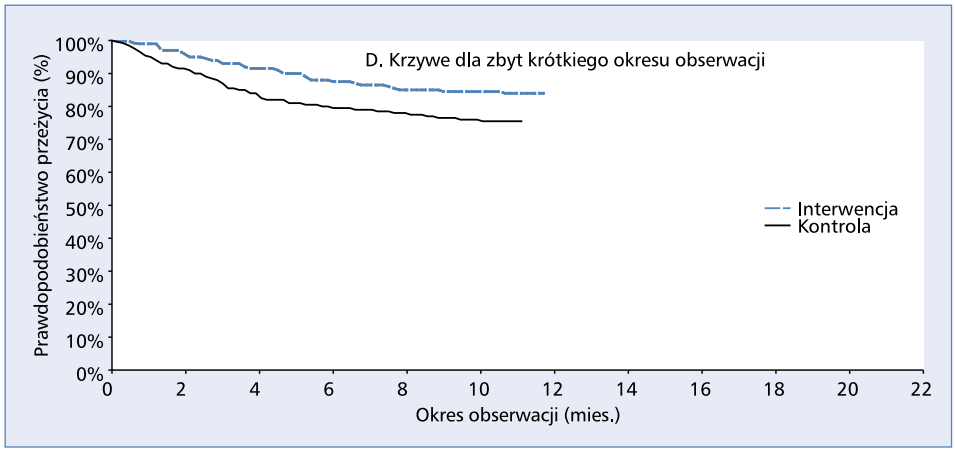
Stwierdzenie znamiennych różnic między interwencjami w OS zgodnie z założoną mocą statystyczną może umożliwiać chorym po wystąpieniu progresji przechodzenie z grupy kontrolnej na ocenianą interwencję (***cross-over, treatment switching***), co działa w kierunku konserwatywnym, zawyżając skuteczność interwencji kontrolnej i zmniejszając szacowany efekt analizowanego leku. Oczywiście ma to znaczenie tylko dla oceny OS, sama zaś ocena PFS jest niezaburzona. W takim przypadku istnieje możliwość zastosowania odpowiednich metod korekcji wpływu *cross-over* na OS, z których najprostsza to cenzorowanie obserwacji zmieniających leczenie pacjentów [24, 25].

Wyniki z dopasowaniem i w podgrupach

Analiza wyników z dopasowaniem do charakterystyk wyjściowych w ramach regresji logistycznej (prezentowane są wówczas wartości OR opisane jako *adjusted* lub



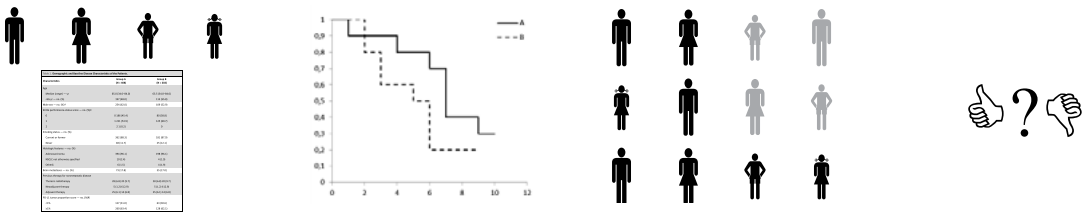
Rycina 2 C. Krzywe Kaplana–Meiera przecinające się w dalszym okresie obserwacji. Początkowa przewaga interwencji ((1)) zanika niespodziewanie w trakcie obserwacji ((2)) i przeżycie pozostaje wyższe w ramieniu komparatora do końca okresu obserwacji. W tym przypadku nie jest spełniony warunek proporcjonalności hazardów i nie można wnioskować o różnicach między grupami. Prawdopodobnie w badaniu wystąpił nieznany czynnik zakłócający, który odwrócił wnioskowanie i konieczna jest szczegółowa analiza w podgrupach w celu jego identyfikacji



Rycina 2 D. Krzywe Kaplana–Meiera dla zbyt krótkiego okresu obserwacji. Śmiertelność utrzymuje się na względnie niskim poziomie i trudno wnioskować na temat dalszych losów pacjentów i różnic między grupami (ocena we wczesnych stadiach nowotworów, o oczekiwanym długookresowym dalszym przeżyciu). Mimo początkowych różnic przy dłuższym okresie obserwacji, gdy zbierze się więcej obserwowanych zdarzeń, przebieg tych krzywych może się zbliżyć do każdej z poprzednio opisywanych sytuacji

multivariate) lub modelu hazardów proporcjonalnych Coxa (w przypadku analizy przeżycia) może być prezentowana jako główna bądź jako analiza wrażliwości. Warto zwrócić uwagę, że dopasowanie powinno obejmować przede wszystkim czynniki rokownicze i — o ile zazwyczaj nie wpływa na precyzję oszacowania — może modyfikować miarę centralną oszacowania. Uwzględnianie natomiast charakterystyk nieodnoszących się do rokowania, chociaż nawet obejmujących czynniki stratyfikacyjne (np. lokalizacja geograficzna), nie wpływa znacząco na wyniki. Dobór czynników dopasowania dokonywany *post-hoc* (niepredefiniowany w planie statystycznym) może rodzić podejrzenie takiego doboru danych, by uzyskać oczekiwany efekt; w takiej sytuacji zawsze należy oczekiwać prezentacji wyników z dopasowaniem i bez dopasowania [23].

Również **analiza w podgrupach** (*subgroup analysis*) powinna być predefiniowana w planie statystycznym. Biorąc pod uwagę zróżnicowanie populacji ogólnej badania, pozwala ona ocenić, czy wyniki ogólne odnoszą się do wszystkich pacjentów, czy też istnieją różnice w skuteczności interwencji. W przypadku takich wyników należy pamiętać, że tracimy moc statystyczną, by wiarygodnie oceniać różnice statystyczne. Z drugiej



Metodyka	Główne wyniki	Analizy dodatkowe	Inne
• Charakterystyka populacji (*baseline characteristics*) • Liczba badanych w porównywanych grupach (*allocation, intervention, control*) • Komparator a lokalna praktyka kliniczna • Hipoteza badawcza • Metoda randomizacji i zaślepienie (*randomization, blind, open label, unmasked*) • Utrata pacjentów z badania (*drop-out, discontinuation*) • Wiarygodność wewnętrzna	• Istotność statystyczna HR dla PFS i OS — wartość p i przedział ufności (*statistical significance, hazard ratio, progression-free survival, overall survival*) • Przebieg krzywych Kaplana–Meiera (*Kaplan-Meier curve*) • Wartości median OS i PFS • Inne punkty końcowe — zbieżność wnioskowania z punktami głównymi — spójność wewnętrzna (*time to treatment discontinuation, objective response, complete response*) • Jakość życia, bezpieczeństwo (*quality of life, safety, adverse events*)	• Analiza zgodnie z protokołem badania (*intention-to-treat, per protocol analysis*) • Analiza w podgrupach — zbieżność wnioskowania z populacją ITT (*subgroup analysis*) • Istotność statystyczna wyników w subpopulacjach, test interakcji (*interaction test*) • Analizy wstępne, obecność *cross-over* (*interim analysis, cross-over*)	• Dalsze linie leczenia (*subsequent treatment*) • Wnioski autorów — jednoznaczne czy zachowawcze (*conclusions*) • Zgodność wyników z innymi badaniami (spójność zewnętrzna) • Możliwość przeniesienia wniosków do praktyki klinicznej (wiarygodność zewnętrzna)

Rycina 3. Aspekty, na które należy zwrócić uwagę, dokonując krytycznej oceny metodyki i wyników badania klinicznego w onkologii

strony, w przypadku zdefiniowania licznych podgrup i wielokrotnego testowania zwiększamy ryzyko zupełnie losowego wystąpienia wyników istotnych statystycznie. Przede wszystkim zaś istotność statystyczna lub jej brak w którejś z podgrup nie jest wystarczająca do wnioskowania o rzeczywistych różnicach w skuteczności interwencji. Zazwyczaj w poszczególnych podgrupach obserwujemy spójny efekt, choć pewną zmienność miary centralnej, a w tych mniejszych liczebnościowo przedziały ufności stają się szersze i w pewnych przypadkach mogą przekraczać wartość 1 (utrata znamienności statystycznej). W takiej sytuacji powinniśmy zwrócić uwagę na znamienność **testu interakcji** (statystyczna analiza, czy wpływ interwencji na obserwowany wynik zależy od innych czynników), by ocenić, czy w danej podgrupie rzeczywiście występują różnice w skuteczności interwencji (pamiętając jednak o możliwości uzyskania fałszywych wyników przy wielokrotnym testowaniu). Analiza podgrup może być również pomocna w poszukiwaniu zawężenia populacji docelowej przy braku znamienności wyniku w populacji ogólnej. Należy jednak pamiętać, że analiza podgrup ma charakter bardziej badawczy (eksploracyjny) i służy raczej tworzeniu dalszych hipotez niż stawianiu ostatecznych wniosków [23].

Podsumowanie

Przekładanie wyników badań klinicznych na codzienną praktykę jest uznaną metodą praktykowania EBM. W tym celu jednak należy ocenić wiele elementów, które łącznie świadczą o wiarygodności badania i znaczeniu jego wyników (ryc. 3). Nie jest to łatwe, zwłaszcza że większość publikacji powstaje w języku angielskim, a autorzy często zakładają z góry, że odbiorca jest biegły w zakresie statystyki i szersze wyjaśnienia są zbędne. Oprócz oceny metodyki i wiarygodności próby klinicznej należy się upewnić, czy populacja poddana ocenie jest reprezentatywna, tj. o charakterystyce zbliżonej do tej, dla której ma zostać podjęta decyzja terapeutyczna, a jeżeli występują rozbieżności (np. inny wiek pacjentów czy obecność chorób współistniejących), to jakie może to mieć znaczenie. Następnie, jeżeli komparator w badaniu nie jest powszechnie stosowany lub dostępny, ale ma podobny mechanizm działania do obecnego

standardu leczenia, należy ocenić, czy są dowody na ich zbliżoną skuteczność, co pozwoliłoby przenieść wnioskowanie z badania na praktykę kliniczną w tym aspekcie. Istotne znaczenie ma liczebność grup w badaniu — jeżeli była niewielka i nie było to spowodowane niskim rozpowszechnieniem jednostki chorobowej, to należy ocenić, czy badanie miało moc statystyczną do wykazania różnic. Trzeba także zwrócić uwagę na ewentualne różnice w charakterystykach wyjściowych między grupami — jeśli były istotne, mogą świadczyć o zaburzonym doborze do grup. Nieprawidłowa metoda randomizacji może powodować nierównomierne rozłożenie czynników zakłócających. Przy braku zaślepienia ocena różnic w subiektywnych punktach końcowych jest obarczona ograniczeniami. W przypadku znaczących różnic w częstości utraty chorych ważne jest, czy mogło to mieć związek z zastosowanym leczeniem. Warto sprawdzić również rodzaj testowanej hipotezy badawczej. W przypadku onkologii głównym punktem końcowym będzie najczęściej analiza przeżycia — PFS i OS. Jeżeli zaobserwowano spójne, istotne statystycznie i klinicznie różnice na korzyść interwencji, to istnieją silne przesłanki o wyższości ocenianej terapii. Jeśli natomiast znamienność wyników zaobserwowano tylko dla PFS, należy sprawdzić, czy jest planowana dalsza (końcowa) ocena OS, w której przy bardziej dojrzałych danych wynik mógłby osiągnąć istotność statystyczną. W przypadku niektórych nowotworów, zwłaszcza przy ocenie ich wczesnych stadiów, wykazanie różnic w przeżyciu może być utrudnione ze względu na oczekiwany okres obserwacji, wynoszący czasami dziesiątki lat. Tak długa obserwacja stanowi dodatkowe wyzwanie, gdyż w tym czasie pacjenci mogą być poddawani np. wielu różnym liniom dalszego leczenia i ocena ostatecznych różnic w przeżyciu jest ograniczona (w takiej sytuacji znaczenie zyskują takie punkty końcowe, jak np. DFS czy patologiczna odpowiedź na leczenie). Istotne jest, czy dopuszczano możliwość zmiany leczenia po progresji (*cross-over*), co mogłoby prowadzić do zawyżenia wyniku OS w kontroli. Jeżeli zanotowano brak istotności HR jednocześnie dla OS i PFS, należy zadać pytanie, czy nie wynika to z braku mocy statystycznej badania, niedojrzałości opublikowanych wyników (analiza *interim*) lub dużej utraty chorych z badania. Jeżeli w grę nie wchodzi żaden z tych czynników, prawdopodobnie nie ma różnic między interwencjami. Bardzo pomocna może być ocena spójności wyników poprzez ich porównanie z innymi publikacjami dotyczącymi podobnej populacji. Jeżeli mediany PFS/OS różnią się od analogicznych median w podobnych badaniach, należy dokładnie przeanalizować charakterystyki populacji. Cennym źródłem informacji jest też analiza w podgrupach. Niekiedy wynik dla subpopulacji staje się istotny statystycznie, mimo jej braku w populacji łącznej — może to stanowić przesłankę o wyższej skuteczności leczenia tylko w określonej podgrupie, ale z drugiej strony należy mieć na uwadze eksploracyjny charakter takiej analizy. Jeśli podobny trend obserwujemy w wynikach w zależności od obecności lub braku danego kryterium, ale w jednej z podgrup brak jest istotności, warto zweryfikować, czy liczebność tej podgrupy nie jest zbyt mała, a także sprawdzić wynik testu interakcji. Jeżeli są dostępne jakiekolwiek opcje terapeutyczne w analizowanym wskazaniu, kwestie bezpieczeństwa ocenianej terapii są niezwykle ważne. W onkologii leczenie o wyższej skuteczności wiąże się często ze zwiększeniem toksyczności; taka sytuacja może być akceptowalna przy wyraźnym zysku klinicznym, jak np. wydłużenie przeżycia.

Jak widać, na ocenę badania klinicznego składa się wiele elementów, omówionych w niniejszym opracowaniu w wielkim skrócie. Autorzy mają jednak nadzieję, że poruszyli najważniejsze aspekty oceny wyników badań klinicznych oraz stosowanej terminologii, a także, że w artykule udało się pokazać złożoność procesu ich interpretacji. W drugiej części pracy zostaną przedstawione przykłady badań klinicznych wraz z oceną ich wiarygodności i wpływu na praktykę kliniczną.

Piśmiennictwo

1. Govani SM, Higgins PD. How to read a clinical trial paper: a lesson in basic trial statistics. Gastroenterol Hepatol (N Y). 2012; 8(4): 241–248. PubMed PMID: 22723755; PubMed Central PMCID. PMCPMC3380258.
2. Bothwell LE, Greene JA, Podolsky SH, et al. Assessing the Gold Standard — Lessons from the History of RCTs. N Engl J Med. 2016; 374(22): 2175–2181, doi: 10.1056/NEJMms1604593, indexed in Pubmed: 27248626.
3. Pocock SJ, Clayton TC, Stone GW. Design of Major Randomized Trials: Part 3 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(24): 2757–2766, doi: 10.1016/j.jacc.2015.10.036, indexed in Pubmed: 26700838.
4. 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(7644): 601–605, doi: 10.1136/bmj.39465.451748.AD, indexed in Pubmed: 18316340.
5. Poolman RW, Struijs PAA, Krips R, et al. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors matter? J Bone Joint Surg Am. 2007; 89(3): 550–558, doi: 10.2106/JBJS.F.00683, indexed in Pubmed: 17332104.
6. Jadad A, Moore R, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinical Trials. 1996; 17(1): 1–12, doi: 10.1016/0197-2456(95)00134-4.
7. Higgins JPT GSe. Cochrane Handbook for Systematic Reviews of Interventions Version 5.2.0 [updated June 2017]. The Cochrane Collaboration, 2017. Available from: http://handbook.cochrane.org.
8. Pocock SJ, McMurray JJV, Collier TJ. Making Sense of Statistics in Clinical Trial Reports: Part 1 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(22): 2536–2549, doi: 10.1016/j.jacc.2015.10.014, indexed in Pubmed: 26653629.
9. Health Technology Assessment Guidelines. The Agency for Health Technology Assessment and Tariff System. 2016. Version 3.0:[Available from: http://www.aotm.gov.pl/www/wp-content/uploads/wytyczne_hta/2016/20161104_HTA_Guidelines_AOTMiT.pdf.
10. Li L, Pan Z. Progression-Free Survival and Time to Progression as Real Surrogate End Points for Overall Survival in Advanced Breast Cancer: A Meta-Analysis of 37 Trials. Clin Breast Cancer. 2018; 18(1): 63–70, doi: 10.1016/j.clbc.2017.07.015, indexed in Pubmed: 28818493.
11. Panasiuk AWR, Budasz-Świderska M, Kaczor M. Cancer immunotherapy in second-line treatment of non-small cell lung cancer — is there a need to change the approach to the assessment of clinical

benefits? Journal of Helath Policy & Outcomes Research. 2017; 2(2): 65–77, doi: 10.7365/JHPOR.2018.1.9.
12. Wilson MK, Collyar D, Chingos DT, et al. Outcomes and endpoints in cancer trials: bridging the divide. Lancet Oncol. 2015; 16(1): e43–e52, doi: 10.1016/S1470-2045(14)70380-8, indexed in Pubmed: 25638556.
13. Wilson MK, Karakasis K, Oza AM. Outcomes and endpoints in trials of cancer treatment: the past, present, and future. Lancet Oncol. 2015; 16(1): e32–e42, doi: 10.1016/S1470-2045(14)70375-4, indexed in Pubmed: 25638553.
14. Pocock SJ, Clayton TC, Stone GW. Challenging Issues in Clinical Trial Design: Part 4 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(25): 2886–2898, doi: 10.1016/j.jacc.2015.10.051, indexed in Pubmed: 26718676.
15. Mauri L, D'Agostino RB. Challenges in the Design and Interpretation of Noninferiority Trials. N Engl J Med. 2017; 377(14): 1357–1367, doi: 10.1056/NEJMra1510063, indexed in Pubmed: 28976859.
16. West CP, Dupras DM. 5 ways statistics can fool you — tips for practicing clinicians. Vaccine. 2013; 31(12): 1550–1552, doi: 10.1016/j.vaccine.2012.11.086, indexed in Pubmed: 23246309.
17. Elwood JM. Interpreting clinical trial results: seven steps to understanding. Can Med Assoc J. 1980; 123(5): 343–345, indexed in Pubmed: 7260774.
18. Case LD, Kimmick G, Paskett ED, et al. Interpreting measures of treatment effect in cancer clinical trials. Oncologist. 2002; 7(3): 181–187, indexed in Pubmed: 12065789.
19. Pocock SJ, Travison TG, Wruck LM. How to interpret figures in reports of clinical trials. BMJ. 2008; 336(7654): 1166–1169, doi: 10.1136/bmj.39561.548924.94, indexed in Pubmed: 18497415.
20. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010; 1(4): 274–278, doi: 10.4103/0974-7788.76794, indexed in Pubmed: 21455458.
21. Fendler WCJ, Młynarski W. Methods of survival analysis applied in oncology — assumptions, methods and common pitfalls. Onkologia w Praktyce Klinicznej. 2011; 7(2): 89–101.
22. Barraclough H, Simms L, Govindan R. Biostatistics primer: what a clinician ought to know: hazard ratios. J Thorac Oncol. 2011; 6(6): 978–982, doi: 10.1097/JTO.0b013e31821b10ab, indexed in Pubmed: 21623277.
23. Pocock SJ, McMurray JJV, Collier TJ. Statistical Controversies in Reporting of Clinical Trials: Part 2 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(23): 2648–2662, doi: 10.1016/j.jacc.2015.10.023, indexed in Pubmed: 26670066.
24. Panasiuk AHM, Pawlik D, Prząda-Machno P, et al. Approach to uncertainty in health technology assessment in a Central and Eastern European country: appraisal of cancer drugs by a Polish HTA agency in presence of high crossover rates in clinical trials. Journal of Helath Policy & Outcomes Research. 2016; 10(2): 17–34, doi: 10.7365/JHPOR.2016.2.2.
25. Haslam A, Prasad V. When is crossover desirable in cancer drug trials and when is it problematic? Ann Oncol. 2018; 29(5): 1079–1081, doi: 10.1093/annonc/mdy116, indexed in Pubmed: 29648572.