Alergia Astma Immunologia, 2003, 8(1), 25-32

Istotne znaczenie statystyki w badaniach medycznych

Statistical issues – significantly important in medical research

M. GELLERSTEDT

Department for Studies of Work, Economics and Health, University of Trollhättan, Uddevalla

Reprinted from: **Allergy 2002; 75: 76-82**

Naukowcy zajmujący się medycyną mogą traktować badania głównie jako projekty badawcze realizowane z zastosowaniem metod medycznych i prowadzone przez personel medyczny. Niemniej, w razie konieczności, badania te mogą zawierać elementy z innych dziedzin nauki. Na przykład często wymagają one obliczeń statystycznych i jeżeli brak jest dostępu do prostego w użytkowaniu programu statystycznego, istnieje konieczność pomocy ze strony statystyka. Wiara w to, że statystyka jest tożsama z analizą prowadzi do zjawiska, w którym ze statystykiem kontaktujemy się na samym końcu badania, kiedy nadchodzi czas na „matematyczne akrobacje". Niestety, bardzo często jest już zdecydowanie za późno; nawet najbardziej skomplikowane metody statystyczne nie mogą wyeliminować błędów popełnionych na pierwszych istotnych etapach np. podczas projektowania badania lub ustalania wielkości próby. W konsekwencji wiele publikacji w czasopismach medycznych jest słabych lub wręcz nieprawidłowych pod względem statystycznym [1-3], a wiele badań nie jest publikowanych ze względu na nieprawidłowe zastosowanie metod statystycznych. Krótko mówiąc, badacze medyczni często nie są świadomi istotności i szerokiego zastosowania statystyki w swoich badaniach, mimo, że wiadomym jest, jak istotną rolę pełni ona w badaniach medycznych. W 1983 roku 70% artykułów opublikowanych w *New England Journal of Medicine* wykorzystywało analizę statystyczną [4]. Od tej pory wykorzystywanie metod statystycznych wzrosło i obecnie stanowi pewien standard.

Statystycy mogą postrzegać badania kliniczne raczej jako badania oparte na koncepcjach statystycznych, w których wykorzystuje się klinikę niż jako badania medyczne z zastosowaniem statystyki. Z tego punktu widzenia frustrujący może być brak ich uczestnictwa na etapie planowania i stawiania pytań dotyczących projektu, które w opinii statystyka powinny być postrzegane jako problemy statystyczne. Statystycy mogą w wielu przypadkach bardzo łatwo wskazać błędy popełnione w badaniach, a czasami brak statystycznej precyzji budzi ich niepokój.

Nie będziemy tu dyskutować dlaczego naukowcy i statystycy różnią się punktem widzenia. Jako statystyk mogę jedynie przyznać, że statystycy, co jest zrozumiałe, ponoszą dużą odpowiedzialność za jakość badania, zwłaszcza w kontekście marketingu i prezentacji całości teorii. Nadszedł czas na zacieśnienie współpracy i stworzenie wspólnego punktu widzenia. Zarówno lekarze, jak i statystycy muszą zrozumieć, że każdy etap badania, jeżeli ma mieć ono wysoką wartość naukową, wymaga zarówno opracowania od strony medycznej, jak i statystycznej. W poniższym artykule zostaną przedstawione i wyjaśnione podstawowe koncepcje statystyczne wykorzystywane w badaniach. Projektowanie badań zostanie omówione w skrócie, natomiast dokładnie omówione zostanie pojęcie istotności statystycznej. Główny nacisk położony zostanie na doświadczalne kliniczne badania potwierdzające (*experimental confirmatory clinical trial*), ale uwzględnione zostaną także badania opisowe np. kliniczno-kontrolne (*case-control studies*). Moim celem jest udowodnienie tezy, że statystyka jest istotna na wszystkich etapach prowadzenia badania. Co więcej, chcę pomóc czytelnikowi lepiej zrozumieć statystykę, jej podstawowe założenia, dać pewne wskazówki i wytyczne, aby pomóc w wykonywaniu badań o wysokiej jakości.

* Opublikowano w Allergy, 2002; 75: 76-82 i przedrukowano za pozwoleniem i dzięki uprzejmości Blackwell Munksgaard
* Reprinted from Allergy, 2002; 75: 76-82 with kind permission of Blackwell Munksgaard
Tłumaczenie: lek. med. Joanna Makowska

Znaczenie dobrego projektu badania

Uznaje się powszechnie, że dobre badanie odpowiada na istotne i interesujące pytania, dając precyzyjną odpowiedź. W badaniach potwierdzających (*confirmatory studies*) naukowcy chcą wykorzystać wyniki otrzymane w pewnej grupie pacjentów w celu uogólnienia wyciągniętych wniosków na wszystkich pacjentów. Można powiedzieć, że jest to zamiar zmiany czasu, z czasu przeszłego na czas teraźniejszy. Chcemy przenieść wyniki z tego jak było (w grupie badanych pacjentów) na to jak jest (w całej populacji pacjentów). Aby móc robić takie uogólnienia – zmieniać czasy – konieczne jest odpowiednie zaprojektowanie badania. Uzasadnienie analizy zależy od sposobu zbierania danych [5]. Niestety element ten jest często zaniedbywany w badaniach medycznych. Być może zależy to od błędu pedagogicznego, jakim jest położenie niedostatecznie silnego nacisku na projektowanie badań podczas zajęć ze statystyki medycznej [6]. Zaprojektowanie badania nie jest łatwym zadaniem: „Istnieje tylko kilka sposobów, aby dobrze zaprojektować badanie, natomiast źle można je zaprojektować na tysiące sposobów" [7] i wymaga wiele wysiłku. Ale jeżeli będziemy pamiętać, że wystarczy jeden błąd, aby wypaczyć wnioski, warto jest włożyć trochę wysiłku w ten etap. Przedyskutujemy teraz pewne podstawowe problemy.

Główne cele

Istotną sprawą jest jasne i dokładne sformułowanie celów badania. Można to wykonać stawiając jasną hipotezę lub opisując zjawisko jakie ma być badane. Zgodnie z zaleceniem *International Conference of Harmonization* [8] „badanie potwierdzające jest badaniem, w którym postawiona na początku hipoteza jest następnie weryfikowana". Oczywiście jest to istotne zarówno z praktycznego, jak i statystycznego punktu widzenia. Jeżeli główne cele i hipotezy zostaną sformułowane na samym początku, to wyniki przeprowadzonej analizy będą miały dużo większa wartość naukową. Problem ten będzie przedyskutowany szerzej w rozdziale „Inne dylematy dotyczące istotności statystycznej". Sprecyzowanie głównego celu pociąga za sobą zdefiniowanie głównych zmiennych, które będą mogły dostarczyć najbardziej istotnych klinicznie i przekonujących dowodów bezpośrednio związanych z głównym celem badania. Mówiąc o głównej zmiennej (zmienna efektywna, zmienna wynikowa, główny efekt końcowy) musimy dokładnie rozważyć co i w jaki sposób zamierzamy zmierzyć.

Załóżmy, że chcemy zbadać wpływ nowego sposobu leczenia na obniżenie ciśnienia krwi. Czy powinniśmy mierzyć ciśnienie rozkurczowe po leczeniu czy też jako badany parametr wykorzystać zmiany ciśnienia rozkurczowego w stosunku do wartości wyjściowych? Mierzony na stojąco czy leżąco? Na którym ramieniu? Po jakim czasie stosowania leczenia? Czy pomiary powinny być wykonywane za pomocą manometru rtęciowego czy aparatami elektronicznymi? Czy pomiary powinny być powtarzane u jednego pacjenta? A jeżeli tak, to ile razy? Nawet tak zwykła zmienna jak ciśnienie krwi stwarza wiele problemów. Statystyk, nawykły do oceny jakości zmiennej np. jej mocy i wiarygodności [9], może na tym etapie badania okazać się cennym partnerem w dyskusji. W wielu badaniach wykorzystywane są zastępcze efekty końcowe (*surrogate endpoint*) [10,11]. Na przykład wysokie ciśnienie krwi samo w sobie nie jest problemem, ale może stać się przyczyną powikłań. Dlatego też, w tych przypadkach, wykorzystujemy ciśnienie krwi jako zastępczy efekt końcowy (*surrogate endpoint)*, po prostu badanie to jest prostsze i szybsze do wykonania i nie wymaga tak wielu pacjentów niż gdy za pierwotne zmienne przyjmiemy zawał mięśnia sercowego czy śmiertelność w grupie pacjentów.

Wiarygodność zmiennej i oczekiwane efekty lecznicze są również istotnymi podstawowymi punktami służącymi zaprojektowaniu badania, np. określenie liczebności wielkości próby. W badaniach opisowych (*observational studies*), głównie w badaniach opartych na kwestionariuszach, spotykamy się z tym samym typem problemów operacyjnych. Musimy użyć wystandaryzowanego kwestionariusza [12] lub stworzyć nowe wiarygodne, uzasadnione pytania i skale oceny [13]. Podsumowując, na początku badania istotną rzeczą jest zdefiniowanie pierwotnego celu i sformułowanie go w postaci hipotezy testującej wyniki oraz oszacowanie i zdefiniowanie odpowiednich zmiennych pierwotnych. Statystyka odgrywa w tej pierwszej fazie rolę po pierwsze w uzasadnieniu wyboru zmiennej i po drugie w przewidywaniu możliwych analiz i oszacowaniu wielkości próby.

Wykorzystanie grupy kontrolnej

Wielu badaczy myśli o wyniku jako różnicy pomiędzy wartościami wyjściowymi a odpowiednimi wartościami zaobserwowanymi po zastosowaniu leczenia. Ale dotyczy to tylko jednego przypadku, sytuacji kiedy możemy zagwarantować, że nic poza leczeniem nie może wpłynąć na badaną zmienną. Tak więc wynik powinien być traktowany jako różnica pomiędzy wartością wyjściową i wartością po zastosowaniu leczenia, porównaną z różnicą, jaką obserwujemy bez zastosowania leczenia [14]. Oczywiście nie jest możliwe w tym samym czasie leczyć i nie leczyć pacjenta. Zadaniem grupy kontrolnej jest oszacowanie, co się może wydarzyć, jeżeli nie zastosujemy leczenia. Prawdziwy efekt leczniczy jest więc określany na podstawie porównania wyników uzyskanych w grupie leczonej z wynikami z grupy kontrolnej. Tak więc uzasadnieniem zastosowania grupy kontrolnej jest fakt, że zmiany dotyczące zmiennej mogą pojawić się bez żadnego leczenia. Zmiany te mogą być związane ze zjawiskiem równania do średniej [15-17]. Zazwyczaj istnieje tendencja, aby rozpoczynać leczenie

kiedy zmienna osiąga najwyższe wartości. Co więcej, zazwyczaj w projekcie badania zawarte są pewne ograniczenia wartości wyjściowych. Na przykład, aby zostać włączonym do badania ciśnienia krwi, pacjent musi mieć ciśnienie rozkurczowe przynajmniej 90mmHg. Ciśnienie jest zmienną randomizowaną (jej wartość zmienia się), naturalnie podnosi się i opada [18]. Jeżeli średnia wartość ciśnienia rozkurczowego pacjenta wynosi 85mmHg nie jest wykluczone, że od czasu do czasu ciśnienie u pacjenta przekroczy 90mmHg. Tak więc pacjent taki mógłby zostać włączony do badania, gdyby ciśnienie zostało zmierzone tylko w tym momencie. Spadek ciśnienia rozkurczowego (DBP) po terapii może być częściowo tłumaczony faktem, że DBP powróciło do wartości średniej (a nawet dolnej). Rozważmy duże przedsiębiorstwo, w którym nagle odnotowano, że wielu pracowników jest nieobecnych w pracy z powodu choroby. W tym momencie przedsiębiorstwo podejmuje pewne strategie zdrowotne. Sześć miesięcy później liczba osób nieobecnych z powodu choroby wraca do normalnego poziomu. Czy było to związane z zastosowanymi środkami czy też po prostu nastąpił powrót do średniej? Istnieją możliwości analizy czy zmiana była związana z wartościami początkowymi czy nie [19,20]. Zmiany wartości mogą być również związane z naturalnym trendem rozwoju choroby (historią naturalną). Rozważmy grupę pacjentów z obturacyjną chorobą płuc. Załóżmy, że badamy tę grupę chorych przez 10 lat. Oczekujemy, że pojemność płuc będzie się progresywnie obniżała. Jeżeli pojemność płuc się nie zmieni po 10 latach leczenia, różnica pomiędzy wartościami początkowymi a wartościami po leczeniu wyniesie zero. Ale nie byłoby dobrze zakładać, że nie osiągnięto efektu leczniczego, ponieważ gdyby pacjent został pozostawiony bez leczenia doszłoby do obniżenia wartości pojemności płuc. Wartości badane mogą ulec zmianom pod wpływem efektu *placebo* np. podświadomego wpływu związanego z oczekiwaniami pacjenta i obserwatora. Problem ten zostanie przedyskutowany w dalszym rozdziale „ślepe próby".

W badaniach doświadczalnych łatwo jest znaleźć grupę kontrolną, poprzez np. losowy dobór (randomizację) połowy pacjentów do grupy kontrolnej. Niemniej w niektórych badaniach wybór grupy kontrolnej nie jest tak oczywisty. W badaniach kliniczno-kontrolnych (*case-control study)* wybór kontroli jest kluczowym momentem, który wymaga analizy statystycznej (patrz „badania opisowe – badania kliniczno-kontrolne"). Istnieją inne możliwości np. kontrola historyczna [21,22], która również musi zostać oceniona pod kątem tendencyjności.

Podsumowując, zmiany wartości mogą pojawić się z kilku różnych powodów; wykorzystanie grupy kontrolnej pozwala nam różnicować pomiędzy zmianami związanymi z leczeniem i zmianami spowodowanymi innymi czynnikami.

Randomizacja (losowy dobór badanych)

Istnieją trzy istotne przesłanki, aby dokonywać losowego doboru badanych (randomizację). Po pierwsze, randomizacja jest najbardziej obiektywną metodą umieszczenia pacjenta w grupie aktywnej (poddanej nowej terapii) lub grupie kontrolnej (*placebo* lub alternatywna metoda leczenia). Jeżeli badacze ingerują w proces umieszczania pacjentów w grupach można popełnić błąd selekcji (*selection bias)* np. tendencja do umieszczania pacjentów w sposób, który faworyzuje grupę leczoną nową metodą [23,24]. Po drugie, randomizacja jest obiektywna w tym sensie, że w większości przypadków tworzone grupy są równe sobie pod względem średnich, zarówno czynników znanych jak i nieznanych – porównywalne grupy. Analiza statystyczna nie wymaga idealnie identycznych grup, gdyż uwzględnia zmienność występującą zarówno wewnątrz, jak i pomiędzy grupami. W gruncie rzeczy, większość wnioskowania statystycznego, takiej jak testy, przedziały ufności ma znaczenie tylko wtedy, gdy została wykonana randomizacja – i to jest trzeci powód dla jej przeprowadzenia. Nierandomizowane badania wymagają specjalnych technik statystycznych [25]. Prosta randomizacja oznacza, że każda osoba jest niezależnie, losowo przydzielana do jednej z grup, najczęściej z jednakowym prawdopodobieństwem umieszczenia w którejś z nich. W sytuacji kiedy ma znaczenie otrzymanie grup jednakowej wielkości w różnych ośrodkach można zastosować randomizację blokową (*blocked randomization*) [26]. Aby osiągnąć równowagę pomiędzy grupami, biorąc pod uwagę niektóre istotne cechy charakterystyczne pacjentów, można wykorzystać rzut monetą, randomizację warstwową (*stratified randomization*) lub minimalizację *(minimalization*) [27-30].

Zaślepienie próby

Badania kliniczne często są podwójnie ślepe, co oznacza, że ani badający ani badany nie wie, czy otrzymuje preparat leczniczy czy *placebo*. Celem tego zabiegu jest zminimalizowanie wpływu podświadomości, np. optymistycznego nastawienia badającego (błąd obserwatora). Jest również wiadome, że pacjenci mogą osiągnąć korzyści wiedząc lub wierząc, że są aktywnie leczeni. Efekt ten znany jest jako efekt *placebo* [31]. Przykładem jest badanie wpływu kwasu askorbinowego na przeziębienie [32]. Po leczeniu wykazano korzystny efekt, ale okazało się, że badani otwierali kapsułki i sprawdzali czy otrzymują *placebo* czy lek. Przeprowadzono analizę biorącą pod uwagę „złamanie zasady ślepej próby" i okazało się, że leczenie kwasem askorbinowym nie przynosi efektów.

Badania opisowe – badanie kliniczno-kontrolne

Opisywane powyżej koncepcje mają szczególnie znaczenie w badaniach doświadczalnych. Istnieje niemniej wiele sytuacji, w których nie jest możliwe zastosowanie projektu eksperymentalnego badania. Randomizacja grup na „palącą" i „niepalącą" nie byłaby możliwa z przyczyn etycznych (prawie tak niemożliwe, jak randomizacja względem płci). Tak więc, aby zbadać czy taka zmienna jest związana swoiście z pewnymi uwarunkowaniami zdrowotnymi należy wykorzystać badania opisowe (*observational studies*). Istnieje kilka możliwych projektów takich badań [33]. Jednym z najbardziej popularnych jest badanie kliniczno-kontrolne (*case-control study*) [34,35]. Słabością tego projektu jest fakt, że zidentyfikowano 35 różnego rodzaju błędów, które mogą być popełnione w trakcie jego realizacji [36]. Podstawowym problemem jest wybór właściwej grupy kontrolnej. Generalną zasadą jest wybór do grupy kontrolnej osób, które mogłyby być również włączone do badania [37]. Podążając za tą zasada, często używa się kontroli kojarzonych (*match control*), grupowo lub indywidualnie. Wykorzystywanie jako kontroli pacjentów z tego samego szpitala, ale hospitalizowanych z innych przyczyn, może być wygodne, ale może prowadzić do niedoszacowania związku pomiędzy możliwymi czynnikami ryzyka i warunkami zdrowotnymi w badaniu. Inną alternatywą jest wykorzystanie randomizowanej próbki populacji jako grupy kontrolnej. Teoretycznie jest to dobry pomysł, ale trudny do realizacji w praktyce, zwłaszcza jeżeli oczekujemy swoistego rozłożenia niektórych cech np. wieku czy płci. W badaniu kliniczno-kontrolnym istotne może być wykorzystanie ślepej próby odnośnie obserwatora. Jeżeli np. badamy możliwy związek pomiędzy nietolerancją pokarmową a pracą w dzień lub w nocy, może okazać się istotne wykonanie diagnostyki nietolerancji pokarmowej bez znajomości faktu czy badany pracuje w dzień czy w nocy. W tym przypadku zaślepienie próby chroni przed błędem obserwatora. Celem badania kliniczno-kontrolnego często jest ocena czynników ryzyka. W tym przypadku badanie kliniczno-kontrolne może być bardzo cenne. Niemniej interpretacja wyników tych badań, ze względny na możliwość popełnienia wielu błędów, powinna być bardzo ostrożna [38]. Co więcej, związek pomiędzy czynnikiem ryzyka a czynnikiem zdrowia wykazany w badaniu kliniczno-kontrolnym nie może być postrzegany sam przez się jako dowód związku przyczynowo-skutkowego. Aby stwierdzić związek przyczynowo-skutkowy należy spełnić jeszcze kilka warunków, a do momentu jego udowodnienia bardziej właściwe byłoby używanie określenia „wskaźnik ryzyka" a nie „czynnik ryzyka" [39-42].

Podsumowując, badanie opisowe np. badanie kliniczno-kontrolne, w wielu przypadkach jest jedyną możliwą opcją. Istotne jest oszacowanie możliwego błędu i przeprowadzenie badania w ten sposób, że wyeliminuje się tyle błędów, ile to możliwe. Ten aspekt pracy jest kluczowy zarówno dla statystyka, jak i badacza.

Wybór projektu badania

Kontrolowane badania randomizowane z podwójnie ślepą próbą (*a controlled randomized double-blind study*) są często postrzegane jako złoty standard i są najbardziej akceptowane naukowo. Oczywiście istnieje pewna hierarchia różnych projektów [43]. Uważam jednak, że wszystkie projekty badań mogą dostarczyć cennych informacji, jeżeli wykorzystuje się je ostrożnie. Ostatnio przeprowadzone badanie wykazało, że informacje uzyskane z kilku różnych projektów badań mogą być porównywalne [44]. Rozważenie wszystkich za i przeciw poszczególnych projektów jest dobrym punktem startu i służy opracowaniu różnych aspektów statystycznych, medycznych, etycznych, praktycznych i finansowych. Jeżeli wybraliśmy już konkretny projekt, trzeba zastanowić się, jak wyeliminować potencjalne błędy. Chociaż może to zabierać wiele czasu, należy pamiętać, że nawet jeden błąd może wypaczyć wyniki.

Istotność statystyczna – raczej puste określenie

Załóżmy, że wyzwałeś kolegę na 10 gier w turnieju szachowym. Jeżeli jesteście równie dobrzy spodziewać się można wyniku 5 do 5 lub podobnego. Załóżmy, że turniej kończy się wynikiem 3-7. Nie ma wątpliwości, że w tym turnieju twój kolega był lepszy – aby to stwierdzić nie potrzeba żadnej statystyki. Ale pozostaje wątpliwość czy wynik ten można uogólniać. Czy możemy zmienić czas przeszły na teraźniejszy? Kolega był lepszy ale czy jest lepszy*?* Aby odpowiedzieć na to pytanie musimy przeanalizować, jakie jest prawdopodobieństwo przypadkowego ukończenia gry z wynikiem 7 lub więcej wygranych, zakładając, że obaj gracze są jednakowo dobrzy. Jak wspomnieliśmy wyżej, w takim przypadku oczekiwalibyśmy wyniku 5-5 pomiędzy graczami na tym samym poziomie, ale w rzeczywistości istnieje prawdopodobieństwo 0,34, że jeden z graczy (jednakowo dobrych) wygra 7 lub więcej gier, w wyniku zadziałania ślepego losu. Tak więc tłumaczyłbym ten wynik jako przypadek a nie dowód na to, że któryś gracz był gorszy. Ale jeżeli zmienimy wynik na 9-1, co odpowiada prawdopodobieństwu 0,02, wtedy tłumaczenie gorszego gracza (takiego jak ja), że jest to wynikiem ślepego losu lub braku szczęścia byłoby żałosne. W tym wypadku musiałbym przyznać, że kolega rzeczywiście jest lepszy.

W tym przykładzie hipoteza zerowa jest taka, że obaj gracze są jednakowo dobrzy. Dla danego wyniku można obliczyć wartość P, która określa prawdopodobieństwo osiągnięcia równie dobrego lub lepszego wyniku, zakładając, że hipoteza zerowa jest prawdziwa. Na przykład przy wyniku 1-9 wartość P wynosi 0,02 co oznacza, że prawdopodobieństwo wystąpienia tak dużej różnicy pomiędzy

dwoma graczami na jednakowym poziomie (hipoteza zerowa) wynosi zaledwie 0,02. W badaniach medycznych standardowo odrzuca się hipotezę zerową, jeżeli wartość P jest niższa niż 0,05. Co oznacza, że ryzyko błędnego odrzucenia hipotezy zerowej wynosi 5%, np. poziom istotności wynosi 5%.

Przeanalizujmy dwa inne przykłady. Jeżeli testujemy różnicę pomiędzy dwoma sposobami leczenia i wartość P wynosi 0,02 prowadzi to do wniosku, że te dwa sposoby lecznicze różnią się efektywnością. Jeżeli badamy potencjalny czynnik ryzyka w badaniu kliniczno-kontrolnym i wartość P jest mniejsza niż 0,05 możemy wnioskować o wystąpieniu związku (ale nie przyczynowości). Proszę zauważyć, że wnioski wyrażane są w czasie teraźniejszym. Podstawy logiczne testowania hipotezy mogą wydawać się proste. Ale koncepcja ta prowadzi często do niewłaściwego zrozumienia i błędnej interpretacji wyniku. Wartość P równa 0,05 jako graniczna, poniżej której nie ma racjonalnych wątpliwości, że hipoteza zerowa jest prawdziwa, jest wartością ustaloną arbitralnie. Gdy P wynosi 0,049 powiedziałbym, że wartość dowodowa nie jest większa niż w przypadku P równego 0,051, ale ponieważ standardowo wartością graniczną jest P 0,05 w jednym przypadku będziemy mówić o istotności statystycznej, a w drugim nie. W rzeczywistości może to być różnica pomiędzy wartością istotną statystycznie i nieistotną opisaną w artykule, co gorsza może taka różnica wystąpić pomiędzy wynikami opublikowanymi i nieopublikowanymi. Faktem jest, że jeżeli wyniki nie wykazują istotności statystycznej mają dużo mniejsze szanse na opublikowanie jest to tzw. tendencyjne publikowanie [45,46]. Co więcej, istotny statystycznie efekt leczenia nie daje nam informacji na temat wielkości efektu, tak więc nie możemy wnioskować, czy jest on istotny klinicznie czy nie. Załóżmy, że badane jest ciśnienie rozkurczowe i porównywane są dwie metody leczenia. Wyobraźmy sobie badanie, w którym różnica ciśnienia po zastosowaniu dwóch sposobów leczenia wynosi 0,6mmHg i mamy wąski przedział ufności z błędem standardowym 0,2. Oznacza to, że prawdziwa różnica wartości waha się od 0,4 do 0,8mmHg. Ponieważ nie występuje tu wartość 0, oznacza to, że obserwowana różnica jest istotna statystycznie. Ale różnica 0,4 do 0,8 mmHg, nie jest zbytnio istotna klinicznie, tak więc obie metody leczenia można uznać za równorzędne pod względem skuteczności. Jeżeli wynik nie jest istotny często jest błędnie interpretowany, jako nie mający wpływu klinicznego, ale jest to błędna konkluzja. Wróćmy do turnieju szachowego: jeżeli zakończy się on wynikiem 3-7 nie jest to wynik istotny statystycznie. Tak więc nie możemy wnioskować, że kolega jest lepszym graczem, ale nie możemy wnioskować również o równym poziomie szachistów. Jedynym możliwym wnioskiem jest stwierdzenie, że nie mamy dość dowodów, aby wykazać różnicę pomiędzy graczami. Analizując 71 opublikowanych badań, w których nie stwierdzono istotnych statystycznie wyników wykazano (poprzez obliczenie przedziałów ufności zamiast wartości P), że w prawie połowie badań nastąpiła poprawa o 50% pod wpływem leczenia [47].

Tak więc nieistotny statystycznie wynik nie wyklucza efektu istotnego klinicznie. Można to elegancko opisać w zdaniu „Brak dowodów nie jest dowodem braku związku" [48]. Błędną interpretacją wartości P jest postrzeganie jej jako odzwierciedlającej siłę efektu opisywanego w badaniu. Powszechnie się uważa, że bardzo niskie wartości P, powiedzmy niższe niż 0,001 świadczą o większym efekcie niż badania, w których P było niższe od 0,05. Bardzo duże badania mogą dać małe wartości P, nawet jeżeli obserwowana różnica jest nieduża. Tak więc bardzo niskie wartości P nie świadczą o dużym zaobserwowanym efekcie. Ta błędna interpretacji wartości P jest przyczyną, dla której zaleca się wykorzystywanie przedziałów ufności zamiast wartości P w prezentacji wyników [49]. Przedział ufności zawiera informacje o wielkości obserwowanego efektu, jak i informacje czy wynik był istotny statystycznie. Tak więc wyniki istotne statystycznie mogą być znaczące lub nieznaczące klinicznie, a nawet klinicznie równoważne (rzadka sytuacja ale możliwa). Co więcej, brak istotności statystycznej nie może wykluczyć możliwej skuteczności klinicznej. Tak więc wyrażenie „istotny statystycznie" jest raczej pustym wyrażeniem. Zalecane jest użycie przedziałów ufności w prezentacji, ponieważ pozwalają one na interpretowanie wartości statystycznych w perspektywie klinicznej.

Wielkość próby

Podczas testowania hipotezy można popełnić 2 rodzaje błędów. Po pierwsze możemy odrzucić hipotezę zerową mimo, że jest prawdziwa (błąd I rodzaju). Wykorzystanie poziomu istotności 0,05 oznacza, że prawdopodobieństwo popełnienia błędu I typu wynosi 0,05. Po drugie, możemy zaakceptować hipotezę zerową pomimo, że jest błędna (błąd II rodzaju). Zamiast rozważać prawdopodobieństwo popełnienia błędu II rodzaju powszechnie wykorzystuje się moc testu, czyli prawdopodobieństwo prawidłowego odrzucenia błędnej hipotezy zerowej. W naszym przykładzie turnieju szachowego, moc testu jest prawdopodobieństwem osiągnięcia istotnego statystycznie wyniku (wyciągnięcia wniosku, że jeden gracz jest lepszy), przy założeniu, że jeden z graczy jest rzeczywiście lepszy. Jeżeli porównujemy dwie metody lecznicze, moc oznacza prawdopodobieństwo otrzymania istotnego statystycznie wyniku, jeżeli prawdziwa jest różnica pomiędzy dwiema metodami leczenia. Krótko mówiąc, jest to prawdopodobieństwo udowodnienia istotności statystycznej, jeżeli różnica rzeczywiście istnieje.

Jest jasne, że moc jest zależna od wielkości obserwowanej różnicy. Na przykład, jakie jest prawdopodobieństwo, że turniej zakończy się istotnym statystycznie wynikiem, jeżeli jeden z graczy jest lepszy. Zależy to od

tego czy gra on dużo lepiej od przeciwnika. Jeżeli kolega jest jednym z 10 najlepszych szachistów świata a ty dopiero poznałeś zasady gry, to moc próby jest duża nawet jeżeli zostanie rozegranych tylko 10 partii. Ale jeżeli kolega jest tylko nieznacznie lepszy od ciebie prawdopodobieństwo udowodnienia tego w 10 partiach jest nieduże (na szczęście dla ciebie).

Moc zależna jest także od poziomu istotności, zmienności zmiennej cechy i wielkości próby. Duże próby mogą prowadzić do istotnych wyników, nawet przy niewielkiej różnicy. Ale należy sobie zadać zawsze pytanie, czy warto jest czynić wysiłki, żeby udowodnić tak niewielką różnicę. Nie powinno się także wykorzystywać zbyt małych prób, jeżeli obserwowany efekt ma istotne znaczenie kliniczne, ponieważ nie wykażą one istotności statystycznej. Jak wspomnieliśmy wcześniej analiza badań klinicznych pokazała, że prawie połowa badań, które nie wykazały istotności statystycznej potencjalnie wskazywało na 50% efekt poprawy po leczeniu. Inne badanie wykazało, że większość badań nieistotnych statystycznie nie miało próby odpowiedniej wielkości do wykazania względnych różnic wielkości 25-50% pomiędzy poszczególnymi metodami leczenia [50].

Przed rozpoczęciem badania należy dokładnie rozważyć moc badania. Warto wspomnieć, że kalkulacje mocy [51,52] w większości przypadków opiera się na szacowaniu różnicy i zmienności, co pociąga za sobą to, że moc i wielkość próby są również przybliżone. Tak więc dobrze jest przedyskutować otrzymane wyniki w kontekście wielkości próby. Na przykład, zmienność może być wyższa niż oczekiwana, co może tłumaczyć fakt, że otrzymane wyniki nie są istotne statystycznie, nawet jeżeli obserwowana różnica ma znaczenie kliniczne i jest tak duża jak oczekiwana.

Inne dylematy dotyczące istotności statystycznej

Jeżeli przebadamy kilka różnych zmiennych i wykonamy test istotności dla każdej z nich, to ryzyko fałszywego odrzucenia przynajmniej jednej z nich jest całkiem duże. Załóżmy, że wykonujemy randomizowane badanie 2 grup i stosujemy w nich dokładnie takie samo leczenie. Jeżeli analizujemy zmienne niezależne, prawdopodobieństwo otrzymania przynajmniej jednej różnicy istotnej statystycznie jest większe niż 50%, nawet jeżeli wiemy, że leczenie jest *de facto* takie samo. Dlatego też trudno stwierdzić obiektywnie, czy istotność statystyczna zmiennej odkryta podczas analizy wielu zmiennych jest wynikiem przypadku czy też rzeczywistego skutku działania leku. Jeżeli badanie potwierdzające (*confirmatory study)* dotyczy kilku zmiennych pierwotnych możliwe jest dostosowanie się do wielokrotnego testowania [53,54]. W badaniu skuteczności (*explanatory study)* nie potrzeba definiować na początku żadnych hipotez ani pierwotnych zmiennych. Takie podejście może dostarczyć wielu cennych informacji. Ale wyniki te należy bardzo ostrożnie interpretować ze statystycznego punktu widzenia. W związku z kłopotami dotyczącymi wielokrotnego testowania, istotność statystyczną należy traktować jako stworzoną hipotezę, sygnał mówiący nam, że w otrzymanych wynikach może być coś interesującego. Oczywiście wartość dowodu wzrasta, jeżeli istotność statystyczna jest umotywowana medycznie lub jeżeli została ona potwierdzona w innych niezależnych badaniach (potwierdzających lub badaniach skuteczności).

Patrząc na sprawę z innego punktu widzenia, wydaje się, że hipoteza powinna być postawiona na początku. Weźmy przykład małego przedsiębiorstwa zatrudniającego elektryków, w którym zaobserwowano, że proporcja pomiędzy liczbą chłopców i dziewczynek wśród dzieci zatrudnionych osób jest bardzo wysoka (p=0,04). Trudno uznać to za dowód, że elektrycy mają więcej synów niż córek. Jeżeli rozważymy wszystkie przedsiębiorstwa na świecie, będą wśród nich takie z przewagą zarówno jednej jak i drugiej płci, w większości przedsiębiorstw będzie jednak równowaga między dziećmi obu płci, w takim przedsiębiorstwie nikt nie wpadnie na pomysł sprawdzania proporcji. Tak więc wykorzystanie przypadkowego zjawiska w celu stworzenia hipotezy i traktowania jej jako udowodnioną nie ma sensu. Ale jeżeli hipoteza (o związku między elektrykami i płcią dzieci) została stworzona na podstawie wcześniejszych prób i jeżeli przedsiębiorstwo zostało wybrane losowo, wtedy istotność statystyczna miałaby większą wartość dowodową. Tak więc postawienie hipotezy i wybór badanej zmiennej powinny być umotywowane i dokonane przed rozpoczęciem badania.

Wnioski

Po pierwsze, stwierdziliśmy, że statystyka ma zastosowanie już na początku, w fazie planowania badania. Po drugie, na każdym etapie badania można popełnić wiele brzemiennych w skutki błędów. Po trzecie wyjaśniłem, że poszczególne etapy badania są zależne od siebie, tak więc siła analizy zależy od zaprojektowania badania. Te trzy argumenty potwierdzają tezę, że statystyka jest istotna w badaniach medycznych.

Wykazano, że statystycy powinni pełnić aktywną rolę w zespole badawczym, a nie grać drugie skrzypce jako doradcy w rzadkich przypadkach. Współpraca pomiędzy klinicystami a statystykami gwarantuje, że badanie zostanie zaplanowane i wykonane w sposób, który pozwoli na owocną analizę i wyciągnięcie wartościowych wniosków. Idealnie byłoby, gdyby statystyk był obecny w fazie planowania każdego nowego badania, niestety zdaję sobie sprawę, że nie zawsze jest to możliwe.

Dostępnych jest wiele pomocnych materiałów dotyczących metod statystycznych, np. polecane są książki Altmana, Blanda i Campbella i wsp oraz Pococka przeznaczone dla „niestatystyków" [55-58]. Książka Senna dotyczy metod

statystycznych wykorzystywanych w badaniach lekowych i dodatkowo wykazuje, że statystyka i poczucie humoru nie muszą się nawzajem wykluczać [59]. Dostępne są także wytyczne. Opisując badania powinno się zajrzeć do zestawienia CONSORT [60]. Dokładne wytyczne, które zwierają spis wykorzystywany przez kilka czasopism medycznych jest zawarty w książce opublikowanej przez Gardnera i Altmana [61]. Aby wykonać metaanalizę [62] musi być możliwa ocena szczegółów dotyczących każdego projektu; dlatego też opisywanie badania zgodne z wytycznymi ma szczególne znaczenie.

Nawet jeżeli odpowiedzi na niektóre pytania są dostępne w literaturze, mam nadzieję, że przyszłość przyniesie ściślejszą współpracę między statystykami i badaczami medycznymi. Skorzystałyby na tym badania medyczne i statystyczne, jakość nauki i pacjenci.

Piśmiennictwo

1. Altman DG. Statistics in medical journals. Statistics Med 1982; 1: 59-71.
2. Altman DG. Statistics in medical journals. Development in the 1980s. Statistics Med 1991; 10: 1897-1913.
3. Editorial: The scandal of poor medical research. Br Med J 1994; 308: 283-284.
4. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983; 309: 709-713.
5. Schoolman HM, Becktel JM, Best WR, Johnson AF. Statistics in medical research: principals versus practice. J Laboratory Clin Med 1968; 71: 357-367
6. Noller KL, Melton LJ. Study design in perinatal medicine. Am J Perinatol 1985; 2: 250-255.
7. Sacket DL. Rational therapy in the neurosciences: the role of the randomized trial. Stroke 1986; 17: 1323-1329.
8. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Guidelines: Statistical Principals for Clinical Trials (ICH E-9). Available from URL: http://www.ifpma.org/ich5e.html #Design.
9. Campbell MJ, Machin D. Medical Statistics. A common sense approach, 3rd edn. Chichester: Wiley, 1999; 28-31.
10. Fleming T. Surrogate endpoints in clinical trials. Drug Information J 1996; 30: 545-551.
11. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Statistics Med 1989; 8: 431-440.
12. McDowell I, Newell C. Measuring Health. A Guide to Rating Scales and Questionnaires. Oxford: Oxford University Press, 1987.
13. Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to Their Development and Use. Oxford: Oxford University Press, 1989.
14. Senn S. Statistical Issues in Drug Development. Chichester: Wiley, 1997; 28-29.
15. Bland M, Altman DG. Regression towards the mean. Br Med J 1994; 308: 1499.
16. Bland M, Altman DG. Some examples of regression towards the mean. Br Med J 1994; 309: 780.
17. Chuang-Stein C. The regression fallacy. Drug Information J 1993; 27: 1213-1220.
18. Armitage P, Fox W, Rose GA, Tinker CM. The variability of measurements of casual blood pressure II. Survey experience. Clin Sci 1966; 30: 337-344.
19. Yudkin P, Stratton IM. How to deal with regression to the mean in intervention studies. Lancet 1996; 347: 241-243.
20. Hayes RJ. Methods for assessing whether change depends on initial value. Statistics Med 1988; 7: 915-927.
21. Pocock SJ. Randomised clinical trials. Br Med J 1977; i: 1661.
22. Sacks HS, Chalmers TC, Smith H. Sensitivity and specificity of clinical trials: randomized vs historical controls. Arch Intern Med 1983; 143: 753-755.
23. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. New Eng J Med 1983; 309: 1358-1361.
24. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Bias due to non-concealment of randomization and non-double-blinding. J Am Med Assosiation 1995; 273: 408-412.
25. Rothman KJ. Statistics in nonrandomized studies. Epidemiology 1990; 1: 417-418.
26. Pocock SJ. Clinical Trials. Chichester: Wiley 1983; 73-89.
27. Efron B. Forcing a sequential experiment to be balanced. Biometrika 1971; 58: 403-417.
28. Atkinson A. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika 1982; 69: 61-67.
29. Cochran WG. Sampling Techniques. New York: Wiley 1977.
30. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. Clin Pharmacol Therapeutics 1974; 15: 443-453.
31. Byerly H. Explaining and exploiting placebo effects. Pers Biol Med 1976; 19: 423-435.
32. Karlowski TR, Chalmers TC, Frenkel LD et al. Ascorbic acid for the common cold. A prophylactic and therapeutic trail. JAMA 1975; 231: 1038-1042.
33. Bailar JC, Louis TA, Lavori PW, Polansky M. A classification for biomedical research reports. N Engl J Med 1984; 311: 1482-1487.
34. Breslow NE, Day NE. Statistical Methods in Cancer Research, Vol. 1. The Analysis of Case-Control Studies. Oxford University Press, 1993.
35. Schlesselman JJ. Case-control studies. Design Conduct, Analysis. Oxford University Press 1982.
36. Sacket DL. Bias in analytical research. J Chron Dis 1979; 32: 51-63.
37. Rothman KJ, Greenland S. Modern Epidemiology. Philadelphia, PA: Lippincott-Raven 1998.
38. Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. Int J Epidemiol 1988; 17: 680-685.
39. Beck JD. Risk revisited. Community Dent Oral Epidemiol 1998; 26: 220-225.
40. Kleinbaum D, Kupper L, Morgenstern H. Epidemiologic research: principals and quantitative methods. Belmont, CA: Lifetime Learning Publications 1982.
41. Hill B. Principals of medical statistics, 9th edn. New York: Oxford University Press 1971; 309-323.

42. Koch GG, Beck JD. Statistical methodologies useful for the analysis of data from risk-assesment studies. J Public Health Dent 1992; 52: 146-167.
43. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000; 342: 1887-1892.
44. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000; 342: 1878-1886.
45. Easterbrook PJ, Berlin JA. Publication bias in clinical research. Lancet 1991; 337: 867-872.
46. Berlin JA, Colin BB, Louis TA. An assessment of publication bias using a sample of published clinical trials. J Am Stat Assosiation 1989; 84: 381-392.
47. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: survey of 71 'negative' trials. N Engl J Med 1978; 299: 690-694.
48. Altman DG, Bland M. Absence of evidence is not evidence of absence. Br Med J 1995; 311: 485.
49. Gardner MJ, Altman DG. Confidence intervals rather than P-values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746-750.
50. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994; 272: 122-124.
51. Machin D, Campbell MJ, Fayers PM, Pinol APY. Sample Size Tables for Clinical Studies. Oxford: Blackwell Scientific 1997.
52. Desu MM, Rhaghavarao D. Sample Size Methodology. Boston: Academic Press 1990.
53. Bauer P. Multiple testing in clinical trials. Stat Med 1991; 10: 871-890.
54. Pocock SJ, Geller N, Tsiatis A. The analysis of multiple endpoints in clinical trials. Biometrics 1987; 43: 487-498.
55. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall 1991.
56. Bland M. An introduction to medical statistics, 3rd edn. Oxford: Oxford University Press 2000.
57. Campbell MJ, Machin D. Medical Statistics: A common sence approach, 3rd edn. Chichester: Wiley 1999.
58. Pocock SJ. Clinical Trials. Chichester: Wiley 1983.
59. Senn S. Statistical Issues in Drug Development. Chichester: Wiley 1997.
60. Begg C, Cho M, Eastwood S et al. Improving the quality of reporting of randomized controlled trials. CONSORT Statement. JAMA 1996; 276: 637-639.
61. Gardner MJ, Altman DG, eds. Statistics, with confidence. Confidence intervals and statistical guidelines. London: British Medical Journal 1989.
62. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. New Eng J Med 1987; 316: 450-455.