

Artykuł dyskusyjny

Jak często piśmiennictwo lekarskie może być źródłem błędnych decyzji klinicznych?

Krzysztof Bujko¹, Wojciech Michalski², Marek P. Nowacki¹

W praktyce klinicznej, na zmianę postępowania wpływają przede wszystkim badania pozytywne – wyższość badanej metody jest potwierdzona statystycznie istotnym wynikiem testu – podczas gdy badania negatywne są często ignorowane. Z przeprowadzonej symulacji komputerowej wynika, że odsetek badań fałszywie pozytywnych wśród wszystkich pozytywnych, może być duży – nawet około 30% – 50%. Zależy on przede wszystkim od odsetka doświadczeń, w których jedna z metod jest rzeczywiście lepsza; w mniejszym stopniu zależy od poziomu błędu II rodzaju (liczebności chorych w badaniu). Wysoki odsetek badań fałszywie pozytywnych wśród wszystkich pozytywnych, oznacza również wysokie ryzyko podjęcia błędnej decyzji wyboru sposobu leczenia na podstawie piśmiennictwa. Dodatkowo, ryzyko to wzrasta na skutek większej skłonności do publikowania prac z pozytywnymi wynikami, w porównaniu z pracami zawierającymi wyniki negatywne (publication bias). A zatem, na wzrost ryzyka podjęcia błędnej decyzji klinicznej w oparciu o piśmiennictwo wpływa: 1. Wybiórcze czytanie piśmiennictwa – przywiązywanie większej wagi do wyników pozytywnych niż negatywnych, 2. Większa skłonność publikowania wyników pozytywnych niż negatywnych. 3. Mały odsetek badań testujących metody leczenia o rzeczywistej przewadze nad dotychczas stosowanymi. 4. Zbyt mała liczebność chorych (duży poziom błędu II rodzaju) w badaniach klinicznych. 5. Zbyt mała liczba badań potwierdzających.

How often may medical literature be a source of incorrect clinical decisions

Clinical practice is influenced mainly by positive trials (superiority of the tested method confirmed by significant result of statistical test), whereas negative trials are unfortunately often ignored. Computer simulation has shown that the rate of false positive trials might be high among all positive trials – even up to 30%-50%. This percentage depends mainly on the rate of trials with a real difference between tested methods; in a lesser extent it depends on a level of type II error (number of patients enrolled in a trial). The high rate of false positive trials among all positive trials indicates that the risk of undertaking an incorrect decision based on literature is also high. In addition, this risk is also increased due to publication bias. Thus, the risk of undertaking an incorrect decision based on literature is increased by: (i) selective reading of literature – paying more attention to positive than to negative results; (ii) publication bias; (iii) low rate of trials with real difference between tested methods; (iv) too few patients enrolled in clinical trials (high level of type II error); (v) too few confirmatory trials.

Słowa kluczowe: wyniki fałszywie pozytywne, randomizowane badania kliniczne

Key words: false positive results, randomised clinical trials

Wstęp

Decyzja o zmianie metody leczenia powinna być podejmowana na podstawie wszystkich opublikowanych wyników badań klinicznych, zarówno wykazujących wyższość nowych sposobów postępowania (potwierdzoną statystycznie istotnym wynikiem testu – badania pozytywne), jak i tych, które tej wyższości nie potwierdzają – badania negatywne.

Jednakże w praktyce klinicznej, na zmianę metody leczenia wpływają przede wszystkim badania pozytywne; natomiast, badania negatywne są często ignorowane, a nierzadko zdarza się, że nie są wcale publikowane [1–6]. Tęgo rodzaju „cenzura” powoduje wzrost ryzyka podjęcia błędnej decyzji o zmianie metody leczenia, gdyż, jak podają niektóre publikacje, odsetek badań fałszywie pozytywnych w odniesieniu tylko do badań pozytywnych może wynosić nawet około 30–50% [1, 5, 7]. Co wpływa na wysokość tego odsetka? Jak postępować, aby zmniejszyć niekorzystny wpływ wyników fałszywie pozytywnych na praktykę kliniczną? Niniejszy artykuł jest próbą odpowiedzi na te pytania

¹ Klinika Nowotworów Jelita Grubego

² Pracownia Biostatystyki

Centrum Onkologii-Instytut im. Marii Skłodowskiej-Curie w Warszawie

Błąd I rodzaju

W badaniu klinicznym z doбором losowym chorych, obserwowana różnica w skuteczności leczenia pomiędzy badanymi metodami może wynikać z dwóch powodów: 1. większej skuteczności jednej z metod, 2. może być dziełem przypadku – otrzymany wynik jest wówczas wynikiem fałszywie pozytywnym, zdefiniowanym jako błąd I rodzaju. Prawdopodobieństwo popełnienia błędu I rodzaju określa poziom istotności statystycznej (α), który najczęściej przyjmuje się jako równy 0,05. Oznacza to, że średnio nie częściej niż w 5 przypadkach na 100, popełniony zostanie błąd I rodzaju, przez uznanie badanej metody za skuteczniejszą.

Poziomu istotności statystycznej (poziomu błędowi I rodzaju) nie należy mylić z „wartością p” czyli wartością krytyczną testu statystycznego. Poziomu istotności statystycznej jest określony arbitralnie przed badaniem, natomiast „wartość p” określa prawdopodobieństwo otrzymania różnicy w wynikach o zaobserwowanej wielkości tylko przez przypadek (czyli zakładając, że badane sposoby leczenia byłyby jednakowo skuteczne) [5]. Decyzja, czy obserwowana różnica w wynikach różnych metod leczenia jest, czy nie jest istotna statystycznie, podejmowana jest po porównaniu przyjętego poziomu istotności statystycznej (zazwyczaj równego 0,05) z otrzymaną „wartością p” – gdy „wartość p” jest mniejsza, to wówczas przyjmuje się, że różnica jest statystycznie istotna. Niestety, rozróżnienie pojęć „wartości p” i poziomu istotności statystycznej (α) utrudnia zamieszanie w nazewnictwie. Pojęcia te są często mylone – „wartość p” (wartość krytyczna testu statystycznego) jest często nieprawidłowo nazywana poziomem istotności.

Błąd II rodzaju

Istnieją dwa powody nie zaobserwowania różnicy pomiędzy badanymi metodami: 1 – metody są równorzędne, 2 – nie zaobserwowanie różnicy jest dziełem przypadku – otrzymany wynik jest wówczas wynikiem fałszywie negatywnym, zdefiniowanym jako błąd II rodzaju. Ryzyko popełnienia tego błędu określone jest przez poziom błędowi II rodzaju (β), któremu odpowiada moc testu równa $1-\beta$. Wartość przyjmowana jest najczęściej na poziomie 0,1 lub 0,2 (moc testu 0,9 lub 0,8). Przyjęcie $\beta=0,1$ zapewnia, że średnio nie częściej niż w 1 przypadku na 10, nie zostanie zaobserwowana rzeczywista różnica w skuteczności badanych metod. Wymagana moc testu statystycznego ($1-\beta$) warunkuje konieczną liczbę chorych uczestniczących w badaniu.

Metoda

W celu oszacowania odsetka wyników fałszywie pozytywnych w stosunku do wszystkich wyników pozytywnych, przeprowadzono symulację komputerową. Założono, że badania te są od siebie niezależne oraz, że poziom istotności statystycznej równy jest 0,05. Obliczeń dokonano w zależności od odsetka badań, w których testowana metoda leczenia jest rzeczywiście lepsza oraz w zależności od po-

ziomu błędowi II rodzaju (β). Sposób wykonania obliczeń przytaczamy za Tannock’iem [1]. Założmy, że przeprowadzono 200 badań, w których przyjęto poziom istotności statystycznej 0,05, a testy statystyczne dysponowały mocą równą 0,9 ($\beta=0,1$). Założmy dalej, że jedynie w 20 (10%) przeprowadzonych badaniach, jedna z metod była rzeczywiście lepsza, natomiast w pozostałych 180 doświadczeniach testowane metody były faktycznie równorzędne. Wówczas, w 18 badaniach ($20 \times 0,9$) wynik będzie prawdziwie pozytywny, a w 9 ($180 \times 0,05$) fałszywie pozytywny. Oznacza to, że wyniki fałszywie pozytywne stanowią jedną trzecią (9/27) spośród wszystkich wyników pozytywnych – ten konkretny rezultat wyliczenia został zaznaczony w tabeli pogrubionym drukiem.

Wyniki

Otrzymane wyniki przedstawiono w Tabeli I. Można zauważyć, że odsetek wyników fałszywie pozytywnych zależy przede wszystkim od odsetka badań, w których testowana metoda jest rzeczywiście lepsza. Dla przykładu, przy założeniu $\beta=10\%$, jeśli tylko w 5% badań jedna z metod jest rzeczywiście lepsza, to wówczas wyniki fałszywie pozytywne stanowią 51% wyników pozytywnych; natomiast, jeśli w 50% badań jedna z metod jest faktycznie lepsza, to wówczas wyniki fałszywie pozytywne stanowią jedynie 5% wyników pozytywnych.

Wpływ założonego poziomu błędowi II rodzaju nie jest już tak duży. Dla przykładu, przy założeniu, że w 10% badań jedna z metod jest rzeczywiście lepsza, jeśli poziom błędowi II rodzaju wynosi 10%, to wówczas wyniki fałszywie pozytywne stanowią 33% wyników pozytywnych; natomiast jeśli poziom błędowi II rodzaju wynosi 50%, to wówczas wyniki fałszywie pozytywne stanowią 47% wyników pozytywnych.

Omówienie

Z przeprowadzonej symulacji wynika, że aby oszacować rzeczywisty procent publikowanych wyników fałszywie pozytywnych w odniesieniu do wyników pozytywnych, należy znać odsetek badań, w których testowana metoda leczenia jest rzeczywiście lepsza. Jednak odsetek ten wydaje się niemożliwy do oszacowania. Jeśli jest on niski, to duży procent wyników jest fałszywie pozytywnych (Tab. I). Czy oznacza to, że ryzyko podjęcia błędnej decyzji wyboru sposobu leczenia na podstawie piśmiennictwa

Tab. I. Symulacja komputerowa wielkości odsetków wyników fałszywie pozytywnych w odniesieniu do wszystkich wyników pozytywnych w hipotetycznych badaniach z losowym doбором chorych – objaśnienia w tekście

Założony % doświadczeń z rzeczywiście lepszą metodą	Poziomu błędowi II rodzaju – β		
	10%	20%	50%
5%	51%	54%	66%
10%	33%	36%	47%
20%	18%	20%	29%
50%	5%	6%	9%

jest również wysokie? Rzeczywiście, ryzyko to może być duże, gdy brane są pod uwagę prace rejestrujące tylko pozytywne wyniki. Może się to zdarzyć, gdy np. szukając nowości, przegląda się kilka czasopism, zapamiętując jedynie prace odnotowujące wyniki pozytywne. Znacznie mniejsze ryzyko podjęcia błędnej decyzji terapeutycznej istnieje wtedy, gdy przegląda się piśmiennictwo w celu znalezienia odpowiedzi na konkretne pytanie kliniczne. Nawet, gdy jeden z wyników badań jest fałszywie pozytywny, to jego znaczenie maleje na tle wyników innych badań na ten sam lub pokrewny temat, które zgodnie ze stanem faktycznym będą negatywne. Z kolei, gdy jedno z badań wykaże statystycznie istotną wyższość nowej metody nad standardową, a inne badanie potwierdzi ten wynik, to przy braku badań negatywnych, zaufanie do tych wyników wzrasta. To stwierdzenie wskazuje na ważną rolę tzw. badań potwierdzających, które niestety rzadko są podejmowane [1].

Ważna jest także liczebność próby chorych, na jakiej wykonywano badania (poziom błąd II rodzaju – β) – im ona jest większa, tym odsetek wyników fałszywie pozytywnych wśród wszystkich badań pozytywnych jest mniejszy (Tab. I) [5]. Niestety, liczba chorych uczestniczących w badaniach jest często zbyt mała. Na przykład, spośród 92 badań randomizowanych, dotyczących radioterapii i zarejestrowanych w Medlin w 1996 roku, mediana liczebności na jedno ramię badania wyniosła średnio 72 chorych [6]. Badanie o tej liczebności, dla mocy testu 90%, można zaplanować tylko dla bardzo dużej spodziewanej różnicy w wynikach np. z wieloletniego przeżycia 35% na ponad 60% [6]. Tylko jedno na dziesięć doświadczeń obejmowało więcej niż 250 chorych w jednym ramieniu badania. Oznacza to, że klinicznie ważna różnica w wynikach, rzędu 10%, tylko w niewielu badaniach miała szansę zostać ujawniona [6]. Przykład ten obrazuje, że zbyt mała liczebność chorych sprawia, że zmniejsza się odsetek badań prawdziwie pozytywnych, co w konsekwencji powoduje wzrost udziału badań fałszywie pozytywnych wśród wszystkich badań pozytywnych.

Należy także dodać, że wzrost ryzyka podjęcia błędnej decyzji klinicznej w oparciu o piśmiennictwo, powoduje znana i udokumentowana tendencja do chętniejszego przesyłania do redakcji czasopism prac z pozytywnymi wynikami, w porównaniu z pracami zawierającymi wyniki negatywne i zarazem chętniejszego akceptowania tych prac do druku (*publication bias*) [1–6]. Zjawisko to dotyczy w większym stopniu badań mało liczebnych [3]. Mechanizm ten powoduje, że dla całościowej oceny może zabraknąć wspomnianego wyżej „tła” wyników prawdziwie negatywnych. Ta sytuacja może spowodować błędną decyzję o zmianie metody leczenia.

Wnioski

Na wzrost ryzyka podjęcia błędnej decyzji klinicznej w oparciu o piśmiennictwo wpływa:

1. Wybiórcze czytanie piśmiennictwa – przywiązywanie większej wagi do wyników pozytywnych niż negatywnych.

2. Większa skłonność do publikowania wyników pozytywnych niż negatywnych.
3. Mały odsetek badań testujących metody leczenia o rzeczywistej przewadze nad dotychczas stosowanymi.
4. Zbyt mała liczebność chorych (duża wartość β) w badaniach klinicznych.
5. Zbyt mała liczba badań potwierdzających.

Podziękowania

Autorzy są wdzięczni Profesorowi Andrzejowi Hliniakowi i Doktorowi Maciejowi Chwalińskiemu za uwagi pomocne w napisaniu pracy.

Dr n. med. Krzysztof Bujko

Klinika Nowotworów Jelita Grubego
Centrum Onkologii-Instytut im. Marii Skłodowskiej-Curie
W.K. Roentgena 5
02-781 Warszawa

Piśmiennictwo

1. Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *J Natl Cancer Inst* 1996; 88: 206–7.
2. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. *J Natl Cancer Inst* 1989; 81: 107–15.
3. Easterbrook PJ, Berlin JA, Gopalan R i wsp. Publication bias in clinical research. *Lancet* 1991; 337: 867–72.
4. De Bellefeuille C, Morrison CA, Tannock IF. The fate of abstract submitted to a cancer meeting: factors, which influence presentation and subsequent publication. *Ann Oncol* 1992; 3: 187–91.
5. Peto R, Pike C, Armitage P i wsp. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976; 34: 565–612.
6. Bentzen SB. Towards evidence based radiation oncology: improving the design, analysis, and reporting of clinical outcome studies in radiotherapy. *Radiother Oncol* 1998; 46: 5–18.
7. Zelen M. Alternatives to classic randomized trials. *Surg Clin N Am* 1981; 61: 1421–32.

Otrzymano: 24 lutego 2000 r.

Przyjęto do druku: 18 kwietnia 2000 r.