PUBLICATION BIAS – JAKI JEST EFEKT CHOWANIA WYNIKÓW BADAŃ W SZUFLADZIE?

Michał Kusy, StatSoft Polska Sp. z o.o.

Na podstawie efektu wyznaczonego w badaniu staram się oszacować i ocenić nieznany rzeczywisty efekt. Jak bardzo ufam wynikom tego badania? Mogę je porównać z innymi podobnymi badaniami. Nie jestem jednak w stanie dotrzeć do wszystkich. Czy zidentyfikowane badania, a w szczególności wyniki opublikowane nie będą w związku z tym obciążone? Zacznijmy od dwóch krótkich historii.

Dwóch badaczy

Publicysta

Rozważmy następującą sytuację. Badacz analizuje pewne schorzenie, które nie jest na tyle znane, żeby był w stanie wskazać jego potencjalne przyczyny. W związku z tym przeprowadza wiele porównań, spośród których jedno wskazuje na istotne powiązanie pewnej cechy pacjenta ze schorzeniem. Ciężko jednak znaleźć uzasadnienie dla takiej zależności. Ponieważ wynik jest istotny statystycznie, badacz mimo wszystko decyduje się go opublikować. Dlatego nazwiemy go dalej Publicystą.

Archiwista

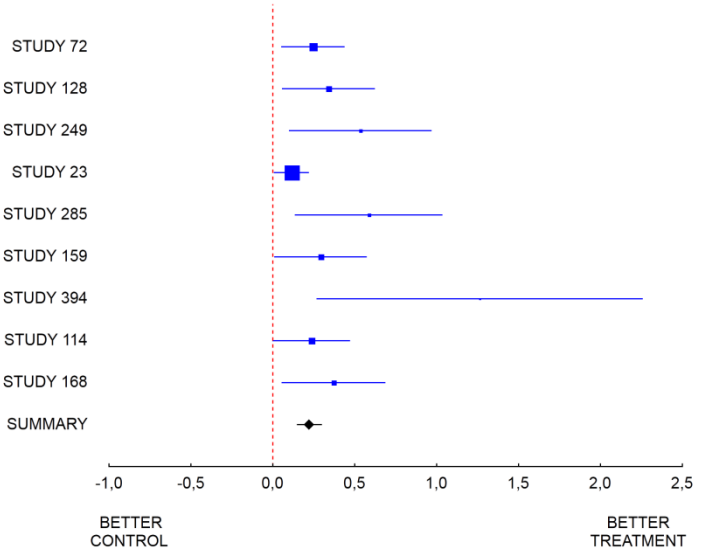
Dla odmiany wyobraźmy sobie innego badacza. Chce wykazać skuteczność terapii pewnego schorzenia określoną substancją. Badaniu poddaje 200 pacjentów losowo przypisanych do dwóch równolicznych grup: leczonej, której podaje analizowaną substancję, oraz kontrolnej, której podaje placebo. Mimo że spodziewa się uzyskać silny efekt, wynik okazuje się jednak nieistotny statystycznie. Ponieważ raport z badania nie zostaje przyjęty do publikacji, badacz z bólem serca chowa go do szuflady biurka. W ten sposób zapracował sobie na tytuł Archiwisty.

Obaj badacze nie znają stanu faktycznego. Gdyby Archiwista znał rzeczywisty efekt leczenia i wiedział, że terapia jest skuteczna, nie musiałby przeprowadzać badania. Niestety może się okazać, że wyniki nie są zgodne z rzeczywistością. Wtedy oba przypadki łączy wspólny problem – zniekształcają ocenę badanego zjawiska. Zwykle skupiamy się na tym, żeby błędu nie popełnił Publicysta. Inaczej mówiąc staramy się kontrolować tzw. błąd

Copyright © StatSoft Polska 2014

I rodzaju, czyli wyniki fałszywie pozytywne. Okazuje się jednak, że brak informacji o nieistotnych wynikach może być równie szkodliwy jak wyniki niezgodne z rzeczywistością, ale istotne statystycznie.

Wynik pojedynczego badania może być obarczony dużym błędem, wynikającym choćby z małej liczby badanych. W związku z tym badacz dokonuje przeglądu dostępnych badań i porównuje uzyskane w nich wyniki. Otrzymuje wykres leśny (*forest plot*), widoczny poniżej. Każde z badań przedstawia wynik istotny. Wyznaczony na ich podstawie efekt łączny jest obarczony mniejszym błędem niż pojedyncze badania i wskazuje na istotną skuteczność terapii. Czy badacz przekonał się, że substancja faktycznie działa?



Rys. 1. Wykres leśny.

Jakość wyników metaanalizy zależy od jakości danych, na których jest przeprowadzona. Przed jej wykonaniem definiuje się kryteria włączenia badań do analizy. Najczęściej analizuje się jednak badania, które zostały opublikowane, lub do których był stosunkowo łatwy dostęp. Nie ma jednak informacji, ile faktycznie przeprowadzono podobnych badań i jakie uzyskano w nich wyniki.

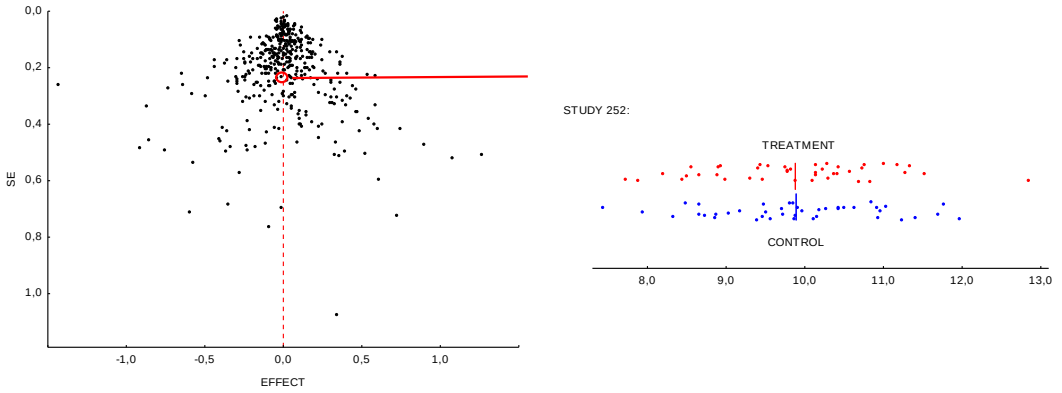
Aby zilustrować możliwe konsekwencje analizy na obciążonych wynikach, przeprowadzimy symulację badań opisujących skuteczność terapii. Symulacja stawia nas w luksusowej sytuacji, na którą zwykle nie możemy sobie pozwolić. Ponieważ znamy całą populację, znamy prawdę. Wiemy, czy w rzeczywistości lek jest skuteczny czy nie.

 Copyright © StatSoft Polska 2014

Nieskuteczny lek

Na początku stworzymy lek nieskuteczny. Dla każdego z pacjentów losujemy wartość, która charakteryzuje stan jego zdrowia. Załóżmy, że niskie wartości odpowiadają pacjentom bardziej schorowanym. Ponieważ chcemy symulować lek zupełnie nieskuteczny, w dwóch porównywanych grupach – leczonej i kontrolnej losujemy wartości z tego samego rozkładu. W związku z tym przeciętny poziom analizowanej zmiennej będzie w obu grupach jednakowy.

Kolejny krok to przeprowadzenie badania, a w zasadzie wielu badań skuteczności naszego leku. Zakładamy, że lek został oceniony w 400 badaniach na próbach o różnej liczności (od 4 do 10 000 pacjentów w każdej z grup). Efekty i błędy standardowe wyznaczone w poszczególnych badaniach przedstawiono poniżej. Po prawej stronie widzimy wyniki pacjentów dla wybranego badania *STUDY 252* i zbliżony średni poziom w porównywanych grupach.



Rys. 2. Lek nieskuteczny – przykładowe badanie.

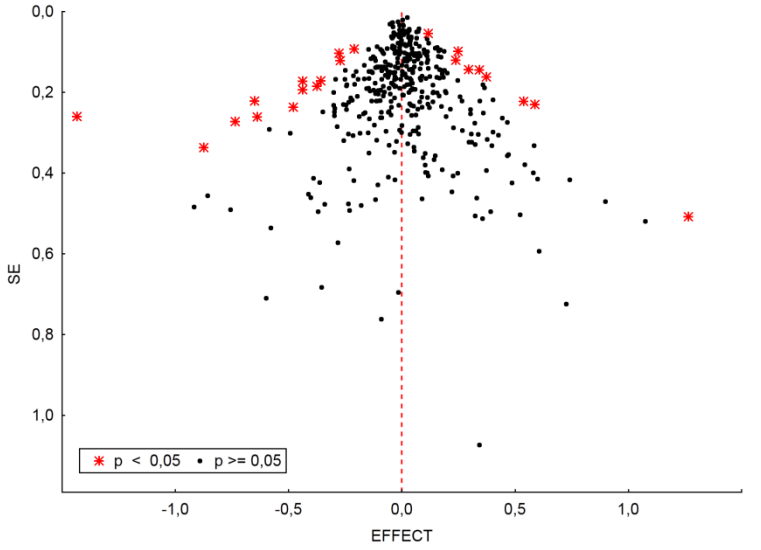
Nad każdym z 400 badaczy mamy niesamowitą przewagę – znamy prawdę. Wiemy, że lek jest nieskuteczny. Większość z nich dochodzi do podobnego wniosku, jednak w 22 badaniach wynik okazuje się istotny statystycznie. 13 badaczy stwierdza, że leczenie jest szkodliwe, a 9, że skuteczne.

Zatrzymamy się na chwilę przy tym problemie. Dlaczego zdarzają się badania, które stwierdzają skuteczność terapii? W rzeczywistych badaniach moglibyśmy tłumaczyć to np. nie do końca losowym przypisaniem pacjentów do grup lub brakiem zaślepienia próby. Tutaj nie ma jednak mowy o źle przygotowanym badaniu. Nasi „pacjenci" nie mają żadnej charakterystyki – tu nie ma kobiet ani mężczyzn, młodych lub starych. Symulacja nie faworyzuje również żadnej z grup.

Takie niezgodne z prawdą wyniki (w tym przypadku informacje o szkodliwości lub skuteczności leku) pojawiają się losowo ze względu na procedurę stosowaną we wnioskowaniu statystycznym. Mimo że nie zawsze zdajemy sobie z tego sprawę, stosując test statystyczny, określamy procent badań, w których taki błąd jesteśmy skłonni zaakceptować.

Copyright © StatSoft Polska 2014

Poziom istotności α, który w tym przykładzie przyjęliśmy jako 0,05, mówi, że w około 5% badań opisujących nasz nieskuteczny lek otrzymamy wynik istotny, ale niezgodny z rzeczywistością. Faktycznie 22 spośród 400 badaczy otrzyma istotne statystycznie wyniki i będą to niestety potencjalni Publicyści. Na poniższym wykresie oznaczyliśmy ich gwiazdkami.



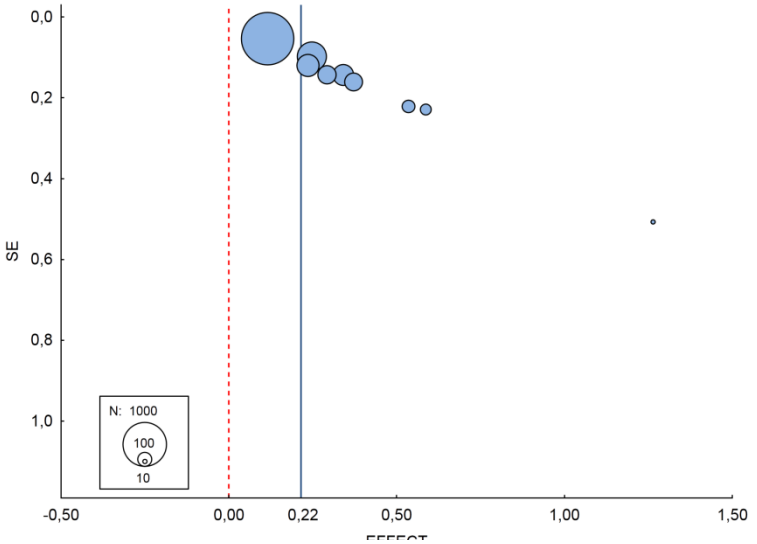
Rys. 3. Lek nieskuteczny – wyniki istotne statystycznie.

Autor pojedynczego badania z istotnym wynikiem może być przekonany o skuteczności leczenia. Jeśli porówna swój wynik z innymi badaniami, w których brak istotności, może zweryfikować ocenę leczenia. Co się jednak stanie, jeśli autorzy badań z nieistotnymi wynikami schowają je w szufladzie, podobnie jak Archiwista?

Załóżmy, że opublikowano jedynie wyniki istotne statystycznie. Są to badania zarówno korzystne, jak i niekorzystne dla leku. Efekt łączny wyznaczony w metaanalizie wynosi wtedy -0,03 i jest nieistotny (p=0,305). Będąc lekarzem, takiego leku nie podalibyśmy swoim pacjentom. Gdyby jednak do szuflady trafiła również część badań niekorzystnych dla leku? Badania są często prowadzone na zlecenie producenta leku i w związku z tym nastawione na wykazanie jego skuteczności. Załóżmy więc zupełnie skrajną sytuację – żadne z 400 badań nie jest rejestrowane i autorzy nie czują się zobligowani do opublikowania wyników niekorzystnych dla leku.

Pozostaje więc 9 Publicystów, którzy sugerują istotną skuteczność terapii. Zostali przedstawieni poniżej. Badania mają różną wielkość – najmniejsze uwzględnia 10 pacjentów, największe 1410. Efekt łączny wyznaczony na ich podstawie wynosi 0,22 i jest istotny statystycznie (p=0,000). Poniższe wyniki widzieliśmy już wcześniej w innej formie na wykresie leśnym (rys. 1.). Przypomnijmy, że w przeciwieństwie do autorów znamy rzeczywisty efekt leczenia - wynosi 0.

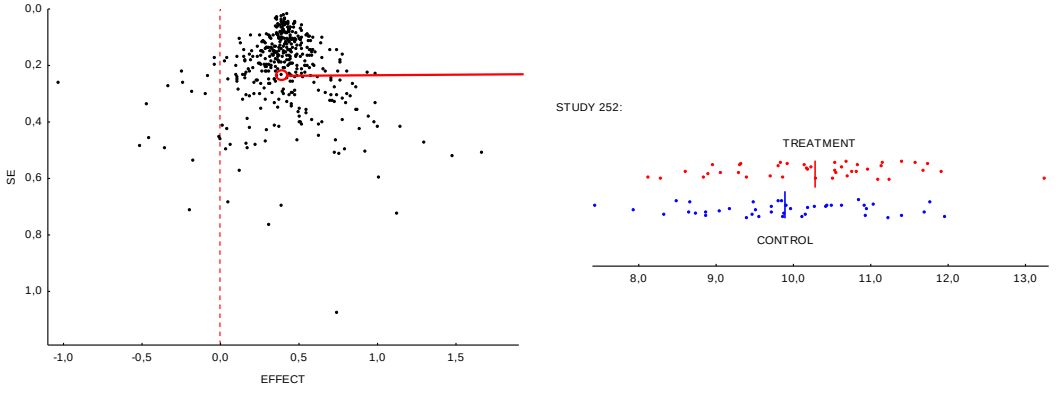
 Copyright © StatSoft Polska 2014



Rys. 4. Lek nieskuteczny – badania z istotną skutecznością.

Skuteczny lek

Dla odmiany opracujemy teraz lek skuteczny. Tym razem przeciętny poziom analizowanej zmiennej wśród wszystkich leczonych jest większy o 0,4 od przeciętnego poziomu w grupie kontrolnej. Poniżej widzimy wyniki uzyskane w poszczególnych badaniach oraz wyniki pacjentów dla wybranego badania *STUDY 252*.

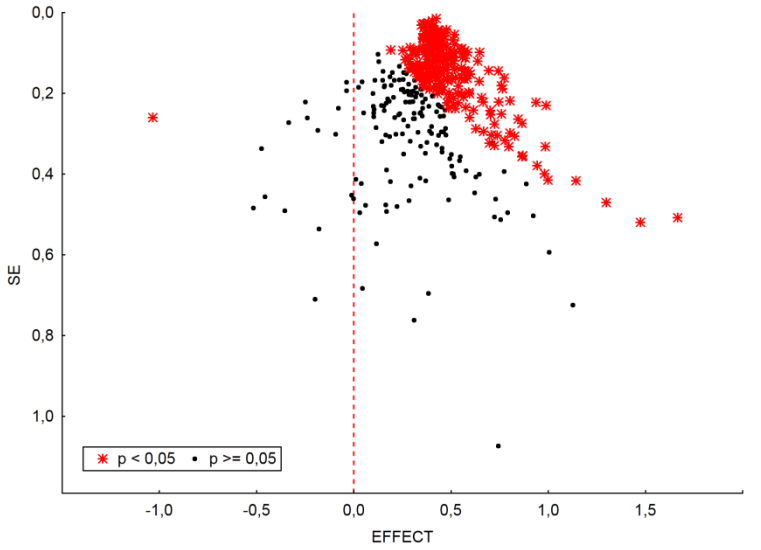


Rys. 5. Lek skuteczny – przykładowe badanie.

Widząc wszystkie badania, stwierdzamy, że średni efekt jest rzeczywiście większy od zera. Również w tym przypadku żaden z 400 badaczy nie zna prawdy. Większość (242) z nich

Copyright © StatSoft Polska 2014

otrzyma wynik istotny statystycznie wskazujący na skuteczność leczenia. Jeden badacz uzna lek za istotnie szkodliwy.



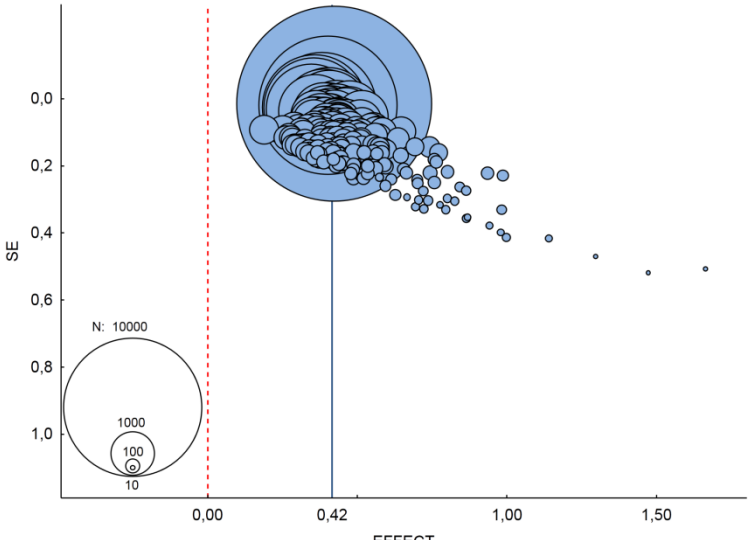
Rys. 6. Lek skuteczny – wyniki istotne statystycznie.

Ponieważ wiemy, że lek jest skuteczny, możemy powiedzieć, że w pozostałych 158 badaniach popełniono błąd. Nazywa się go błędem II rodzaju. Powyżej gwiazdkami zostały zaznaczone badania z wynikami istotnymi statystycznie. Jak widać błąd II rodzaju popełniano głównie w lewej, dolnej części wykresu. Na brak istotności mogła mieć wpływ stosunkowo mała różnica wartości między grupami, duża zmienność wyników w grupach lub zbyt mała liczba pacjentów.

Jeśli zestawimy wyniki wszystkich badań, otrzymamy istotny statystycznie wynik równy 0,40. Pamiętamy jednak, że wiele nieistotnych wyników pada ofiarą Archiwistów. Sprawdzimy zatem, jaki będzie efekt analizy opartej jedynie na wynikach istotnych statystycznie i dodatkowo korzystnych dla leku.

Istotność statystyczna wiąże się silnie z wielkością próby. Wśród wyników istotnych statystycznie znajduje się większość dużych badań, m.in. badanie z udziałem 20 000 pacjentów. Wielkość próby w poszczególnych badaniach możemy porównać poniżej. Metaanaliza przeprowadzona na tak wybranych 241 badaniach wskazuje na efekt równy 0,42, istotny statystycznie (p=0,000). Wiemy jednak, że jest on zawyżony w stosunku do rzeczywistego efektu leku na poziomie 0,4.

 Copyright © StatSoft Polska 2014



Rys. 7. Lek skuteczny – badania z istotną skutecznością.

Błąd publikacji (*publication bias*)

Przegląd systematyczny wymaga określenia kryteriów doboru badań do analizy. W rzeczywistości najczęściej nie ma jednak możliwości wskazania wszystkich badań, które spełniają zdefiniowane kryteria. Jeśli metaanalizę wykonamy na obciążonej próbie badań, jej wynik będzie również obarczony błędem. Problem ten nie dotyczy wyłącznie metaanalizy, lecz każdego przeglądu literatury.

Spośród możliwych źródeł błędu można wskazać m.in.:

- Błąd publikacji – stosunkowo silne efekty mają większą szansę na publikację od efektów słabych lub nieistotnych. Dodatkowo wyniki publikacji naukowych częściej pojawiają się w przeglądach systematycznych od innych, niepublikowanych wyników.
- Język – najczęściej przeszukiwane są bazy danych i czasopisma anglojęzyczne.
- Dostępność – badania trudnodostępne mają mniejszą szansę znaleźć się w przeglądzie.
- Koszty – tylko część badań jest dostępna bezpłatnie lub za niską opłatą.
- Powtórzenia – wyniki istotne statystycznie mają większą szansę być publikowane więcej niż raz.
- Cytowanie – wyniki istotne statystycznie łatwiej jest znaleźć ze względu na częste cytowanie.

Nasze rozważania skupiają się na pierwszym z wymienionych źródeł błędu. Błąd publikacji wiąże się zarówno z siłą i istotnością statystyczną efektu w opublikowanych badaniach, jak i stosunkowo małą liczbą odwołań do tzw. szarej literatury (*grey literature*),

czyli wyników niepublikowanych. Jest to m.in. efekt często nieuzasadnionego przekonania o niskiej jakości takich badań.

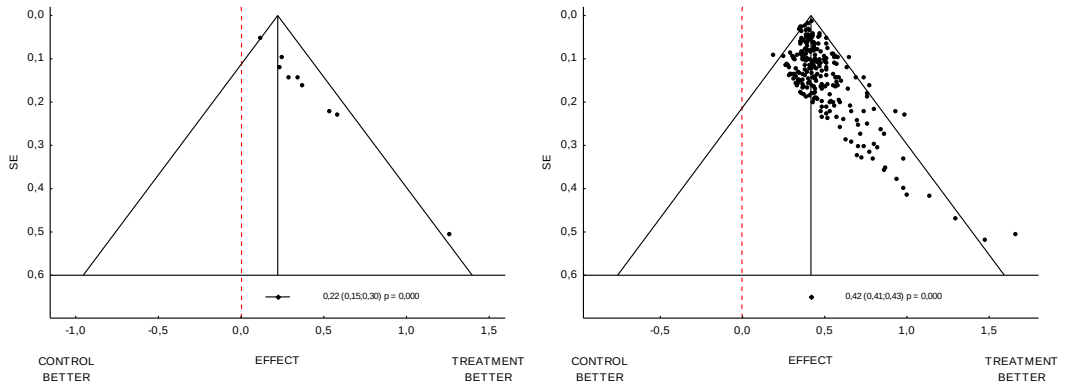
Aby dokładnie wyznaczyć błąd publikacji, musielibyśmy zestawić wyniki wszystkich Publicystów i Archiwistów. W rzeczywistości do wielu badań nie mamy jednak dostępu, a nawet nie zdajemy sobie z nich sprawy. Możemy natomiast skorzystać z pewnych narzędzi, które oceniają występowanie błędu oraz jego wpływ na wyniki analizy. Opierają się one na następujących założeniach:

- Ryzyko pominięcia badania maleje wraz ze wzrostem wielkości próby (precyzji wyników).
- Ryzyko pominięcia badania maleje wraz ze wzrostem obserwowanego w nim efektu.

Inaczej mówiąc, zakładamy, że duże badania są częściej publikowane, bez względu na istotność statystyczną wyników, w związku z ich rejestracją oraz zaangażowaniem czasu i środków. Największym ryzykiem pominięcia obarczone są z kolei badania na małych próbach, w których obserwowano efekt umiarkowany lub słaby.

Grzebanie w szufladach

Pora na przegląd szuflad. Co może zrobić badacz, który nie wie, ile łącznie przeprowadzono badań i do jakich wyników nie dotarł? Dysponuje jedynie wynikami dostępnych badań. Na ich podstawie ma wyobrazić sobie szuflady i sprawdzić ich zawartość. Zadanie jest niestandardowe i wymaga niestandardowych narzędzi.



Rys. 8. Wykres lejkowy - lek nieskuteczny (L) i skuteczny (P).

Jedną z najpopularniejszych metod wykorzystywanych do oceny błędu publikacji jest wykres lejkowy (*funnel plot*). To wykres rozrzutu, na którym oś pozioma przedstawia mierzony efekt, a pionowa miarę precyzji badań (np. błąd standardowy). Gdy nie ma obciążenia publikacji, spodziewamy się równomiernego rozrzutu badań wokół efektu łącznego. Jeżeli błąd publikacji występuje, to w miarę przesuwania się w dół wykresu symetria powinna być coraz bardziej zaburzona. Powyżej widzimy wykresy lejkowe odpowiednio dla

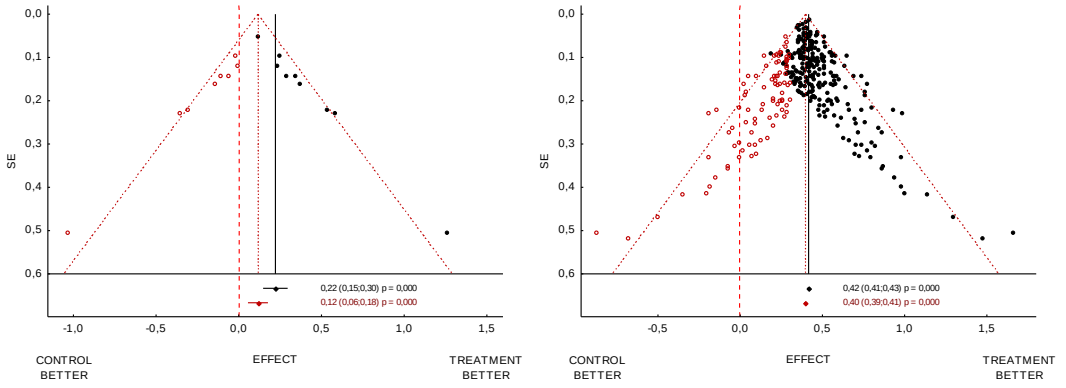
 Copyright © StatSoft Polska 2014

nieskutecznego i skutecznego leku. W obu przypadkach obserwujemy asymetrię wykresu. Brakuje badań o stosunkowo słabym efekcie i dużym błędzie.

Ocena wykresów jest jednak subiektywna. Możemy więc skorzystać ze specjalnych testów asymetrii. Do najpopularniejszych należą test Begga i Mazumdar [1] i test Eggera [4]. Testy w obu przypadkach wskazują na istotną zależność między błędem standardowym efektu a jego wartością. Sugerują zatem, że efekt łączny jest obarczony błędem systematycznym.

W obu wariantach leku badacz zorientuje się zatem, że ma do czynienia z błędem publikacji. Czy dowie się jednak, w jakim stopniu wpłynął on na uzyskany wynik? Do odtworzenia brakujących badań może wykorzystać procedurę „trim & fill" (Duval i Tweedie [3]). Dwa etapy procedury – „odcinanie" i „wypełnianie" pozwalają skorygować efekt łączny i jego błąd. „Odcinanie" badań trwa, dopóki wykres nie stanie się symetryczny względem przeliczonej miary efektu. W kolejnym etapie wykres jest ponownie „wypełniany" usuniętymi badaniami i ich lustrzanymi odbiciami. Teoretycznie procedura powinna doprowadzić do nieobciążonego oszacowania miary efektu i jej zmienności.

Analiza wyników leku nieskutecznego prowadzi do usunięcia 8 badań, czyli wszystkich, oprócz badania o najsłabszym efekcie (*STUDY 23*). W związku z tym metoda skoryguje efekt na tyle, na ile pozwoli jej ten najsłabszy wynik. Efekt skorygowany wynosi 0,12. Jest równy efektowi z badania *STUDY 23*, ma jednak mniejszy błąd standardowy. Wynik jest nadal istotny statystycznie. Badacz niestety nie zna faktycznej liczby badań i rzeczywistego efektu. Może się posiłkować jedynie wynikami widocznymi na wykresie lejkowym. W tym przypadku, mimo że wynik jest nadal istotny, powinien jednak zwrócić uwagę na liczbę usuniętych badań.



Rys. 9. Metoda „trim & fill" - lek nieskuteczny (L) i skuteczny (P).

Jeśli badacz skoryguje wyniki dla leku skutecznego, okaże się, że otrzyma efekt zgodny z rzeczywistym (0,4). Zanim osiągnięta zostanie symetria wykresu, procedura usuwa znaczną część badań. W szacowaniu korekty biorą udział głównie badania z najmniejszym błędem standardowym, czyli mające najsilniejszy wpływ na wyniki analizy. Mimo, że

Copyright © StatSoft Polska 2014

badacz nie dysponuje badaniami schowanym w szufladzie, na podstawie reszty wyników udało mu się odtworzyć rzeczywisty efekt leczenia.

Jak z tym żyć

Sprawdziliśmy, jak może zachować się efekt wyznaczony na obciążonej próbie badań. Oba przykłady zostały jednak celowo przerysowane. W rzeczywistości prowadzone są rejestry badań, których wyniki są publikowane, nawet mimo braku istotności statystycznej. W przeglądach systematycznych pojawiają się nieistotne efekty, a autorzy coraz częściej docierają do wyników niepublikowanych.

Można znaleźć prace, w których podjęto próbę oceny skali zjawiska. Sutton [5] pokazał, że błąd publikacji występuje w większości uwzględnionych przez niego metaanaliz. W 50% metaanaliz nie powoduje znacznych rozbieżności w wynikach. W 45% zmienia wielkość efektu, ale nie ma wpływu na ostateczne ustalenia (np. odnośnie skuteczności leczenia). Natomiast istnienie błędu publikacji w pozostałych 5% metaanaliz podważa ich kluczowe ustalenia. Wyniki analizy oparto głównie na pracach z bazy Cochrane. Ponieważ Cochrane Collaboration przywiązuje dużą wagę do szerokiego przeglądu literatury, w ich pracach pojawia się zazwyczaj więcej badań niż w innych publikowanych metaanalizach. W związku z tym paradoksalnie błąd publikacji mógł się wkraść do analizy, która próbowała go ocenić i niewykluczone, że skala zjawiska jest w rzeczywistości większa.

Na koniec przydałaby się jakaś pozytywna konkluzja. Wskazaliśmy dolegliwość, ale pacjent czeka na lekarstwo. Przedstawiliśmy kilka metod, które wyglądają obiecująco. Warto wspomnieć, że są dostępne w kolejnej odsłonie (3.0) programu *STATISTICA Zestaw Medyczny*. Narzędzia te pomagają diagnozować problem i walczą z jego skutkami. Trzeba jednak sięgnąć głębiej, do samego źródła. Najlepszą profilaktyką wydaje się w tym przypadku uświadamianie problemu. Być może dzięki temu w niedalekiej przyszłości wyniki prowadzonych badań staną się powszechnie dostępne, niezależnie od głębokości szuflady badacza.

Literatura

1. Begg C.B., Mazumdar M., *Operating characteristics of a rank correlation test for publication bias*. Biometrics. 1994; 50, 1088-1101.
2. Borenstein M., Hedges L.V., Higgins J.P.T., Rothstein H.R., *Introduction to Meta-Analysis*. John Wiley and Sons Ltd. 2009.
3. Duval S.J., Tweedie R.L., *A non-parametric 'trim and fill' method of accounting for publication bias in meta-analysis*. Journal of the American Statistical Association. 2000; 95,89-98.
4. Egger M., Davey Smith G., Schneider M., Minder C., *Bias in meta-analysis detected by a simple, graphical test*. BMJ. 1997; 315, 629-634.

 Copyright © StatSoft Polska 2014

5. Rothstein H.R., Sutton A.J., Borenstein M., *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. John Wiley and Sons Ltd. 2005.
6. StatSoft, Inc. (2013). *STATISTICA* (data analysis software system), version 12. www.statsoft.com.
7. StatSoft Polska Sp. z o.o. 2014. *STATISTICA Zestaw Medyczny* wersja 3.0. www.statsoft.pl.