BJPsych

Na ile wiarygodne są badania naukowe?

Marcus R. Munafò, Jonathan Flint

The British Journal of Psychiatry (2010) 197, 257–258.

Coraz większy niepokój wzbudza przypuszczenie, że znacząca część badań naukowych może w rzeczywistości się mylić. Przyczyn występowania w piśmiennictwie dużej liczby wyników fałszywie dodatnich może być wiele, jedną z nich są błędy związane z publikacjami. Autorzy omawiają dowody empiryczne na ten temat. The British Journal of Psychiatry (2010) 197, 257–258.

„Jedna z mocnych stron nauki polega na tym, że nie jest konieczne, aby naukowcy nie popełniali błędów, tylko żeby różni naukowcy popełniali różne błędy."

David Hull, *Science as a Process*

Odkrycia naukowe a przypadek

Podczas II wojny światowej fizyk Enrico Fermi zapytał generała Lesliego Grovesa o to, ilu generałów można określić mianem „wielkich" i na jakiej podstawie. Groves odpowiedział, że każdy generał, który wygrał pięć dużych bitew pod rząd może być nazwany wielkim i że mniej więcej 3 na 100 generałów spełnia ten warunek. Fermi obliczył, że, jeżeli walczące ze sobą siły są mniej więcej równe, to szansa, że generał wygra jedną bitwę wynosi 1 do 2, szansa, że wygra po kolei dwie bitwy wynosi 1 do 4, dla trzech kolejnych wygranych bitew 1 do 8, dla czterech – 1 do 16, zaś dla pięciu bitew pod rząd – 1 do 32. „Zatem, generale, ma pan rację, mniej więcej trzech na stu. Prawdopodobieństwo matematyczne, a nie geniusz." Innymi słowy, wyraźnie uderzająca zgodność może być jedynie następstwem nieubłaganych zasad prawdopodobieństwa. W tym wstępnym artykule sugerujemy, że dzięki takiej samej nieubłaganej logice wiele odkryć naukowych można nazwać wielkimi.

Analogiczne do zdefiniowanego przez Fermiego „wielkiego generała" może być „wielkie odkrycie naukowe" – pozornie ekscytujące wyniki często trudne do powtórzenia w kolejnych badaniach, które początkowo być może uzyskano tylko przez przypadek, będący skutkiem jedynie liczby przeprowadzanych badań naukowych. Poniżej posłużymy się przykładem prac badaczy zajmujących się zależnościami między podatnością na chorobę a zmiennością sekwencji DNA, ustalaną na podstawie genetycznych badań sprzężeń.

Dla osób postronnych szansa zaobserwowania korelacji (w tym przypadku sprzężenie genetyczne), która w rzeczywistości nie istnieje, wynosi 1 do 20 (przy założeniu, że czasopisma naukowe akceptują próg p na poziomie 0,05 jako wystarczający dowód umożliwiający publikację), zaś szansa, że dane odkrycie zostanie powtórzone przez przypadek wynosi 1 do 400, co oznacza wiarygodny poziom pewności, że większość powtórzonych wyników będzie odpowiadać rzeczywistości. Programy do analizy statystycznej umożliwiają badaczom przeprowadzanie wielu testów statystycznych z zaskakującą szybkością, co stało się elementem rutynowego postępowania. Na podstawie jednego z niedawnych realistycznych badań symulacyjnych, w którym wykorzystywano 10 wariantów sekwencji w szeroko badanym genie dla enzymu O-metylotrasferazy katecholowej (COMT) oraz pakiet analiz podobnych do wykorzystywanych w praktyce, opisano, że odsetek wyników fałszywie dodatnich wynosił 96,8% przy poziomie istotności $p=0{,}05$.[1] Co więcej, przy przyjęciu obszernej definicji replikacji, w większości przypadków uzyskano „powtórzenie" błędnych wyników, znowu z wykorzystaniem danych losowych.

Czy zdarza się tak w praktyce? Chociaż empiryczne dowody dotyczące „nadmiarowych" wartości p, poniżej progu 5%, wskazują, że badacze często przeprowadzili wiele testów w celu sprawdzenia swoich danych,[2] autorzy są przekonani, że wyniki fałszywie dodatnie przenikają do piśmiennictwa z innych przyczyn. Zwrócili uwagę, że jedno najbardziej zna-

Marcus R. Munafò, PhD, Department of Experimental Psychology, University of Bristol; Jonathan Flint, FRCPsych, Wellcome Trust Centre for Human Genetics, University of Oxford, UK

Adres do korespondencji: Marcus R. Munafò, Department of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, Wielka Brytania; e-mail: marcus.munafo@bristol.ac.uk

Marcus R. Munafò jest starszym wykładowcą Biological Psychology na Univerisity of Bristol; zajmuje się badaniem biologicznego podłoża cech behawioralnych.

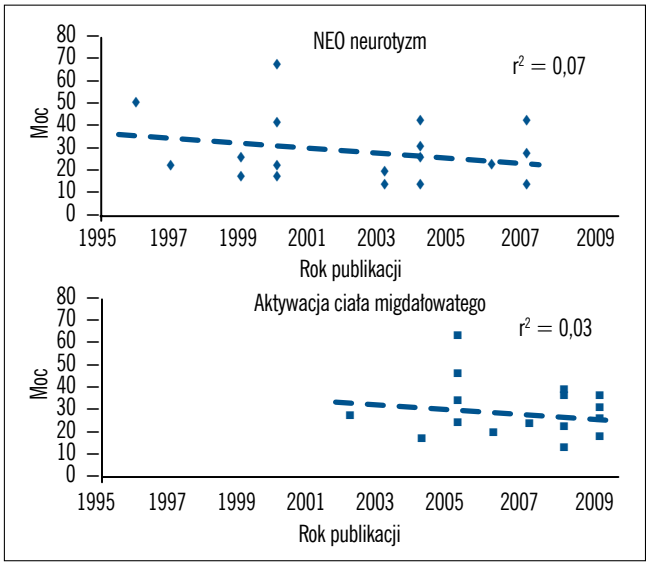
Jonathan Flint, Michael Davys Professor, University of Oxford; zajmuje się modelami tłumaczącymi genetyczne uwarunkowania cech behawioralnych.

Finansowanie: M.R.M. otrzymuje wsparcie od Higher Education Funding Council for England (HEFCE). J.F. otrzymuje wsparcie od Wellcome Trust.
Konflikt interesów: obydwaj autorzy podzielają poglądy, które mogą wpływać na interpretację dowodów, również tych zaprezentowanych w tym artykule.

czących i często cytowanych doniesień z dziedziny genetyki behawioralnej, zgodnie z którym przypuszcza się, że podatność na depresję zależy od występowania określonego allelu genu transportera serotoniny, najprawdopodobniej wynika z przypadku.[3] Analiza różnych metod, na podstawie których interakcje między wariantami genetycznymi a stresorami życiowymi uznano za powtarzalne, wykazała, że w badaniu replikacyjnym często ignorowano charakter interakcji. Wskutek tego replikacje w większości przypadków nie stanowiły dokładnego powtórzenia uzyskanych początkowo wyników.

Co więcej, wydaje się, że w wielu dziedzinach endemicznie występuje mała moc statystyczna stosowanych testów. Autorzy przeanalizowali badania sprzężeń genetycznych,[4] fenotypów ustalonych na podstawie badań neuroobrazowych[5] i zestawów badań laboratoryjnych do oceny znaczenia czynników środowiskowych u osób używających substancji psychoaktywnych.[6] We wszystkich przypadkach stwierdzono średnią moc statystyczną (na podstawie mediany próby badań w każdej metaanalizie), w przybliżeniu między 15 a 25% (rycina). Jeżeli te wartości są reprezentatywne, oznacza to, że 90% hipotez to w rzeczywistości hipotezy zerowe i utrzymuje się poziom alfa rzędu 5%, wówczas okaże się, że większość istotnych statystycznie (i dlatego przypuszczalnie opublikowanych) wyników może w rzeczywistości być fałszywa.[7]

Rycina. Moc statystyczna badań sprzężeń genetycznych dotyczących neurotyzmu i aktywacji ciała migdałowatego



Przestawiono moc statystyczną poszczególnych badań w zależności od roku publikacji. Badania dotyczyły wariantu genetycznego 5-HTTLPR. Oceniano w nich neurotyzm (oceniany na podstawie kwestionariusza osobowości NEO) i aktywację ciała migdałowatego, na podstawie wielkości efektu w odpowiedniej metaanalizie. W obu przypadkach moc statystyczna przez kilka lat utrzymywała się na niskim poziomie, pomimo coraz większej liczby dowodów na to, że badania charakteryzują się zbyt małą mocą statystyczną. Mała moc statystyczna zwiększa proporcję wyników fałszywie dodatnich do wyników prawdziwie dodatnich w tych badaniach, które osiągnęły nominalny poziom istotności statystycznej. Dane zaadaptowano i zaktualizowane z Munafò i wsp.[4,5]

Co podważa wiarygodność badań?

Dlaczego tak wiele badań naukowych może dawać fałszywe wyniki? Z doświadczenia wiadomo, że wiele czynników może być przyczyną zafałszowań w piśmiennictwie i przyczyniać się do ryzyka uzyskania wyników fałszywie dodatnich: błędy w publikacjach, dłuższy czas do opublikowania wyników, które nie osiągają poziomu istotności statystycznej, tendencja do zmniejszania wielkości efektu wraz z rokiem publikacji, słaba wartość predykcyjna wstępnych doniesień; badania *post hoc* podgrup wyróżnianych na podstawie płci lub czynników środowiskowych oraz sposób finansowania badania. Są dowody, że wszystkie te czynniki są częste.

Wśród czynników społecznych wpływających na tworzenie nauki są jednak także inne źródła błędów, które są gorzej opisane i słabiej przebadane. Autorzy wykorzystywali na przykład dane z trzech przeglądów mających charakter metaanaliz, które dotyczyły zależności między genem a chorobą w psychiatrycznym piśmiennictwie genetycznym i oszacowali stopień, w jakim każde pojedyncze badanie przeceniało rzeczywistą wielkość efektu lub jej nie doceniało (z odpowiednich metaanaliz). Stwierdzili, być może paradoksalnie, że badania opublikowane w czasopismach o niskim wskaźniku cytowań (impact factor) częściej prezentują wielkość efektu dokładnie taką jak oszacowana, w porównaniu z badaniami opublikowanymi w czasopismach o wysokim wskaźniku cytowań.[8] Autorzy znaleźli również dowody na związek między miejscem, w którym zostało przeprowadzone badanie, a stopniem, w jakim przeszacowano rzeczywistą wielkość efektu. Badania przeprowadzone w Ameryce Północnej przeszacowywały rzeczywistą wielkość efektu mniej więcej o 10% w porównaniu z badaniami przeprowadzonymi w Europie i innych częściach świata.[9]

Prawdopodobnie na publikowanie wyników badań naukowych wpływają pewne subtelne czynniki.[10] W „gorących" dyscyplinach naukowych, w których istnieje znaczna elastyczność schematu badania, znaczenie może mieć większy zakres tych czynników.[7] Wiele z zaprezentowanych dowodów pochodzi z badań obserwacyjnych genetyki molekularnej, nie ma jednak powodu, aby podejrzewać, że ta dziedzina jest wyjątkowa pod tym względem. Raczej duża liczba stosunkowo porównywalnych badań umożliwia bardziej dogłębne zbadanie czynników pozanaukowych niż w badaniach z innych dziedzin, gdzie próby replikacji są podejmowane rzadziej. Brak powtórzeń badań w niektórych dziedzinach stanowi problem sam w sobie.

Co można zrobić

Czy można zrobić cokolwiek, aby poprawić tę sytuację? Recenzenci, redaktorzy i osoby tworzące zasady funkcjonowania nauki mogliby narzucić wyższe standardy, traktując piśmien-

nictwo opisujące badania kliniczne jako przykład dobrej praktyki. Na przykład takie czynniki, jak opracowywane przed publikacją protokoły badań, zniechęcanie do odchyleń od planowanych analiz, a także zbieranie danych i przeprowadzanie analizy metodą potrójnie ślepej próby, służą ograniczeniu niepotrzebnych testów statystycznych, zniechęcaniu do „zaminowywania" danych i ułatwiają ich przejrzyste opisywanie, natomiast rutynowe wykorzystywanie analizy mocy statystycznej w celu określenia wielkości próby zmniejsza stosunek wyników fałszywie dodatnich do wyników prawdziwie dodatnich. Być może istnieje zapotrzebowanie na naukę opartą na dowodach, podobnie, jak na medycynę opartą na dowodach.

Czytelnicy czasopism naukowych być może powinni ufać tylko dużym badaniom, które opisują wyniki w uznanym piśmiennictwie (w przeciwieństwie do wczesnych wyników w nowej dziedzinie); kładą mniejszy nacisk na nominalną istotność statystyczną, a zamiast tego koncentrują się na wielkościach efektu i przedziałach ufności oraz są publikowane w czasopismach o niskim wskaźniku cytowań. Wiele z problemów, na które zwrócono powyżej uwagę, jest coraz częściej rozpoznawanych w piśmiennictwie dotyczącym genetycznych uwarunkowań w psychiatrii, co znajduje odzwierciedlenie w wykorzystywaniu o wiele większych prób w celu uzyskania wystarczającej mocy statystycznej, wymaganiu wielokrotnych powtórzeń, zanim wyniki, chociaż na wstępnym poziomie, zostaną uznane za wiarygodne, oraz szerszym omawianiu zagadnień statystycznych, szczególnie przy zastosowaniu podejścia Bayesa.[11] Jest to krok w dobrym kierunku, który świadczy o tym, że nauka dzięki identyfikowaniu problemów może się sama korygować. Możemy uczyć się na błędach i ulepszać wykorzystywane metody. Na bardziej ogólnym poziomie trzeba mieć świadomość, że błędy mogą występować w wielu postaciach, poza najczęstszym przypadkiem konfliktu interesów i sposobem finansowania, oraz mieć znaczenie na wszystkich etapach tworzenia nauki. Trzeba zaakceptować, że ostateczne odpowiedzi wymagają precyzyjnych badań (co przeważnie oznacza nie tylko dużych, ale również dobrej jakości), a być może również skoncentrowania się na tym, aby przeprowadzać mniej badań naukowych, ale robić je lepiej.

Podziękowania

Jako źródło informacji zawartych w tym artykule wykorzystano opublikowane artykuły naukowe. M.R.M. jest również gwarantem tego artykułu.

piśmiennictwo na str. 39

KOMENTARZ

Dr n. med. Piotr Świtaj

I Klinika Psychiatryczna Instytutu Psychiatrii i Neurologii w Warszawie

Artykuł Munafò i Flinta, choć ilustrowany zabawną anegdotą o „wielkich" generałach, porusza jednak bardzo poważną w swej istocie kwestię rzetelności i wiarygodności publikowanych wyników badań naukowych. Problem ten jest w ostatnim okresie przedmiotem coraz większej troski znacznej części środowiska naukowego. Kilka lat wcześniej podobna krytyka została sformułowana w głośnym, cytowanym zresztą przez autorów, artykule Ioannidisa,[1] który na podstawie przeprowadzanych przez siebie analiz i symulacji doszedł do wniosku, że właściwie większość publikowanych obecnie wyników badań może być fałszywa.

Przyznać należy, iż stwierdzając, że fałszywe wyniki są wszechobecne we współczesnym piśmiennictwie naukowym, Munafò i Flint nie są gołosłowni. Co prawda w krótkim artykule redakcyjnym, mającym raczej pobudzić do refleksji i dyskusji, niż zdać szczegółowo sprawę z analiz, które stanowiły podstawę dla wyrażonych opinii, autorzy byli w stanie jedynie zasygnalizować niektóre zagadnienia związane z rzetelnością badań naukowych. W pracach, na które się powołują, zainteresowany czytelnik znajdzie jednak bardziej rozbudowaną argumentację i dość przekonujące dowody empiryczne na poparcie stawianych w artykule tez.

Oprócz czynników omawianych przez Munafò i Flinta, szczególnie istotna wydaje mi się jeszcze jedna okoliczność mogąca rzutować na rzetelność publikacji naukowych, mianowicie zdecydowana dominacja określonego paradygmatu w danej dziedzinie – wiele wskazuje na to, że odwołujące się do takiego dominującego paradygmatu i wspierające go badania mogą być mniej krytycznie oceniane i chętniej publikowane. Powoli staje się jasne, że takie zjawisko miało (i być może nadal ma) miejsce, na przykład w przypadku psychiatrii biologicznej, stanowiącej główny model wyjaśniający w psychiatrii w ostatnich kilku dekadach. Zwrócił na to uwagę m.in. obecny prezes Światowego Towarzystwa Psychiatrycznego Mario Maj,[2] według którego nadmierny liberalizm w ocenie badań biologicznych na przestrzeni ostatnich dziesięcioleci doprowadził do tego, że wiele prac

z tej dziedziny, które ukazały się w renomowanych międzynarodowych czasopismach psychiatrycznych zaledwie 10-15 lat temu, wydaje się dziś w sposób oczywisty przestarzała i nieaktualna. Jego zdaniem, przyczyną takiego stanu rzeczy nie jest bynajmniej szybki rozwój nowych technologii, ale raczej zaniżone kryteria oceny, których rezultatem było niejednokrotnie publikowanie wyników pozbawionych istotnego znaczenia naukowego i klinicznego (w typowych przypadkach opierających się na stwierdzeniu niewielkich, choć formalnie istotnych statystycznie, różnic w średnich wartościach jednej lub więcej zmiennych biologicznych między grupą osób z zaburzeniami psychicznymi a grupą kontrolną osób zdrowych).

Na marginesie warto zauważyć, że czasopisma naukowe publikujące w dominującym akurat nurcie teoretycznym mają zwykle najwyższy impact factor. Wydaje się więc oczywiste, że wskaźnik ten, który w ostatnich latach jest wręcz fetyszyzowany, należałoby raczej traktować z odpowiednią dozą krytycyzmu. Analizy Munafò i Flinta dowodzą zresztą niezbicie, jak zawodna jest to miara oceny rzeczywistej wartości publikowanych danych naukowych.

Nasuwa się też refleksja, że choć sformułowany przez autorów postulat, aby uprawiać mniej nauki, ale za to lepszej jakości, wydaje się ze wszech miar słuszny, to jednak zachodzi obawa, że niektóre dominujące obecnie w nauce instytucjonalnej tendencje i mechanizmy mogą skłaniać badaczy do działania w dokładnie przeciwnym kierunku, tzn. do przedkładania ilości nad rzetelność i wartość poznawczą. Zauważalne niekiedy wśród decydentów bezkrytyczne dążenie do „urynkowienia" nauki i promowanie w społeczności naukowej mechanizmów bezwzględnego konkurowania o środki finansowe i stanowiska, przy równoczesnym zredukowaniu kryteriów oceny dorobku naukowego do czysto formalnych wskaźników, takich jak sumaryczny impact factor, wręcz wymusza na uczonych dążenie do maksymalizacji liczby publikacji za wszelką cenę. Tego rodzaju presja z pewnością wydatnie zwiększa prawdopodobieństwo publikowania danych nierzetelnych, wtórnych, a w skrajnych przypadkach nawet sfałszowanych. Fakt, że jest to tendencja ogólnoświatowa, nie zmniejsza wcale zagrożeń z nią związanych. Wydaje się, że należy mieć ich świadomość przy planowaniu przedsięwzięć określanych jako „reforma nauki".

Na koniec wypada poczynić jeszcze jedno ważne zastrzeżenie. Munafò i Flint skupili się na niedostatecznej rzetelności badań naukowych jako przyczynie tak znacznego rozpowszechnienia w piśmiennictwie fałszywych wyników. Jest to jednak tylko jedna strona medalu, gdyż również badania przeprowadzone rzetelnie, z zachowaniem nawet bardzo wyśrubowanych rygorów metodologicznych, mogą dawać fałszywe, a w najlepszym razie, mało istotne z teoretycznego i praktycznego punktu widzenia wyniki, jeśli będą badać „nieprawdziwe", nietrafne konstrukty teoretyczne. Jak to ujął Van Praag (1999, cyt. za: Ghaemi[3]), „badania nietrafnych konstruktów prawdopodobnie dadzą nietrafne wyniki, jak bardzo wyrafinowana byłaby zastosowana metodologia". Badania naukowe nie mają bowiem charakteru czysto empirycznej obserwacji i zawsze zakładają, w mniejszym lub większym stopniu, jeśli nie explicite, to implicite, pewną teorię. Można wręcz powiedzieć, że istotą pracy naukowca jest formułowanie i sprawdzanie teorii.[4] Jeśli więc teoria ma poważne wady czy braki, to opierające się na niej badania empiryczne same z siebie nie są w stanie doprowadzić do znaczącego rozwoju danej dziedziny nauki. W psychiatrii dobrą ilustracją tego problemu może być przypadek schizofrenii, której ciągle niezadowalające zrozumienie nie da się przypisać niedostatkowi zgromadzonych danych empirycznych. Jak dowodzą Tandon i wsp.[5] w swoim erudycyjnym przeglądzie stanu współczesnej wiedzy o schizofrenii, niewątpliwie udało się dotychczas ustalić liczne „fakty" na temat tej choroby, w znacznej części bardzo rzetelnie udokumentowane. Mimo to ta akumulacja ogromnej liczby „danych" lub „dowodów" nie jest już postrzegana jako wskaźnik postępu „wiedzy", ale raczej jako oznaka niepewności i zamieszania pojęciowego.[6] Zasadne wydaje się więc przypuszczenie, że skoro tysiące danych nie przyczyniają się istotnie do lepszego zrozumienia natury schizofrenii, to przyczyna trudności może leżeć raczej w niewłaściwej jej konceptualizacji. A jeśli tak, to badania empiryczne nad schizofrenią są prowadzone niejako po omacku i nie mogą przynieść zadowalających rezultatów. Chociaż więc problem trafności teorii stanowiących inspirację i przesłankę dla badań empirycznych wykraczał poza zakres komentowanego artykułu, trzeba mieć świadomość że jest on co najmniej równie istotny jak omawiane przez autorów zagadnienie rzetelności badań naukowych. Kwestie te są zresztą ze sobą ściśle związane.

W każdym razie tekst Munafò i Flinta uważam za ważny i ożywczy głos w dyskusji nad problemami współczesnej nauki, który powinien skłaniać naukowców i klinicystów do odpowiednio krytycznej lektury publikowanych prac badawczych, ale także do ostrożności i staranności przy projektowaniu własnych badań i interpretacji ich wyników.

Piśmiennictwo:

1. Ioannidis JPA. Why most published research findings are false. *PLoS Medicine* 2005;2(8):e124.
2. Maj M. The new impact factor of World Psychiatry. *World Psychiatry* 2010;9(3):129-130.
3. Ghaemi SN. *The concepts of psychiatry: A pluralistic approach to the mind and mental illness.* The Johns Hopkins University Press: Baltimore; 2003.
4. Popper KR. *Logika odkrycia naukowego.* Wydawnictwo Naukowe PWN: Warszawa; 2002.
5. Tandon R, Keshavan MS, Nasrallah HA. Schizophrenia, „just the facts": What we know in 2008. Part 1: Overview. *Schizophrenia Research* 2008;100(1-3):4-19.
6. Maj M. Understanding the pathophysiology of schizophrenia: Are we on the wrong or on the right track? *Schizophrenia Research* 2011;127(1-3):20-21.