Mirosław Szreder

Błędy ludzkie i losowe w badaniach naukowych

W badaniach statystycznych najtrudniej uniknąć ludzkich błędów, takich jak pomyłki w raportach o zrealizowanych testach na Covid-19. Dlatego nacisk kładzie się współcześnie na działania, które ułatwiają identyfikację tego rodzaju błędów, a dodatkowo mogą się okazać pomocne w ustaleniu prawdziwych wartości liczbowych.

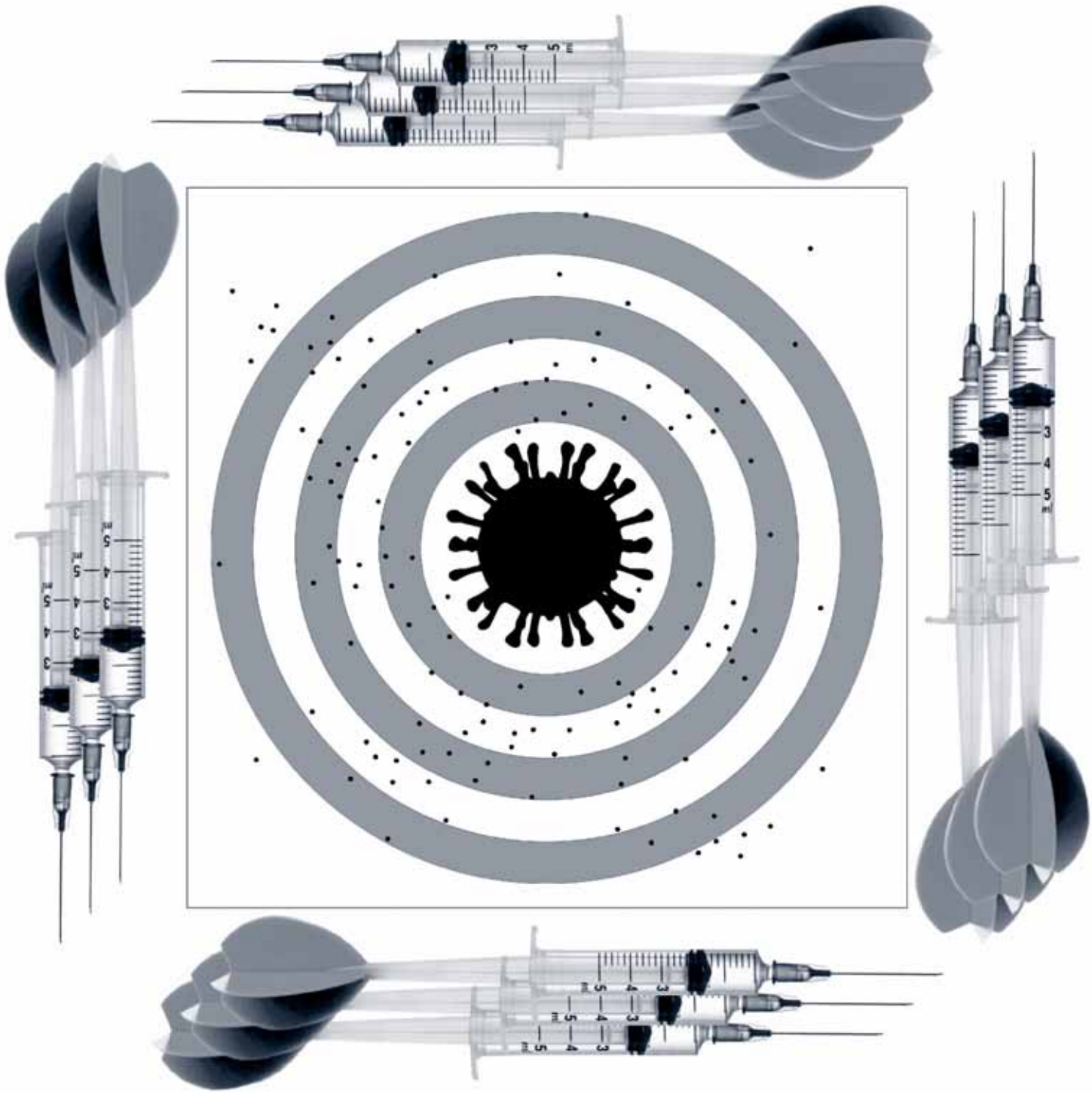
Od ogólnej puli przeprowadzonych badań odejmujemy 230 tys. testów – poinformowało Ministerstwa Zdrowia w komunikacie z 8 sierpnia br. Praktycznie więc od początku pandemii otrzymywaliśmy nieprawdziwe dane o liczbie wykonanych w Polsce testów na koronawirusa. Powodem tego, jak się okazało, był błąd w raportowaniu liczby testów zrealizowanych w laboratorium Wojewódzkiego Szpitala Zespolonego w Kielcach. Konsekwencje tej pomyłki są poważne, bo oznaczają konieczność korekty nie tylko codziennych danych o liczbie wykonanych w całym kraju testów przez ostatnie kilka miesięcy, ale także obliczanych na ich podstawie wskaźników, takich jak: liczba testów przypadających na mieszkańców Polski czy liczba stwierdzonych zakażeń (pozytywnych wyników) do liczby wykonanych testów. To są konsekwencje bezpośrednie. Ale obok nich warto dostrzec – chyba co najmniej tak samo ważne – implikacje pośrednie, skłaniające do ogólnej jakości danych statystycznych w badaniach naukowych.

Epidemia, która od marca w szybkim tempie obejmowała poszczególne obszary globu, wzbudzała rosnące zainteresowanie reprezentantów rożnych dyscyplin naukowych. Najpilniej śledzono postępy w badaniach nad szczepionką i nad lekami ułatwiającymi leczenie zakażonych na Covid-19. Ale podobne zainteresowanie budziło i nadal budzi poszukiwanie skutecznych działań prewencyjnych hamujących rozwój zakażeń. Te z kolei wymagają wcześniejszego poznania wzorców i prawidłowości statystycznych w transmisji epidemii. Nie da się tego dobrze zrobić bez odpowiedniej jakości danych liczbowych. I mimo że informacji o liczbach testów, zachorowań i zgonów przybywało każdego dnia, to ich jakość wzbudzała obawy naukowców w rożnych krajach i co rusz prowadziła do ostrych polemik w kręgach epidemiologów i statystyków. Najszerszym echem odbiła się w środowisku naukowym opinia profesora Uniwersytetu Stanforda, lekarza i naukowca Johna Ioannidisa, który już 17 marca ostrzegał, że wiele decyzji w pierwszych tygodniach pandemii podejmowane jest na podstawie niepełnych i mało wiarygodnych danych. Ze strony tych, którzy akcentowali konieczność szybkich działań w obawie przed utratą kontroli nad rozwojem epidemii, postawa taka musiała się oczywiście spotkać z surową krytyką. Po kilku miesiącach jednak, gdy zasoby danych statystycznych wzrosły wielokrotnie, nie maleje wcale grono naukowców, którzy skłonni są podkreślać przede wszystkim słabą jakość danych liczbowych na temat pandemii. Niektórzy gotowi są już teraz odpowiedzieć twierdząco na wyrażoną przez prof. Ioannidisa wątpliwość, czy w kwestii danych o koronawirusie nie doświadczymy wszyscy „wiekowej porażki" (once-in-a-century data fiasco).

Wyrazem tej niezadawalającej jakości danych w USA są – podobne jak w Polsce – przypadki błędów w raportowaniu (w lipcu stan Teksas wykreślił ze statystyk prawie 3,5 tys. pozytywnych wyników, które stanowiły jedynie „prawdopodobne zakażenia" u osób faktycznie niezbadanych), błędne łączenie w jedną kategorię testów serologicznych na obecność wirusa z testami na występowanie przeciwciał, niejasne kryteria oceny przyczyn zgonów, prowadzące najprawdopodobniej do zaniżenia w wielu krajach liczby zmarłych z powodu koronawirusa. Dziennikarze „New York Timesa" przytaczają oszacowania wskazujące na to, że niedoszacowana liczba zgonów spowodowanych pandemią tylko w 28 krajach (w tym w USA) przewyższa 161 tys. (wydanie z 31 lipca br.). Jak widać, źródeł błędów w danych liczbowych jest co najmniej kilka. Do najczęstszych zaliczyć trzeba: nieprecyzyjne definicje kategorii i pojęć występujących w badaniach, pośpiech w realizacji badań oraz zwykłe ludzkie pomyłki.

Nieco zaskakującym przykładem pośpiechu, który potencjalnie może doprowadzić do brzemiennych w skutkach błędów statystycznych, jest nagłe skrócenie okresu zbierania danych w realizowanym obecnie w Stanach Zjednoczonych spisie ludności i mieszkań (*Population and Housing Census 2020*). Wykonywany raz na dziesięć lat spis zostanie w swojej najważniejszej fazie skrócony o miesiąc – o czym informował m.in. „The Washington Post" z 4 sierpnia br. – sugerując polityczne powody tej decyzji. Przedstawiciele mniejszości dopatrują się w niej celowego dążenia administracji prezydenta Donalda Trumpa do niedoszacowania, czyli faktycznego zniekształcenia rzeczywistej liczby imigrantów i kolorowej ludności USA, co w przyszłości może przynieść korzyści partii republikańskiej. Zaska-

kujące jest przede wszystkim to, że decyzję o skróceniu czasu zbierania danych w tak poważnym badaniu podjęto niespodziewanie. Na ogół bowiem spisy ludności są przygotowywane i realizowane ze szczególną pieczołowitością oraz troską o jakość rejestrowanych danych. Wymagają wielomiesięcznych przygotowań, skomplikowanej logistyki i dobrze zorganizowanej pomocy respondentom, w tym pracy rachmistrzów spisowych. Warto w tym miejscu wspomnieć, że od kilkunastu mie-



Rys. Sławomir Makal

sięcy trwają w Polsce przygotowania do przyszłorocznego Narodowego Spisu Powszechnego Ludności i Mieszkań 2021. Po raz pierwszy będzie to samospis internetowy, co także może mieć wpływ na jakość pozyskanych danych.

Równoległe źródła danych

W badaniach statystycznych najtrudniej uniknąć ludzkich błędów, takich jak wspomniane wyżej pomyłki w raportach o zrealizowanych testach na Covid-19. Dlatego nacisk kładzie się współcześnie na działania, które ułatwiają identyfikację tego rodzaju błędów, a dodatkowo mogą się okazać pomocne w ustaleniu prawdziwych wartości liczbowych. Największe znaczenie wśród tych działań ma poszukiwanie lub organizowanie innego, równoległego źródła danych. Konfrontowanie liczb z co najmniej dwóch rożnych źródeł staje się jedną z podstawowych zasad współczesnych badań statystycznych, w których coraz większą rolę odgrywają dane pochodzące z rejestrów administracyjnych.

Pierwszym wyzwaniem w takich sytuacjach jest ustalenie jednakowych kategorii i definicji, które będą czyniły dwa lub więcej zbiorów danych wzajemnie spójnymi. Nieraz najbardziej podstawowe, wydawałoby się, kategorie, takie jak zgon lub fakt pozostawania bezrobotnym, wymagają precyzyjnych określeń, aby nie stały się powodem błędów statystycznych. Sama definicja „bezrobotnego" została w Wielkiej Brytanii zmieniona 31 razy w latach 1979-1996. Z kolei kategoria „zgonu" jest w USA – mimo obowiązującej od 1981 r. wspólnej dla całego obszaru Stanów Zjednoczonych koncepcji stwierdzania i określania zgonu (*Uniform Declaration of Death Act*) – różnie definiowana w poszczególnych stanach. Ktoś uznany za zmarłego w Alabamie może, przynajmniej co do zasady, nie zostać tak samo zakwalifikowany w sąsiednim stanie Floryda, gdzie zarejestrowanie zgonu wymaga potwierdzenia tego przez dwóch lekarzy – stwierdza brytyjski statystyk D. Spiegelhalter w książce *The Art of Statistics. Learning from Data*. Jednoznaczne definicje pojęć i precyzyjne słownictwo w formularzach ankiet są, obok dobrej organizacji badania, istotnymi warunkami uniknięcia poważnych błędów o charakterze nielosowym.

Na potrzebę równoległego gromadzenia danych z dwóch rożnych źródeł wskazują niespójne statystyki przestępstw w wielu krajach. Na przykład w Wielkiej Brytanii i w USA, niezależnie od policyjnych rejestrów przestępstw, prowadzone są regularnie sondażowe badania osób i gospodarstw domowych na temat przestępczości. Wyniki zaś, a nawet odczytane tendencje, okazują się niekiedy rozbieżne. Podczas gdy policyjne statystyki dla Anglii i Walii wskazywały na wzrost o 13% liczby przestępstw w latach 2016-2017, to z badania reprezentacyjnego 38 tys. respondentów wynikało, że skala przestępstw zmniejszyła się o 9%. Częściej jednak różnice te mają przeciwny znak, tj. większa jest liczba przestępstw wynikająca z badania ludności niż ze statystyk policyjnych. Powodów tych rozbieżności może być wiele. Nie wszystkie wykroczenia i przestępstwa są zgłaszane policji, nie zawsze osoba jest świadoma tego, że stała się ofiarą przestępstwa (np. dziecko ulegające przemocy) albo usilnie pragnie uniknąć kontaktu z policją (osoby z nieuregulowanym statusem pobytu w danym kraju). W Stanach Zjednoczonych szacuje się, że policja rejestruje jedynie 55% poważnych przestępstw kryminalnych oraz 35% przestępstw dotyczących mienia (włamania, kradzież pojazdów). Przy czym, jeśli chodzi o samą kradzież samochodów, to statystyki te sięgają 75%, co wynika z obowiązku przedłożenia towarzystwu ubezpieczeniowemu przez poszkodowanego zaświadczenia o zgłoszeniu policji faktu kradzieży.

O tym, że statystyki policyjne są niepełne, wiadomo z badań realizowanych corocznie na reprezentatywnej próbie ok. 95 tys. gospodarstw domowych (*National Crime Victimization Survey*). Wywiady przeprowadzane są ze wszystkimi członkami wylosowanego gospodarstwa domowego w wieku co najmniej 12 lat. Od kilku dekad krzywe prezentujące dane o liczbie przestępstw przebiegają na osi czasu równolegle, z tym że ta prezentująca dane z policyjnych statystyk położona jest poniżej krzywej skonstruowanej na podstawie danych z sondażu.

Długie ciągi liczb słabej jakości

W powszechnym odbiorze dane statystyczne uzyskane z sondażu – reprezentacyjnego badania statystycznego – są traktowane jako mniej wiarygodne od danych administracyjnych lub danych ze spisu powszechnego, tzw. badania pełnego. Być może na takie postrzeganie badań próbkowych ma wpływ to, że ich wyniki opatruje się zwykle komentarzem informującym o maksymalnym błędzie wnioskowania (często wynoszącym 3% lub mniej). U niektórych odbiorców powstaje podwójnie błędne przekonanie. Po pierwsze, że błąd ten (poprawnie nazywany błędem losowym albo błędem losowania) odpowiada za wszystkie możliwe rodzaje błędów, jakie mogły wystąpić w badaniu. Tak oczywiście nie jest, bo nie odnosi się on w ogóle do błędów o charakterze nielosowym, o których wspominaliśmy wyżej. Po drugie, że badania statystyczne obejmujące wszystkie jednostki zbiorowości (spisy powszechne, rejestry urzędowe) nie są obciążone żadnym błędem. To także nie jest prawdą, bo wspomniane błędy ludzkie, błędy definicji pojęć, przetwarzania danych, braków odpowiedzi mogą poważnie zniekształcić każde badanie. Prawdziwym kłopotem dla statystyków nie są błędy losowe, których wielkość maleje wraz ze wzrostem liczebności próby badawczej, lecz błędy nielosowe, niepodlegające tej prawidłowości. Tym błędom poświęcać się będzie coraz więcej uwagi. I nawet w gorącej w ostatnich latach dyskusji na temat kryzysu replikowalności eksperymentów w naukach przyrodniczych i społecznych trudno byłoby zupełnie abstrahować od znaczenia błędów nielosowych. Istotą tego kryzysu są co prawda kwestie rozstrzygania i komunikowania o prawdziwości lub nieprawdziwości hipotez badawczych, a więc odnoszące się do samego modelu wnioskowania statystycznego. Trzeba w nich jednak dostrzec – szczególne w dziedzinie nauk społecznych – takie kwestie, jak nielosowy dobór próby (np. w drodze autoselekcji, dobrowolnego zgłoszenia się do badania z wykorzystaniem internetu) oraz błędy ludzkie, w tym popełnione w rejestrowaniu i przetwarzania danych, które mają swój udział w szerokiej gamie źródeł tego kryzysu.

Co to oznacza dla teraźniejszości i przyszłości badań statystycznych? Przede wszystkim, jak sądzę, zmagania z pandemią koronawirusa uwydatniły potrzebę posiadania nie tylko odpowiedniej ilości, ale głównie dobrej jakości danych liczbowych w podejmowaniu ważnych decyzji. Ale obok tego, coraz bardziej oczywiste i zrozumiałe staje się to, że duże zbiory danych i długie ciągi liczb nie są w stanie zrekompensować słabej ich jakości. Błędy o charakterze nielosowym, będące wciąż dużym wyzwaniem dla statystyków, nie są funkcją liczebności próby. W ważnych rozstrzygnięciach, nie tylko medycznych czy epidemicznych, badacze polegać będą na danych liczbowych pochodzących z kilku źródeł o sprawdzonej wcześniej wiarygodności bądź utrwalonej reputacji.

Prof. dr hab. Mirosław Szreder, Uniwersytet Gdański