How sure are you of your result? Put a number on it

Any scientist publishing a claim should quantify their confidence in it with a probability, argues **Steven N. Goodman**.

Picture our bafflement if weather forecasts said, "There is a non-statistically significant chance of rain tomorrow," rather than, "There is a 60% chance of rain." Or the confusion among Florida residents if, instead of being told, "Tallahassee has an 85% chance of a direct hit by Hurricane Michael," they heard: "Tallahassee will be hit, $P = 0.03$."

When the stakes are high, we need accurate and understandable risk estimates to make informed decisions. We demand that weather services clearly convey the chance of rain or hurricane, because lives and livelihoods are at stake. That's why forecasts distil immensely complex models into one number.

Scientists should do the same.

Let's require that any researcher making a claim in a study accompany it with their estimate of the chance that the claim is true — I call this a confidence index. As well as, "This drug is associated with elevated risk of a heart attack, relative risk (RR) = 2.4, $P = 0.03$," investigators might add: "There is an 80% chance that this drug raises the risk, and a 60% chance that the risk is at least doubled."

Analyses using Bayesian statistical methods, which generate the probability of a hypothesis being true, go part way down this path. For example, a 2017 study calculated that induced hypothermia has a 76% chance of benefiting newborn babies who have brain damage from oxygen deficiency, even though the test and control groups were not statistically different — I calculated the *P* value for this study to be 0.6 (A. R. Laptook *et al. J. Am. Med. Assoc.* **318**, 1550–1560; 2017).

A confidence index would formally incorporate the impact of previous evidence (as some Bayesian analyses do) and investigators' judgement about the plausibility of a claim's explanation. Importantly, a confidence index should capture the limitations of the study that are currently addressed only qualitatively. It would apply whether or not researchers also calculate confidence intervals, a separate metric.

Many scientists assume that the *P* value is a confidence index; a widespread, mistaken belief is that a *P* value under 5% implies a 95% or greater probability of the effect. But the *P* value does not measure the probability that the null hypothesis is true. This stubborn misconception so distressed the American Statistical Association that in 2016 it issued a rare public statement to dispel it, and to discourage the use of 'bright line' *P*-value thresholds (usually 0.05) to justify claims (see go.nature.com/2p9hcxn). More reliance on confidence intervals has been proposed as a remedy, but these are also often used in 'bright-line' fashion and share many of the limitations of *P* values. Most crucially, a confidence interval from a reliable study can be identical to one from a study in which we have zero confidence.

Claims are often communicated so obliquely that it is hard to know what to make of them. If it is statistically significant, the existence of a relationship is asserted as if it is definitively true, as in "Fibre intake reduces cancer risk by 18% (confidence interval 2% to 34%), $P = 0.02$". If the result is not statistically significant, an array of statements is possible, from "there is no difference …" to "there is a trend …" and various other creative circumlocutions.

This fuzziness makes clear communication difficult between scientists, and all but impossible with others, from journalists to doctors, policymakers and the public.

Some people will say that they cannot translate all the nuances of research into one number, but the practice has ample precedent. The Intergovernmental Panel on Climate Change puts confidence levels on its statements. Crowdsourcing techniques such as prediction markets, in which people place bets on outcomes, have been used to estimate with decent accuracy the chance that scientific studies will be replicated. Some people have proposed that scientists bet their own money on their claims, in part to discourage over- or under-confidence and other cognitive pitfalls.

FUZZINESS MAKES CLEAR COMMUNICATION DIFFICULT BETWEEN SCIENTISTS, AND ALL BUT IMPOSSIBLE WITH OTHERS.

A confidence index could help in other ways. Some researchers manipulate analyses or selectively report outcomes to achieve statistical significance (a process called *P* hacking) because publication, recognition and funding are most likely to flow from statistically significant studies. This year, a survey of 390 biostatisticians found that at least 20% had been asked by a collaborator to manipulate their data or analysis to exaggerate their results' importance (M. Q. Wang *et al. Ann. Intern. Med.* **169**, 554–558; 2018). With a confidence index, because there is no 'bright line' to aim for, the incentive to hack it might be replaced by an incentive to get it right.

Of course, the reasoning and method used to calculate a confidence index should be reported. The foundations for such methods already exist — in Bayesian statistics, sensitivity analyses and more — although they need further development. And then there is scientific judgement: the same judgement behind the words currently used. But numbers always speak louder than words, and it is beyond time to convert those words into numbers with clearer meaning.

Although simple on paper, requiring a confidence index would entail a profound overhaul of scientific and statistical practice. But the crisis of reproducibility and credibility in research demands no less. Crises should not be wasted; if there was ever a time for transformation, it is now. ■

Steven N. Goodman *is professor of medicine and of epidemiology at Stanford University in California.*
e-mail: steve.goodman@stanford.edu