

TOM SIEGFRIED } **RANDOMNESS**

Medicine needs a sensible way to measure weight of the evidence



Let's talk about evidence-based medicine.

Suppose you're in the hospital and a nurse takes your temperature to find out whether you have a fever. Providing that the thermometer is working properly, it will give you a number that answers the question. It's all the evidence you need. It doesn't matter how many other patients in the hospital have had their temperature taken lately.

But suppose you need to gather more sophisticated medical evidence: You want to know whether your version of a specific gene predisposes you to a certain disease. Let's say scientists have just finished a study of people with that disease to see which gene variants turn up more often than usual.

Again, you get a numerical answer — not a temperature, but a “P value.” It's the probability that even if there was no real link, the study would turn up the same number of people (or more) with your version of the gene. A low P value should mean that chances of a bogus result are small. But unlike with temperature, a P value is not all the evidence you need. It turns out that the strength of a P value's evidence depends on other things, such as how many different gene variants the study examined. (The more things you study, the more low-probability flukes you'll find.)

Plenty of other objections can (and have) been raised about the worth or lack thereof of drawing medical or other scientific conclusions on the basis of P values. As a measure of evidence, P values have more flaws than college football's system for ranking the top teams. It's too bad P values aren't more like temperatures.

Or, to make the point more generally, it's too bad that statistical evidence cannot be quantified directly, with the same sort of logical consistency as thermometer readings.

“It is obvious to us that measurement devices lacking fixed units and constancy of scale across applications are problematic, yet we seem oddly laissez-faire in our approach to measurement of one critically important quantity: statistical evidence,” writes Veronica Vieland of the Battelle Center for Mathematical Medicine at the Research Institute at Nationwide Children's Hospital in Columbus, Ohio.

As measures of evidence, Vieland points out, P values have several questionable properties. For one thing, P values can't be used

“The P value is in some very fundamental way simply not tracking with the evidence at all.”

to rank-order genes suspected of contributing to disease risk — a given P value’s strength of evidence differs from gene to gene.

What’s more, statistical tests cannot be logically combined to reflect the strength of the evidence, either. Suppose a statistical test suggests that a coin is weighted to come up heads more often than tails, for instance. A second test indicates the same bias. Logically, the two tests together constitute stronger evidence than either test alone. But the P value for the combined tests might come out at an intermediate value between the two.

“This means that the P value is in some very fundamental way simply not tracking with the evidence at all,” Vieland writes. In short, measuring medical evidence today is about as advanced as measuring temperature with uncalibrated thermometers.

In the 19th century, physicists solved the temperature problem by establishing an absolute temperature scale that removes all the ambiguity about measuring heat. That scale grew from understanding the interplay of temperature, pressure and volume as heat flowed in and out of a system (such as a steam engine), as quantified by mathematical formulas expressing the laws of thermodynamics. Perhaps, Vieland argues in a recent issue of *Human Heredity*, similar math could be devised to account for the flow of information into an assessment of evidence.

It’s not that far-fetched. Information theory, after all, relies on a concept called entropy that is precisely analogous to the entropy measured in thermodynamics. “Evidence,” however it is ultimately quantified, ought to have some relationship to information, as Vieland and collaborator Susan Hodge of Columbia University argue in another recent paper, in *Statistical Applications in Genetics and Molecular Biology*.

Evidence and information are not identical, of course. Lots of information could be collected that yields very little evidence for or against a hypothesis. But perhaps some sort of “evidential energy” is conveyed by the information and could be measured much the same way temperature is, Vieland and Hodge propose.

“Whether the mathematics of thermodynamics can be directly harnessed to yield an absolute measure of evidence ... remains to be seen,” they write. “But one way or another, we cannot meaningfully discuss measurement of statistical evidence without being clear on fundamental issues of measurement.”

In the meantime, medical statistics will continue to be full of a lot of hot air. ▣