Data-driven hypotheses

Paul van Helden

Scientists can be passionately critical of different approaches to a given problem. Over the years, I have found myself crossing swords frequently with epidemiologists with whom I am otherwise good friends ‘off the pitch’. To some extent, this difference in approach seems to me to be grounded in the logistics of gathering or generating data, and the analysis thereof. The epidemiological approach is often incredibly labour-intensive with regards to data gathering—consider the interview approach and dietary studies that require subjects to recall meals eaten—and therefore requires careful thought and planning to achieve a meaningful outcome with limited resources. By contrast, many areas of molecular biology generate an enormous amount of data easily and quickly, and therefore lend themselves to a different way of thinking and heated debate about the ‘correct way to do science’. Such arguments are futile at best and possibly damaging to science at worst. What is regarded as the correct approach might be, in part, due to limitations in generating data, but it might not be necessary or correct in the future, as the history of science has shown.

The accepted scientific method consists of formulating a hypothesis and then testing it by experimentation which, at least in theory, attempts to disprove the hypothesis. Experiments generate data that can be analysed by various means to test the hypothesis. This is a common approach in epidemiological studies—for example, for identifying a risk factor for a given disease. This approach necessitates that one proposes a risk factor for the condition or disease—alcohol consumption or dietary deficiency—and then collects data pertaining to that factor. Of course, studies of social or nutritional risk factors often simultaneously collect data for testing multiple parameters.

The main risk of such studies, in my opinion, is that the researcher must first predict or guess the risk factor and then collect data

pertaining to it. Of course, previous work might have suggested or even identified a risk factor in a separate study. Nevertheless, I suggest that such an approach can be fundamentally deficient, as the researcher approaches the study with his or her inherent limitations of knowledge and experiential bias, merely by choosing one possible risk factor out of perhaps hundreds of possibilities. There is a good chance that the risk factor will be confirmed as such and give rise to a publication, as a result of the bias in publishing positive studies and, particularly, studies that verify previous findings.

By contrast, the new ‘omics’ technologies allow us to generate massive quantities of data rapidly and thereby enable us to take a far less biased approach to tackle a given problem. There are already about one million transcriptome data sets available [1], or single-nucleotide polymorphism (SNP) chips that can analyse one million SNPs each. One might easily use these to investigate a large number of individuals, generating one billion separate pieces of information per 1,000 individuals, with no prejudice or preconceived ideas.

Equipped with such huge data sets, we can perform data mining in an objective way. For some purists, this approach to data acquisition is anathema, as it is not ‘hypothesis-driven’. However, I submit that it is. In this case, the original hypothesis is broad or generic—we generate data, assess it and probably find something useful for elucidating our research problem. The broad hypothesis states that we use the results to generate models that identify differences, for example between experimental subjects and controls, without specifying what those differences are and without collecting specific and limited data sets. The ‘old-style stickler’ might find this approach unacceptable; however, it might be the best way to avoid bias. Contrary to what some have suggested to me, this approach is not simply playing

with data to generate a hypothesis, which would violate the principle that one should not look for a primary hypothesis in results. The hypothesis is that one will design an algorithm and find a pattern, which allows us to distinguish between cases and controls.

By using this approach, the examination of large data sets might generate useful and specific leads for further study and validation. These follow-up studies can be governed by the traditional hypothesis-driven approach: ‘biomarker X’ is a risk factor for ‘condition Y’. Such a combination of data-gathering and hypothesis-driven approaches might be the only way to understand complex diseases, even infectious diseases, in which the invading pathogen might be necessary for disease, but in itself is inadequate as a single risk factor.

Of course, if we examine large data sets to find interesting patterns or biomarkers that might correlate with a given condition, we will probably identify false positives at a rate of at least 1:20, simply by chance. This is why replication studies that use independent sample sets are important. As long as we have proper controls in place and use statistics appropriately, this approach to science should yield wonderful new results and massively increase our knowledge of the world, instead of merely proving or refuting notions that we already suspect.

CONFLICT OF INTEREST

The author declares that he has no conflict of interest.

REFERENCE

1. Baker M (2012) *Nature* **487**: 282–283

Paul van Helden is a professor and Head of the Department of Biomedical Science at Stellenbosch University, Faculty of Health Science, Tygerberg, Cape Town, South Africa. E-mail: pvh@sun.ac.za

EMBO reports (2013) **14**, 104; published online 21 December 2012; doi:10.1038/embo.2012.207