Commentary: Evidence synthesis and evidence consistency

AE Ades

The paper by van Valkengoed is a critique of the way in which investigators have chosen parameter values for cost effectiveness analyses (CEA), but it also suggests a way of looking at evidence that is seriously underused in both epidemiology and medical decision making.

The essential problem is set out schematically in Figure 1. The adult female population is at risk of chlamydial infection; a proportion of those infected develop pelvic inflammatory disease (PID), and PID may in a proportion of cases lead to ectopic pregnancies and/or infertility. Other causal pathways can lead to these outcomes, although it may be possible to determine approximately what proportion of them can be attributed to chlamydia.

Figure 1 is of course a simplification, it ignores the potential effects of time, of age, and severity of infection. However, it sets out the parameters on which we require information in order to conduct CEA of potential screening programmes. The cost effectiveness of any screening programme compared with no screening will be strongly influenced by what we assume the values of these parameters to be. In particular, cost effectiveness will be very sensitive to the absolute numbers of PID and ectopic pregnancies prevented, and hence to the absolute numbers we assume are occurring in the absence of screening.

Valkengoed et al. show that if one takes the estimated progression rate parameters \( b, c, \) and \( d \) assumed in published CEA, and applies these to the prevalence of chlamydia \( a \), as observed in studies in the Netherlands, the predicted numbers of PID, ectopic pregnancies, and tubal factor infertility cases are far in excess of the numbers actually observed in local registration data.

The situation can be characterized as follows: we have cross-sectional survey evidence on chlamydia prevalence, \( a \); we have prospective study data to inform progression parameters \( b, c, \) and \( d \). The register- or possibly hospital-based evidence on the numbers of women with PID, ectopic pregnancies, and infertility due to chlamydia are, in effect, a third type of evidence that ‘closes the loop’, and tells us about products of the basic parameters. The number of PID cases for example gives us information on the product \( ab \), and the number of ectopic pregnancies on the product \( abc \). It may be necessary to make adjustments that subtract a proportion of the outcomes that appear to result from other causes, but nevertheless, the logic of the situation is clear. (And, in any case, if the predicted number of PID exceeds the numbers observed before adjustment, the mismatch will only become worse after the adjustment is made.)

There has to be a reason why sources of information should disagree. Either the model, which tells us how the data sources should be related, is mis-specified, or the data are not giving us an unbiased estimate of the parameters specified by the model. If we consider PID, for example, we have two parameters, \( a \) and \( b \), and three sources of information. The model (Figure 1) does not seem controversial, so the data must be at fault. Valkengoed et al. identify the data informing parameter \( b \) as the culprits. Earlier investigators had taken their estimates of sequelae rates from high risk populations with clinical disease, and plugged them into CEA intended to apply to the same kind of lower risk, general population being considered by the Amsterdam group. Proposed screening programmes might then promise to ‘prevent’ more illness than was actually occurring!

If we somehow ‘knew’ that all the evidence sources were unbiased, i.e. appropriately representative for the target population of interest, then no consistency problem could arise. We would not expect the estimates for \( a \) and \( b \) exactly equal what is estimated by our data on \( ab \), but we would expect any differences to result only from statistical sampling, and to eventually diminish if all sources of data were based on very large numbers. In this situation, the ‘best’ estimate of the parameters \( a \) and \( b \) would be obtained by statistically combining all three sources of evidence. This, of course, could be achieved by a simple weighted average, but the situation in Figure 1 is more complex: there are several sources of information that have a bearing on some of the parameters.

There are formal methods for combining evidence structured in this—or still more complex—ways. Best known in the health field is the Confidence Profile Method of David M Eddy and colleagues, but strikingly similar ideas have cropped up repeatedly in different fields, although employing different technical machinery. These include estimates of the physical constants, applications in environmental health risk assessment under the heading Bayesian Monte Carlo, and combinations of evidence to inform deterministic models. All these examples combine information on model parameters with information on model outputs.

The idea of ‘borrowing strength’ across potentially complex networks of evidence has a Bayesian flavour. The freely available Bayesian software WinBUGS is becoming a popular way of carrying out the complex computations underlying a wide range of evidence synthesis problems.

However, almost wholly absent from this earlier work was an appreciation of the importance of consistency between different sources of evidence. Before combining information, we need to determine whether, given a particular model, the different sources of evidence are telling us the same things about the parameters. Of course, we can never get to examine ‘consistency’ unless we have more data points than parameters. Statistical combination of all the available data, including data on combinations of parameters, should not only give us better...
estimates, but also give us the best chance of revealing inconsistencies in the evidence base—in effect a statistically-based version of model validation.

Valkengoed’s paper is based on a ‘deterministic’ version of the Figure 1 model. Here, each parameter is fixed at its ‘best’ value. The formal methods cited above compute a ‘probabilistic’ version of the model, taking account of the statistical uncertainty in each item of data, and propagating uncertainty correctly through the model. However, the paper is none the worse for having assessed consistency informally and without the aid of complex statistical methods.

But what should be the response to inconsistency? Also, how should the totality of evidence be used when there is no clear statistical evidence for inconsistency, but we are unsure whether consistency can be assumed? Valkengoed seems to imply that we simply drop the biased—or more exactly inappropriate—evidence on disease progression, and infer the disease progression parameters by combining the high-quality prevalence data with the register data.

More generally, questions on the quality of evidence, or the causes and extent of bias are, of course, routinely raised and debated by epidemiologists, but rational and transparent decision making cannot generally be delayed while perfect evidence is gathered. We have to wrap-up the debate and agree on a provisional ‘best’ answer, making explicit the degree of uncertainty at each point. (No coincidence that most of the methodological work cited above, as well as Valkengoed’s work, was motivated by the need for rational policy choice.)

The stance we take on different types of evidence is not really a technical or statistical issue, but a question of expert judgement. When evidence does not ‘add-up’ there has to be a reason. It is up to experts who know the data and the epidemiology to determine whether the model is wrong, whether the data are so biased as to be irrelevant, or whether the information is usable after bias adjustment, or whether some form of implicit or explicit re-weighting of evidence is appropriate. A model based on all available evidence, including evidence on complex functions of parameters, backed by an expert consensus on quality and relevance of data, is a strong basis for credible and transparent decision making, and may sometimes reveal something interesting about the science too. Use of all available data also gives us a more realistic assessment of uncertainty, and thus a more objective basis for a rational research agenda.

References