GETTING THE BEST OUT OF EXPERTS: A REVIEW

Mark Burgman

Centre of Excellence for Biosecurity Risk Analysis, School of Botany, University of Melbourne

INTRODUCTION

Skills are abilities to execute particular tasks efficiently and effectively, acquired through training and practice, repetition and feedback (Welke et al. 2009). Expertise, in contrast, refers to judgement and prediction, rather than repetitive, concrete action, and is not necessarily supported by verifiable actions and outcomes. Experts are usually professionals who are considered by their peers or by society at large to have specialist knowledge in a particular domain, and who are consulted to make a judgement or prediction.

There is a continuum between skill and expert judgement. An engineer's skill may be to design a particular kind of bridge. Circumstances may be such that we consult them on related matters in which they have no direct experience, such as building other kinds of bridges. Beyond that, they may also appear to be expert in more distantly related topics such as other structures, but have no repeated exposure beyond the things they have seen in textbooks or heard about from colleagues. At what point does their ability to judge or predict become no better than that of a random person from the street? Do they know, themselves, when their knowledge becomes too thin? Do their peers know?

We rely on experts when we have to make decisions and we do not have enough information. Our reliance is greatest when circumstances are unique, the consequences of the decision are significant, the decision is imminent and we have to make judgements about future and uncertain situations. We find someone with the right training and experience about the topic at hand, someone whom we trust and can understand (Meyer & Booker 1990; Gullet 2000). Often, expert judgement is all we have.

Broadly, experts help with three kinds of questions (French 2012). We seek their judgements of simple, verifiable facts, such as:

- What is the disease rate in the population?
- What is the maximum weight this bridge can carry?

Alternatively, we ask them to predict events, such as:

- Will the president still be in office next year?
- How much rain will fall next week?

Quite often, we ask more comprehensive questions about a best course of action, such as:

- What is the best way to manage this problem?
- Is this the best portfolio of investments for me?

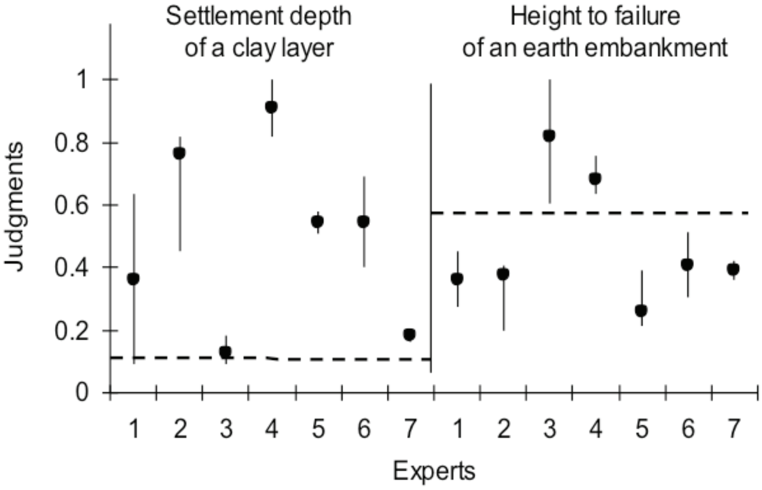
When experts answer questions about verifiable facts, we want them to draw on the storehouse of data they have accumulated through training and experience. In the case of predicting outcomes of events, we want them to use mental models together with their treasure trove of data and experience. We are especially demanding when asking about a course of action, because we expect the expert to have data and models on hand, and to understand our context and sensitivities. We trust them to have our best interests in mind. We will see below, often this is not the case. Society generally accepts that scientific and technical experts provide a unique and valuable resource. The US National Research Council, for instance, asserts that scientific experts have indispensible knowledge, methodological skills and experience (Stern and Fineberg 1996).

WHAT IS THE PROBLEM?

Despite such optimism, expert judgements can be worryingly frail. Geophysicist Ellis Krinitzsky spent many years working on earthquake risk, a notoriously difficult scientific problem. In an early review on the reliability of expert judgements, he described an experiment in which seven geotechnical experts predicted the height of fill at which an embankment would fail, and the depth to which sediment would settle in specific circumstances (Krinitzsky 1993). These questions were typical of the kinds of problems geotechnical experts were expected to assess reliably. The experts were provided with the data necessary to make calculations. They used a variety of methods.

The results were not heartening. There are six important things to note about the results of this simple experiment, shown in Figure 1. The dashed lines represent the correct answers to the two questions. The dots are the experts' best guesses and the vertical lines connect their 'minimum' and 'maximum' estimates.

Figure 1: The results of an experiment in expert judgement described by Krinzsky (1993) in which seven geotechnical experts estimated two facts: the height of fill at which an embankment would fail, and the depth to which sediment would settle.



The correct (measured) value for settlement depth was 1.5 cm and for height to failure was 4.9 m. The x-axis for both in Figure 1 is rescaled so the maximum value is 1. Correct values are shown as dashed horizontal lines. The intervals join the 'minimum' and 'maximum' values reported by the experts.

First, the experts were reasonably sure that the truth lay within the interval shown by the lines connecting their minimum and maximum guesses. In the first case only two people's intervals enclosed the truth. In the second case, no-one's interval enclosed the truth. If their estimates of uncertainty were generally reliable, we would expect most of the intervals to enclose the horizontal dashed lines. Because they did not, it means that in both cases, the experts were overconfident when they assessed the reliability of their own knowledge.

Second, geophysicists conducted the study in the 1970s. That is, technical experts have been aware of these kinds of phenomena for at least 40 years.

Third, it is possible for everyone to be wrong in the same direction. In the left-hand panel, all the experts overestimated the truth. That is, experts may be biased.

Fourth, the fact that someone did well on one question does not mean that they will do well on another. Expert 4 did best in the right-hand panel and worst in the left-hand.

Fifth, the width of the intervals between the minimum and maximum values tells us how confident they were. In the left-hand panel, expert 3 was confident (their interval was narrow) and accurate (their best guess was close to the truth), whereas expert 5 was confident and inaccurate. More generally, there was no clear relationship between confidence and accuracy.

Lastly, these were all credible professionals. They would have passed muster as expert scientists in a court or serving on a government panel dealing with the safety of earth embankments. All were well-credentialled members of scientific societies, attending an international scientific conference. No doubt each had a confident and plausible story to tell about how they arrived at their estimate and could defend the interval that they gave with their answers.

Misjudgements such as those reflected in the geophysicists' judgements above may seem relatively benign, but experts' mistakes may have important consequences. They include the mistaken interpretation of fingerprint evidence (Ulery et al. 2011; Dror 2005) and diagnostic errors in clinical medicine (Elstein 1995; Berner & Graber 2008).

Motivational bias is a conscious or subconscious adjustment of an expert's objective judgement attributable to the expert's values or their prospects for personal reward (Kunda 1990). Scientists, for example, are taught to believe in the objectivity of the scientific method. They find it very difficult to imagine that another scientifically trained person, equally clever and with access to the same data and models, could come to a different conclusion. As a result, they can be very credible because they believe in their own objectivity.

Decisions involve both facts and values. Values are statements about what we want or what we think is important (Gregory et al. 2012). Facts or technical judgements are statements about quantities or events that could be verified with independent information, at least in theory. Experts are not entirely objective and independent. Their judgements are compromised by perceptions, values and conflicts of interest (Shrader-Frechette 1996). In most practical situations, the pool of potential experts is small, composed of people with overlapping experiences, so their judgements are not independent. Values are inescapable because measures of consequence and impact are inherently value-laden (Slovic 1999).

When the NASA Space Shuttle *Challenger* exploded soon after launch in 1985, the public wondered how it could have happened. After all, Christa McAuliffe, the schoolteacher who died on board the rocket together with the astronauts, had been told the risk of a failure was 1 in 100,000 launches. Up to that point, there had been only about 100 launches, and there had been no failures.

Physicist Richard Feynman was part of the team that investigated the accident. He noted (Feynman 1986) that the range safety engineer had studied all previous rocket flights and found that out of a total of nearly 2900 flights, 121 failed (1 in 25). The engineer noted that with special safety systems, a figure of below 1 in 100 might be achieved but '1 in 1,000 is probably not attainable with today's technology.'

This judgement was in stark contrast to the NASA manager, who argued that, since the shuttle was a manned vehicle, 'the probability of mission success is necessarily very close to 1.0' (Feynman 1986). Managers believed that the probability of failure should be as low as 1 in 100,000. They could estimate this level of safety only by ignoring their own records that showed difficulties, near accidents and accidents, all giving warning that the probability of flight failure was not so very small. Feynman concluded: 'It would appear that, for whatever purpose, be it for internal or external consumption, the management of NASA exaggerates the reliability of its product, to the point of fantasy'. NASA management had a vested interest in a safe system, and convinced themselves and others that it was so, despite the data. In hindsight, we can see there were 135 missions in the Space Shuttle program between 1981 and 2011, and 2 catastrophic failures, much closer to the range engineer's assessment than to NASA management's estimate.

WE RELY ON EXPERTS, EVEN WHEN WE SHOULD NOT

Interestingly, our propensity to ignore evidence is quite pervasive. In 1954, psychologist Paul Meehl published a book that summarised about 20 studies comparing the clinical diagnoses of doctors with the predictions of simple statistical models (Meehl 1954). The models outperformed expert diagnoses consistently. Meehl's book caused a controversy (Meehl 1986) that is still not resolved. It was confronting for clinicians to be told that a simple statistical model would make fewer mistakes than an experienced professional.

Meehl and his colleagues repeated the analysis in 1996 (Grove & Meehl 1996). They found 136 medical and mental health studies comparing clinical and statistical prediction. Yet, despite decades of consistent research findings in favour of the statistical method, most professionals continue to use subjective, clinical judgements and do not use quantitiative tools, even when they are available. For example, expert cardiologists generally do worse than the predictive equations recommended by the American College of Cardiology (Lipinski et al. 2002). In their 2008 review of clinical misdiagnoses, Berner and Graber (2008) noted: 'Decision-support tools have the potential to improve care and decrease variations in care delivery, but, unfortunately, clinicians disregard them, even in areas where care is known to be suboptimal and the support tool is well integrated into their workflow'. The superiority of simple models over expert judgement has also been demonstrated for legal opinion (Martin et al. 2004).

THE CAUSE OF THE ADDICTION

Why do we persist in ignoring data and trusting experts, even when the data are available? After all, neither patients nor doctors want doctors to use statistical tools. People listen to political pundits, scientists and financial advisers, despite the data.

Grove and Meehl (1996) suggested that experts are motivated to ignore data through fear of becoming redundant. Experts value esteem and status. Consider how unhappy senior partners in a law firm would be, say Grove and Meehl, to learn that paralegals with a few years of experience could predict the opinions of an appellate court as accurately as a partner can. Experts often hold a fondness for a personal theory. Mathematical prediction is often seen as dehumanising, especially when people lack education in quantitative methods.

From the perspective of the user of expert advice, it may be that it is mentally difficult to make carefully reasoned decisions. We tend to 'offload' the effort to someone we believe is better equipped to perform the task (Engelmann et al. 2009).

Forecasting specialist Amstrong (1980) speculates that many clients are mainly interested in avoiding responsibility, and do not care about accuracy. A client who calls in the best expert available at the time avoids blame if the forecasts are inaccurate. Thus, decisions may be affected by the desire of officials to avoid the possibility of being held to blame.

Psychologist Daniel Kahneman had read Meehl's book when he got a job assessing who would make good military leaders. Disarmingly, Kahneman (2011) remembered that the statistical evidence should have shaken his confidence in his judgements of particular candidates, but it did not. He was reminded of visual illusions, which remain compelling even when you know that what you see is false, and coined the term the illusion of validity.

The reason we rely on expert judgement lies in a deep-rooted human need to believe that certainty exists, and that if just consult the right oracle, we will discover it. Society creates hierarchies of technical and scientific status that pander to this need (Evatts et al. 2006). These edifices resist data and criticism and generate self-fulfilling pronouncements.

In practice, evidence pertaining to any real-world decision is compiled from a myriad of sources. Experts may draw on data and personal experience. They may also consult friends and colleagues, and information in books, papers and reports. Experts dredge information from these sources and combine it in unstructured ways, filtering it through their memories and their personal psychological baggage. Rarely are their judgements cross-examined or verified.

A key problem is that experts often assume a position of authority, reinforced by professional status. It can intimidate people who wish to examine expert judgements critically, leading to a culture of technical control in which expert opinions are rarely challenged successfully (Walton 1997). Critically, the traditional view of experts excludes people with useful knowledge.

Expert judgement is unavoidable. Experts and expert advisory panels are enshrined in legal and administrative frameworks. Yet legislation rarely defines expertise or specifies the composition of expert panels. We seek and provide expert judgement in very primitive ways using an unrealistic mental model of scientific objectivity.

Scientists consider it to be unconscionable to manipulate data. It is unethical to weight data arbitrarily or filter them to suit personal goals. Yet we know expert judgement is prey to strong social and psychological forces that lead to inadvertent or overt weighting and filtering. And we do virtually nothing about it.

Another philosophy is that expert judgement should be treated with the same reverence as data. That is, we should use repeatable methods to acquire expert judgements. We should strive to avoid bias and error. We should test our methods and validate our expert predictions with data, and adjust both accordingly as we learn about them.

The questions then arise, how is expert status decided and validated? Who qualifies as an expert, are their judgements any good, and can we find ways to improve their reliability and accuracy?

Different disciplines have evolved different ways of dealing with uncertainty. Most remain mired in the conventional, unwarranted belief in unaided, unstructured expert judgement. Others admit uncertainty. They develop methods to provide instant personal feedback. They train, and use models to assist prediction and data to validate judgements. Meteorologists took this path. The consequence is that they make better predictions, at the same time as they lose the mantle of scientific authority.

I do not mean to imply that data-driven models give perfect forecasts. The world walked into the global financial crisis of 2007 with the help of quantitative models. Rather, when it comes to estimating narrowly defined facts, we know we can do a lot better than unaided subjective judgement. One of the strengths of meteorologists is that they know their own limits; they do not claim to make reliable predictions more than about five days in advance.

A few general tools may be useful in achieving these goals. They include structured question formats, interval judgements and Delphi group interactions (Burgman et al. 2011; McBride et al. 2012). The details of these approaches are beyond the scope of this review.

Unfortunately, there is very little relationship between an expert's status, their own or their peers' expectations of their performance on questions of fact, and how they actually perform (Burgman et al. 2011). There is no way to distinguish an accurate and well-calibrated expert from an incompetent one, other than by testing them on real predictons or judgements.

It is helpful to think of them as advocates. They spend their time trying to convince others of their position, even if they are unaware of it. It may be that they advocate a scientific position based on an accepted range of data and methodologies. They may do so on behalf of a client, such as a proponent for a particular project or decision. Advocacy is especially strident when issues are emotionally or politically charged.

Reviews of scientific authority suggest that what counts as expertise depends on context. Expert performance is likely to be affected in subtle and unpredictable ways by motivations and psychology. If experts are tested, then expertise from all domains may be considered, including what may be considered lay knowledge.

A hallmark of an assessment that attempts an honest evaluation is that it exposes experts to unfettered, critical evaluation. If we see scientists as advocates, valid questions from any source should be considered. That is, it should not only be experts (or well-informed lawyers) who can put critical questions to an expert. Anyone with a stake in the outcome should be able to question an expert's opinion. Decision-makers should avoid arbitrary, sharp, conventional delineations of expertise. Instead they should develop processes to examine knowledge claims critically (Gregory et al. 2006).

Expert judgement is about more than estimation and prediction. It fills a social role. People need to be able to share responsibility for decisions. Sometimes, they need to claim an evidentiary basis or retain a semblance of objectivity. Experts fulfill this function, and, in doing so, their status, connections, memberships, publications and so on become important.

I do not mean to trivialise this important social role. If, however, the decision we confront depends on the veracity of facts, then we need to be aware that the most revered expert available to us may be frail, error-prone and emotional, irrespective of how they appear. When facts matter, we need to employ smarter strategies to engage with experts.

ADVICE FOR DECISION-MAKERS

Policy-makers are concerned with ensemble judgements. They are obliged to consider all potential sources of uncertainty, including those not examined formally by experts. Some aspects of a decision may be concrete, but almost always other aspects will be political or intangible. Some may affect policy-makers personally. Decision-makers may be more interested in robust strategies that avoid the worst outcomes than in trying to maximise the expected benefits (Simon 1959).

Generally, forecasting is the business of making statements about events that have not yet occurred. It involves creating models of systems that include ideas about underlying processes, allowing us to anticipate changes. For example, large numbers of atmospheric scientists, physicists, glaciologists, earth scientists, oceanographers and biologists have been working for many years to forecast the climate outcomes of increasing carbon dioxide in the earth's atmosphere. Their work is the basis for global policy decisions.

This review is concerned only with the relatively simple problems of how to ask experts about well-defined, unique facts (numbers, quantities, rates, outcomes of events). The facts may exist in the present or they may be realised in the future. Even in these relatively simple circumstances, we need to combat the pervasive weaknesses and deeply buried, unacknowledged myopia exhibited by most experts (Armstrong 1980). To do so, we recommend the following.

1. Be clear about what you want from experts: judgements of simple facts, predictions of the outcomes of events, or advice on a best course of action.
2. Be clear about the domains of expertise that will help, and choose people whose skills, training or verified experience (where it exists) are squarely in those domains.
3. Choose as many experts as possible; do not be concerned about their age, number of publications, peer status, technical qualifications or apparent impartiality.
4. If the matter at hand is politically sensitive or socially or emotionally charged, ensure the experts have divergent relevant opinions or positions. Work to diversify the culture, context and perspectives of the participants. Try to include people who are less self-assured and assertive, and who integrate information from diverse sources.
5. Compose questions to avoid arbitrary linguistic misunderstandings and psychological trip wires such as framing, anchoring, availability bias and so on.
6. Use structured question formats to counter tendencies towards overconfidence, and oblige participants to consider counter-factuals and alternative theories.
7. Use structured, facilitated group interactions to counter dominance effects, anchoring and other factors that lead to group-think.
8. Provide opportunities for participants to see the opinions of other participants, only after they have made an initial private judgement. Give the group the opportunity to reconcile misunderstandings and to introduce new information. Ensure that the experts actively seek and consider evidence and arguments that disagree with their position. Then, ask for a second, private opinion from each participant.
9. Weight opinions equally, unless you have proven, documented, unambiguous measures of performance on similar questions, in which case, weight by performance history.
10. If the opinions coincide, use the average of the group. If opinions diverge, consider ways of combining or summarising their judgements that retain the breadth of opinions. In both cases, retain and consider the ranges of opinions and uncertainties. Give the experts feedback on their estimates and any weights you applied. Make the full set of information about the process and the estimates available for peer review.

Acknowledgements

I am grateful to the editor for helpful comments. This work is an excerpt from a book on expert judgement forthcoming from Cambridge University Press.

References

Armstrong, J.S., 1980. The seer-sucker theory: the value of experts in forecasting. *Technology Review* 82, June/July 1980:16–24.

Berner, E.S. & Graber, M.L., 2008. Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine* 121: S2–S23.

Burgman, M.A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. & Twardy, C., 2011. Expert status and performance. *PLoS ONE* 6, e22998. Doi: 10.1371/journal.pone.0022998

Dror, I.E., Peron, A.E., Hind, S.L. & Charlton, D., 2005. When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology* 19: 799–809.

Elstein A.S., 1995. Clinical reasoning in medicine. In Clinical Reasoning in the Health Professions, J.J.M. Higgs, ed. Butterworth-Heinemann Ltd, Oxford, England, pp. 49–59.

Engelmann, J.B., Capra, C.M., Noussair, C. & Berns, G.S., 2009. Expert financial advice neurobiologically 'offloads financial decision-making under risk. *PLoS ONE* 4(3): e4957. doi:10.1371/journal.pone.0004957

Evatts, J., Mieg, H.A. & Felt, U., 2006. Professionalization, scientific expertise, and elitism: a sociological perspective. In *The Cambridge Handbook of Expertise and Expert Performance*, K.A. Ericsson, N. Charness, P.J. Feltovitch & R. R. Hoffman, eds. Cambridge University Press, Cambridge, pp. 105–123.

Feynman, R.P., 1986. Personal observations on the reliability of the Shuttle. In Roger's Commission *Report of the Presidential Commission on the Space Shuttle* Challenger *Accident*, Appendix F. http://science.ksc.nasa.gov/shuttle/missions/51-l/docs/rogers-commission/Appendix-F.txt

French, S., 2012. Expert judgment, meta-analysis and participatory risk analysis. *Decision Analysis* 9: 119–127.

Gregory, R., Failing L., Ohlson D. & McDaniels T., 2006. Some pitfalls of an overemphasis on science in environmental risk management decisions. *Journal of Risk Research* 9: 717–735.

Gregory, R., Failing, L., Harstone, M., Long, G., McDamiels, T. & Ohlson, D., 2012. Structured decision making. Wiley-Blackwell, Chichester, pp. 26–27.

Grove, W.M. & Meehl, P.E., 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy. *Psychology, Public Policy, and Law* 2: 293–323.

Gullet, W., 2000. The precautionary principle in Australia: policy, law and potential precautionary EIAs. *Risk: Health, Safety and Environment* 11: 93–124.

Kahneman, D., 2011. *Thinking, fast and slow.* Farrar, Straus, and Giroux, New York.

Krinitzsky, E.L., 1993. Earthquake probability in engineering – Part 1: the use and misuse of expert opinion. *Engineering Geology* 33: 257–288.

Kunda, Z., 1990. The case for motivated reasoning. *Psychological Bulletin* 108: 480–498.

Lipinski, M., Froelicher, V., Atwood, E., Tseitlin, A., Franklin, B., Osterberg, L., Do, D. & Myers, J., 2002. Comparison of treadmill scores with physician estimates of diagnosis and prognosis in patients with coronary artery disease. *American Heart Journal* 143: 650–658.

McBride, M.F., Garnett, S.T., Szabo, J.K., Burbidge, A.H., Butchart, S.H.M., Christidis, L., Dutson, G., Ford, H.A., Loyn, R.H., Watson, D.M. & Burgman, M.A., 2012. Structured elicitation of expert judgments for threatened species assessment: a case study on a continental scale using email. *Methods in Ecology and Evolution* 3: 906–920.

Martin, A.D., Quinn, K.M., Ruger, T.W. & Kim, P.T., 2004. Competing approaches to predicting supreme court decision making. *Perspectives on Politics* 2: 761–767.

Meehl, P.E., 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence.* University of Minnesota Press, Minneapolis.

Meehl, P.E., 1986. Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50: 370–375.

Meyer, M.A. & Booker, J.M., 1990. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Office of Nuclear Regulatory Research, Division of Systems Research, US Nuclear Regulatory Commission, Washington, DC.

Shrader-Frechette, K., 1996. Value judgments in verifying and validating risk assessment models. In *Handbook for Environmental Risk Decision Making: Values, Perceptions and Ethics*, C.R. Cothern, ed. CRC Lewis Publishers, Boca Raton.

Simon, H., 1959. Theories for decision-making in economic and behavioral science. *American Economic Review* 49, 253–283.

Slovic, P., 1999. Trust, emotion, sex, politics, and science: surveying the risk-assessment battlefield. *Risk Analysis* 19: 689–701.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G. & Burgman, M., 2010. Reducing overconfidence in the interval judgments of experts. *Risk Analysis* 30, 512–523.

Stern, P.C. & Fineberg, H., editors. 1996. *Understanding Risk: Informing Decisions in a Democratic Society*, National Academy Press, Washington, DC.

Ulery, B.T., Hicklin, R.A., Buscaglia, J. & Roberts, M.A., 2011. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences* 108: 7733–7738.

Walton, D., 1997. *Appeal to Expert Opinion: Arguments from Authority*, Pennsylvania State University Press, Pennsylvania.

Welke, K.F., O'Brien, S.M., Eric D., Peterson, E.D., Ungerleider, R.M., Jacobs, M.L. & Jacobs, J.P., 2009. The complex relationship between pediatric cardiac surgical case volumes and mortality rates in a national clinical database. *Journal of Thoracic and Cardiovascular Surgery* 137: 1133–1140.