A simpler way to understand the results of risk assessment instruments

Eileen Munro

A simpler way to understand the results of risk assessment instruments

Introduction

An eminent surgeon at a conference on evidence-based practice confessed:

> I myself chose to be trained as a surgeon in order to avoid two things: statistics and psychology. Now I realize they are both indispensable (Gigerenzer, 20002, p.94).

While his hostility to psychology would be unusual in child protection work, his desire to avoid statistics will strike a chord with many. Once we have left school, we may have little to do with numbers and, for many, there is heart-felt relief that this is so. Unfortunately, numeracy is becoming a crucial skill. Professionals using an evidence-based approach need to be able to understand statistics and how to apply them to a particular client or family. Risk assessment instruments, which are becomingly increasingly common, pose a particular mathematical issue: if this family is assessed as at high risk of abuse, how likely is this result to be accurate?

Without specific training, people given the necessary figures about an instrument's reliability find it surprisingly hard to work out the correct answer. Research shows a high level of misunderstanding among medical practitioners, with a persistent bias towards grossly over-estimating the accuracy of results (Casscells et al, 1978; Gigerenzer, 2002). These mistakes can have devastating effects on patient care. A similar level of ignorance in child protection work could have an equally damaging impact on children and parents as these instruments are more widely used in daily practice.

Psychologists have repeatedly shown that people have a particularly poor ability to reason intuitively about probabilities. Many have taken these findings as evidence of the frailty of the human intellect and its vulnerability to cognitive illusions and biases (Tversky and Kahneman, 1974). However, research at the Max Planck Institute for Human Development in Germany has found that people's reasoning is significantly affected by the way the information is presented and that, given the right methods, their intuitive grasp of probabilities can be significantly improved (Gigerenzer and Selten, 2001). This article will show how these lessons can be applied to dealing with the key probability calculations in judging the reliability of a risk assessment instrument.

The power of a diagnostic instrument

Professionals need to know how accurate a result is because it should effect how they use it in working with a family. If they know, for instance, that most positive results are, in fact, false positives, then they should treat a specific result as, at best, a tentative hypothesis to be tested further. Equally, if a family is assessed as low risk using an instrument where a lot of the negatives will be false, then, again, professionals should stay alert to the possibility that this benign assessment may be wrong.

Practitioners face two obstacles in working out the reliability of a risk assessment instrument, or any other diagnostic procedure. First, the authors of the instrument rarely inform the reader of all the statistics needed to evaluate its accuracy. Secondly, when they do present statistics, they usually do it in a way that is hard for the average person to understand.

To evaluate accuracy, we need to know the values of three variables:

(1) The sensitivity of the instrument: how many high risk families it will identify accurately (the percentage of true positives);

(2) The specificity: how many low risk families it will identify accurately (percentage of true negatives);

(3) The base rate or prevalence of the phenomenon (in this case, abuse) being measured.

Each of these three variables play a distinctive part in working out the overall usefulness of an instrument but it is the final one – the base rate – that is most often overlooked or misunderstood. Put briefly, the rarer the phenomenon being assessed, the harder it is to develop an instrument with a clinically useful level of accuracy. Conversely, the higher the prevalence, the easier it is. Hence, researchers face a harder task trying to develop a risk assessment instrument to screen the general population, where the incidence of abuse is relatively low, than if their target population were specifically families known to child protection agencies, where the base rate will be much higher.

When the full set of relevant figures for an instrument are provided, they are usually presented in a way that is hard for anyone to follow intuitively. (This criticism applies also to my own work e.g. Munro, 2002).

This is the way the calculation is usually presented:

> The probability of a family being abusive is 20%. If a family is abusive, the probability of being identified by the instrument is 90%. If the family is not abusive, the probability that they will be correctly identified is 80%. In a case where the instrument produces a positive result, what is the probability that it is a true positive?

Extensive research among medical practitioners has consistently found a high error rate in answering this type of question. The vast majority of respondents gave a reply between 70 and 90%, which, as you will see, is a gross over-estimate of the probability (Hoffrage and Gigerenzer, 1998).

Intuitive efforts to deal with problems like this are clearly very defective. We can get a more accurate answer by using formal probability theory but the following section will probably seem even more daunting for most child protection workers. The answer can be calculated by using Bayes Theorem:

Figure 1

In formal terms, where p(a/t) means the probability of a positive result being a true positive:

Figure 2

Readers with some statistical background will be able to follow this formula but the chances are that they will be few in number. However, if we present the same information in terms of frequencies, it becomes much easier to understand and people are much more successful in working out the answer without using a formal theorem:

> Twenty out of every 100 families in this population are abusive (the base rate). Of these 20 families, 18 will get a positive result on using the instrument (the sensitivity). Of the other 80 families, some 16 will also get a positive result (the specificity). Imagine the instrument has given a positive result for a group of families. How many of these families with a positive result will actually be abusive?

The calculation can be made even clearer by using a tree diagram:

Tree diagram 1

It is easy to work out that 34 families will get a positive result, of which 18 will be true positives. Thus the probability of a positive result being a true positive =

$$\frac{18}{18 + 16} = 0.53.$$

The crux of the problem is that people forget, or do not know, the importance of the base rate in determining the level of false positives. This was shown in a famous study: the Harvard Medical School Test (Casscells et al, 1978). Staff and students at Harvard Medical School were told of a diagnostic test that had a high specificity of 95% and a superb sensitivity of 100% (no-one who had the disease would test negative). They were asked the probability of someone who tested positive actually having the disease. As the reader of this article will now know, they cannot answer this question without the additional information about the prevalence of the specific disease being tested for. However, few respondents realised this and the majority said the probability was 0.95 – the rate of true positives.

HIV tests provide a particularly vivid and tragic illustration of the consequences of this level of professional ignorance about how to interpret test results. Getting a positive HIV result is so serious that it is crucial to know how many are true and how many are false positives. Yet studies of doctors' and counsellors' knowledge reveal that they have similar difficulties to the staff and students at the Harvard Medical School and many are giving patients misleading advice (Gigerenzer, 2002, p.127).

HIV testing typically involves an initial ELISA test designed to detect antibodies against HIV in blood samples. If the result is negative then the patient is notified that he or she is HIV-negative. If it is positive, then at least one more ELISA test is done and, if this is also positive, a Western blot test is performed. If this is positive then the patient is told that he or she is HIV-positive. This very thorough sequence of tests produces high sensitivity (99.9%) and specificity rates (99.99%). Unfortunately, most doctors and counsellors involved in the testing tell patients with a positive result that the probability of a false positive is so low that it can be discounted. However, this is not necessarily true, depending on what group the patient belongs to. The base rates in the following account, taken from Gigerenzer (2002), are based on German statistics.

If the patient comes from the high-risk group of homosexual men, then the base rate of infection is 1.5%. Therefore the calculation is:

> Take a group of 10,000 homosexual men. We expect 150 to be infected with the virus, and most likely all of them will test positive (sensitivity). Of the 9,850 men who are not infected, we expect that 1 will test positive (specificity). Thus, we have 151 men who test positive, of who 150 have the virus. A patient's chances of not having the virus, given a positive result, are 1 out of 151, that is, less than 1%.

Tree diagram 2

However, if the patient is from a low risk group, with none of the known risk factors applying, then the conclusion is very different. For this group, the base rate was 0.01% - one in 10,000:

> Take a group of 10,000 men who are not in any known risk category. We would expect 1 to be infected and he will test positive with practical certainty. Of the 9,999 men who are not infected, we expect that another one will also test positive. Thus, we have 2 men with positive results, only one of whom is infected. If a patient has a positive result, his chances of not having the virus are 50/50.

Tree diagram 3

Gigerenzer (2002, p.127) reports that one of his students volunteered to check what information was given to patients before being tested for HIV. He visited twenty public health centres in Germany, presenting himself as someone who had no known risk factors for HIV, and had pre-test counselling. This is mandatory in Germany and staff are responsible for explaining the reliability of the result before the test is performed. The student specifically asked what were the chances of men in his risk group (very low risk) actually having HIV given a positive result. Ten counsellors told him it was absolutely certain that it was a true positive. Five said it was 99.9% certain. Two avoided answering his questions and three gave estimates above 90 but below 99.9%. None gave anything like the correct answer of 50%.

Implications in child protection work

Increasingly, child protection agencies are adopting risk assessment instruments (Wald and Woolverton, 1990; Lyons et al, 1996; Baird et al, 1999). They may be used at several different stages. Some are designed to screen the general population to identify potentially abusive parents so that preventive measures can be offered. Some help agency staff decide how to respond to an initial allegation of abuse, i.e. how seriously and urgently the referral is dealt with (Johnson and Clancy, 1988; Wells and Anderson, 1992; Zuravin et al, 1995). Others assess which of the families already known to the agency are high or low risk and the findings inform the allocation of resources or other aspects of the management of the case (Lyons et al, 1996).

The existing literature on risk assessment instruments does not give sufficient information to let the reader work out the accuracy of an instrument. Some merely give a single statistic on accuracy, saying, for instance, an instrument has a 65% degree of accuracy (e.g. Johnson, 1996). Others provide more detail and tell the reader the sensitivity and the specificity (Lyons, et al, 1996). However, information about the base rate is usually missing. Yet, as the earlier discussion showed, base rates are a crucial part of working out how reliable any result is. How big a difference they can make was seen in the HIV example but let me also illustrate it in child protection work. Because of the difficulties of getting accurate statistics on the incidence of abuse, I have made an educated estimate of base rates in the following examples.

Zuravin et al (1995) report an instrument for screening initial referrals – deciding whether or not they warrant further investigation - with a sensitivity of 69% and a specificity of 74%. Besharov (1990) estimates that the number of referrals that are substantiated as cases of abuse is between 35 and 45%. Taking the average, let us assume a base rate of 40% and calculate the probability that a positive result on this instrument will be a true positive.

In a group of 1,000 families, we would expect that 400 would be abusive. Of these 400 families, 276 will get a positive result on using the instrument (sensitivity 69%). Of the other 600 families, 156 will also get a positive result (specificity 74%). Therefore, there will be a total of 432 positive results, of which 276 will be true positives.

Tree diagram 4

The probability of a positive result being a true positive =

$$\frac{276}{276+156} = 0.64$$

Let us now consider how an instrument with the same sensitivity and specificity would perform if used on a sample with a much lower base rate – say, as a screening of the general population. Corby (1993, p.53), using the numbers of children on child protection registers in England, estimated a prevalence of 4 per thousand, 0.4%.

If we take a group of 1,000 families, we expect that 4 will be abusive. Of these 4 families, 3 approximately will get a positive result on using the instrument (sensitivity 69%). Of the other 996 families, 259 will also get a positive result (specificity 74%). Therefore, there will be a total of 262 positive results, of which 3 will be true positives.

Tree diagram 5

The probability of a positive result being a true positive =

$$\frac{3}{3+259} = 0.12$$

In both cases, the accuracy is far lower than most people expect intuitively from being told just the sensitivity and specificity and it has serious practice implications. If we are screening the general population, an instrument with this degree of reliability would identify a very large group of families as potentially abusive, but only a tiny percentage would be true positives. Therefore, if an agency used such an instrument as the basis for providing preventive services, it would need generous resources to meet the identified need. At the same time, it would have to recognise that about a quarter of dangerous families were being overlooked.

In screening initial referrals, the probability of a positive result being accurate is considerably higher (0.64) but this is still far below certainty. This has implications for how professionals should use it. If it is mistakenly taken as highly reliable, then it will affect how professionals work with a family, what information they seek, and how they interpret it. There is overwhelming evidence from psychology that, once people have formed a judgement about a person or family, they are slow to change their minds. They tend to be selective in what they notice and in how they interpret new information to fit with their existing judgement (Kahneman et al, 1990). Therefore, professionals who put too much trust in the result of an instrument will be biased towards confirming it, paying attention to the evidence, for example, that supports the claim that the family is high risk and being slow to see the significance of counter evidence. Conversely, if they conclude firmly that a family is low risk, they will be biased against new evidence that should cause concern. Many of the deaths of children known to child protection agencies have arisen in circumstances where those in close contact with the parents had decided that all was going well and were blind to evidence that, with hindsight, looks overwhelming (Munro, 1999).

A critical mindset is needed in using the results of risk assessment instruments. Whatever judgement is made about the level of risk to a child, professionals need to remember this is a hypothesis, not a certain truth, and to keep it under constant, critical review.   The individual professional, however, needs to work in an agency culture that allows and encourages this approach. Supervisors play a key role in helping front line workers stand back to reflect on and critique their reasoning. Unfortunately, access to good casework supervision is becoming scarce (Rushton and Nathan, 1996). Managers' priority can be to supervise the administrative process, ensuring that all legal and procedural rules have been followed.

This is, in part, due to the rise of defensive practice. In both USA and the UK, the public and media have played a large part in shaping child protection priorities. Outrage at children's deaths at the hands of their carers has been directed at the professionals, especially the social workers, who have the duty of protecting children (Waldfogel, 1998; Parton, Thorpe & Wattam, 1997). This pressure has led agencies to increase the level of investigations to minimise the chances that a child will mistakenly be left in danger. However, in an area with such imperfect knowledge, our ability to predict which parents will seriously harm their children is extremely limited. The public desire for safety cannot be met.

Agencies, realising that the risk of child deaths cannot be wholly removed, may take steps to protect themselves. They may try to transfer or dissipate blame by engaging in ‘blame prevention re-engineering’ (Hood, Rothstein and Baldwin, 2000). Hood et al’s research found that one strategy for doing this is ‘protocolozation’: the organisation introduces more and more formal procedures to guide practice so that they create a ‘correct’ way to deal with a case. Then, if a tragedy occurs, they can claim the defence of ‘due diligence’ and show that their employees followed these correct procedures in working on the case. A child may have died but the agency staff can show a clear audit trail of what they did and cannot be faulted for the tragic outcome. This defence certainly seems to be operating in the UK where recent inquiries into the deaths of children known to child protection agencies appear to focus more on whether procedures were followed than whether competent professional judgements and decisions were made (Cambridgeshire County Council, 1997; Norfolk Health Authority, 2002).

Unfortunately, if such a defensive culture exists, it is in the agency’s interests to act as if the results of a formal assessment instrument are certain. They provide a clear basis for management of the case and removes problems associated with individual professional competence. There is a ‘right’ way to respond which can be proven to anyone investigating the agency. In a defensive culture, the protection of the agency can start to dominate over the protection of children.

Conclusion

In the past, professionals generally relied upon their own expert judgement in judging the level of risk but child protection agencies are increasingly introducing risk assessment instruments to inform and support decisions about service response. Some way of categorising cases into high risk and low risk groups is necessary.   Intruding into the privacy of the family can only be justified in a liberal society by serious concern for the rights and welfare of the child. Also, agency resources are limited and all families cannot be offered a service even if they wanted to receive help.

Introducing formal instruments as an aid has many advantages. There is evidence that it leads to a more consistent and, so, more equitable response (Baird et al, 1999). It is also likely that actuarial instruments that use formal statistics to compute the answer are more accurate than clinical judgement. There is little evidence on this specifically in relation to child protection but all the evidence from other fields where comparison has been made between actuarial and clinical judgement is strongly weighted in favour of actuarial (Grove and Meehl, 1996). This should not surprise us since people generally have difficulty in working out probability calculations and these instruments produce a large set of data to weigh. Paul Meehl, the psychologist, gives the analogy:

> Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up (Meehl, 1986, p.372).

Just as the clerk does not hesitate to use an adding machine, the child protection worker should not be reluctant to use an actuarial method to calculate the overall weight of the various risk factors.

However, the number obtained at the end of the calculation needs to be viewed with realistic caution. Numbers have an air of authority and objectivity that can mislead people into crediting them with more accuracy than they deserve. The aim of this article has not been to criticise the use of instruments but to caution against over-confidence in their results. If the rate of true and false positives is understood, then clearly, the conclusion should be treated tentatively and practitioners should continue to keep an open mind about its accuracy. The examples used in this article show how even an instrument with reasonably good sensitivity and specificity can produce large numbers of false positives and false negatives, depending on the base rate in the population in question. An additional cautionary note is that the relevant statistics are not known for many instruments or, if available, are to some degree estimates rather than confident assertions. The formal probability calculation can be impressive and inspire confidence but the reader should remember that it is being used on imperfect statistics.

Human beings seem to have a persistent yearning for certainty (Gigernenzer, 2002). Risk assessment instruments that make mathematical calculations based on the best empirical evidence may well be the best way of assessing the level of danger to a child but their conclusions carry no magical guarantee of truth. Professionals need to know how to ask questions about the level of accuracy and to understand the implications the answers have in terms of ratios of true and false positives and true and false negatives. Agencies and the public need to accept that professional judgements and decisions, even when based on the best evidence and the best way of computing the evidence, are fallible. On-going work with a family needs to be undertaken knowing that the current risk assessment is only a best estimate and may need to be revised in the light of new information.

The shift to evidence-based practice represents a move from an appeal to authority or tradition to an appeal to evidence gained from empirical research (Gambrill, 2001). Unless practitioners have sufficient numeracy skills to assess the evidence themselves, there is a danger of replacing the authority of the clinical expert with a misguidedly exaggerated deference to the scientific expert.

Figures for ‘A simpler way to understand the results of risk assessment instruments’

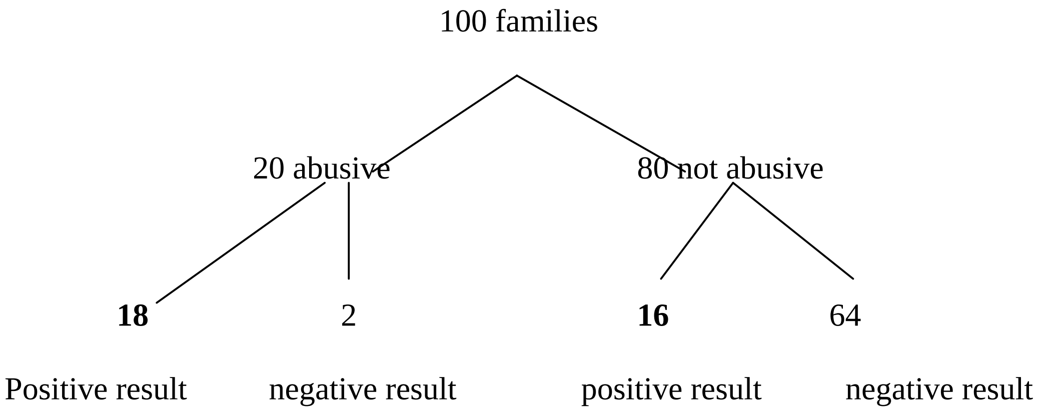
Figure 1

$$\text{Probability of a True positive} = \frac{\text{sensitivity x base rate}}{\text{probability of a positive result}}$$

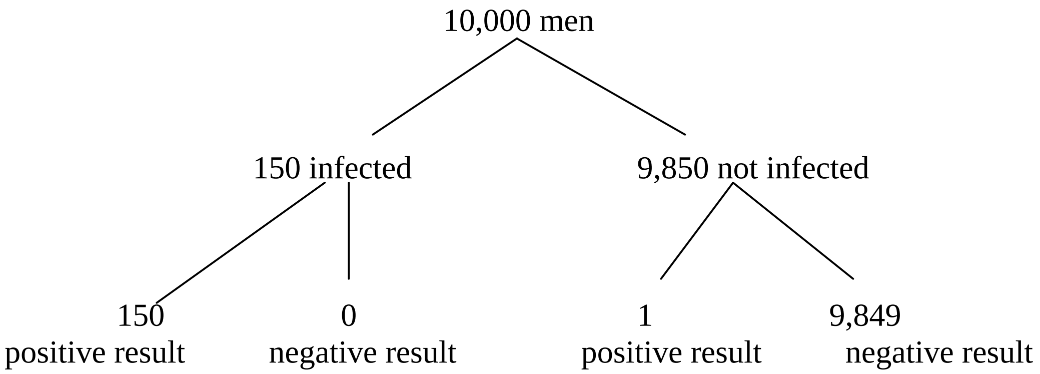
Figure 2

$$P(a/t) = \frac{p(t/a)\ p(a)}{P(t)}$$

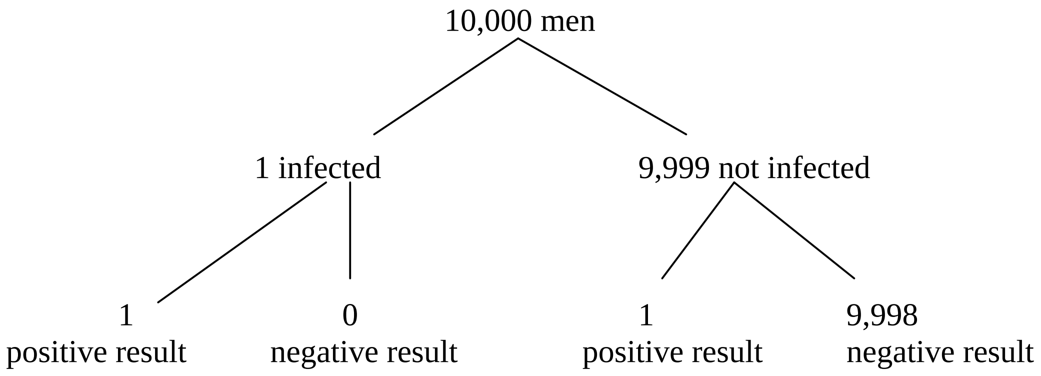
Tree diagram 1



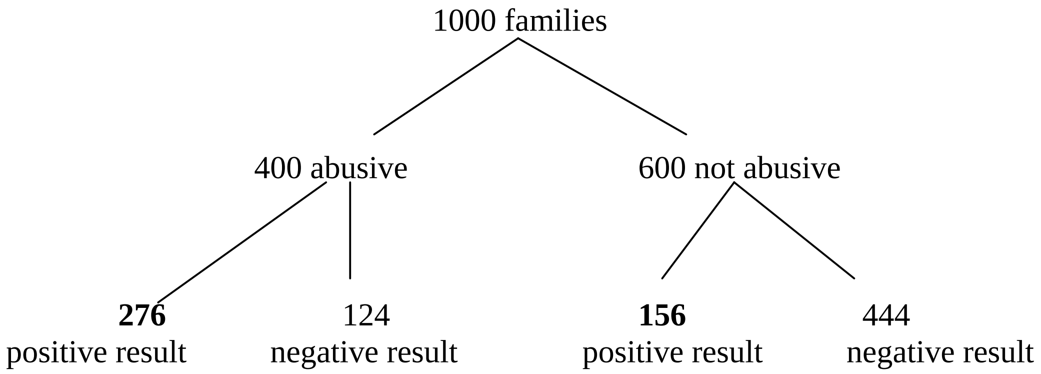
Tree diagram 2



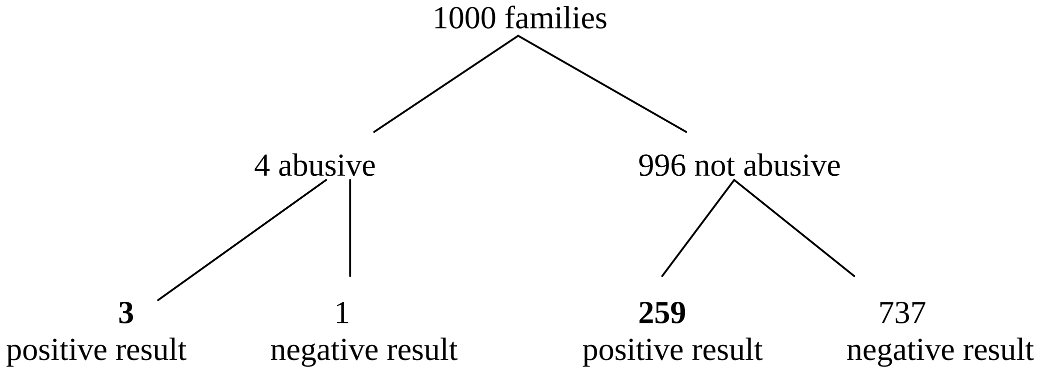
Tree diagram 3



Tree diagram 4



Tree diagram 5



References

Baird C., Wagner D., Healy T. & Johnson K. (1999). Risk Assessment in Child Protective Services: Consensus and Actuarial Model Reliability. *Child Welfare,* LXXV111, 723-748.

Besharov D. (1990). *Recognisiing Child Abuse: A Guide for the Concerned.* New York: The Free Press.

Cambridgeshire County Council (1997). *Bridge Report and Action Plan.* Cambridge: Cambridgeshire Country Council.

Casscells W., Schoenberger A. & Grayboys T. (1978). Interpretation by Physicians of Clinical Laboratory Results. *New England Journal of Medicine*, 299, 999-1000.

Corby B. (1993). *Child Abuse: Towards a Knowledge Base*. Buckingham, England: Open University Press.

Gambrill E. (2001). Social Work: An Authority-Based Profession. *Research in Social Work Practice*, 11, 166-175.

Gigerenzer G. & Selten R. (eds.) (2001). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, Mass: The MIT Press.

Gigerenzer G. (2002). *Reckoning with Risk*. London: Allen Lane, The Penguin Press.

Grove W. & Meehl P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures. *Psychology, Public Policy and Law*, 2, 293-323.

Guyatt G. & Rennie D (eds.). (2002). *Users' Guides to the Medical Literature*. Chicago: AMA Press.

(Hood C., Rothstein H. & Baldwin R. (2000). *The Government of Risk: Understanding Risk Regulation Regimes*. Oxford: Oxford University Press.

Hoffrage U. & Gigerenzer G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, 538-540.

Johnson W. & Clancy T. (1988). *A study to find improved methods of screening and disposing of reports of child maltreatment in the emergency program in Alameda County, California.* Oakland, CA: Alameda County Social Services.

Kahneman D., Slovic P. & Tversky A. (1990). *Judgements under uncertainty: Heuristics and Biases*. Cambridge, England: Cambridge University Press.

Keller L. & Ho J. (1988). Decision problem structuring: generating options. *Systems, Man and Cybernetics*, 18, 715-728.

Lyons P., Doueck H. & Wodarski J. (1996). Risk assessment for child protective servies: A review of the empirical literature on instrument performance. *Social Work Research*, 20, 43-155.

Meehl P. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50, 370-375.

Munro E. (1999). Common errors of reasoning in child protection work. *Child Abuse and Neglect*, 23, 745- 758.

Munro E. (2002). *Effective Child Protection.* New York: Sage Publications.

Norfolk Health Authority (2002). *Summary Report of the Independent Health Review*. Norfolk, England: Norfolk Health Authority.

Parton N., Thorpe D. & Wattam C. (1997). *Child Protection: Risk and the Moral Order*. London: Macmillan.

Rushton A. & Nathan J. (1996). The Supervision of Child Protection Work. *British Journal of Social Work,* 26, 357-374.

Sackett D., Straus S., Richardson W., Rosenberg W. & Haynes R. (2000). *Evidence-Based Medicine; How to Practice and Teach EBM.* New York: Churchill Livingstone.

Trocme N., McPhee D., Tam K. & Hay T. (1994). *Ontario Incidence Study of Reported Child Abuse and Neglect: Final Report.* Toronto: Institute for the Prevention of Child Abuse.

Tversky A. & Kahneman D. (1974). Judgement under uncertainty: heuristics and biases. *Science*, 185, 1124-1131.

Wald M. & Woolverton M. (1990). Risk assessment: The emperor's new clothes? *Child Welfare,* 69(6), 483-511.

Waldfogel J. (1998) *The Future of Child Protection* Cambridge, Mass: Harvard University Press.

Wells S. & Anderson T. (1992). *Model building in child protection services intake and investigation. Final report to the National Center on Child Abuse and Neglect for Grant #90-CA-1407.* Washington, DC: American Bar Association Centre on Children and Law.

Zuravin S., Orme J. & Hegar R. (1995). Disposition of Child Physical Abuse Reports: Review of the Literature and Test of a Predictive Model. *Children and Youth Services Review*¸17, 547-566.