

Correspondence

Contents

- Shall evidence-based risk assessment be abandoned?

Shall evidence-based risk assessment be abandoned? Comments on "Precision of actuarial risk assessment instruments" by Hart et al.

Hart, Michie, and Cooke³ reminded readers of a BJP journal supplement that, "Predicting the future is very difficult" (p. s63). All physicians are acutely aware of the difficulty of prognosis. But does this mean it should not be attempted? Competent practice, especially for serious conditions and therapies carrying risks, is impossible without some evaluation, one patient at a time, as to the likelihood of various outcomes as a function of various contemplated interventions (including no intervention), or as a function of various diagnostic tests. The advice from Hart and colleagues seems to call for clinicians to eschew empirical data about outcomes among groups of similar patients, but they failed to advise readers about what to do instead.

Statistical and Technical Matters

Hart and colleagues made a statistical argument that the results of widely replicated actuarial systems for forensic risk assessment (the *Violence Risk Appraisal Guide*, VRAG, and the Static-99) must be "virtually meaningless" (p. s60). Unfortunately, they were led into statistical error by conflating test reliability and validity -- precision of measurement must be treated separately from a test's association with an outcome. The first error resulting from this conflation was using confidence intervals to assess the "precision" or "margin of error" for an individual test result; in fact, confidence intervals were not designed for this purpose. The appropriate statistic is the standard error of measurement -- the margin of error associated with a single person's true score (an aspect of reliability). The VRAG's standard error of measurement has been reported both for the development sample and independent replications⁵ consistently indicating that any single score has at most a .05 probability of yielding misclassification by more than one VRAG category. The analysis by Hart and colleagues was correct in one sense -- the amount of misclassification to be expected does vary as a function of the score -- those at the extremes exhibit greater risk of misclassification. Again, however, confidence intervals are not the way to compute this error; conditional standard error of measurement is the statistic for this purpose.

Hart and colleagues' analysis of confidence intervals did legitimately "prove" statistically that one usually cannot learn much from a single case -- one observation usually conveys only a little scientific information. But is it true, as they imply,

Grant T. Harris, Director of Research,
Mental Health Centre Penetanguishene, Ontario, Canada
Assoc Professor of Psychology, Queen's University at Kingston
Assoc. Professor of Psychiatry, University of Toronto

Marnie E. Rice, Director of Research Emerita,
Mental Health Centre Penetanguishene
Professor of Psychiatry and Behavioural Neuroscience, McMaster University
Professor of Psychiatry, University of Toronto
Assoc. Professor of Psychology, Queen's University at Kingston

Vernon L. Quinsey, Professor of Psychology, Biology, and Psychiatry,
Queen's University at Kingston, Ontario, Canada

Published online, *British Journal of Psychiatry*, January, 2008

that a single observation conveys absolutely no information? Readers will recognize that most research findings are simply the aggregation of many single observations. The fact that some research findings yield consistent replication inevitably means that single observations do convey valid scientific information. It's just that we must often aggregate the single observations in order to evaluate and learn from them.

Hart and colleagues' second mistake related to aggregated findings about the accuracy of actuarial tools. They slipped from "precision" to "accuracy" as though these are formally synonymous. They are not. As most medical professionals know, test accuracy (i.e., validity) is assessed in terms of sensitivity, specificity, and the tradeoff between these two. We are aware of more than 45 independent tests of the accuracy of the VRAG (and its allied tool the *Sex Offender Risk Appraisal Guide*, SORAG) in predicting violent recidivism in a total of approximately eight thousand released correctional inmates, sex offenders, forensic patients, civil psychiatric patients, and other clinical samples (<http://www.mhcr-research.com/ragrepr.htm>). These tests have been conducted in at least seven countries and have employed mean follow-up periods ranging from a few months to ten years. By conventional standards, average predictive effects (in terms of the sensitivity-specificity tradeoff) are large and are distributed as expected by psychometric principles and the laws of probability. Contrary to the assertions of Hart and colleagues, VRAG/SORAG scores have been shown to predict the speed and severity of violent recidivism. If recalculated using all available cases, confidence intervals for category outcomes would be considerably smaller than those calculated for the development sample alone. Similarly, we are aware of approximately 40 replications, involving more than 13,000 cases, of the Static-99. The statistical argument by Hart and colleagues does not and cannot refute these empirical results supporting the accuracy of actuarial risk assessments.

It is instructive to consider the argument by Hart and colleagues in a broader medical context. Predicting violent recidivism with actuarial instruments is, in principle, no different than using diagnostic tests to predict development or outcome for such disorders as cancer. The accepted measure of predictive and diagnostic accuracy is the area under the

Relative Operating Characteristic (ROC)⁶ which indexes the tradeoff between sensitivity and specificity as a function of test score. Under conditions of good measurement reliability, equal follow-up duration, and few missing items, the VRAG produces ROC values that compare favorably with widely used diagnostic tests⁵. This is true even though the accuracy of actuarial instruments is artifactually lowered by error in measuring the outcome (violent reoffending recorded in official records) whereas the accuracy of diagnostic tests for cancer prediction is generally less affected by such measurement error (for example, using death or autopsy results as the predicted outcome). Because ROC analyses are the standard for accuracy, the advice of Hart and colleagues would seem to require that many diagnostic tools also be abandoned.

Finally, classification accuracy is the standard in assessing the kind of “precision” attempted by Hart and colleagues. In most tests of the VRAG, there have been no statistically significant differences between the observed rates and those expected on the basis of the proportions provided as norms², especially given known variation predicted by Bayes’ Rule. Thus, classification accuracy has also been successfully replicated. In essence, Hart et al. have attempted, but failed, to gainsay an empirical result with a statistical argument.

The notion that it is somehow wrong to base individual decisions on “group data” has been thoroughly refuted^{1,5}. Consider the example offered by Hart and colleagues themselves – betting on whether a card other than a diamond (probability = .75, 3 to 1 odds) will be drawn from an ordinary deck of playing cards. Hart and colleagues assert that one can have little confidence in winning in a single trial. What do they then advise -- bet on a diamond?! A careful reading of their paper yields only one piece of advice – refuse to bet. Yet consistently betting against a diamond is the winning strategy and all rational gamblers would make that bet. In the context of violence risk assessment over long durations, offenders in the highest two VRAG categories have generally exhibited probabilities of officially detected violent recidivism greater than 75 percent. And the lowest four categories have consistently exhibited rates below 25 percent. Surely forensic clinicians should not refuse to provide this information to those making decisions about violent offenders.

Clinical Decisions about One Case

What should a forensic clinician do when deciding to release or detain one previously violent forensic patient? Hart et al. imply that the clinician should make no release decision, presumably leaving it up to the unaided judgment of others.

We disagree. An actuarial tool (such as the VRAG or Static-99) is simply an efficient, available distillation of relevant empirical evidence. An actuarial tool does not afford certainty, of course, but, as Hart and colleagues fully acknowledged, it affords more accuracy than any other known method for making such decisions.

In conclusion, the undeniable superiority in accuracy of actuarial systems over all known alternatives means they must be used where available. Except for refusing to make risk-related decisions, Hart and colleagues offer no alternative for actual forensic practice. Taken seriously, their advice is likely to worsen the practice of clinicians who must make decisions about the risk of violent recidivism. Reluctance to make risk-related decisions based on actuarial methods may well have a motivation in addition to (misguided) concerns about accuracy, however. These concerns relate to clinical and philosophical objections to civil commitment for sex offenders in the U.S. and the dangerous severe personality disorders legislation in the UK⁴. Although we have some sympathy here, it is important to understand that the reliability and validity of actuarial instruments are independent of their use in particular schemes for sentencing and managing offenders. Further, if forensic clinicians refuse to make risk-related decisions, decisions will be made by others using less accurate means: less accurate decisions inexorably accumulate in more avoidable harm to victims, more unnecessary restriction of offenders, or both.

- 1 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.
- 2 Harris, G.T. & Rice, M.E. (2007) Characterizing the value of actuarial violence risk assessment. *Criminal Justice and Behavior*, 34, 1638-1656.
- 3 Hart, S. D., Michie, C., & Cooke, D.J. (2007) Precision of actuarial risk assessment instruments. *British Journal of Psychiatry*, 190 (suppl. 49), s60-s65.
- 4 Monahan, J. (2006) A jurisprudence of risk assessment: Forecasting harm among prisoners, predators, and patients. *Virginia Law Review*, 92, 391-435.
- 5 Quinsey, V.L., Harris, G.T., Rice, M.E., & Cormier, C.A. (2006) *Violent offenders: Appraising and managing risk* (Second Edition). Washington, DC: American Psychological Association.
- 6 Swets, J., Dawes, R., & Monahan, J. (2000) Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1-26.