The Unreliability of Clinical and Actuarial Predictions of Dangerous Behavior

I would not say that the future is necessarily less predictable than the past. I think the past was not predictable when it started – Donald Rumsfeld

September 23, 2016

Topic Areas

1) The clinical (i.e., expert) or actuarial (i.e., statistical) prediction of dangerous behavior

2) The 2×2 contingency table specification of the prediction problem

3) An example of clinical prediction – the Kozol et al. (1972) study

4) The Meehl and Rosen notion of "clinical efficiency"

5) The MacArthur Study of Mental Disorder and Violence

6) A validation study for the Classification of Violence Risk (COVR) instrument

7) Barefoot v. Estelle (1983) – prediction by labeling someone a "sociopath"

8) The Goldwater Rule

Introduction to the Problem

An ability to predict and treat dangerous or violent behavior in criminal offenders is important to the administration of the criminal justice system in the United States.

This prediction might be in the context of preventive detentions, parole decisions, forensic commitments, or other legal forms of restriction on personal liberty.

Behavioral prediction might rely on clinical judgement (usually through trained psychologists or other medically versed individuals) or by actuarial (statistical) assessments.

In any case, concern should be on the reliability of such predictions, and more pointedly, on the state of clinical and actuarial prediction.

The General Conclusion

So, the question: are we at such a level of predictive accuracy that as a society we can justify the necessary false positives that would inappropriately restrict the personal liberty of those who would prove to be neither dangerous or violent.

Unfortunately, the conclusion reached in this module is that for both clinical or actuarial prediction of dangerous behavior, we are quite far from a level that could sanction routine use.

The Familiar 2×2 Contingency Table and Event Specification

Evidence on prediction accuracy can typically be presented in the form of a 2×2 contingency table defined by a cross-classification of individuals according to the events A and $\bar{A}$ (whether the person proved dangerous (A) or not ($\bar{A}$)); and B and $\bar{B}$ (whether the person was predicted to be dangerous (B) or not ($\bar{B}$)):

Prediction:

B (dangerous)
$\bar{B}$ (not dangerous)

Outcome (Post-Prediction):

A (dangerous)
$\bar{A}$ (not dangerous)

The Generic 2×2 Table

A generic 2×2 table presenting the available evidence on prediction accuracy might then be given in the following form (arbitrary cell frequencies are indicated using the appropriate subscript combinations of A and $\bar{A}$ and B and $\bar{B}$):

		Outcome		
		A (D)	$\bar{A}$ (ND)	row sums
Prediction	B (D)	n_{BA}	$n_{B\bar{A}}$	n_B
	$\bar{B}$ (ND)	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
	column sums	n_A	$n_{\bar{A}}$	n

D: Dangerous

ND: Not Dangerous

Clinical Prediction

The 2×2 contingency table given immediately below illustrates the poor prediction of dangerous behavior when based on clinical assessment.

These data are from Kozol, Boucher, and Garofalo (1972), "The Diagnosis and Treatment of Dangerousness":

		Outcome		
		A (D)	$\bar{A}$ (ND)	row sums
Prediction	B (D)	17	32	49
	$\bar{B}$ (ND)	31	355	386
	column sums	48	387	435

For these data, 2 out of 3 predictions of "dangerous" are wrong (.65 = 32/49 to be precise).

Also, 1 out of 12 predictions of "not dangerous" are wrong (.08 = 31/386).

Barefoot v. Estelle (1983)

In his dissent opinion in the Barefoot v. Estelle case, Justice Blackmun quotes the American Psychiatric Association *amicus curiae* brief as follows:

" [the] most that can be said about any individual is that a history of past violence increases the probability that future violence will occur."

In other words, the best we can say is that "past violence" (B) is facilitative of "future violence" (A) but the error in that prediction can be very large as it is here for the Kozol et al. data: $P(A|B) = \frac{17}{49} = .35$ is greater than $P(A) = \frac{48}{435} = .11$.

But this implies that 2 out of 3 such predictions of "dangerous" are wrong (or, 1 out of 3 are correct).

To us, the accuracy of these behavioral predictions is insufficient to justify any incarceration policy based on them –

Clinical Efficiency

In Module 4 on diagnostic testing, the Meehl and Rosen (1955) notion of "clinical efficiency" is formally discussed, or when a diagnostic test is more accurate than just predicting using base rates.

For these data, prediction by base rates would be to say everyone will be "not dangerous" because the number of people who are "not dangerous" (387) is larger than the number of people who are "dangerous" (48).

Here, we would be correct in our predictions 89% of the time ($.89 = 387/435$).

Based on clinical prediction, we would be correct a *smaller* 86% percentage of the time ($.86 = (17 + 355)/435$).

So, according to the Meehl and Rosen characterization, clinical prediction is *not* "clinically efficient" because one can do better by just predicting according to base rates.

Monahan (1973) Comments on Kozol, et al.

In commenting on the Kozol, et al. study, Monahan (1973) takes issue with the article's principal conclusion that "dangerousness can be reliably diagnosed and effectively treated" and notes that it "is, at best, misleading and is largely refuted by their own data."

Mohahan concludes his critique with the following quotation from Wenk, Robison, and Smith (1972):

Confidence in the ability to predict violence serves to legitimate intrusive types of social control. Our demonstration of the *futility* of such prediction should have consequences as great for the protection of individual liberty as a demonstration of the utility of violence prediction would have for the protection of society. (p. 402)

Actuarial Prediction

Paul Meehl in his iconic 1954 monograph, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of data combination, such as multiple regression, outperform (expert) clinical prediction.

The enormous amount of literature produced since the appearance of this seminal contribution has uniformly supported this general observation;

It appears that individuals who are conversant in a field are better at selecting and coding information than they are at actually integrating it.

Combining such selected information in a more mechanical manner will generally do better than the person choosing such information in the first place.

A Robyn Dawes Observation on Actuarial Prediction

A 2005 article by Robyn Dawes in the *Journal of Clinical Psychology* (*61*, 1245–1255) has the intriguing title "The Ethical Implications of Paul Meehl's Work on Comparing Clinical Versus Actuarial Prediction Methods."

Dawes' main point is that given the overwhelming evidence we now have, it is unethical to use clinical judgment in preference to the use of statistical prediction rules. We quote from the abstract:

Whenever statistical prediction rules . . . are available for making a relevant prediction, they should be used in preference to intuition. . . . Providing service that assumes that clinicians "can do better" simply based on self-confidence or plausibility in the absence of evidence that they can actually do so is simply unethical. (p. 1245)

The MacArthur Study of Mental Disorder and Violence

The MacArthur Research Network on Mental Health and the Law was created in 1988 by a major grant to the University of Virginia from the John D. and Catherine T. MacArthur Foundation.

The avowed aim of the Network was to construct an empirical foundation for the next generation of mental health laws, assuring the rights and safety of individuals and society.

New knowledge was to be developed about the relation between the law and mental health; new assessment tools were to be developed along with criteria for evaluating individuals and making decisions affecting their lives.

The major product of the Network was the MacArthur Violence Risk Assessment Study; its principal findings were published in the very well-received 2001 book, *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence* (John Monahan, et al.).

COVR: Classification of Violence Risk

The major analyses reported in *Rethinking Risk Assessment* are based on constructed classification trees — these are branching decision maps for using risk factors to assess the likelihood that a particular person will commit violence in the future.

All analyses were carried out with an SPSS classification-tree program, called CHAID, now a rather antiquated algorithm

Moreover, these same classification tree analyses have been incorporated into a proprietary software product called the Classification of Violence Risk (COVR) that is available from the Florida-based company PAR (Psychological Assessment Resources).

The program is to be used in law enforcement/mental health contexts to assess "dangerousness to others," a principal standard for inpatient or outpatient commitment or commitment to a forensic hospital.

Cross-validation of the COVR

There is one small cross-validation study done to justify this actuarial software COVR: "An Actuarial Model of Violence Risk Assessment for Persons with Mental Disorders" (John Monahan, et al., *Psychiatric Services*, 2005, 56, 810–815).

The complete 2×2 table from the COVR validation study follows:

		Outcome		
		A (D)	$\bar{A}$ (ND)	row sums
Prediction	B (D)	19	36	55
	$\bar{B}$ (ND)	9	93	102
	column sums	28	129	157

Some Statistics From the Cross-validation Study

As seen in the table, a high prediction of "dangerous" is wrong 65% ($= 36/55$) of the time.

A prediction of "not dangerous" is incorrect 9% ($= 9/102$) of the time (again, this is close to the 1 out of 12 incorrect predictions of "not dangerous" typically seen for purely clinical predictions).

The accuracy or "hit-rate" is $(10 + 93)/157 = .71$.

If everyone were predicted to be nondangerous, we would be correct 129 out of 157 times, the base rate for $\bar{A}$:
$P(\bar{A}) = 129/157 = .82$.

Obviously, the accuracy of prediction using base rates (82%) is better than for the COVR (71%), making the COVR not "clinically efficient" according to the Meehl and Rosen terminology.

Barefoot v. Estelle (1983) Documents

The Module Two discussion on probabilistic reasoning concerns the unreliability of clinical and actuarial behavioral prediction, particularly for violence

Module Two readings include two extensive redactions in appendices: one is the majority opinion in the Supreme Court case of *Barefoot v. Estelle* (1983) and an eloquent Justice Blackmun dissent; the second is an *amicus curiae* brief in this same case from the American Psychiatric Association on the accuracy of clinical prediction of future violence.

Both of these documents are detailed, self-explanatory, and highly informative about our current lack of ability to make clinical assessments that lead to accurate and reliable predictions of future behavior.

Dr. Death: James Grigson

The psychiatrist featured so prominently in the opinions for *Barefoot v. Estelle* and the corresponding American Psychiatric Association *amicus* brief, James Grigson, played the same role repeatedly in the Texas legal system.

For over three decades before his retirement in 2003, he testified when requested at death sentence hearings to a high certainty as to "whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society."

An affirmative answer by the sentencing jury imposed the death penalty automatically, as it was on Thomas Barefoot;

he was executed on October 30, 1984.

Prediction by Labeling

The two psychiatrists mentioned in *Barefoot v. Estelle*, James Grigson and John Holbrook, appeared together repeatedly in various capital sentencing hearings in Texas during the later part of the 20th century.

Although Grigson was generally the more outrageous of the two with predictions of absolute certitude based on a sociopath diagnosis, Holbrook was similarly at fault ethically.

This pair of psychiatrists of Texas death penalty fame might well be nicknamed "Dr. Death" and "Dr. Doom."

They were both culpable in the famous exoneration documented in the award winning film by Errol Morris, *The Thin Blue Line*.

Federal Rules of Evidence: Rule 702 on the Admissibility of Expert Witnesses

Rule 702, Testimony by Experts, states:

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

The Goldwater Rule

The offering of a professional psychiatric opinion about an individual without direct examination is an ethical violation of the Goldwater Rule, named for the Arizona Senator who ran for President in 1964 as a Republican.

Promulgated by the American Psychiatric Association in 1971, it delineated a set of requirements for communication with the media about the state of mind of individuals.

The Goldwater Rule was the result of a special September/October 1964 issue of *Fact:* magazine, published by the highly provocative Ralph Ginzburg.

The issue title was "The Unconscious of a Conservative: Special Issue on the Mind of Barry Goldwater," and reported on a mail survey of 12,356 psychiatrists, of whom 2,417 responded: 24% said they did not know enough about Goldwater to answer the question; 27% said he was mentally fit; 49% said he was not.

Goldwater's Supreme Court Case

Goldwater brought a $2 million libel suit against *Fact:* and its publisher, Ginzburg.

In 1970 the United States Supreme Court decided in Goldwater's favor giving him $1 in compensatory damages and $75,000 in punitive damages.

More importantly, it set a legal precedent that changed medical ethics forever.

For an updated discussion of the Goldwater Rule, this time because of the many psychiatrists commenting on the psychological makeup of the former chief of the International Monetary Fund, Dominique Strauss-Kahn, after his arrest on sexual assault charges in New York, see Richard A. Friedman's article, "How a Telescopic Lens Muddles Psychiatric Insights" (*New York Times*, May 23, 2011).

Goldwater Incidentals

His most famous quote:

I would remind you that extremism in the defense of liberty is no vice! And let me remind you also that moderation in the pursuit of justice is no virtue!

One of most famous political attack ads of all time was run against Barry Goldwater by Lyndon Johnson – google: daisy lyndon johnson

Texas Defender Service Resources

A good resource generally for material on the prediction of dangerous behavior and related forensic matters is the Texas Defender Service (`www.texasdefender.org`), and the publications it has freely available at its web site:

A State of Denial: Texas Justice and the Death Penalty (2000)

Deadly Speculation: Misleading Texas Capital Juries with False Predictions of Future Dangerousness (2004)

Minimizing Risk: A Blueprint for Death Penalty Reform in Texas (2005)