From Group Data to Useful Probabilities: The Relevance of Actuarial Risk Assessment in Individual Instances

Douglas Mossman, MD

Probability plays a ubiquitous role in decision-making through a process in which we use data from groups of past outcomes to make inferences about new situations. Yet in recent years, many forensic mental health professionals have become persuaded that overly wide confidence intervals render actuarial risk assessment instruments virtually useless in individual assessments. If this were true, the mathematical properties of probabilistic judgments would preclude forensic clinicians from applying group-based findings about risk to individuals. As a consequence, actuarially based risk estimates might be barred from use in legal proceedings. Using a fictional scenario, I seek to show how group data have an obvious application to individual decisions. I also explain how misunderstanding the aims of risk assessment has led to mistakes about how, when, and why group data apply to individual instances. Although actuarially based statements about individuals' risk have many pitfalls, confidence intervals pose no barrier to using actuarial tools derived from group data to improve decision-making about individual instances.

J Am Acad Psychiatry Law 43:93–102, 2015

Over the past two decades, forensic mental health professionals have developed several actuarial tools for assessing the risk that an individual will engage in future criminal or aggressive behavior.[1–4] An actuarial risk assessment instrument (ARAI[5]) implements a procedure for obtaining, weighting, and combining a relatively small number of prespecified items to yield a numerical judgment concerning the probability of future violence. The empirical underpinnings of these algorithms and probability judgments come from studies of reference groups in which the same data items and outcomes were gathered and evaluated.

ARAIs have received much criticism. By their very design, they depend on relationships established in specific populations at specific times in the past, and these relationships may not apply, or may not apply in exactly the same way, to future populations living in different social contexts and circumstances.[6,7] The creators of some ARAIs recommend that evaluators accept their risk estimates rigidly (e.g., Ref. 8, p 182), without allowing for the potential presence of other factors with clear relationships to risk that "a prudent evaluator will always consider" (Ref. 9, p 3). Practicing clinicians can be tempted by the apparent definitiveness of numerical values to apply ARAIs uncritically or beyond their limited areas of established application, with results that can be misleading and prejudicial in legal contexts.[10–12]

The criticism of ARAIs that has aroused the most professional consternation in recent years involves a "controversy [that] relates to the applicability of group-derived risk estimates to an individual case" (Ref. 7, p 180). The controversy stems from mathematical claims set out in three publications by Hart, Cooke, and Michie (HCM), that the confidence intervals (CIs) for individual risk estimates are so wide "as to render risk estimates virtually meaningless" (Ref. 5, p s60). HCM initially made their case[5] using previously published data for the Violence Risk Assessment Guide (VRAG)[8] and the STATIC-99.[9]

Dr. Mossman is Professor of Clinical Psychiatry and Program Director of the Forensic Psychiatry Fellowship, University of Cincinnati College of Medicine, Department of Psychiatry and Behavioral Neuroscience, Cincinnati, OH. Address correspondence to: Douglas Mossman, MD, Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati College of Medicine, 260 Stetson Street, Suite 3200, Cincinnati, OH 45219. E-mail: douglas.mossman@uc.edu.

Disclosures of financial or other potential conflicts of interest: None.

More recently, Hart and Cooke used logistic regression methods to conclude that ARAIs cannot "estimate the specific probability or absolute likelihood of future violence with any reasonable degree of precision or certainty" (Ref. 13, p 81). If correct, this conclusion would represent a "brick wall limiting predictive accuracy at the individual level" (as one commentator put it[14]). Hart and Cooke concluded that "it is difficult to understand how ARAIs can be found legally admissible under *Daubert* or similar criteria . . . when the margins of error for individual risk estimates made using the tests are large, unknown, or incalculable" (Ref. 13, p 97).

Critics[15] have pointed out that the assertions of Hart and colleagues imply that people are mistaken when they do things that seem perfectly logical and rational. Yet the HCM argument has perplexed or persuaded many psychologists and psychiatrists. For example, DeClue and Zavodny have advised forensic mental health professionals not to report estimates of individual risk because "Hart and Cooke persuasively show that the lack of precision is not a limitation in one sample or one tool, but is endemic to attempts to make such predictions about individuals" (Ref. 16, p 149).

There are many good reasons for not making ARAI-based statements about individuals' risk of recidivism, but the mathematical argument offered by HCM is not one of them. The HCM argument errs in assuming implicitly that the purpose of risk assessment and probabilistic judgment is to make a prediction of something. Usually, however, we assess probabilities and risks to decide what to do, given the information we have, when the outcome is uncertain. Once this main purpose is clarified, the problems with the HCM argument become easier to see.

In this article, I summarize Hart and Cooke's most recent publication,[13] which they regard as an improvement on their earlier statements of the HCM thesis. Then, using a data set discussed by HCM, I describe a hypothetical betting scenario to convince readers that useful risk estimates (or probabilities) for individual instances flow naturally and obviously from information about groups of outcomes. Having convinced readers that it is practical and sensible to use group-derived probabilities for decisions about individual instances, I examine several key assertions by HCM to explain where their notions were valid and where their mathematical assertions led them astray.

The HCM Argument

As Hart and Cooke explain, ARAIs are tools "designed to estimate the likelihood of future criminal or violent behavior" (Ref. 13 13, p 81) and to "make individual risk estimates" (p 83) of the form "the risk that Jones will commit future violence is similar to the risk of people" (p 82) in a group with characteristics similar to those of Jones. Not all members of a group look or behave alike, however. HCM therefore "tried to distinguish between the precision of risk estimates at the aggregate or group level versus precision at the individual level" (p 83).

Suppose one draws a random sample of size n from a much larger group of persons, some of whom carry a particular trait. One can then estimate the proportion of persons in the group who have the trait by counting the number of persons in the sample with the trait, then dividing by n. "Like all sample statistics," state Hart and Cooke, "the proportion estimated is associated with a degree of uncertainty, or 'margin of error'" (Ref. 13, p 83). The size of the error depends, in part, on the size of n; the larger n is, the smaller the calculated margin of error, and vice versa.

To make statements about whether an individual from the group has the trait, say Hart and Cooke, one might want "to calculate the margin of error for individual propensities inferred from group risk estimates" (Ref. 13, pp 84–5). In their first article, HCM "employed an *ad hoc* procedure" to estimate a confidence interval for this individual propensity: they chose a formula developed by Wilson[17] and set $n = 1$ to calculate the precision of risk estimates for individuals. The resulting intervals were so broad as to encompass most of the possible 0-to-1 probability range, leading HCM to conclude that ARAIs "appeared to have some (albeit weak) predictive validity at the group level," but "the margins of error for individual risk estimates made using ARAIs are either large, unknown, or incalculable" (Ref. 13, p 85).

In response to what HCM interpreted as criticisms of their "*ad hoc* procedure," Cooke and Michie[18] used multivariate logistic regression "to predict the probability of a categorical outcome variable" (Ref. 13, p 86). They produced what they interpreted as prediction intervals and found that "the corresponding precision of individual probability estimates for offenders . . . was very low" (Ref. 13, p 86), with values again spanning nearly the entire possible 0-to-1 probability range.

In the third article, Hart and Cooke[19] used four item scores on the Sexual Violence Risk-20 administered to 90 sex offenders as the independent variables in a logistic regression "to evaluate the precision of individual risk estimates [here, of sex offense recidivism] made using ARAIs" (Ref. 13, p 88). They then generated 90 "individual risk estimates and their margins of error" (i.e., a 95 percent confidence interval for each individual's risk estimate). These intervals "overlapped completely" within the low- and high-risk groups, "and almost completely across groups save for a handful of cases . . . . These findings clearly illustrate that it was virtually impossible to make meaningful distinctions among subjects based on individual risk estimates made using ARAI scores" (Ref. 13, p 93). Hart and Cooke concluded that ARAIs are mathematically incapable of "estimat[ing] the specific probability or absolute likelihood that an individual person will commit violence in the future with any reasonable degree of precision or certainty" (Ref. 13, p 95).

Were HCM correct? To answer this, let us place some group data used by HCM[5] in a concrete (but hypothetical) context to see whether such data can yield probabilities precise enough to prove useful in individual instances.

Aunt Dorothy's Bequest

The morning after their Aunt Dorothy's funeral, her surviving kin—nephews Jim and Steve, and nieces Kathy and Mary—sat at Dorothy's kitchen table as Jim, the executor of Dorothy's estate, read provisions of their late aunt's will. After covering disposition of major financial assets and other items, Jim read this paragraph:

> I hereby bequeath my penny collection to my nieces Kathy and Mary, on condition that, during the six months after my death, they will honor my memory and amuse themselves by using the collection to engage in low-stakes games of chance.

Dorothy's penny collection occupied nine large jars with labels indicating that the contents of each were collected in one of the nine years before Dorothy's death. Kathy and Mary decided that for the next half year, they would make small bets on whether individual pennies drawn from the collection came from the Philadelphia or the Denver mint. (Denver pennies bear a D just below the year of mintage; pennies minted in Philadelphia have no letter below the year.)

Each evening over the next six months, Steve, Kathy, and Mary held three-way phone calls during which Steve took a jar, mixed its contents thoroughly, reached in blindly, drew a penny, and held it while Kathy and Mary used the following betting process:

1. The sisters took turns naming a price for a ticket like the one shown in Figure 1 that paid $1.00 for a Denver penny, but nothing otherwise. (Prices in fractional cents were allowed.)

2. After one sister set the price, the other sister announced whether she would buy or sell the ticket. Then Steve announced the outcome (D or no D).

3. Six such bets occurred each evening, with Mary and Kathy each naming three prices, and

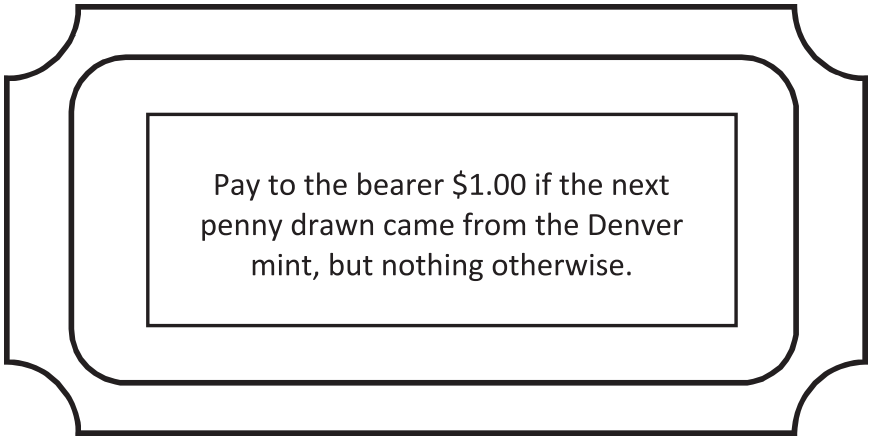

Figure 1. Ticket of the type bought and sold by Mary and Kathy as they bet on outcomes of penny drawings.

after determining the result of each bet, Steve returned the penny to the jar from which he had drawn it.

From the outset, Kathy and Mary played only for bragging rights; rather than keep the money, they planned to use their betting proceeds to buy dinner for everyone after six months.

Before the betting started, Jim had told Steve, "You know, Dorothy mentioned something about the penny collection a few months ago. When she started collecting the pennies, almost all of them had Denver mint marks. But over the nine years, Denver pennies became less and less common. When you draw each penny, tell Kathy and Mary which jar it came from before they place their bets. It might affect the betting odds that they agree on." Steve told Kathy what Dorothy had said, but he forgot to tell Mary. So, although both sisters heard which jars each of the pennies came from, only Kathy knew about the trend her aunt had observed.

Fifteen weeks and 103 phone calls later, 618 bets had taken place. From the outset, Kathy and Mary kept track of the drawings' results and accumulated the data shown in the first three columns of Table 1. The first column, labeled *y*, indicates how many years ($y = \{1, 2, \ldots, 9\}$) before Dorothy's death the jar's pennies were collected. The second column, labeled *r*, shows the number of Denver pennies observed, and the third column, labeled *n*, shows the number of draws from each jar. (By amazing coincidence, the numbers of Denver pennies and draws for each year equal the seven-year violent recidivism rates and numbers of recidivists in each risk category as reported by Quinsey and colleagues (Ref. 20, p 240) and by HCM.[5])

Throughout the betting, both sisters examined their data to see what they could reasonably say about the proportion of Denver pennies in Jar *y* ($y = \{1, 2, \ldots, 9\}$), given the outcomes thus far. Before we consider each sister's analysis after 618 bets, let's think about their aims. Both sisters sought to determine as precisely as possible what proportion of each jar's pennies came from the Denver mint, because this proportion would equal the probability of drawing a Denver penny. Notice also that this probability would equal the price in dollars at which each sister would be indifferent between buying and selling the ticket and the price that she would propose for the ticket. Why?

Suppose it was Mary's turn to name a price for a penny from Jar 4, and suppose that her best estimate (based on the data she had accumulated so far) was that 20 percent of a jar's pennies came from the Denver mint. If Mary named a price above $0.20, she would know that by selling the ticket, Kathy would gain a small advantage; if Mary named a price below $0.20, Kathy could buy the ticket and gain an advantage. So, unless Mary knew that Kathy was making systematic errors in estimating the proportion of Denver pennies in the jars, she could avoid giving Kathy an advantage only by naming a price equal to her best estimate of the proportion. Because Kathy was in exactly the same position as Mary, she also adopted the strategy of price equals best estimate of the proportion.

Now, neither sister knew exactly how the other was estimating proportions, but they had in fact used

Table 1 Results and Inferences About π_y, the Proportion of Denver Pennies in Jar *y, After n* Draws (With Replacement) From Each Jar in the Penny Collection

			$\hat{\pi}_y$ and 95% Credible Intervals			
y	*r*	*n*	Mary's Estimates		Kathy's Estimates*	
1	0	11	0.042	[0.000–0.200]	0.039	[0.022–0.063]
2	6	71	0.090	[0.036–0.166]	0.068	[0.044–0.010]
3	12	101	0.123	[0.067–0.192]	0.118	[0.086–0.155]
4	19	111	0.174	[0.110–0.249]	0.197	[0.159–0.237]
5	41	116	0.355	[0.271–0.443]	0.310	[0.270–0.352]
6	42	96	0.438	[0.341–0.537]	0.453	[0.401–0.505]
7	41	74	0.553	[0.441–0.663]	0.603	[0.535–0.670]
8	22	29	0.750	[0.584–0.885]	0.736	[0.656–0.806]
9	9	9	0.950	[0.762–1.000]	0.835	[0.761–0.895]

r, number of Denver Pennies; *y*, number of years before Dorothy's death.
* Inferences are based on 25,000 iterations of two WinBUGS chains that converged rapidly (after 100 updates). The first 5,000 values from each chain were discarded, and inferences were made on the final 20,000 values in each chain, thinned at an interval of 20.

different procedures. Because Mary had not heard Dorothy's statements to Jim about the declining frequency of Denver pennies, Mary treated the results from each jar as independent of one another. Using y subscripts to designate the jars (again, $y = \{1, 2, \ldots, 9\}$), Mary could simply have estimated each year's Denver-penny proportion π_y as $\hat{\pi}_y = \frac{r_y}{n_y}$, the number of Denver pennies, r_y, divided by the total number of drawings, n_y, from Jar y, in the belief that this unbiased value[21] should represent her expectation for future outcomes.

Mary took a different approach, however. She realized first that having observed no Denver pennies in 11 draws from Jar 1 did not necessarily mean that the jar contained no Denver pennies, so that setting a price of $0.00 seemed illogical. Similarly, getting nine Denver pennies in the first nine draws from Jar 9 did not imply that all the pennies in that jar came from Denver and that the fair ticket price must be $1.00. Mary also knew, however, that the data from each jar placed some numerical restrictions on what she should believe about the jar's likely contents. Therefore, Mary was interested in both the best single-number estimates for π_y and in knowing what her data should lead her to conclude about the plausible range of values for π_y. The technical background for Mary's reasoning and calculations appears in Appendix I, and her estimates and 95 percent credible intervals appear in the fourth and fifth columns of Table 1.

A different perspective informed Kathy's Bayesian analysis. She (based on Dorothy's statement) believed that the proportions of Denver pennies in each jar would be correlated with and therefore linked mathematically to y. Appendix II describes the details of Kathy's calculations, and her estimates and 95 percent credible intervals results appear in the sixth and seventh columns of Table 1.

Because the sisters began with different prior assumptions about the same data, their estimates of π_y for each jar were not identical. After 111 drawings from Jar 4, for example, Mary's estimates left her indifferent about buying or selling a $0.17 ticket that paid $1.00 upon her drawing a Denver penny. Kathy would have bought such a ticket gladly, however, because she was fairly confident that π_4 was more than .17.

An important conclusion from this discussion is that Mary and Kathy have posited different, subjective probabilities[22–25] regarding the next draw from Jar 4. Realizing that probability is subjective makes it reasonable to utter a phrase such as "Kathy's probability of drawing a Denver penny from Jar 4," because the phrase refers not to the contents of Jar 4, but to Kathy's degree of belief about the likelihood of drawing a Denver penny from that jar. The sisters' subjective probabilities determined how they proposed and accepted wagers, and their betting behavior represented a concrete illustration of the following general principle: "probability . . . is a rate at which an individual is willing to bet on the occurrence of an event. Betting rates are the primitive measurements that reveal your probabilities or someone else's probabilities, which are the only probabilities that really exist" (Ref. 26, p 90).

The preceding paragraphs let us distinguish among related but different quantities that the sisters might describe, based on their assumptions and the available data:

1. Asked to describe π_y, the proportion of Denver pennies in the jar from year y, each sister might respond with her single-number estimate of π_y listed in Table 1 (that is, her expectation based on the methods described in the appendices), or she might instead say that she was 95 percent sure that π_y lay within the intervals shown in Table 1.

2. Asked to describe the proportion of pennies that would have Denver mint marks were a large number of subsequent drawings to occur, the sisters might give similar answers (i.e., either reporting the single-number values for $\hat{\pi}_y$ or the 95% ranges that describe their beliefs).

3. Asked about the price of a Figure 1–type ticket that would leave them indifferent between buying and selling it, the sisters would use the single-number values for $\hat{\pi}_y$ listed in Table 1. These values are the sisters' expectations for each of the nine jars, the bases for their decisions about bets, and their single-event probabilities that the next penny from Jar y will bear a Denver mint mark.

4. Using her data in Table 1, Kathy might say, "The probability of drawing a Denver penny from Jar 4 is 16 to 24 percent." Though this assertion seems to refer to a single instance (the next drawing), Kathy's statement implicitly refers either to (a) a proportion of the jar's contents, or to (b) a plausible frequency for a certain

type of event over the long run. If (a) is what Kathy intends, the assertion means, "Based on the data and my background assumptions, I'm pretty sure that 16 to 24 percent of pennies in Jar 4 bear Denver mint marks." If (b) is Kathy's intention, her assertion means, "In a very large series of drawings, I'm 95 percent sure that 16 to 24 percent of the pennies will bear Denver mint marks."

Responding to HCM

In evaluating the penny data, the sisters used Bayesian statistical methods that are known to produce results very close to those yielded by the traditional, frequentist methods used by Hart, Cooke, and Michie in their 2007 and 2013 publications. If the mathematical arguments of HCM are correct, however, then the sisters, as Hart and Cooke suggest, "should consider whether it is best to give up altogether on the idea of calculating probability estimates of" drawing Denver pennies from each jar (see Ref. 13, p 98).

One point of telling the penny-betting story was to recast the calculations of Hart, Michie, and Cooke[5] in a context that makes it easy to see the relevance of group data to individual decisions. After 618 drawings, Mary would be foolish to think she knew next to nothing about the jars' contents or about how to establish a price for a ticket, the payoff of which depended on the next draw from Jar *y*. Mary's Bayesian analyses of the penny data yielded virtually the same group data (i.e., the same 95% intervals) as those that Hart, Michie, and Cooke reported in their Table 1,[5] and Mary felt 95 percent sure that her intervals contained the true proportion of Denver pennies in each jar. For purposes of making a betting decision, however, the relevant probability for Mary was the value at which she would be indifferent about buying or selling a ticket like the one shown in Figure 1. This value should equal her best point estimate of proportion of pennies that came from Denver.

Thinking about probabilities as expressions of beliefs that can be the basis for decisions helps one to avoid a mistake that HCM make in their discussion of group data. They ask readers to imagine a game of chance and write, "Suppose that Dealer, from an ordinary deck of cards, deals one to Player. If the card is a diamond, Player loses; but if the card is one of the other three suits, Player wins. After each deal, Dealer replaces the card and shuffles the deck" (Ref. 5, p s62). Over 10,000 games, say Hart and colleagues, Player can be 95 percent confident he will win 74 to 76 percent of the games. But if Dealer and Player play this game just once, "the estimated probability of a win is still 75 percent but the 95 percent CI is 12 to 99 percent. The simplest interpretation of this result is that Player cannot be highly confident that he will win—or lose—on a given deal" (Ref. 5, p s62).

Indeed, Player should not be confident, but not because of the interval that Hart and colleagues provided, which is what Player would calculate (using Wilson's method) for plausible values of the deck's nondiamond proportion if Player began knowing nothing about decks of cards and learned that in a single draw, one-fourth of the cards was a diamond. Leaving aside the impossibility of such a draw, HCM elided the distinction between one's confidence about a single yes-or-no outcome and one's probability concerning that outcome. Player should know that in a standard 52-card deck, three-fourths of the cards are not diamonds. If the payoff for a nondiamond is $1.00, the fair price for each round of the Dealer-Player game described above is $0.75, because probability of getting a nondiamond on any instance is .75. The problem with Hart and colleagues evaluation of ARAI data[5] is not their "*ad hoc* procedure" for interpreting information about a group of outcomes, but their misuse of Wilson's method. Their interval calculations effectively throw out all the information about each group, just as Player would be doing if he threw out his knowledge of 52-card decks, drew one card, and learned (somehow) that it was one-fourth a diamond.

Kathy's Bayesian analysis used the statistical procedure (logistic regression) that Hart and Cooke[13] employed to model probabilities for individuals. Kathy's credible intervals for each jar's Denver-penny proportions were narrower than those that Hart and Cooke[13] described, in part because Kathy used a larger data set. Had her data set been smaller, Kathy's credible intervals would have been wider, but her single-value probability estimates still would have helped her to make decisions about bets.

The HCM articles often refer to a predicted probability or an estimated probability, and they produce calculations which, they asserted, show that these entities are too imprecise to be useful. But what HCM tried to calculate is puzzling. To see why,

imagine a lunchtime discussion among Kathy, Mary, and their longtime friend Jane, who questioned whether they could apply their group data to individual bets.

"For the jars from which Steve has drawn lots of pennies," said Jane, "you know within fairly narrow intervals what fraction of Denver pennies you'd get in a very large number of future drawings from the jar. Those intervals are group results, however. On any individual penny drawing, you cannot predict the outcome with much confidence."

"We aren't trying to predict what will happen on any particular draw," replied Mary. "We're simply setting prices and deciding how to bet on each drawing."

"The probabilities we've calculated aren't predictions," added Kathy. "They represent degrees of belief based on rational mathematical strategies and our knowledge and experience. We don't know how many pennies are in each jar or how many came from Denver, but we're implementing the best possible strategy based on what we do know."

"But you still can't tell me the precise probability you predict for the next penny, nor can you give me a confidence interval for your prediction," protested Jane.

"We aren't trying to predict a probability, or anything else!" responded Mary. "I don't even understand what you mean."

"Maybe this will help," offered Kathy. "We aren't trying to predict anything: neither the outcome of the next penny drawing, nor our probability for that drawing. Probabilities aren't something we predict; they are degrees of belief that we ascribe to possible outcomes. Based on our data, we have formed beliefs about the intervals within which each jar's Denver proportion probably lies. But for purposes of making bets, our single-value estimates of the jars' proportions of Denver pennies are our probabilities for drawing a Denver penny. If I say my probability for getting a Denver penny is .20, that means I'm indifferent between selling or buying a $0.20 ticket that pays $1.00 if the next penny has a D mint mark."

Hart and Cooke believe that "the state of knowledge is arguably more advanced" in medicine than in psychology, yet "it is not common for physicians to give individual risk estimates" (Ref. 13, p 98) for outcomes. They quoted Henderson and Keiding,[27] who believe that while "'models and statistical indices can be useful at the group or population level, . . . human survival is so uncertain that even the best statistical analysis cannot provide single-number predictions of real use for individual patients'" (Ref. 13, p 99, quoting Ref. 27, p 703).

But it's easy to think of counterexamples. Suppose a 50-year-old man learns that half of people with his diagnosis die in five years. He would find this information very useful in deciding whether to purchase an annuity that would begin payouts only after he reached his 65th birthday. When you purchase insurance coverage, you may not tell yourself that you're making a bet, but that's what insurance is, and insurers find actuarial data very useful in deciding whether to offer you coverage and what your premium will be.

Hart and Cooke also state that "the definition of individual risk estimate used by ARAIs assumes that every person has a propensity for violence that is stable, dispositional, or trait-like" (Ref. 13, p 87). This characterization is incorrect for reasons that the penny story helps us understand. Each penny is unique: it has characteristics (e.g., its position in the jar) that make it theoretically distinguishable from other pennies and that influence whether it is the next one drawn by Steve. Other characteristics (e.g., mintage year) affect a penny's likelihood of coming from Denver. When the sisters set prices and made bets, however, the only information they had about each penny was its source jar, so the pennies' other characteristics could not affect their probability estimates. Similarly, if all one knew about an individual was his Static-99R score and that he came from a population for which the Static-99R data and rates were relevant, the individual's Static-99R score would be the best and the only basis for making a probabilistic judgment about his future behavior. This is true even though many factors not considered by the Static-99R (e.g., employment status, substance use, and family relationships) affect a sex offender's likelihood of recidivism.

Making Predictions Versus Assessing Risks

Making individual predictions is neither the aim of ARAIs nor the purpose for which they are designed. As the term actuarial risk assessment instrument suggests, a validated actuarial tool provides a numerical value for the risk of an event.

In states that authorize civil commitment of so-called sexual predators (see, for example, Ref. 28), courts typically solicit expert testimony relevant to

whether an individual is "likely to engage in acts of sexual violence" if not confined. Courts disagree about exactly what numerical value is entailed by the word "likely."[29–31] Yet under a plain-English interpretation as well as the interpretations that U.S. courts have provided, this phrase requests a statement regarding an individual's probability of engaging in a certain kind of behavior, not a prediction.

When Kathy and Mary bet on Dorothy's pennies, the results of previous drawings from a particular jar were obviously relevant to the likelihood that the next penny from that jar would bear a Denver mint mark. In providing so-called norms—for example, rates of recidivism or violent acts—for translating ARAI scores to probabilities of recidivism, ARAI designers are saying that their source data (the bases for the norms) are as relevant to any evaluee as are the source data used by Mary and Kathy to calculate probabilities of drawing Denver pennies.

This claim is questionable. An ARAI may do equally well at ranking the risks of individuals from two populations, yet have probabilities associated with particular scores that differ because the populations' overall base rates differ.[6,32] Helmus and colleagues[33] and Singh and colleagues[7] have shown that the offending rates associated with particular ARAIs scores differ across locales, but this should not surprise anyone. Social, economic, and political conditions in different places are likely to influence interpersonal behaviors such as acting violently or committing a sex offense.

Thus, one can disagree with the HCM mathematical argument, yet agree with Hart and Cooke that "it is arbitrary and therefore inappropriate to rely solely on a statistical algorithm . . . professionals [should] recognize that their decisions ultimately require consideration of the totality of circumstances—not just the items of a particular test" (Ref. 13, p 98). Sensible exponents of actuarially validated risk assessment know that factors besides those considered by the instrument may influence risk, but because actuarially based risk assessment methods typically outperform other judgment methods (especially unstructured clinical judgment), the onus rests with those who propose adjusting estimates to prove that their adjustments yield results that are superior to those based on actuarial judgment alone.

Final Comments

The term probability causes confusion because it has many uses. Readers interested in exploring these uses would do well to start with the recent article by Buchanan,[34] which provides a short, elegant discussion of probability in the context of forensic risk assessment.

Thinking carefully about probability involves thinking carefully about numbers, and many people, including judges and jurors, have trouble understanding numerical information and using it rationally. Even numerically sophisticated people can get confused by the statistics that describe probabilities, estimates of proportions, and risks of events, and also by the relationships between these mathematical quantities, people's predictions about individual events, and optimal decisions about what to do in uncertain individual circumstances.

To make these relationships clearer and to dispel the misunderstandings generated by the well-intended efforts of HCM, I have explicated the relationship between rational use of data and probabilities with a story about betting. Some people object to betting on moral grounds,[35] and some mental health professionals may disapprove of describing psycholegal matters as though the clinicians involved were making bets,[36] but since the 17th century, the arguments for deriving and illustrating basic principles of probability have used gambling as a standard metaphor for explaining those principles.[37]

I have used the same metaphor to give readers an intuitive feel for why HCM's mathematical analyses must contain errors. Some readers may take offense at legal schemes that impose confinement based on principles that govern rational betting. If you are one of those readers, I agree with you, but such opinions reflect moral or legal positions about the proper basis for confining people, not mathematical arguments about the precision of risk assessments.

Appendix I

Mary approached the problem of estimating from a Bayesian perspective. She sought to establish a probability distribution $p(\pi_y|y, r_y, n_y)$ for each jar, based on her data. She treated the series of penny drawings as a set of Bernoulli trials—that is, as independent, random experiments with exactly two possible outcomes—in which the probability of a Denver penny was the same every time a drawing occurred. Starting with the Jeffreys' prior for binomial data from Bernoulli trials, Mary's posterior distributions for $p(\pi_y|y, r_y, n_y)$, after having observing r_y Denver pennies in n_y draws from each jar, were Beta($r_y + ½, n_y - r_y + ½$). The Jeffreys' prior for beta

distributions produces intervals with good coverage properties from a traditional (frequentist) statistical standpoint.[38] (For further explanations of the rationale for using this prior and examples applied to psychiatric contexts, see Refs. 39,40.)

The expectation for a Beta(α,β) distribution on the interval [0,1] is $\alpha/(\alpha+\beta)$. In Table 1, Mary's Bayesian estimates for π_y reflect this calculation, so they differ a bit from what calculating estimates of π_y as $\hat{\pi}_y = \frac{r_y}{n_y}$ would yield. Because Mary considered and used the data to assign Bayesian probability distributions to π_y, then having observed (for example) that $r_2 = 6$ and $n_2 = 71$, she could say, "I am 95 percent sure that π_2 lies between .036 and .166."

Appendix II

From among several candidates for link functions (see Ref. 41, § 6.5), Kathy chose this logistic regression model to fit the data:

$$r_y \sim \text{Binomial}(\pi_y, n_y), \quad \text{logit}(\pi_y) = \alpha + \beta(y - \bar{y})$$

The first part of Kathy's model states that r_y, the number of Denver pennies out of n_y drawings from Jar y, results from a series of Bernoulli trials that follow (or are distributed as) a binomial distribution, where π_y is the actual (but unobserved) proportion of Denver pennies in Jar y. (If taken by itself, this part of Kathy's model would be the same as Mary's.) The second part posits a standard logistic model in which y (years before death) is the independent variable and $\bar{y}$ (the mean years in the data sample; here, $\bar{y} = 3.74$) aids in convergence because it reduces the dependence of α on β (see Ref. 41, p 115). Kathy's model was implemented using WinBUGS, a free statistical software program for effecting Bayesian analyses using Markov chain Monte Carlo methods.[41] For additional details, see the footnote to Table 1.

References

1. Harris GT, Rice ME, Quinsey VL: Violent recidivism of mentally disordered offenders: the development of a statistical prediction instrument. Crim Just & Behav 20:315–35, 1993
2. Hanson RK, Thornton D: Improving risk assessments for sex offenders: a comparison of three actuarial scales. Law & Hum Behav 24:119–36, 2000
3. Helmus L, Thornton D, Hanson RK, *et al*: Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: revised age weights. Sex Abuse 24:64–101, 2012
4. Monahan J, Steadman H, Robbins P, *et al*: An actuarial model of violence risk assessment for persons with mental disorders. Psychiatr Serv 56:810–15, 2005
5. Hart SD, Michie C, Cooke DJ: Precision of actuarial risk assessment instruments: evaluating the "margins of error" of group v. individual predictions of violence. Br J Psychiatry 190:s60–5, 2007
6. Mossman D: Another look at interpreting risk categories. Sex Abuse 18:41–63, 2006
7. Singh JP, Fazel S, Gueorguieva R, *et al*: Rates of violence in patients classified as high risk by structured risk assessment instruments. Br J Psychiatry 204:180–7, 2014
8. Quinsey VL, Harris GT, Rice ME, *et al*: Violent Offenders: Appraising and Managing Risk (ed 2). Washington, DC: American Psychological Association, 2006
9. Harris AJ, Phenix A, Hanson RK, et al: STATIC-99 coding rules revised. Ottawa: Public Safety and Emergency Preparedness Canada, 2003
10. Rogers R: The uncritical acceptance of risk assessment in forensic practice. Law & Hum Behav 24:595–605, 2000
11. Silver E, Miller LL: A cautionary note on the use of actuarial risk assessment tools for social control. Crime & Delinq 48:138–61, 2002
12. Campbell TW, DeClue G: Flying blind with naked factors: problems and pitfalls in adjusted-actuarial sex-offender risk assessment. Open Access J Forensic Psychol 2:75–101, 2010
13. Hart SD, Cooke DJ: Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. Behav Sci & L 31:81–102, 2013
14. Franklin K: Showdown looming over predictive accuracy of actuarials: large error rates thwart individual risk prediction. In the News by Karen Franklin, PhD, January 27, 2013. Available at http://forensicpsychologist.blogspot.com/2013/01/showdown-looming-over-predictive.html. Accessed November 17, 2014
15. Skeem JL, Monahan J: Current directions in violence risk assessment. Curr Dir Psychol Sci 20:38–42, 2011
16. DeClue G, Zavodny DL: Forensic use of the Static-99R: Part 4. Risk communication. J Threat Assess Manage 1:145–61, 2014
17. Wilson EB: Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22:209–12, 1927
18. Cooke DJ, Michie C: Limitations of diagnostic precision and predictive utility in the individual case: a challenge for forensic practice. Law & Hum Behav 34:259–74, 2010
19. Boer DP, Hart SD, *et al*: Manual for the Sexual Violence Risk-20: Professional Guidelines for Assessing Risk of Sexual Violence. Vancouver, British Columbia, Canada: Institute Against Family Violence, 1997
20. Quinsey VL, Harris GT, Rice ME, *et al*: Violent Offenders: Appraising and Managing Risk. Washington, DC: American Psychological Association, 1998
21. Jaynes ET. Probability Theory: The Logic of Science. Cambridge, UK: Cambridge University Press, 2003
22. Jeffrey R: Subjective Probability: The Real Thing. Cambridge, UK: Cambridge University Press, 2004
23. de Finetti B: Foresight: its logical laws, its subjective sources. (La prévision: ses lois logiques, ses sources subjectives.) In French. Ann Instit Henri Poincaré (Annales de l'Institut Henri Poincaré) 7:1–68, 1937
24. Ramsey FP: Truth and probability, in The Foundations of Mathematics and Other Logical Essays. Edited by Braithwaite RB. London: Rutledge & Kegan Paul, Ltd., 1929, pp 156–98
25. Savage LJ: The Foundations of Statistics. New York: John Wiley and Sons, 1954 (ed 2, 1972, New York: Dover)
26. Nau RF: De Finetti was right: probability does not exist. Theory Decis 51:89–124, 2001
27. Henderson R, Keiding N: Individual survival time prediction using statistical models. J Med Ethics 31:703–6, 2005
28. N.J.S.A. § 30:4-27.26
29. In re Commitment of W. Z., 801 A.2d 205 (N.J. 2002)
30. Brooks v. Franklin, 36 P.3d 1034 (Wash. 2001)
31. People v. Ghilotti, 44 P.3d 949 (Cal. 2002)
32. Singh J: Predictive validity performance indicators in violence risk assessment: a methodological primer. Behav Sci & L 31:8–22, 2013
33. Helmus L, Hanson RK, Thornton D: Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: a meta-analysis. Crim Just & Behav 39:1148–71, 2012
34. Buchanan A: Violence risk assessment in clinical settings: being sure about being sure. Behav Sci & L 31:74–80, 2013

35. Borna S, Lowry J: Gambling and speculation. J Bus Ethics 6:219–24, 1987
36. Rogers R, Salekin RT: Beguiled by Bayes: a reanalysis of Mossman and Hart's estimates of malingering. Behav Sci & L 16:147–53, 1998
37. Hald A: A History of Probability and Statistics and Their Applications Before 1750. Hoboken, NJ: John Wiley & Sons, 2003
38. Brown LD, Cai TT, DasGupta A: Interval estimation for a binomial proportion. Stat Sci 16:101–33, 2001
39. Mossman D, Berger JO: Intervals for post-test probabilities: a comparison of five methods. Med Decis Mak 21:498–507, 2001
40. Berger JO: The case for objective Bayesian analysis. Bayesian Anal 3:385–402, 2006
41. Lunn D, Jackson C, Best N, *et al*: The BUGS Book: A Practical Introduction to Bayesian Analysis. Boca Raton, FL: Chapman & Hall/CRC, 2013