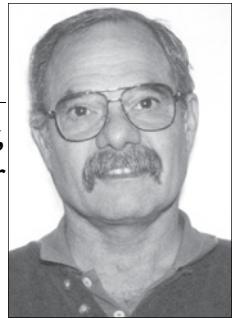


Visual Revelations

Howard Wainer,
Column Editor



Until Proven Guilty: False Positives and the War on Terror

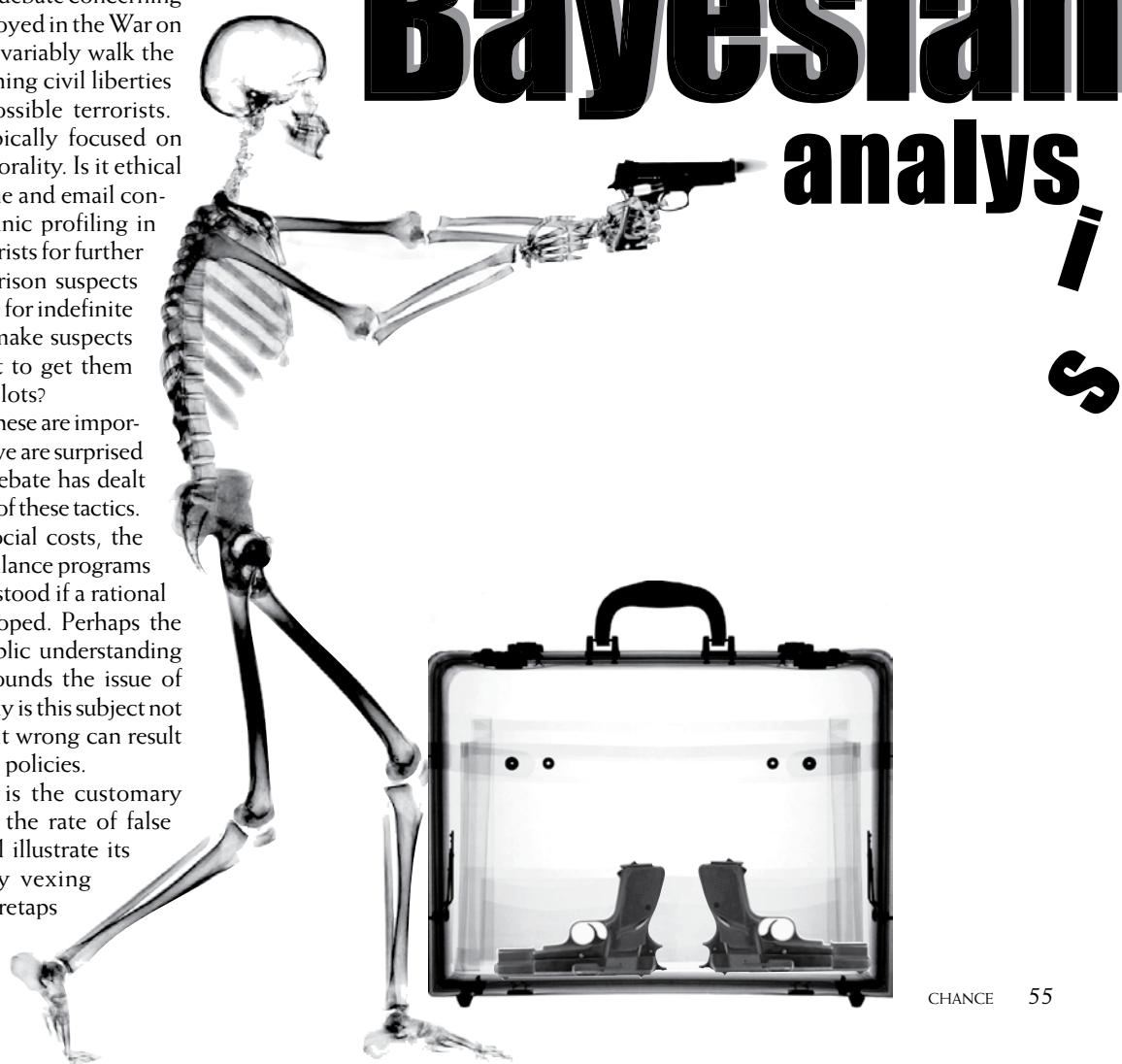
Sam Savage and Howard Wainer

Print and electronic media have been filled with debate concerning the tactics employed in the War on Terror. These must invariably walk the line between maintaining civil liberties and screening for possible terrorists. Discussions have typically focused on issues of ethics and morality. Is it ethical to eavesdrop on phone and email conversations; to use ethnic profiling in picking possible terrorists for further investigation; to imprison suspects without legal recourse for indefinite amounts of time; to make suspects miserable in an effort to get them to reveal cabals and plots?

While we believe these are important questions to ask, we are surprised by how little of the debate has dealt with the likely success of these tactics. Given the obvious social costs, the efficacy of such surveillance programs must be clearly understood if a rational policy is to be developed. Perhaps the biggest barrier to public understanding of this problem surrounds the issue of false positives. Not only is this subject not intuitive, but getting it wrong can result in counter-productive policies.

Bayesian analysis is the customary tool for determining the rate of false positives, and we will illustrate its use on a particularly vexing problem: the use of wiretaps to uncover possible

Bayesian analysis



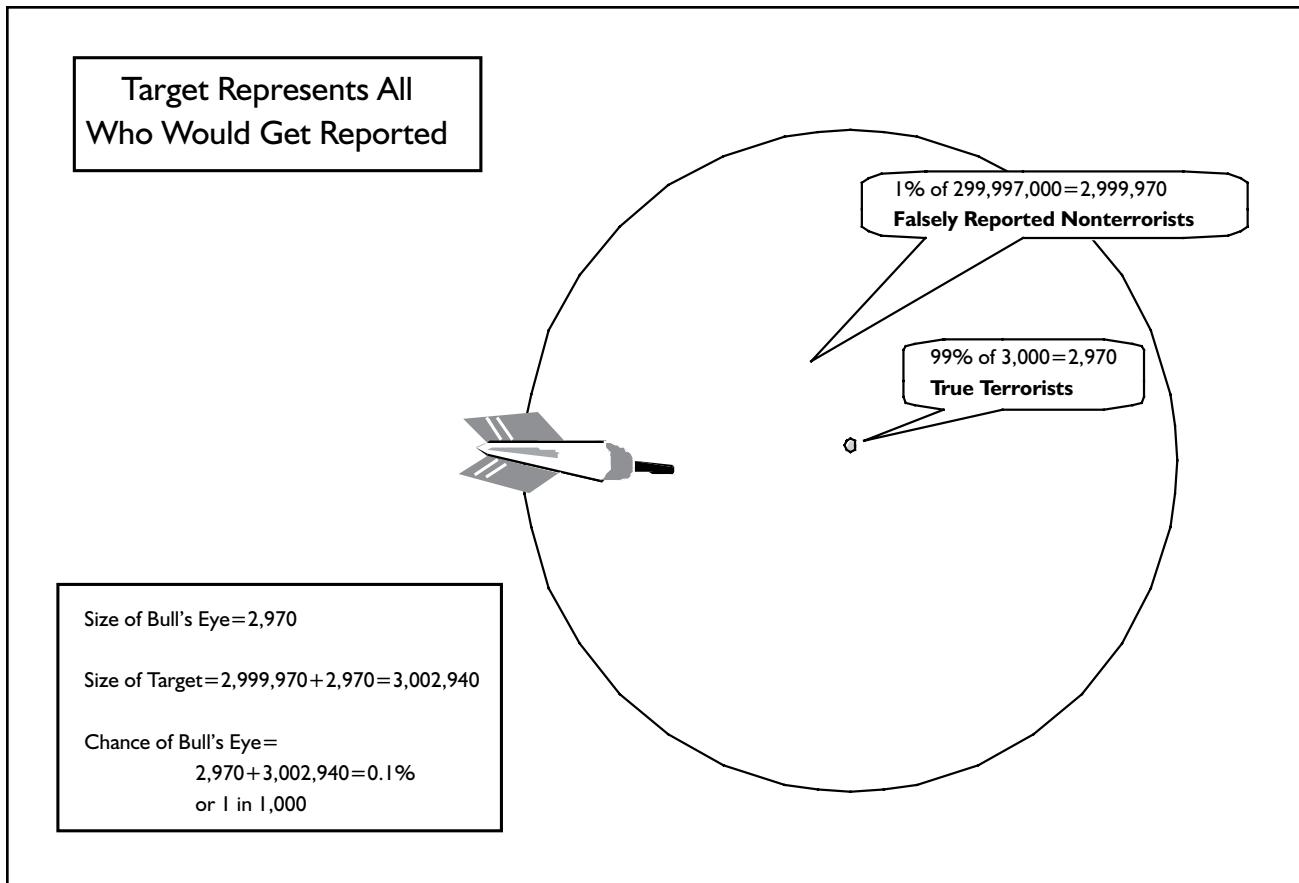


Figure 1. A target whose total area is $2,999,970 + 2,970$ and whose bull's eye has an area of 2,970, representing those terrorists who would be identified by wire tapping. (From *The Flaw of Averages* by Sam Savage, in preparation for John Wiley & Sons, all rights reserved.)

terrorists living in our midst. The use of such wiretaps has been hotly debated. To evaluate their effectiveness, we must consider both the chance we will correctly identify a terrorist if we have one on the other end of the line and the overall prevalence of terrorists in the population we are listening in on.

We shall start with the latter. How many terrorists are currently in the United States? (By terrorists, we mean hard-core extremists intent on mass murder and mayhem.) Let us assume for argument's sake that among the 300,000,000 people living in the United States, there are 3,000 terrorists. Or, in other words, the prevalence is one terrorist per 100,000. Once this case is understood, it will be easy to generalize for other numbers.

Now, consider a magic bullet for this threat: unlimited wiretapping tied to advanced voice analysis software on everyone's line that could detect would-be terrorists within the utterance of their

first three words on the phone. The software would automatically call in the FBI, as required, to arrest and question those who triggered the terror detector. Let's assume the system was 99% accurate. That is, if a true terrorist was on the line, it would notify the FBI 99% of the time, while for nonterrorists, it would call the FBI (in error) only 1% of the time. Although such detection software probably could never be this accurate, it is instructive to think through the effectiveness of such a system if it did exist.

The False Positive Problem

When the FBI gets a report from the system, what is the chance it has identified a true terrorist? There are two possibilities. Either there has been the correct report of a true terrorist or the false report of a nonterrorist. Of the 3,000 true terrorists, 99%—or 2,970—would be correctly identified. Of the 299,997,000 nonterrorists (300 million

minus the 3,000 terrorists), only 1%—or 2,999,970—would be falsely reported. This may be thought of as a target whose total area is $2,999,970 + 2,970$ and whose bull's eye has an area of 2,970, as shown in Figure 1.

Thus, the chance of correctly identifying a true terrorist is analogous to hitting the bull's eye with a dart—roughly one in a thousand, even with a 99% accurate detector. If there were fewer than 3,000 terrorists, this probability would decrease still further. And even if the number of terrorists went up ten-fold to 30,000, the chances of a correct identification would still be only 1 in 100. What looked at first like a magic bullet doesn't look as attractive once we realize that about 999 out of 1,000 suspects are innocent. This is the "false positives" problem.

But, of course, we started with an extreme case for dramatic effect. In reality, the bad news is that your terrorist detector would be nowhere near 99%

effective. But the good news is that you would be much more selective in who you wire tapped in the first place. So, in reality, the accuracy would be lower, but the prevalence could be expected to be higher.

Suppose you can sufficiently narrow your target population to the point that the prevalence is up to one actual terrorist per 100 people wiretapped. Also, assume a 90% effective test. That is, a true terrorist will be correctly identified 90% of the time and an innocent person 10% of the time. The chance of a false positive is still 91.7%. That is, even when someone triggers an arrest, the odds are 11 to 1 they are not a terrorist.

Table 1 shows the probability that someone implicated by a terror screening system is actually innocent, based on the prevalence of terrorists in the screened population and the accuracy of the test. In this table, the accuracy is defined as both the probability a true terrorist is identified as such and an innocent person is identified as innocent. There is no need for these two probabilities to be the same, and later we provide a calculator that will solve the general case of asymmetric accuracy.

Upon calculating this table, even the authors were surprised to find that if 1 person in 10 were an actual terrorist, and if the screen were 90% accurate, you would still have as many innocent suspects as guilty ones.

For example, consider a battle with insurgents who make up 10% of a population. Table 1 implies that if you call in air strikes on suspected enemy positions based on targeting intelligence that is 90% accurate, you will be killing as many civilians as combatants.

This result suggests arresting and prosecuting terrorists in our midst is a real challenge. A fascinating web site at Syracuse University provides real statistics in this area. According to the Transactional Records Access Clearinghouse at <http://trac.syr.edu/tracreports/terrorism/169>, in recent federal criminal prosecutions under the Justice Department Program of International Terrorism, roughly 90% of the cases brought in have been declined for further action.

The Cost of False Positives

It is tempting for politicians to play off our fears of horrific, but extremely unlikely, events. When they do, it is

Table 1—Percentage of False Positives by Prevalence and Indicator Accuracy

Prevalence Number Screened per Actual Terrorist		Accuracy of Screen				
		50%	60%	70%	80%	90%
100,000	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
10,000	100.0%	100.0%	100.0%	100.0%	99.9%	
1,000	99.9%	99.9%	99.8%	99.6%	99.1%	
100	99.0%	98.5%	97.7%	96.1%	91.7%	
10	90.0%	85.7%	79.4%	69.2%	50.0%	

easy for the nonstatistically trained to fall for faulty logic. For example, a few years ago, a supporter of an anti-missile system for protecting the United States from rogue states argued that the specter of a nuclear weapon destroying New York was so horrible that the U.S. government should stop at nothing to deter it. Oddly, he didn't bring up the fact that it is much more likely such a weapon would be delivered to New York by ship, and that a missile attack—aside from being much more expensive and difficult—is the only delivery method providing a definitive return address for our own nuclear response. Interestingly, recent studies have shown that an effective program for detecting weapons of mass destruction smuggled on ships would cost about as much as an anti-missile system. So, here, the question is easy. Should we defend against a likely source of attack, or a rare one? But, in general, the decision will be how much incursion of our personal freedom we are willing to endure per life saved.

In making this calculation, we must be mindful of the extent to which the harsh treatment of innocents can create terrorists. For example, in a 2003 memo, then U.S. Defense Secretary Donald Rumsfeld asked, "Are we capturing, killing, or deterring and dissuading more terrorists every day than the madrassas and the radical clerics are recruiting, training, and deploying against us?" We must add to our cost function the chance that the prosecution of the innocent will bolster the recruiting efforts of our enemies.

Is this probability algebra limited to the War on Terror? Consider what would be the likely outcome if we had universal

AIDS testing. Because the test is far from perfect, we would surely find that the number of false positives would dominate the number of AIDS cases uncovered. And how much of the agony associated with receiving such an incorrect diagnosis would compensate for finding an otherwise undetected case?

Our point is not that all testing, whether for disease or terrorism, is fruitless; only that we should be aware of the calculus of false positives and use whatever ancillary information is available and suitable to shrink the candidate population and probability of errors enough so that the false positive rates fall into line with a reasonable cost function.

The False Positive Probability Calculator

Calculating the rate of false positives involves comparing the ratio of areas in the target of Figure 1. This is usually done with Bayes' formula, which is known to all statisticians, but apparently few policymakers. Thus, as a service to mankind, we have created a false positive calculator in Microsoft Excel, available for free download at www.FlauOfAverages.com.

This calculator was created with XLTree (see www.AnalyCorp.com). First, a probability tree was generated based on the prevalence and probabilities shown in Figure 2. This tree was then inverted (flipped) to perform what is known as Bayesian Inversion. Figure 3 displays the formula section of the calculator, along with the other outputs in the dark boxes and intermediate calculations in grey boxes.

Last, we are certainly not the first statisticians to use the tools of our trade

False Positive Probability Calculator

Inputs

Prevalence of Trait X in Screened Population
1 in 100

Probability That Trait X Is Correctly Detected
90%

Probability That Lack of Trait X Is Correctly Detected
95%

Probability of False Positive \Rightarrow 84.62%

Figure 2. False positive calculator, www.flawofaverages.com

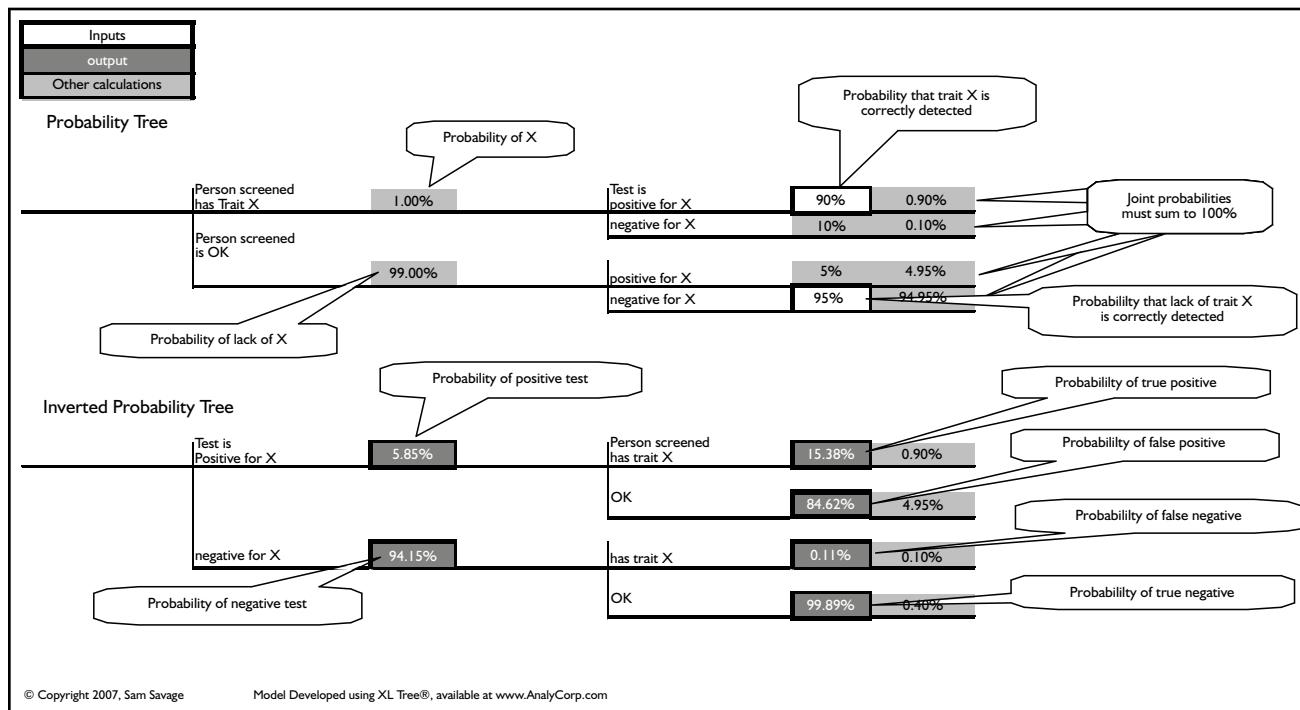


Figure 3. Trees underlying the false positive calculator. Courtesy of Sam Savage.

to look into the topic of terrorism. Nor even the first to use the pages of *CHANCE* to do so. The challenge is not to convince other statisticians, but to get decisionmakers to take both Type I and Type II errors into consideration when making policy. Toward this end, we believe embodying these ideas in a widely available calculator increases the chances of holding decisionmakers accountable. So, the next time you become aware of a politician or bureaucrat about to make a decision that may

bring more harm through false positives than benefit though true ones, we suggest you send them the link to the calculator. **C**

Further Reading

Banks, D. (2002). "Statistics for Homeland Defense." *CHANCE*, 15(1):8–10.
Rumsfeld, D. (2005). "Rumsfeld's War-on-Terror Memo." *USA Today*, www.usatoday.com/news/washington/executive/rumsfeld-memo.htm.

Savage, S. (In preparation for John Wiley & Sons). *The Flaw of Averages*. www.FlauOfAverages.com.

Stoto, M.A., Schonlau, M., and Mariano, L.T. (2004). "Syndromic Surveillance: Is It Worth the Effort?" *CHANCE*, 17:19–24.

Wein, L.M., Wilkins, A.H., Baveja, M., and Flynn, S.E. (2006). "Preventing the Importation of Illicit Nuclear Materials in Shipping Containers." *Risk Analysis*, 26(5):1377–1393.