Screening Tests for Terrorism

Does the TSA's latest procedure make us safer?

BY G. STUART MENDENHALL AND MARK SCHMIDHOFER

Humans are notoriously bad at judging risk of low-probability events. We may fear getting struck by lightning, which kills approximately 40 people per year nationwide, but think little of the risk of driving, where annual deaths are approximately 35,000. Even when provided with observed statistics of an occurrence, the results may be confusing and occasionally nonintuitive, as the relative calculation and assessment of minute occurrences is difficult to grasp. This may lead to poor personal decisions, such as the overestimation of risk of theft or injury leading to overpaying for insurance. However, in determination of public policy, a significant error in judgment of risks can have wide-reaching effects, harming and inconveniencing large numbers of people over an inappropriate concern.

In the medical field, we routinely use statistical analysis to assess the advisability of screening for diseases in our patients, acknowledging the deficiencies of intuition and estimation of even the most expert in a field. Many conditions only affect a small number of persons, and it is up to doctors, public health officials, and statisticians to devise a way to look for these diseases so that they may be addressed and treated. On the surface, the problem may seem simple: Why not screen everybody? What possible negative effect could a screening test have for a person or population?

G. Stuart Mendenhall is a cardiologist and cardiac electrophysiologist, and is an assistant professor of medicine at the University of Pittsburgh Medical Center. Mark Schmidhofer is a clinical cardiologist and an associate professor of medicine at the University of Pittsburgh Medical Center.

Unfortunately, in real life no diagnostic test—medical or otherwise—can be free. By "free" we mean not just without monetary cost, but without the costs of discomfort, hassle, or risk of harm to the individual being tested or undergoing the procedure. As an example, a screening CAT scan to look for cancer may ironically contribute to cancer, as those same X-rays have a possibility—albeit low—of damaging DNA. If the cancer is infrequent enough, or there is no benefit to early detection, the low risk of screening may not be justified. Similarly, an invasive test with needles, pain, or discomfort may be subjectively worse to a patient than a small probability of a disease. We must judge low risks carefully, balancing one risk against another, weighted by their relative desirability or undesirability.

Policy decisions such as the implementation of broad screening measures are difficult to make with absence of emotion, yet these are precisely the types of decisions that are best made by appealing to facts to prevent distortions based on incorrect assumptions or perceptions. For instance, prostate cancer has a yearly incidence (diagnoses made per year) of around 180,000 in the United States, and about 29,000 die from this disease each year. Many of these cancers can be detected by a rectal exam or simple blood test, so, on cursory examination, it seems to follow naturally that doctors should issue a recommendation to screen all males. Indeed, a few groups have heavily invested in public service announcements to recommend screening, coupled with highly emotional appeals that "if you love your partner, you will get him screened." However, the low specificity of the test means that many men who would test positive may not have prostate cancer, or would have a very benign form of it. They may then be subject to unneeded and invasive treatment, including biopsy and surgery, and suffer consequences including incontinence and impotence. Many men will die from something else, old age or another medical condition, not a slowly growing, yet now discovered, prostate cancer. The indignity and suffering resulting from the screening itself may indeed be worse than the disease. After closely looking at the outcomes of widespread screening, the U.S. Preventative Services Task Force has not recommended routine prostate screening in men, citing the insufficient evidence that it overall favorably affects outcomes.

Screening tests are never perfect in real life. How accurate each test is, and the types of failures or misdiagnoses that occur, may be closely characterized in well-defined methods. The World Health Organization formalized these in the 1968 publication "Principles and Practice of Screening for Disease," a landmark formulation of concepts that remain at the cornerstone of detection of rare diseases in a population. The core metrics outlined for analysis of the performance, and thus fitness for widespread implementation, of a screening test are validity, reliability, yield, cost, and acceptance by the population. We will discuss each of these in detail in this article, as they apply to another type of screening that affects much of the public today: airport screening procedures for terrorist activity.

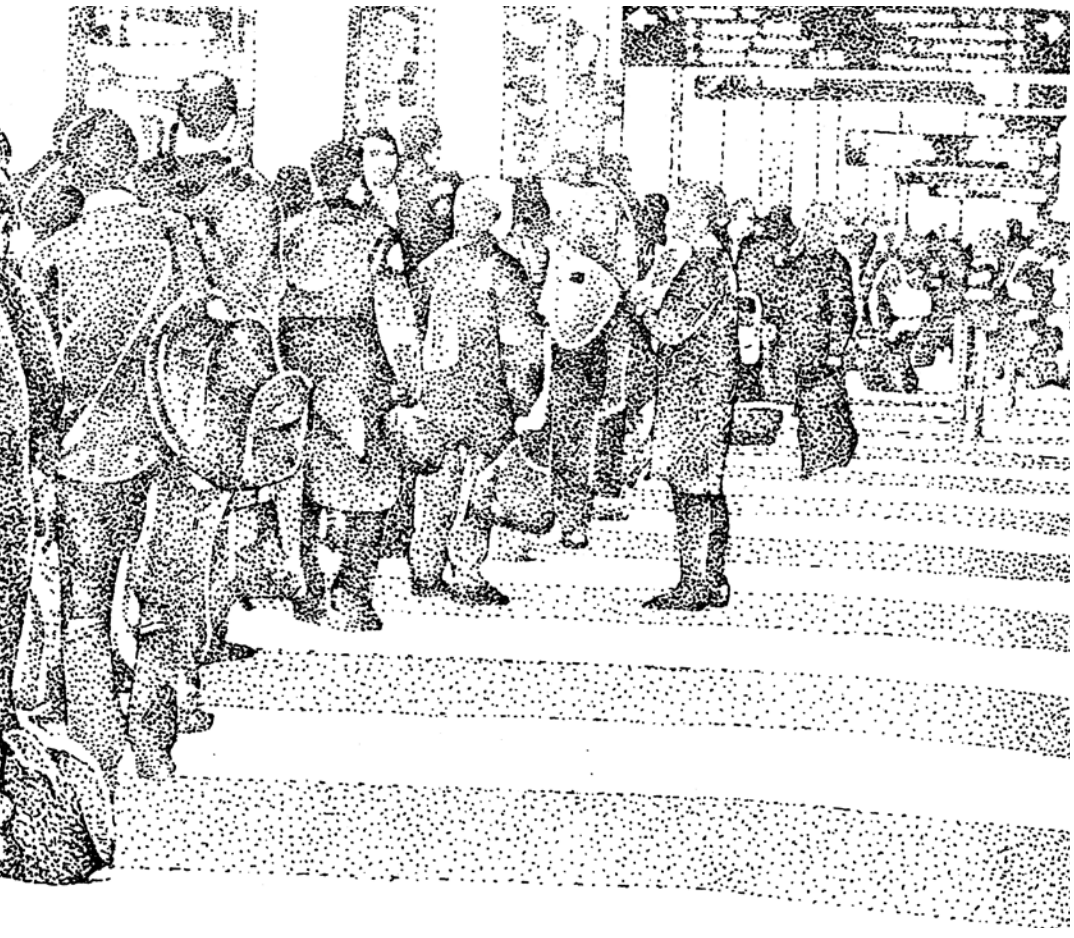

ILLUSTRATION BY MORGAN BALLARD

The U.S. Transportation Security Administration has recently started adding an additional "layer of screening" designed to detect terrorism, based on "behavior detection." The new screening regimen, which goes by the clever acronym "SPOT" (for "Screening Passengers by Observation Techniques"), uses seconds-long interviews with passengers to look for "micro-expressions" that will give "evidence of deception." Subjects who fail this screen will then be sent for higher levels of screening. The program is currently slated for Boston's Logan Airport and a few others, but if the program shows value, there are advertised plans for it to be implemented more widely.

Could a program like this help find terrorists? Just as a medical test is designed to detect a rare condition, this interview is designed to "diagnose" terrorists. We can thus use the principles of evaluating medical screening

to theorize what to expect from the TSA's new terrorism screen.

Terrorism And Thomas Bayes

As outlined, the first question we should ask of any test is, "Is the test valid?" That is, what is its performance? How many times does it miss disease (false negative), and how many times does it over-diagnose (false positive)? Just as in medical testing, no real-world test can perform perfectly and there are always tradeoffs that must be made in implementation. A test can be more or less sensitive (meaning that it catches/flags most of what it is designed to detect), and more or less specific (meaning it rarely flags inappropriately), but it is not possible for any real life test to have 100 percent accuracy.

However, even with extremely good tests, when attempting to detect small probabilities, a problem emerges. This was first characterized in the 18th century by Thomas Bayes, and gave rise to the mathematical theorem that is named for him. The theorem tells us how sure we can be of the presence or absence of a condition after application of an imperfect test when applied to a rare condition, be it a disease or a terrorist. Surprisingly, when evaluated using this theorem, we discover that many tests may not add any net benefit and may indeed be harmful on net, wasting time and resources.

Bayes' Theorem states that the probability of a condition actually being present in the face of a positive result of a test is equal to the probability of the test being positive when the condition is actually present (a measure of the sensitivity of the test), multiplied by the probability of the condition actually being present, divided by the overall probability of a positive test result.

To better flesh out this equation, we first need to define a few terms. The "true positive" rate is the fraction of people who actually have a condition that will have a positive test. The "false positive" rate is the fraction of people who don't have a condition but will have an abnormal (positive) test. The "prevalence" of a condition is the fraction of the tested population who actually have the condition. Mathematical shorthand for a conditional probability is $P(A|B)$, read as "probability of A given B." Concerning terrorism, we can express Bayes' Theorem for the probability of identifying a terrorist using the TSA's new behavioral testing screen as follows:

$$P(\text{terrorist} \mid \text{abnormal behavior}) = \frac{P(\text{abnormal behavior} \mid \text{terrorist}) \times P(\text{terrorist})}{P(\text{abnormal behavior} \mid \text{terrorist}) \times P(\text{terrorist}) + P(\text{abnormal behavior} \mid \text{nonterrorist}) \times P(\text{nonterrorist})}$$

As the prevalence of terrorists goes to zero, the test rapidly diminishes in value, as the probability of any actual positive result (a good detection by the TSA interviewer) also rapidly approaches zero. Similarly, if the false negative rate (inappropriate positive result of the test) is much higher than the prevalence, the test again rapidly diminishes in value. This can explain why the results of many screening tests are not very useful

Let's analyze SPOT using Bayes' Theorem and some numerical approximations and conservative assumptions. There are 2 million domestic air travelers each day. Over the past 11 years, there have been 19 September 11th terrorists, 1 underwear bomber that was not detected by screening, and 1 shoe bomber, also undetected. That means that there have been 21 persons actively engaged in terrorism who boarded U.S. flights in 11 years. To make the math easier, let's substantially increase the number of terrorists and say that 50 active terrorists board U.S. flights in a decade. Thus, the probability of an airline flyer, chosen at random, being a terrorist on a mission is:

$$50 \div (2{,}000{,}000 \times 365 \times 10) = 0.000000007$$

Now, assume that our screeners are really good and in a few seconds of a conversation they are able to correctly spot 99 percent of actual terrorists that they would ever see. The true positive rate is thus 0.99. Also assume that an unrealistically low rate of normal individuals—say 1 percent—are "nervous fliers" who will flunk their interview even though they are perfectly innocuous, or the screener's preconception causes him to assume they are terrorists based on certain characteristics, and thus he flags them. The false positive rate is thus 0.01. Doing the math, what is the probability that an individual is a terrorist if he is flagged by behavior detection? Using Bayes' Theorem, it would equal:

$$(0.99 \times 0.000000007) \div \{(0.99 \times 0.000000007) + (0.01 \times 0.999999993)\} = 0.00000069$$

So, even with unrealistically good interviewers, if somebody flunks a behavior detection test, he or she has a 1in 1.5 million chance of being an actual terrorist. Playing with these equations tells us that this test adds large amounts of value only when the probability of a non-terrorist flunking the test (being flagged inappropriately) is near to the probability of being a terrorist—that is, extremely low and near zero.

We can infer some more realistic performance estimates of the SPOT test by looking at observed performance of similar procedures. Known to many people through portrayal in film and television, and occasionally used in actual investigative work, the polygraph "lie detector" test measures numerous physiologic parameters including heart and respiratory rate, blood pressure, and skin conductance in order to determine if the subject is being deceptive. These parameters are reviewed during the examination and after data collection. Despite this extensive analysis, the efficacy is undoubtedly low and the accuracy of the test itself remains quite controversial. Many experts feel that the overall accuracy of the test, as implemented, is no better than chance. Generous estimates of testing would give an overall sensitivity and specificity of 80 percent.

A lie detector test requires numerous sensors and takes several hours. If the TSA's short-term behavior detection program has the specificity of a full polygraph, the most favorable assumptions, and if the interviewers correctly catch 100 percent of true terrorists, the chance of a flagged person actually being a terrorist is 1 in 115 million.

The above analysis is extremely generous in its assumptions regarding any actual sensitivity and specificity of the SPOT test, which in truth is most likely not any better than chance. In 2008,

the National Research Council of the National Academy of Sciences noted that "there is not a consensus within the relevant scientific community nor on the committee regarding whether any behavioral surveillance or physiological monitoring techniques are ready for use at all in the counterterrorist context." A Government Accountability Office report assessing the TSA's own notes reaches a similar conclusion:

> According to TSA, anecdotal examples of [interviewer] actions at airports show the value added by SPOT to securing the aviation system. However, because the SPOT program has not been scientifically validated, it cannot be determined if the anecdotal results cited by TSA were better than if passengers had been pulled aside at random, rather than a consequence of being identified for further screening by [interviewers].

Applying Bayes | Police detectives generally understand the concepts behind Bayes' Theorem, even if they do not know the mathematical or quantitative formulation. When looking for a murderer, the first thing police do is narrow down the list of suspects, using intelligence, investigation, and old-fashioned police work. They use common sense and do not waste time on people who can be rapidly and logically excluded. No detective would line up every citizen in the county and give them a polygraph or canned interview—the pre-test probability is so low, even with this involved test, that the "positive" results would be dominated by false positives. This is smart; many innocent people are nervous around the police, and many criminals are cool, smooth talkers. Similarly, "screening" numerous individuals using a lie detector, even among suspects, is a clear folly. The test has imperfect sensitivity: the guilty may "beat the test" and some of the innocent will flunk it. Any "high-risk" results may not contain the murderer after all, and may falsely exonerate the culprit, causing the investigators to focus their attention on the positive group, which now only includes innocent people. Thus, relying to any degree on that imperfect test may significantly decrease the probability of identifying a criminal or terrorist.

Even with unrealistically good interviewers, if somebody flunks a behavior detection test, he or she has a 1 in 1.5 million chance of being an actual terrorist.

Reliability And Performance

After validity, the second quality of a good screening test is reliability. This means that the test is repeatable and largely gives similar results each time it is applied. Given the nature of the interviews described by the TSA, it is difficult to know if all screeners have similar results, as there is little way of "standardizing" the inputs and responses of passengers to the interviews. Reliability is difficult to judge in this situation, but given the subjective nature of the SPOT test, it is not likely to be very high.

The third requirement for a screening test is that it actually shows some performance in the real world, yielding successful results. For a medical example, there are numerous anecdotes of patients who have been saved when their colon cancers are detected early by colonoscopy. When these reports are grouped and systematically analyzed, they give solid data supporting the use of early detection of colonic cancer by invasive colonoscopy—more patients are helped than harmed by this intervention. Large-scale data examination confirms that the test is actually useful in addition to the heavily advertised cases that put a "face" on the outcome. In contrast, the detection of terrorism by interviews or routine deployment of body scanners has not yielded a single terrorist, giving a yield of zero. Of course, there are limitations to this direct comparison—colonic cancers are not "chased away" by colonoscopy, but presumably invasive airport detection routines may have a deterrent effect on terrorism that is difficult to evaluate precisely.

Cost and benefit | The fourth tenet of screening addresses the issue of cost. Beginning students of economics hear of the "broken window fallacy." This is a thought experiment, introduced by an 1850 essay by political theorist Frédéric Bastiat, of an economically stagnant town in which a child carelessly breaks the window of a shopkeeper. The window is subsequently repaired, which gives the local window repairman employment, and he in turn buys paint from the paint distributer and hires laborers to clean up the surrounding damage. One might say that the boy should be commended for stimulating the economy and providing employment for his community!

The core of this fallacy deals with the isolated treatment of employment and economic conditions without regard for the whole society; it does not account for the opportunity costs inherent to spending on a single program. The money and time that the individuals spent repairing this window are resources that they will not have for expanding or investing in other places in town. Similarly, as a country, for every dollar that we ineffectually spend to fight terrorism, we take away a dollar from what might be more effective efforts, as well as domestic programs such as the construction and repair of roads, schools, and infrastructure, funding of education or research, or paying down prior U.S. obligations. Of course, benefit may be partially realized due to the efficiency of a form of ready employment for security employees with low barriers to entry (a prospective TSA agent may otherwise

be unemployed and may be unsuited to service, construction, or other productive work), and as a form of economic stimulus and increased employment this program may be moderately effective. Nevertheless, this is only a partial reduction in the tradeoff that is made by the decision to invest in an otherwise demonstratively ineffective project.

Clearly these costs are not trivial, and dwarf other governmental arenas that may benefit from increased funding. The TSA's allocated budget for fiscal year 2011 is $8.1 billion, increased from the previous year's $7.8 billion. For comparison, the Department of Transportation budget for 2011 to modernize the air traffic controller system from ground radar to satellite/GPS-based location, critical to ensuring continued safety in crowded airspace, was $1.14 billion. For comparison to other areas, the National Endowment for the Arts—a perennial target for spending cuts—has a fiscal year 2011 budget of $154 million, down from 2010's $167.5 million.

Trusting authorities | The fifth and final requirement of a successful test is perhaps the most important. A screening test, whether for a tumor, tuberculosis, or terrorist, should be accepted by informed members of the population before it is widely implemented. Here, as a general rule, the TSA has largely benefited from the public's respect for its work. The vast majority of travelers silently comply with security measures because they trust the system and are obedient to societally sponsored authority.

Doctors receive a similar deference when dealing with medical matters, including when advising on the suitability of any medical test. However, in return for the trust of the public, medical professionals have the obligation to conduct deep analysis of applied tests and to disseminate and apply the resulting knowledge. Physicians try to aggregate their data to make larger decisions that are removed from mere anecdote and strive to provide dispassionate analysis when deciding on public health issues. We "earn" and legitimize the trust given by acting as an agent for those who trust us, and by never hiding any findings or data from public, external scientific, or expert evaluation. In stark contrast, the TSA has not reported performance data regarding any form of enhanced or behavior-based screening. Their most recent report from 2006 of carry-on screening showed a 70 percent failure rate of detecting guns and knives passing through luggage screening, after which the agency ceased public release of any testing data.

There have been innumerable complaints in the media regarding long wait times for TSA screening, feelings of violation because of invasive pat-downs, concern regarding the untested effects of irradiating the whole body with ionizing radiation, and lost productivity during the time that one takes to remove shoes and pass through security. With a lack of understanding of the true probabilities involved in their test, SPOT screeners will likely wildly overestimate the probability that a "positive detection"—somebody acting "nervous" or "shifty"—is an actual threat to an airplane. It remains unclear if, given knowledge of the performance of the test, this would be accepted by passengers who must undergo screening.

Conclusion

It makes sense that "layers of security" would be effective in preventing a terrorist attack, and if the tests are independent, then the probability of detection multiplies. However, these layers must not inconvenience massive amounts of people in order to add negligible security benefit. A metal detector is capable of extremely high specificity, while a SPOT interviewer is not. Intelligence detection, coherently acting on tips and observation of known terrorist organizations behind the scenes, may similarly have good specificity at minimal economic and social cost to the business and pleasure traveler.

Utilizing this construct, we believe that Americans should not tolerate the charade of mini-interviews of all passengers. It would add virtually no additional security to our airports, but it would come at great cost. This is modern-day phrenology, with components of mysticism and mind-reading resulting in an avoidance of rational examination. There is a very real risk of systematic bias from the subconscious transference of the "behavior detectors," repeated persecution of "nervous fliers," and degeneration of detection into simple racism or religious-appearance-based screening.

It is easy to criticize a person or institution as we have done in this paper. It is more difficult to offer remedies or explicit methods to follow for improvement. Fortunately, in the case of airport screening, there are many deficiencies that would be most cost-efficient to remedy. On domestic flights, there is no bag–passenger matching prior to takeoff and only a fraction of luggage is screened by any sort of method, according to current TSA proceedings. It is currently possible to pack a large bomb into a suitcase of a domestic flight, check it to the destination, and leave the airport, and there is a very good chance the luggage will then be directly loaded on a plane with hundreds of people. Many flights mix cargo and passengers, without a 100 percent evaluation rate for explosives using readily available X-ray or CAT scan technologies. The TSA has set multiple internal deadlines for the goal of screening all checked luggage, but all have been missed and the agency reports it is currently not accomplished. This remains a gaping hole for security. Screening all checked luggage is a relatively inexpensive fix. In our research for this article, it became immediately clear that travelers are aware of all of the "increased security initiatives" at airports, but not a single person knew of the screening procedures, or lack thereof, for the luggage sitting 10 feet below them in the plane's pressurized cargo hold.

It is very important to note that in medicine, screening tests are never used once a patient has symptoms. Once a patient presents with a cough, tests should focus on diagnosis leading to treatment, not asymptomatic screening. The risk-to-benefit analysis significantly changes and the "pre-test probability" is assumed to be much higher. Similarly, once the reasonable identification of persons of interest has occurred, appropriate testing

is both warranted and necessary, which may include interviews or enhanced searches. Improved use of intelligence-gathering to find those "symptomatic individuals" may pay extremely high dividends. The "underwear bomber" of 2009, which set off a flurry of reactionary measures with body-scanning device implementation, was brought to U.S. intelligence by reports to the Central Intelligence Agency in Nigeria by the suspect's concerned father, yet he was allowed to fly without specific, targeted examination. The routine screening procedures did not identify anything suspicious. It remains uncertain if current "enhanced" screening procedures would have detected his underwear explosive, given low sensitivity and the continued randomness of implementation of screening measures. The identification and examination of a minute number of high-risk individuals, or detection of organized terrorism, is a relatively low-cost, high-efficiency method for thwarting terrorism.

We can never have perfect security; there are simply too many holes to plug them all. It would be trivial for a determined bomber to hide explosives in body cavities, such as "drug mules" routinely manage. Clearly, routine screening will not identify this method of concealment, and the logical method to detect this will most clearly be unacceptable to all but the smallest fraction of the populace. Once one security hole is plugged, the next "easiest" avenue will be exploited. As an extreme example, a terrorist could purchase a small plane and simply fly it into a line of heavy jets lining up to take off, all fully loaded with fuel and passengers. Is there a way to prevent this? Reasonable vigilance at airports and monitoring of suspicious aviation activity is acceptable, but the surest methods—such as banning private aviation from all airports that serve commercial flights or class B airspace (serving major airports)—will have unacceptable side effects.

We must ask ourselves how many resources we are willing to devote to small probability events, and as a nation we should focus on high-security return for cost expenditure. The identification of terrorist affiliates and their plans, increasing the pre-test probability significantly for a few suspicious individuals, or screening of all checked luggage, is an effective way of addressing these issues. Low-yield, ineffective, and costly measures such as instantaneous mind-reading and detection of deception, or measures that similarly have high societal cost such as the broad, untargeted restriction of private planes or highly invasive routine searches of passengers, must be avoided.

In creating rules, guidelines, and state or governmental entities, we must decide what kind of society we want to inhabit. We will omit clichéd dicta from Benjamin Franklin regarding the inability to attain both liberty and security, in hopes that the dispassionate analysis and reasoning herein will convince well-meaning policymakers and force a re-analysis of methods, rather than rushing to decisions through fear, emotion, or anecdote. Thus, we must screen the tests themselves for efficacy prior to implementation, lest the screening tests terrorize the domestic population and inadvertently accomplish the goals of terrorists. Inappropriate tests waste time and money, and hurt the people they were designed to help. R

Regulation

EDITOR
Peter Van Doren

MANAGING EDITOR
Thomas A. Firey

DESIGN AND LAYOUT
David Herbick Design

ARTISTS
Morgan Ballard, Kevin Tuma

CIRCULATION MANAGER
Alan Peterson

PUBLISHER
John A. Allison IV
President, Cato Institute

Regulation was first published in July 1977 "because the extension of regulation is piecemeal, the sources and targets diverse, the language complex and often opaque, and the volume overwhelming." Regulation is devoted to analyzing the implications of government regulatory policy and its effects on our public and private endeavors.