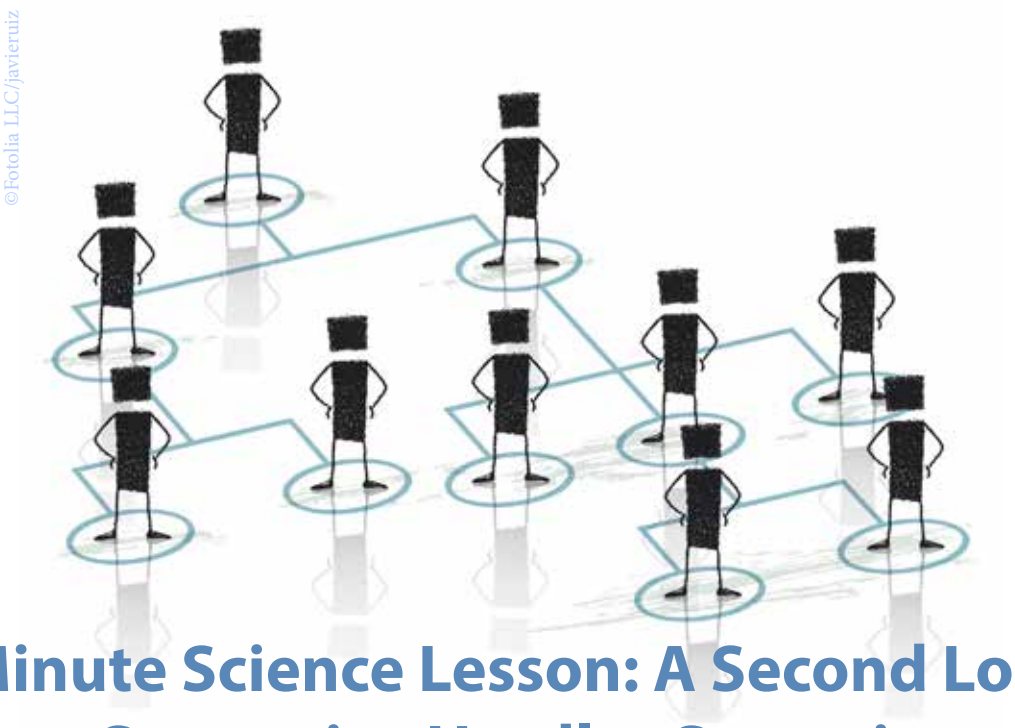
©Fotolia LLC/javieruiz

5 Minute Science Lesson: A Second Look at Successive Hurdles Screening

(Tourists and Troublemakers or the Medical Model)

By Raymond Nelson

A thought experiment: imagine a group of tourists at an airport. Before the tourists can board their airplanes and travel to their final destinations they are required to undergo a security screening procedure – a test – intended to reduce the potential threat of harmful objects or hazardous materials. Then imagine that some of our tourists are actually troublemakers who want to carry their water-bottles through the security screening process and onto the airplanes. (Never-mind the fact that non-dangerous water-bottles may be purchased in most airport terminals; this is simply a contextual example.) In this example, all persons will be classified as either tourist or potential troublemaker by the screening process.

Next imagine that the proportion of troublemakers is 25%, meaning that for each 1000 travelers there are some ordinary tourists and some troublemakers who desire to carry their water-bottles through the security screening process. In practice we never actually know the exact proportion of troublemakers or tourists, but we often have some knowledge from either prior experience or previous studies. This prior knowledge is referred to as the base rate, or incidence rate, and also the prevalence rate, but is more formally referred to as a prior probability. Prior probability refers to our knowledge of the class probability prior to completing the testing process. Remember that a prior probability is a probability, meaning that we do not know the exact proportion of troublemakers and simply use the best evidence-based estimate using the best information, knowledge and experience that is available prior to testing. Our present task is to use the screening test to determine (i.e., predict or classify) the state of each traveler at a rate that is better than that which could be achieved by random guessing or by guessing the class with the largest prior probability (sometime referred to as guessing the base rate).

In practice, we do not know the exact proportions of troublemakers and tourists traveler

screening tests involve subjects for whom we do not know the exact criterions state or class of each person, and so we often consider a range or distribution of several possible prior probabilities. We also do not know the exact criterion state or class of each participant in the screening process. If we knew this, we would not need a screening test. Instead, Bayesian statistical procedures have been developed to help us to refine or improve both our imperfect or uncertain knowledge about the prior probability distributions for each criterion class. Bayesian methods can also be used to combine our prior knowledge with the data or evidence from a test or experiment and improve the proportion of correct decisions/prediction/classification compared to what we could achieve by chance alone or by simply guessing the prior probability (i.e., classify all persons in largest criterion category. In this example 75% of the travelers are tourists and so simply guessing "tourist" for each person would result in greater classification accuracy than random guessing).

Finally, lets imagine that the airport security screening test – primarily a visual analysis task that requires a human observer using imaging technology – may have an error rate that can vary with target prevalence, and is also subject to phenomena involving human attention and cognition, but may converge to something roughly near 15% (Biggs, Adamo, Dowd & Mitroff, 2015; Biggs, Cain, Clark, Darling, & Mitroff, 2013; Biggs & Mitroff, 2013; Biggs & Mitroff, 2014; Wolfe, Brunelli, Rubinstein & Horowitz, 2013; Wolfe, Horowitz & Kenner, 2005; Wolfe, et al., 2007). For the purpose of this example we will make an additional convenience assumption that both test sensitivity and test specificity are 85% and that the false-positive and false-negative rates are both 15%.

Initial screening test

Total time for each person in the screening process can take several minutes, but the screening task itself takes an average of 20 to 30 seconds at most. This means that 2 or more persons can be screened per minute, and 2 persons per minute * 60 minutes * 8 hours = 960 persons can be screened during an eight-hour work shift. For convenience, can round this number upwards to 1000 screenings per day per screening station. With three screening stations it is not difficult to imagine screening 3000 persons per day or over 1,000,000 persons annually. Table 1 shows a 2x2 table of frequencies and conditional probabilities that can be expected to result from the initial screening process with 1000 travelers for which there are 750 ordinary tourists (75%) and 250 troublemakers (25%).

Table 1. Table summary of posterior probabilities for initial screening results, assuming screening accuracy of 85% with a prior probability of 25% for a sample of N=1000.

	Troublemakers	Non-troublemakers	Totals	Correct classifications	Posterior Probabilities
Positive	**250*.85=212 (TP)**	**750*.15=113 (FP)**	*325*	*212*	*.652 PPV*
Negative	**250*.15=38 (FN)**	**750*.85=637 (TN)**	*675*	*637*	*.944 NPV*
-			1000	849	*.849*

Among the group of 250 troublemakers, 212 have been correctly identified (true positive or TP), though 38 have been incorrectly classified as ordinary tourists (false negative or FN) and may proceed through security and onto the aircraft unless there is another layer of security that can identify or deter them. Of the 750 tourists 637 have been correctly identified (true negative or TN). However, 113 tourists have been misclassified as possible troublemakers (false positive or FP), and may not be permitted to proceed to their final destination

unless there is some way of reducing the potential for FP error.

We can calculate the ratio of TP screening results to all positive results (TP + FP) in this manner: TP/(TP+FP) = 212 / (212+113) = 212 / 325 = .652, meaning that 65.2% of positive results are correct, referred to as the positive predictive value (PPV) after the first screening procedure. Complimentary to the PPV is the false positive index (FPI) which is calculated as the ratio FP results to all positive result in this manner: FP / (FP + TP) = 113 / (113+212) = 113 / 325 = .348, meaning that 34.8% of all positive results are incorrect after the initial screening test. We can also calculate the ratio of TN screening results to all negative results (TN + FN) in the same manner: TN / (TN+FN) = 637 / (637+38) = 637 / 675 = .944, meaning that 94.4% of the negative results are correct. This is referred to as the negative predictive value (NPV). Complimentary to the NPV is the false negative index (FNI) which can be calculated in this way: FN / (FN + TN) = 38 / (38+637) = 38 / 675 = .056, meaning that 5.6% of negative results are incorrect after the initial screening test. Overall classification accuracy is effectively 85%, as expected.

Immediately we can observe that very desirable objective is achieved in the reduction of troublemakers from the initial 250 (25% of the total population of travelers) to 38 (3.8%). However, we can also observe a well-known phenomenon in that whenever the prior probability is low the FPI will be higher. Table 2 shows the 2x2 frequencies and proportions when the prior probability is 50%.

Table 2. Table summary of posterior probabilities for initial screening results, assuming screening accuracy of 85% with a prior probability of 50% for a sample of N=1000.

	Troublemakers	Non-troublemakers	Totals	Correct classifications	Posterior Probabilities
Positive	**500*.85=425 (TP)**	**500*.15=75 (FP)**	*500*	*425*	*.85 PPV*
Negative	**500*.15=75 (FN)**	**500*.85=425 (TN)**	*500*	*425*	*.85 NPV*
-	-	-	1000	850	*.85*

When the prior information suggests that the proportion of troublemakers is 50%, we can expect that among the 1000 travelers that are screened daily at each station in this imaginary example there are 500 troublemakers. When prior information is not available, or when prior information is of such low quality that it is uninformative, it is common to simply assume that the prior probability is 50%.

> In Bayesian terms this is sometimes referred to as an uninformative prior or weak prior because prior information is uninformative and provides only weak information about how best to guess whether any particular traveler is a tourist or troublemaker. In contrast, when there is strong information to suggests that one class probability substantially exceeds the other – when the prior probability is either high or low – in which case it also indicates how best to guess the class probability of any individual with the greatest likelihood of success, it is referred to as a strong prior.

Among the expected 500 troublemakers under the uninformative prior probability, 425 TP results will be occur, along with 75 FNs, sometimes referred to as false-misses. Among the 500 ordinary tourists we expect 425 TNs along with 75 FPs, sometimes referred to as false-hits. We can calculate the PPV in the same way as before: PPV = TP / (TP + FP) = 425 / (425+75) = 425 / 500 = .85 or 85%. Similarly, FPI = FP / (FP + TP) = 75 / (75+425) = 75 / 500 = .15 or 15%. The NPV is also calculated as before: NPV = TN / (TN + FN) = 425 / (425+75) = 425 / 500 = .85 or 85%. Finally, the FNI = TN / (TN + FN) = 75 / (75+425) = 75 / 500 = .15 or 15%.

In simple numerical terms, a 15% false negative errors among 1000 travelers will mean that 75 water-bottle carrying troublemakers may succeed at passing through the initial screening process and onto the aircraft unless there are additional layers of security to identify or deter them. Although imperfect, the reduction to 75 (7.5%) from the initial 500 (50%) troublemakers is substantial and important.

A concern to many will be the false positive error rate and the potential that an imperfect screening procedure may interfere with the goals and plans of some individuals. These concern may result in some persons actually questioning the value of a screening process. Although it would be very reckless and potentially dangerous to abandon the screening process altogether, it is sometimes worth considering some form of additional screening for those persons that do not initially pass the screening process.

Subsequent testing

For the purpose of this teaching example, we will consider the use of an addition testing process to further investigate the state of persons who do not pass an initial screening test. At the second round only 500 of the initial group of 1000 travelers will be tested, including 425 TP results and 75 FP results. We can now use our knowledge of the posterior probabilities from the first screening as a basis of information to estimate the prior probability that a person subject to additional screening is actually a troublemaker. Although we have no knowledge of exactly which cases are TP or FP, we now have a strong prior information to suggest that the majority of these persons are troublemakers not tourists. When evaluating any particular individual without any additional information we are forced to recognize that the most likely possibility is that an individual is a troublemaker. In Bayesian terms, the purpose of a test is to develop additional information so that we can update the prior probability into a more precise posterior probability. Table 3 shows the expected posterior results from an additional screening test.

Table 3. Table summary of posterior probabilities for initial screening results, assuming screening accuracy of 85% using the initial screening results as the prior probability estimate.

	Troublemakers	Non-troublemakers	Totals	Correct classifications	Posterior Probabilities
Positive	**425*.85=361 (TP)**	**75*.15=12 (FP)**	*373*	*361*	*.968 PPV*
Negative	**425*.15=64 (FN)**	**75*.85=63 (TN)**	*127*	*63*	*.496 NPV*
-	-	-	500	424	*.848*

For the 425 troublemakers we can expect to observe 361 TP results and 64 FN errors. Among the 75 ordinary tourists who are subject to additional testing, we can expect 63 TN results along with 12 FP errors. Additional calculations show that: PPV = TP / (TP + FP) = 361 / (631+12) = 361 / 373 = .968, meaning that 96.8% of positive results are correct when the initial and subsequent test results concur. Similarly, FPI = FP / (FP + TP) = 12 / (12+361) = 12 / 373 = .032, meaning that 3.2% of positive result are incorrect when the two test results are concordant. The NPV is also calculated as before: NPV = TN / (TN + FN) = 63 / (63+64) = 63 / 127 = .496, meaning that only 49.6% of negative test results are expected to be correct following an initial positive test result. In a similar way: FNI = TN / (TN + FN) = 64 / (64+63) = 64 / 127 = .504 or 50.4% of negative test results are incorrect if the negative results are following an initial positive test result (assuming that the second test is completely independent and in no way influenced by the results of the first test). The overall precision of the test has remained at or near the expected 85% level.

Herein exists an important practical aspect of successive hurdles testing. Very high accuracy can be inferred for positive test results when two test results are concordant (again, assuming that the tests are conducted

independently so that the results of the first test do not in any way influence the results of the second test). However, when the results of two examinations do not concur, such as when negative results follow an initial positive result, acceptance of the negative test result can result in an undesirable increase in the rate of FN errors. In this example the FN rate following the first screening test was 75 (7.5%). Following the second test the FN rate was increased by 64 cases to 139 (13.9%). This is a substantial increase in FN errors if negative test results are accepted following an initial positive test result. This same phenomenon can also be observed under other prior probabilities. Table 4 shows the posterior results using a prior probability distribution that was based on the posterior information from Table 1, which began with a prior probability 25%.

Table 4. Table summary of posterior probabilities for initial screening results, assuming screening accuracy of 85% using the initial screening results as the prior probability estimate.

	Troublemakers	Non-troublemakers	Totals	Correct classifications	Posterior Probabilities
Positive	**212*.85=180 (TP)**	**113*.15=17 (FP)**	*197*	*180*	*.914 PPV*
Negative	**212*.15=32 (FN)**	**113*.85=96 (TN)**	*128*	*96*	*.75 NPV*
-	-	-	325	276	*.85*

In Table 4, 96 of the 113 tourists who were subject to additional testing are shown as correctly classified as tourists. However, 32 troublemakers are also classified as ordinary tourists (in addition to the 38 FN cases from the first test). Of the 212 troublemakers subject to additional testing, 180 are correctly identified. Additionally, the FP rate has been reduced to 17 or 2.3% of the 750 tourists who were present for screening at the first exam. The PPV shown in Table 4 is 91.4%, with a corresponding FPI of 8.6%. NPV shown in Table 4 is 75% with a corresponding FNI of 25%. Overall with all cases remains at the expected 85% level.

> When the prior incidence rate is high (e.g., above 50%) then simply guessing that every case is positive will result in correctly classifying more than 50% of the cases. When the incidence rate is low we could still effectively identify every positive case by simply classifying every case as positive, but the cost for this approach will be a high rate of false-positive errors. If we wish to correctly identify positive cases and also discriminate them from negative cases, then we will need a testing procedure with test sensitivity and test specificity that both exceed what can be achieved by either random chance or by simply classifying all cases in the single category with the greatest incidence rate.

Discussion

This example serves to illustrate that screening accuracy of 85%, though well below perfection, is substantial and useful enough to contribute in strategic and practical ways to a meaningful decrease in the probability that a troublemaker will succeed at getting through the security screening process. When used strategically this level of precision is also sufficient to ensure that ordinary tourists can proceed to their destination with increased safety with a very small probability that they will be incorrectly regarded as troublemakers. However, a single screening procedure with precision as demonstrated in this example may not be sufficient as a basis of information to cancel or delay the travel plans of an individual. Instead, it will be important to engage in additional activity to increase the level of confidence in our knowledge and decisions. One method of increasing the quality and confidence of our knowledge and decisions will be to use a screening process with multiple stages.

The practice of requiring multiple screening test results, sometimes referred to as successive hurdles or multiple hurdles in the polygraph profession, can result in a reduction of FP errors when the results of the

two independent tests concur. It is important to point out that conducting multiple examinations may not increase decision accuracy when the results of one examination are permitted to influence the conduct or subjective interpretation of the results a subsequent examination; when this occurs the second examination results are merely based, in whole or part, on the data and results from first exam. Bayesian methods are a structured analytic process for combining knowledge from one examination with data from another exam. Another interesting aspect of multiple hurdles screening practices is that classification errors can increase when two test results do not concur.

> In other contexts, the term successive hurdles model has been used to describe employee selection process in which multiple favorable outcomes are required before proceeding with a decision to select an employee. This would be analogous to completing additional screening tests only on persons whose initial test results are negative (i.e., passing), and, at each testing stage, retaining only those persons who continue to produce favorable results. The result of this strategy will be to reduce FN errors, with a corresponding increase in the rejection of persons who may appear to be suitable employment candidates based sources of information other than the test results.

To be useful a test must provide with adequate specificity – the ability to rule out a specific problem – in addition to to adequate sensitivity to the issue of concern. A test with inadequate sensitivity and/or specificity will perform no better, and possibly worse, than simply guessing the base rate.

Perhaps the most obvious way to increase the level of confidence in our knowledge and decisions about the tourist or troublemaker state of a traveler will be to physically inspect the baggage, property and person before classifying a person as tourist or troublemaker. Doing so will provide a deterministic (i.e., effectively perfect, and immune to random error or human behavior) confirmation of the presence or absence of a water-bottle in the possession of each traveler. Each traveler either does or does not possess a water-bottle. A deterministic observation of physical and factual information is always desirable and should be used whenever it is practicable. However, there are times when deterministic inspection on every person is either impossible or impracticable.

Physical inspection and deterministic investigation of each traveler will necessarily increase the time, expense, level of intrusiveness for each person, and may also have an effect on travel delays. It is reasonable to use the results of an initial screening test to reduce the number of travelers that must be subject to physical or deterministic inspection, though it can be expected that physical investigation will confirm both TP and FP results from the screening test. Use of multiple testing phases can be used to optimize the ratio or proportion of TP and FP confirmations that can be expected from deterministic inspections that are often more expensive in terms terms of time, travel delays, individualized professional attention and other economic factors. It is also possible to proceed directly to deterministic investigation activities following a single screening test.

Differences between medical screening and security screening

The medical model of screening is one in which members of a population group are subject to a screening test with the goal of identifying positive cases that may otherwise remain unnoticed for a longer period of time. A related goal of medical screening is to optimize the use of resources and minimize the impact on individuals by not imposing medical treatment on persons who do not need it. Multiple hurdles testing strategies can be very useful towards reducing false positive errors, incorrect diagnoses, and unnecessary medical treatment. In other words, multiple hurdles screening in the medical model is intended to reduce the risks associated with FP errors.

Herein exists an important difference between the medical and security risk-management screening contexts. Security screening is sometimes primarily concerned with reducing risks associated with FN errors. (This is in contrast to the medical diagnostic situation, with a symptomatic patient, in which FN errors may present greater risks than FP errors.) Although classification errors and associated risks can be reduced when classifications are based on concordant results from multiple properly conducted tests, problem may arise when the results of successive screening tests do not concur. This is because there may be an increased risk for FN error if negative results are accepted as a basis to proceed following a previously positive test result. A similar risk for increased classification error would exist if negative results were subject to additional testing. We can expect a reduction of FN errors when the two results concur, along with a greater risk for FP error when the second result differs from the first. Of course, this phenomenon will have no practical effect when negative screening results are of no interest – such as in the medical context.

The difference between security or risk-management screening and medical screening is that negative results in the medical screening context require no action and are regarded as not indicative of any increased risk level, whereas negative results in the security screening and risk management context are a basis on which to proceed with a course of action for which there is inherent risk. As shown in Tables 3 and 4, a subsequent screening test following an initial positive result can result in a reduction of screening sensitivity from .85 to .72 under both the uninformative prior and strong prior conditions. In more practical terms this can be thought of as giving the group of troublemakers, who did not pass a screening test, another chance to benefit from a potential testing error.

Take home points and recommendations

- Screening tests conducted when the prior probability of guilt is low can result in a higher proportion of FP errors. Similarly, testing when the prior probability of guilt is high can result in a higher proportion of FN errors. This is as compared to when the base rates are more even.

- Using Bayesian methods, initial test results can be used as a basis of prior information and prior probabilities when analyzing and interpreting the results of a subsequent examination, though it is important that subsequent examinations are conducted in an unbiased manner according to standardized procedures, so that the influence of the first test result on the second test result is non-subjective and limited to Bayesian analysis.

- Probability estimates for testing error is reduced for both FP and FN errors when two test results concur.

- Both medical and security screening activities are intended to identify possible positive cases that can be investigated further before reaching a conclusion. However, screening strategies in the medical context is intended to reduce costs and impacts associated with false positive errors that would result in un-needed medical intervention whereas security and risk-management screening are often primarily concerned with the costs and impacts associated with false negative errors.

- When two results do not concur, a negative classification based on the second test result are associated with decreased FP rate and increased FN rates. Similarly, positive classifications using the second test result when two test results do not concur are associated with a decreased FN rate and increased FP rate.

- Successive hurdles screening strategies may or may not be an ideal solution for all circumstances.

There may be situations in which a single screening test is adequate or desirable. It depends on the end-user's testing desires.

- Probabilistic test results (which is all test results) may or may not be a satisfactory basis of information for all types of decisions. There may be some decisions for which the basis of information should include all available sources including test results, historical, collateral, and physical information. Probabilistic test results may be best used as a basis of information for decisions about where and when to proceed with the investigation and development of more precise physical or deterministic evidence.

Finally, it is important to remind that this example is metaphorical and mathematical, and does not include other human factors that can affect test decision and performance outcomes. The importance of other complex human variables cannot be ignored. Human analysis and human interpretation can introduce vulnerability to a variety of known phenomena including sensitization, habituation, and the tendency to make subjective heuristic adjustments to minimize the kinds of errors are perceived as costliest. Differences can result when the kinds of errors that an individual professional wishes to avoid – are different from the kinds of errors that an agency or community wishes to avoid. While there is no such thing as a perfect test that can provide deterministic perfection with no potential for error, testing error should ideally not be highly vulnerable to subjective decision making processes that may affect how an examination is conducted or interpreted. The importance of structured, and automated whenever possible, testing administration and test data analysis procedures cannot be understated. Similarly, the importance of conscientious human deliberation should never be overlooked when making decisions that can affect the future of other persons.

References

Biggs, A. T., Adamo, S. H., Dowd, E. W., & Mitroff, S. R. (2015) Examining perceptual and conceptual set biases in multiple-target visual search. Attention, Perception, & Psychophysics, 77(3), 844-855.

Biggs, A. T., Cain, M. S., Clark, K., Darling, E. F., & Mitroff, S. R. (2013). Assessing visual search performance differences between Transportation Security Administration Officers and nonprofessional searchers. Visual Cognition, 21, 330-352.

Biggs, A. T., & Mitroff, S. R. (2013). Different predictors of multiple-target search accuracy between non-professional and professional visual searchers. Quarterly Journal of Experimental Psychology, 67, 1335-1348.

Biggs, A. T., & Mitroff, S. R. (2014). Differences in multiple-target visual search performance between non-professional and professional searchers due to decision-making criteria. British Journal of Psychology. Advance online publication.

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. Journal of Vision, 13(3):33, 1–9.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. Nature, 435, 439-440.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. Journal of Experimental Psychology, 136(4), 623-638.