# The Threat Detection System THAT CRIED WOLF:

## Reconciling Developers with Operators

*Shelley M. Cazares*

The Department of Defense and Department of Homeland Security use many threat detection systems, such as air cargo screeners and counter-improvised-explosive-device systems. Threat detection systems that perform well during testing are not always well received by the system operators, however. Some systems may frequently "cry wolf," generating false alarms when true threats are not present. As a result, operators lose faith in the systems—ignoring them or even turning them off and taking the chance that a true threat will not appear. This article reviews statistical concepts to reconcile the performance metrics that summarize a developer's view of a system during testing with the metrics that describe an operator's view of the system during real-world missions. Program managers can still make use of systems that "cry wolf" by arranging them into a tiered system that, overall, exhibits better performance than each individual system alone.

**DOI:** *http://dx.doi.org/10.22594/dau.16-749.24.01*
**Keywords:** *probability of detection, probability of false alarm, positive predictive value, negative predictive value, prevalence*

*Image designed by Diane Fleischer*


350
0
10
20
30
340
330
320
40
50
60
70
80

The Department of Defense (DoD) and Department of Homeland Security (DHS) operate many threat detection systems. Examples include counter-mine and counter-improvised-explosive-device (IED) systems and airplane cargo screening systems (Daniels, 2006; L3 Communications Cyterra, 2012; L3 Communications, Security & Detection Systems, 2011, 2013, 2014; Niitek, n.d.; Transportation Security Administration, 2013; U.S. Army, n.d.; Wilson, Gader, Lee, Frigui, & Ho, 2007). All of these systems share a common purpose: to detect threats among clutter.

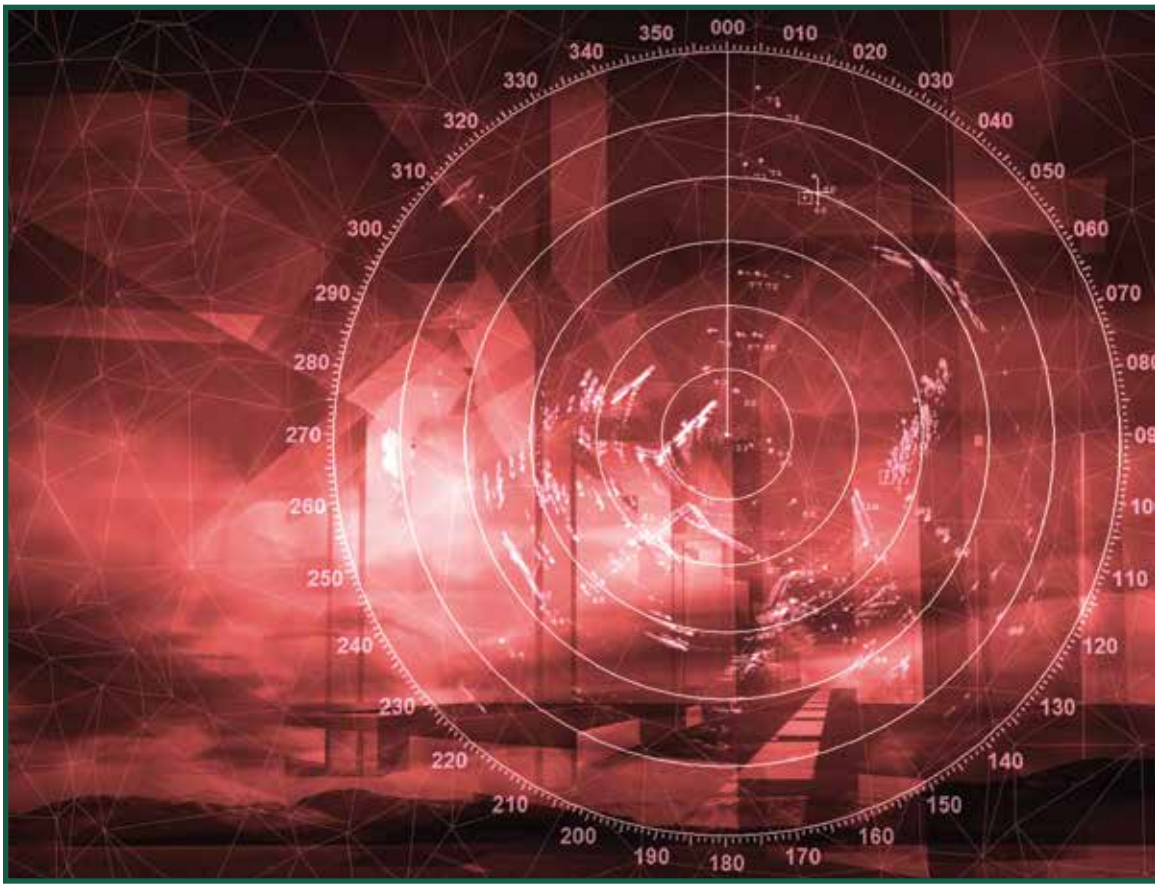
Threat detection systems are often assessed based on their Probability of Detection ($P_d$) and Probability of False Alarm ($P_{fa}$). $P_d$ describes the fraction of true threats for which the system correctly declares an alarm. Conversely, $P_{fa}$ describes the fraction of true clutter (true non-threats) for which the system *in*correctly declares an alarm—a false alarm. A perfect system will exhibit a $P_d$ of 1 and a $P_{fa}$ of 0. $P_d$ and $P_{fa}$ are summarized in Table 1 and discussed in Urkowitz (1967).

**TABLE 1. DEFINITIONS OF COMMON METRICS USED TO ASSESS PERFORMANCE OF THREAT DETECTION SYSTEMS**

| Metric | Definition | Perspective |
|---|---|---|
| Probability of Detection ($P_d$) | The fraction of all items containing a true threat for which the system correctly declared an alarm | Developer |
| Probability of False Alarm ($P_{fa}$) | The fraction of all items *not* containing a true threat for which the system *in*correctly declared an alarm | Developer |
| Positive Predictive Value (PPV) | The fraction of all items causing an alarm that did end up containing a true threat | Operator |
| Negative Predictive Value (NPV) | The fraction of all items *not* causing an alarm that did end up *not* containing a true threat | Operator |
| Prevalence (Prev) | The fraction of items that contained a true threat (regardless of whether the system declared an alarm) | — |
| False Alarm Rate (FAR) | The number of false alarms per unit time, area, or distance | — |

Threat detection systems with good $P_d$ and $P_{fa}$ performance metrics are not always well received by the system's operators, however. Some systems may frequently "cry wolf," generating false alarms when true threats are not present. As a result, operators may lose faith in the systems, delaying their response to alarms (Getty, Swets, Pickett, & Gonthier, 1995) or ignoring

them altogether (Bliss, Gilson, & Deaton, 1995), potentially leading to disastrous consequences. This issue has arisen in military, national security, and civilian scenarios.



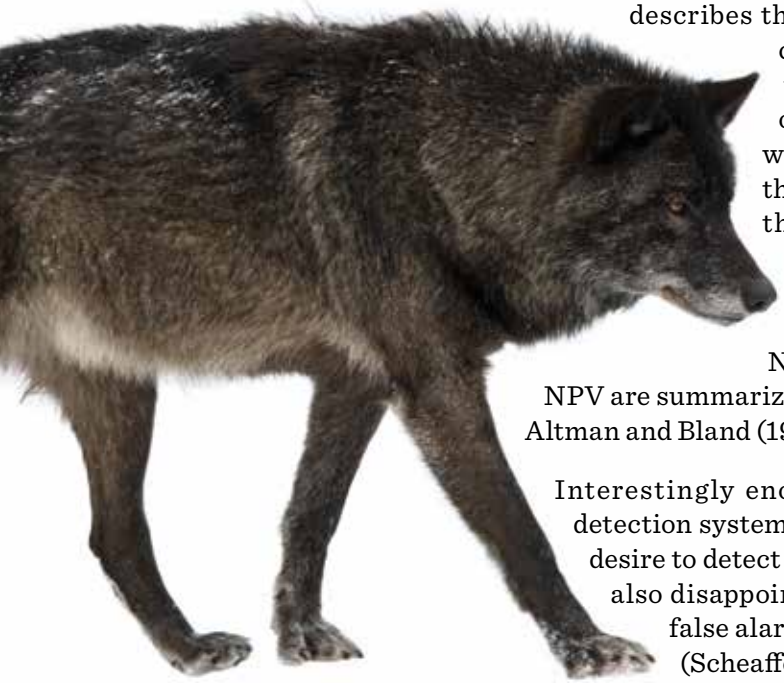

*The New York Times* described a 1987 military incident involving the threat detection system installed on a $300 million high-tech warship to track radar signals in the waters and airspace off Bahrain. Unfortunately, "somebody had turned off the audible alarm because its frequent beeps bothered him" (Cushman, 1987, p. 1). The radar operator was looking away when the system flashed a sign alerting the presence of an incoming Iraqi jet. The attack killed 37 sailors.

That same year, *The New York Times* reported a similar civilian incident in the United States. An Amtrak train collided near Baltimore, Maryland, killing 15 people and injuring 176. Investigators found that an alarm whistle

in the locomotive cab had been "substantially disabled by wrapping it with tape" and "train crew members sometimes muffle the warning whistle because the sound is annoying" (Stuart, 1987, p. 1).

Such incidents continued to occur two decades later. In 2006, *The Los Angeles Times* described an incident in which a radar air traffic control system at Los Angeles International Airport (LAX) issued a false alarm, prompting the human controllers to "turn off the equipment's aural alert" (Oldham, 2006, p. 2). Two days later, a turboprop plane taking off from the airport narrowly missed a regional jet, the "closest call on the ground at LAX" in 2 years (Oldham, 2006, p. 2). This incident had homeland security implications, since DHS and the Department of Transportation are co-sector-specific agencies for the Transportation Systems Sector, which governs air traffic control (DHS, 2016).

The disabling of threat detection systems due to false alarms is troubling. This behavior often arises from an inappropriate choice of metrics used to assess the system's performance during testing. While $P_d$ and $P_{fa}$ encapsulate the *developer's* perspective of the system's performance, these metrics do not encapsulate the *operator's* perspective. The operator's view can be better summarized with other metrics, namely Positive Predictive Value (PPV) and Negative Predictive Value (NPV). PPV describes the fraction of all alarms that correctly turn out to be true threats—a measure of how often the system does not "cry wolf." Similarly, NPV describes the fraction of all *lack* of alarms that correctly turn out to be true clutter. From the operator's perspective, a perfect system will have PPV and NPV values equal to 1. PPV and NPV are summarized in Table 1 and discussed in Altman and Bland (1994b).



Interestingly enough, the very same threat detection system that satisfies the developer's desire to detect as much truth as possible can also disappoint the operator by generating false alarms, or "crying wolf," too often (Scheaffer & McClave, 1995). A system

can exhibit excellent $P_d$ and $P_{fa}$ values while also exhibiting a poor PPV value. Unfortunately, low PPV values naturally occur when the Prevalence (Prev) of true threat among true clutter is extremely low (Parasuraman, 1997; Scheaffer & McClave, 1995), as is often the case in defense and homeland security scenarios. As summarized in Table 1, Prev is a measure of how widespread or common the true threat is. A Prev of 1 indicates a true threat is always present, while a Prev of 0 indicates a true threat is never present. As will be shown, a low Prev can lead to a discrepancy in how developers and operators view the performance of threat detection systems in the DoD and DHS.

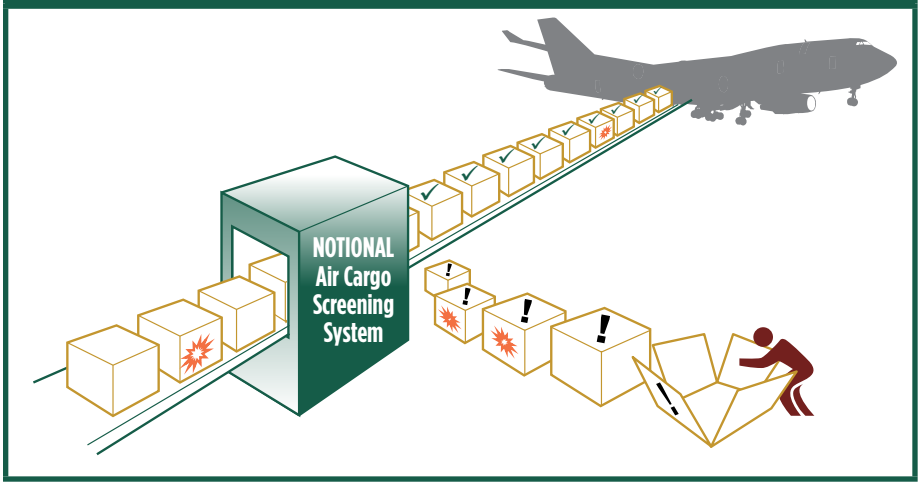
In this article, the author reconciles the performance metrics used to quantify the developer's versus operator's views of threat detection systems. Although these concepts are already well known within the statistics and human factors communities, they are not often immediately understood in the DoD and DHS science and technology (S&T) acquisition communities. This review is intended for program managers (PM) of threat detection systems in the DoD and DHS. This article demonstrates how to calculate $P_d$, $P_{fa}$, PPV, and NPV using a notional air cargo screening system as an example. Then it illustrates how a PM can still make use of a system that frequently "cries wolf" by incorporating it into a tiered system that, overall, exhibits better performance than each individual system alone. Finally, the author cautions that $P_{fa}$ and NPV can be calculated only for threat *classification* systems, rather than genuine threat *detection* systems. False Alarm Rate is often calculated in place of $P_{fa}$.

## Testing a Threat Detection System

A notional air cargo screening system illustrates the discussion of performance metrics for threat detection systems. As illustrated by Figure 1, the purpose of this notional system is to detect explosive threats packed inside items that are about to be loaded into the cargo hold of an airplane. To determine how well this system meets capability requirements, its performance must be quantified. A large number of items is input into the system, and each item's ground truth (whether the item contained a true threat) is compared to the system's output (whether the system declared an alarm). The items are representative of the items that the system would likely encounter in an operational setting. At the end of the test, the True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) items are counted. Figure 2 tallies these counts in a 2 × 2 confusion matrix:

- A TP is an item that contained a true threat, and for which the system correctly declared an alarm.
- An FP is an item that did *not* contain a true threat, but for which the system *in*correctly declared an alarm—a false alarm (a Type I error).
- An FN is an item that contained a true threat, but for which the system *in*correctly did *not* declare an alarm (a Type II error).
- A TN is an item that did *not* contain a true threat, and for which the system correctly did *not* declare an alarm.

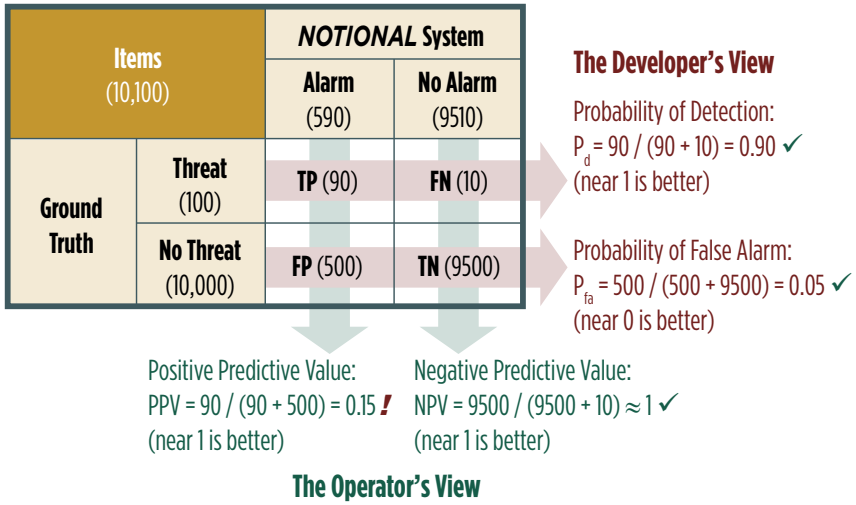
**FIGURE 1. NOTIONAL AIR CARGO SCREENING SYSTEM**




*Note.* A set of predefined, discrete items (small brown boxes) are presented to the system one at a time. Some items contain a true threat (orange star) among clutter, while other items contain clutter only (no orange star). For each item, the system declares either one or zero alarms. All items for which the system declares an alarm (black exclamation point) are further examined manually by trained personnel (red figure). In contrast, all items for which the system does not declare an alarm (green checkmark) are left unexamined and loaded directly onto the airplane.

As shown in Figure 2, a total of 10,100 items passed through the notional air cargo screening system. One hundred items contained a true threat while 10,000 items did not. The system declared an alarm for 590 items and did not declare an alarm for 9,510 items. Comparing the items' ground truth to the system's alarms (or lack thereof), there were 90 TPs, 10 FNs, 500 FPs, and 9,500 TNs.

**FIGURE 2. 2 X 2 CONFUSION MATRIX OF NOTIONAL AIR CARGO SCREENING SYSTEM**




*Note.* The matrix tabulates the number of TP, FN, FP, and TN items processed by the system. $P_d$ and $P_{fa}$ summarize the developer's view of the system's performance while PPV and NPV summarize the operator's view. In this notional example, the low PPV of 0.15 indicates a poor operator experience (the system often generates false alarms and "cries wolf," since only 15% of alarms turn out to be true threats) even though the good $P_d$ and $P_{fa}$ are well received by developers.

## The Developer's View: $P_d$ and $P_{fa}$

A PM must consider how much of the truth the threat detection system is able to identify. This can be done by considering the following questions: Of those items that contain a true threat, for what fraction does the system correctly declare an alarm? And of those items that do *not* contain a true threat, for what fraction does the system *in*correctly declare an alarm—a false alarm? These questions often guide developers during the research and development phase of a threat detection system.

$P_d$ and $P_{fa}$ can be easily calculated from the 2 × 2 confusion matrix to answer these questions. From a developer's perspective, this notional air cargo screening system exhibits good[1] performance:

$$P_d = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.90 \text{ (compared to 1 for a perfect system)} \quad (1)$$

$$P_{fa} = \frac{FP}{FP + TN} = \frac{500}{500 + 9{,}500} = 0.05 \text{ (compared to 0 for a perfect system)} \quad (2)$$

Equation 1 shows that, of all items that contained a true threat (TP + FN = 90 + 10 = 100), a large subset (TP = 90) correctly caused an alarm. These counts resulted in $P_d = 0.90$, close to the value of 1 that would be exhibited by a perfect system.[2] Based on this $P_d$ value, the PM can conclude that 90% of items that contained a true threat correctly caused an alarm, which may (or may not) be considered acceptable within the capability requirements for the system. Furthermore, Equation 2 shows that, of all items that did *not* contain a true threat (FP + TN = 500 + 9,500 = 10,000), only a small subset (FP = 500) caused a false alarm. These counts led to $P_{fa} = 0.05$, close to the value of 0 that would be exhibited by a perfect system.[3] In other words, only 5% of items that did *not* contain a true threat caused a false alarm.

## The Operator's View: PPV and NPV

The PM must also anticipate the operator's view of the threat detection system. One way to do this is to answer the following questions: Of those items that caused an alarm, what fraction turned out to contain a true threat (i.e., what fraction of alarms turned out *not* to be false)? And of those items that did *not* cause an alarm, what fraction turned out *not* to contain a true threat? On the surface, these questions seem similar to those posed previously for $P_d$ and $P_{fa}$. Upon closer examination, however, they are quite different. While $P_d$ and $P_{fa}$ summarize how much of the truth causes an alarm, PPV and NPV summarize how many alarms turn out to be true.

PPV and NPV can also be easily calculated from the 2 × 2 confusion matrix. From an operator's perspective, the notional air cargo screening system exhibits a conflicting performance:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} = \frac{9{,}500}{9{,}500 + 10} \approx 1 \text{ (compared to 1 for a perfect system)} \quad (3)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{90}{90 + 500} = 0.15 \text{ (compared to 1 for a perfect system)} \quad (4)$$

Equation 3 shows that, of all items that did *not* cause an alarm (TN + FN = 9,500 + 10 = 9,510), a very large subset (TN = 9,500) correctly turned out to *not* contain a true threat. These counts resulted in NPV ≈ 1, approximately equal to the 1 value that would be exhibited by a perfect system.[4] In the absence of an alarm, the operator could rest assured that a threat was highly unlikely. However, Equation 4 shows that, of all items that did indeed cause an alarm (TP + FP = 90 + 500 = 590), only a small subset (TP = 90) turned out to contain a true threat (i.e., were not false alarms). These counts unfortunately led to PPV = 0.15, much lower than the 1 value that would be

exhibited by a perfect system.[5] When an alarm was declared, the operator could not trust that a threat was present, since the system generated false alarms so often.

## Reconciling Developers with Operators: $P_d$ and $P_{fa}$ Versus PPV and NPV

The discrepancy between PPV and NPV versus $P_d$ and $P_{fa}$ reflects the discrepancy between the operator's and developer's views of the threat detection system. Developers are often primarily interested in how much of the truth correctly cause alarms—concepts quantified by $P_d$ and $P_{fa}$. In contrast, operators are often primarily concerned with how many alarms turn out to be true—concepts quantified by PPV and NPV. As shown in Figure 2, the very same system that exhibits good values for $P_d$, $P_{fa}$, and NPV can also exhibit poor values for PPV.

Poor PPV values should not be unexpected for threat detection systems in the DoD and DHS. Such performance is often merely a reflection of the low Prev of true threats among true clutter that is not uncommon in defense and homeland security scenarios.[6] Prev describes the fraction of all items that contain a true threat, including those that did and did not cause an alarm. In the case of the notional air cargo screening system, Prev is very low:

$$\text{Prev} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} = \frac{90 + 10}{90 + 10 + 500 + 9{,}500} = 0.01 \quad (5)$$

Equation 5 shows that, of all items (TP + FN + FP + TN = 90 + 10 + 500 + 9,500 = 10,100), only a very small subset (TP + FN = 90 + 10 = 100) contained a true threat, leading to Prev = 0.01. When true threats are rare, most alarms turn out to be false, even for an otherwise strong threat detection system, leading to a low value for PPV (Altman & Bland, 1994b). In fact, to achieve a high value of PPV when Prev is extremely low, a threat detection system must exhibit so few FPs (false alarms) as to make $P_{fa}$ approximately zero.

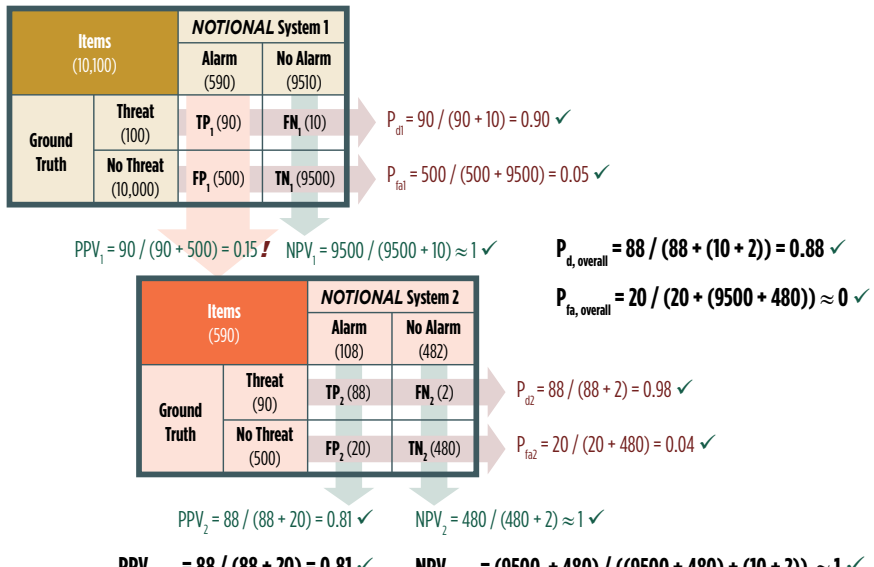
Recognizing this phenomenon, PMs should not necessarily dismiss a threat detection system simply because it exhibits a poor PPV, provided that it also exhibits an excellent $P_d$ and $P_{fa}$. Instead, PMs can estimate Prev to help determine how to guide such a system through development. Prev does *not* depend on the threat detection system and can, in fact, be calculated in the absence of the system. Knowledge of ground truth (which items contain a true threat) is all that is needed to calculate Prev (Scheaffer & McClave, 1995).

Of course, ground truth is not known *a priori* in an operational setting. However, it may be possible for PMs to use historical data or intelligence tips to roughly estimate whether Prev is likely to be particularly low in operation. The threat detection system can be thought of as one system in a system of systems, where other relevant systems are based on record keeping (to provide historical estimates of Prev) or intelligence (to provide tips to help estimate Prev). These estimates of Prev can vary over time and location. A Prev that is estimated to be very low can cue the PM to anticipate discrepancies in $P_d$ and $P_{fa}$ versus PPV, forecasting the inevitable discrepancy between the developer's versus operator's views early in the system's development, while there are still time and opportunity to make adjustments. At that point, the PM can identify a concept of operations (CONOPS) in which the system can still provide value to the operator for an assigned mission. A tiered system may provide one such opportunity.

## A Tiered System for Threat Detection

Tiered systems consist of multiple systems used in series. The first system cues the use of the second system and so on. Tiered systems provide PMs the opportunity to leverage multiple threat detection systems that, individually, do not satisfy both developers and operators simultaneously. Figure 3 shows two 2 × 2 confusion matrices that represent a notional tiered system that makes use of two individual threat detection systems. The first system (top) is relatively simple (and inexpensive) while the second system (bottom) is more complex (and expensive). Other tiered systems can consist of three or more individual systems.

**FIGURE 3. NOTIONAL TIERED SYSTEM FOR AIR CARGO SCREENING**




*Note.* The top 2 × 2 confusion matrix represents the same notional system described in Figures 1 and 2. While this system exhibits good $P_d$, $P_{fa}$, and NPV values, its PPV value is poor. Nevertheless, this system can be used to cue a second system to further analyze the questionable items. The bottom matrix represents the second notional system. This system exhibits a good $P_d$, $P_{fa}$, and NPV, along with a much better PPV. The second system's better PPV reflects the higher Prev of true threat encountered by the second system, due to the fact that the first system had already successfully screened out most items that did not contain a true threat. Overall, the tiered system exhibits a more nearly optimal balance of $P_d$, $P_{fa}$, NPV, and PPV than either of the two systems alone.

The first system is the notional air cargo screening system discussed previously. Although this system exhibits good performance from the developer's perspective (high $P_d$ and low $P_{fa}$), it exhibits conflicting performance from the operator's perspective (high NPV but low PPV). Rather than using this system to classify items as either "Alarm (Threat)" or "No Alarm (No Threat)," the operator can use this system to *screen* items as either "Cue Second System (*Maybe* Threat)" or "Do Not Cue Second System (No Threat)." Of the 10,100 items that passed through the first system, 590 were classified as "Cue Second System (*Maybe* Threat)" while 9,510 were classified as "No Alarm (No Threat)." The first system's extremely high NPV (approximately equal to 1) means that the operator can rest assured that the lack of a cue correctly indicates the very low likelihood of a true threat. Therefore, any item that fails to elicit a cue can be loaded onto the airplane, bypassing the second system and avoiding its unnecessary complexities and expense.[7] In contrast, the first system's low PPV indicates that the operator cannot trust that a cue indicates a true threat. Any item that elicits a cue from the first system may or may not contain a true threat and must therefore pass through the second system for further analysis.

Only 590 items elicited a cue from the first system and passed through the second system. Ninety items contained a true threat, while 500 items did not. The second system declared an alarm for 108 items and did not declare an alarm for 482 items. Comparing the items' ground truth to the second system's alarms (or lack thereof), there were 88 TPs, 2 FNs, 20 FPs, and 480 TNs. On its own, the second system exhibits a higher $P_d$ and lower $P_{fa}$ than the first system, due to its increased complexity (and expense). In addition, its PPV value is much higher. The second system's higher PPV may be due to its higher complexity or may simply be due to the fact that the second system encounters a higher Prev of true threat among true clutter than the first system. By the very nature in which the tiered system was assembled, the first system's very high NPV indicates its strong ability to screen out most items that do *not* contain a true threat, leaving only those questionable items for the second system to process. Since the second system encounters

only those items that are questionable, it encounters a much higher Prev and therefore has the opportunity to exhibit higher PPV values. The second system simply has less relative opportunity to generate false alarms.

> ***The utility of the tiered system must be considered in light of its cost.***

The utility of the tiered system must be considered in light of its cost. In some cases, the PM may decide that the first system is not needed, since the second, more complex, system can exhibit the desired $P_d$, $P_{fa}$, PPV, and NPV values on its own. In that case, the PM may choose to abandon the first system and pursue a single-tier approach based solely on the second system. In other cases, the added complexity of the second system may require a large increase in resources for its operation and maintenance. In these cases, the PM may opt for the tiered approach, in which use of the first system reduces the number of items that must be processed by the second system, reducing the additional resources needed to operate and maintain the second system to a level that may balance out the increase in resources needed to operate and maintain a tiered approach.

To consider the utility of the tiered system, its performance as a whole must be assessed, in addition to the performance of each of the two individual systems that compose it. As with any individual system, $P_d$, $P_{fa}$, PPV, and NPV can be calculated for the tiered system overall. These calculations must be based on *all* items encountered by the tiered system as a whole, taking care *not* to double count those $TP_1$ and $FP_1$ items from the first tier that pass to the second:

$$P_d = \frac{TP_2}{TP_2 + (FN_1 + FN_2)} = \frac{88}{88 + (10 + 2)} = 0.88 \text{ (compared to 1 for a perfect system)} \quad (6)$$

$$P_{fa} = \frac{FP_2}{FP_2 + (TN_1 + TN_2)} = \frac{20}{20 + (9{,}500 + 480)} \approx 0 \text{ (compared to 0 for a perfect system)} \quad (7)$$

$$NPV = \frac{(TN_1 + TN_2)}{(TN_1 + TN_2) + (FN_1 + FN_2)} = \frac{(9{,}500 + 480)}{(9{,}500 + 480) + (10 + 2)} \approx 1 \text{ (compared to 1 for a perfect system)} \quad (8)$$

$$PPV = \frac{TP_2}{TP_2 + FP_2} = \frac{88}{88 + 20} = 0.81 \text{ (compared to 1 for a perfect system)} \quad (9)$$

Overall, the tiered system exhibits good[8] performance from the developer's perspective. Equation 6 shows that, of all items that contained a true threat ($TP_2 + (FN_1 + FN_2) = 88 + (10 + 2) = 100$), a large subset ($TP_2 = 88$) correctly caused an alarm, resulting in an overall value of $P_d = 0.88$. The PM can conclude that 88% of items containing a true threat correctly led to a final alarm from the tiered system as a whole. Although this overall $P_d$ is slightly lower than the $P_d$ of each of the two individual systems, the overall value is still close to the value of 1 for a perfect system[9] and may (or may not) be considered acceptable within the capability requirements for the envisioned CONOPS. Similarly, Equation 7 shows that, of all items that did *not* contain a true threat ($FP_2 + (TN_1 + TN_2) = 20 + (9{,}500 + 480) = 10{,}000$), only a very small subset ($FP_2 = 20$) *in*correctly caused an alarm, leading to an overall value of $P_{fa} \approx 0$. Approximately 0% of items *not* containing a true threat caused a false alarm.

The tiered system also exhibits good[10] overall performance from the operator's perspective. Equation 8 shows that, of all items that did *not* cause an alarm ($(TN_1 + TN_2) + (FN_1 + FN_2) = (9{,}500 + 480) + (10 + 2) = 9{,}992$), a very large subset ($(TN_1 + TN_2) = (9{,}500 + 480) = 9{,}980$) correctly turned out *not* to contain a true threat, resulting in an overall value of NPV ≈ 1. The operator could rest assured that a threat was highly unlikely in the absence of a final alarm. More interesting, though, is the overall PPV value. Equation 9 shows that, of all items that did indeed cause a final alarm ($(TP_2 + FP_2) = (88 + 20) = 108$), a large subset ($TP_2 = 88$) correctly turned out to contain a true threat—these alarms were *not* false. These counts resulted in an overall value of PPV = 0.81, much closer to the 1 value of a perfect system and much higher than the PPV of the first system alone.[11] When a final alarm was declared, the operator could trust that a true threat was indeed present since, overall, the tiered system did not "cry wolf" very often.

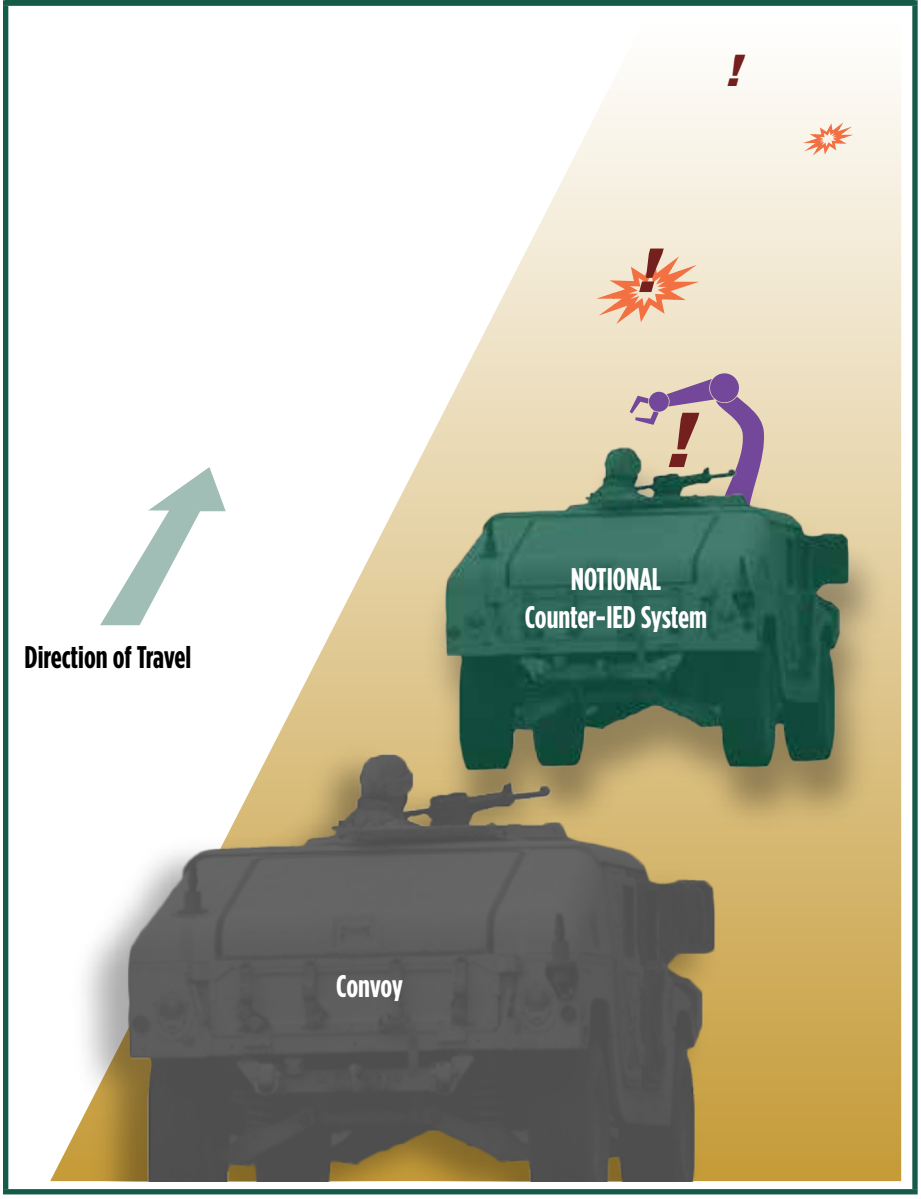
Of course, the PM must compare the overall performance of the tiered system to capability requirements in order to assess its appropriateness for the envisioned mission (DoD, 2015; DHS, 2008). The overall values of $P_d = 0.88$, $P_{fa} \approx 0$, NPV ≈ 1, and PPV = 0.81 may or may not be adequate once these values are compared to such requirements. Statistical tests can determine whether the overall values of the tiered system are significantly less than required (Fleiss, Levin, & Paik, 2013). Requirements should be set for all four metrics based on the envisioned mission. Setting metrics for only $P_d$ and $P_{fa}$ effectively ignores the operator's view, while setting metrics for only PPV and NPV effectively ignores the developer's view.[12] One may argue that only the operator's view (PPV and NPV) must be quantified as capability requirements. However, there is value in also retaining the developer's view

($P_d$ and $P_{fa}$), since $P_d$ and $P_{fa}$ can be useful when comparing and contrasting the utility of rival systems with similar PPV and NPV values in a particular mission. Setting the appropriate requirements for a particular mission is a complex process and is beyond the scope of this article.

## Threat Detection Versus Threat Classification

Unfortunately, all four performance metrics cannot be calculated for some threat detection systems. In particular, it may be impossible to calculate $P_{fa}$ and NPV. This is due to the fact that the term "threat detection system" can be a misnomer, because it is often used to refer to threat detection *and* threat classification systems. Threat classification systems are those that are presented with a set of predefined, discrete items. The system's task is to classify each item as either "Alarm (Threat)" or "No Alarm (No Threat)." The notional air cargo screening system discussed in this article is actually an example of a threat *classification* system, despite the fact that the author has colloquially referred to it as a threat *detection* system throughout the first half of this article. In contrast, genuine threat detection systems are those that are *not* presented with a set of predefined, discrete items. The system's task is *first to detect* the discrete items from a continuous stream of data *and then to classify* each detected item as either "Alarm (Threat)" or "No Alarm (No Threat)." An example of a genuine threat detection system is the notional counter-IED system illustrated in Figure 4.

**FIGURE 4. NOTIONAL COUNTER-IED SYSTEM**




*Note.* Several items are buried in a road often traveled by a U.S. convoy. Some items are IEDs (orange stars), while others are simply rocks, trash, or other discarded items. The system continuously collects data while traveling over the road ahead of the convoy and declares one alarm (red exclamation point) for each location at which it detects a buried IED. All locations for which the system declares an alarm are further examined with robotic systems (purple arm) operated remotely by trained personnel. In contrast, all parts of the road for which the system does not declare an alarm are left unexamined and are directly traveled over by the convoy.

This issue is more than semantics. Proper labeling of a system's task helps to ensure that the appropriate performance metrics are used to assess the system. In particular, while $P_{fa}$ and NPV can be used to describe threat *classification* systems, they cannot be used to describe genuine threat *detection* systems. For example, Equation 2 showed that $P_{fa}$ depends on FP and TN counts. While an FP is a true clutter item that *in*correctly caused an alarm, a TN is a true clutter item that correctly did *not* cause an alarm. FPs and TNs can be counted for threat *classification* systems and used to calculate $P_{fa}$, as described earlier for the notional air cargo screening system.

> ***This issue is more than semantics. Proper labeling of a system's task helps to ensure that the appropriate performance metrics are used to assess the system.***

This story changes for genuine threat *detection* systems, however. While FPs can be counted for genuine threat detection systems, TNs cannot. Therefore, while $P_d$ and PPV can be calculated for genuine threat detection systems, $P_{fa}$ and NPV cannot, since they are based on the TN count. For the notional counter-IED system, an FP is a location on the road for which a true IED is *not* buried but for which the system *in*correctly declares an alarm. Unfortunately, a converse definition for TNs does not make sense: How should one count the number of locations on the road for which a true IED is *not* buried and for which the system correctly does *not* declare an alarm? That is, how often should the system get credit for declaring nothing when nothing was truly there? To answer these TN-related questions, it may be possible to divide the road into sections and count the number of sections for which a true IED is *not* buried and for which the system correctly does *not* declare an alarm. However, such a method simply converts the counter-IED *detection* problem into a counter-IED *classification* problem, in which discrete items (sections of road) are predefined and the system's task is merely to classify each item (each section of road) as either "Alarm (IED)" or "No Alarm (No IED)." This method imposes an artificial definition on the item (section of road) under classification: How long should each section of road be? Ten meters long? One meter long? One centimeter long? Such definitions can be artificial, which simply highlights the fact that the concept of a TN does not exist for genuine threat detection systems.

Therefore, PMs often rely on an additional performance metric for genuine threat detection systems—the False Alarm Rate (FAR). FAR can often be confused with both $P_{fa}$ and PPV. In fact, documents within the defense and homeland security communities can erroneously use two or even all three of these terms interchangeably. In this article, however, FAR refers to the number of FPs processed per unit time interval, or unit geographical area, or distance (depending on which metric—time, area, or distance—is more salient to the envisioned CONOPS):

$$FAR = \frac{FP}{total\ time} \tag{10a}$$

or

$$FAR = \frac{FP}{total\ area} \tag{10b}$$

or

$$FAR = \frac{FP}{total\ distance} \tag{10c}$$

For example, Equation 10c shows that one could count the number of FPs processed *per meter* as the notional counter-IED system travels down the road. In that case, FAR would have units of $m^{-1}$. In contrast, $P_d$, $P_{fa}$, PPV, and NPV are dimensionless quantities. FAR can be a useful performance metric in situations for which $P_{fa}$ cannot be calculated (such as for genuine threat detection systems) or for which it is prohibitively expensive to conduct a test to fill out the full 2 × 2 confusion matrix needed to calculate $P_{fa}$.

## Conclusions

Several metrics can be used to assess the performance of a threat detection system. $P_d$ and $P_{fa}$ summarize the developer's view of the system, quantifying how much of the truth causes alarms. In contrast, PPV and NPV summarize the operator's perspective, quantifying how many alarms turn out to be true. The same system can exhibit good values for $P_d$ and $P_{fa}$ during testing but poor PPV values during operational use. PMs can still make use of the system as part of a tiered system that, overall, exhibits better performance than each individual system alone.

# References


Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 1: Sensitivity and specificity. *British Medical Journal, 308*(6943), 1552. doi:10.1136/bmj.308.6943.1552

Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. *British Medical Journal, 309*(6947), 102. doi:10.1136/bmj.309.6947.102

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behavior in response to alarms of varying reliability. *Ergonomics, 38*(11), 2300–2312. doi:10.1080/00140139508925269

Cushman, J. H. (1987, June 21). Making arms fighting men can use. *The New York Times.* Retrieved from http://www.nytimes.com/1987/06/21/business/making-arms-fighting-men-can-use.html

Daniels, D. J. (2006). A review of GPR for landmine detection. *Sensing and Imaging: An International Journal, 7*(3), 90–123. Retrieved from http://link.springer.com/article/10.1007%2Fs11220-006-0024-5

Department of Defense. (2015, January 7). *Operation of the defense acquisition system* (Department of Defense Instruction [DoDI] 5000.02). Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. Retrieved from http://bbp.dau.mil/docs/500002p.pdf

Department of Homeland Security. (2008, November 7). *Acquisition instruction/guidebook* (DHS Publication No. 102-01-001, Interim, Version 1.9). Retrieved from http://www.it-aac.org/images/Acquisition_Instruction_102-01-001_-_Interim_v1.9_dtd_11-07-08.pdf

Department of Homeland Security. (2016, March 30). *Transportation systems sector.* Retrieved from https://www.dhs.gov/transportation-systems-sector

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1*(1), 19–33. doi:10.1037/1076-898X.1.1.19

L3 Communications Cyterra. (2012). *AN/PSS-14 mine detection.* Orlando, FL: Author. Retrieved from http://cyterra.com/products/anpss14.htm

L3 Communications, Security & Detection Systems. (2011). *Fact sheet: Examiner 3DX explosives detection system.* Woburn, MA: Author. Retrieved from http://www.sds.l-3com.com/forms/English-pdfdownload.htm?DownloadFile=PDF-13

L3 Communications, Security & Detection Systems. (2013). *Fact sheet: Air cargo screening solutions: Regulator-qualified detection systems.* Woburn, MA: Author. Retrieved from http://www.sds.l-3com.com/forms/English-pdfdownload.htm?DownloadFile=PDF-50

L3 Communications, Security & Detection Systems. (2014). *Fact sheet: Explosives detection systems: Regulator-approved checked baggage solutions.* Woburn, MA: Author. Retrieved from http://www.sds.l-3com.com/forms/English-pdfdownload.htm?DownloadFile=PDF-17

Niitek. (n.d.). *Counter IED | Husky Mounted Detection System (HMDS).* Sterling, VA: Author. Retrieved from http://www.niitek.com/-/media/Files/N/Niitek/documents/hmds.pdf

Oldham, J. (2006, October 3). Outages highlight internal FAA rift. *The Los Angeles Times.* Retrieved from http://articles.latimes.com/2006/oct/03/local/me-faa3

Parasuraman, R. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. doi:10.1518/001872097778543886

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Scheaffer, R. L., & McClave, J. T. (1995). Conditional probability and independence: Narrowing the table. In *Probability and statistics for engineers* (4th ed., pp. 85–92). Belmont, CA: Duxbury Press.

Stuart, R. (1987, January 8). U.S. cites Amtrak for not conducting drug tests. *The New York Times.* Retrieved from http://www.nytimes.com/1987/01/08/us/us-cites-amtrak-for-not-conducting-drug-tests.html

Transportation Security Administration. (2013). *TSA air cargo screening technology list (ACSTL)* (Version 8.4 as of 01/31/2013). Washington, DC: Author. Retrieved from http://www.cargosecurity.nl/wp-content/uploads/2013/04/nonssi_acstl_8_4_jan312013_compliant.pdf

Wilson, J. N., Gader, P., Lee, W. H., Frigui, H., and Ho, K. C. (2007). A large-scale systematic evaluation of algorithms using ground-penetrating radar for landmine detection and discrimination. *IEEE Transactions on Geoscience and Remote Sensing, 45*(8), 2560–2572. doi:10.1109/TGRS.2007.900993

Urkowitz, H. (1967). Energy detection of unknown deterministic signals. *Proceedings of the IEEE, 55*(4), 523–531. doi:10.1109/PROC.1967.5573

U.S. Army. (n.d.) *PdM counter explosive hazard: Countermine systems.* Picatinny Arsenal, NJ: Project Manager Close Combat Systems, SFAE-AMO-CCS. Retrieved from http://www.pica.army.mil/pmccs/pmcountermine/CounterMineSys.html#nogo02

## Endnotes

[1] PMs must determine what constitutes a "good" performance. For some systems operating in some scenarios, $P_d$ = 0.90 is considered "good," since only 10 FNs out of 100 true threats is considered an acceptable risk. In other cases, $P_d$ = 0.90 is not acceptable. Appropriately setting a system's capability requirements calls for a frank assessment of the likelihood and consequences of FNs versus FPs and is beyond the scope of this article.

[2] Statistical tests can determine whether the system's value is significantly different from the perfect value or the capability requirement (Fleiss, Levin, & Paik, 2013).

[3] Ibid.

[4] Ibid.

[5] Ibid.

[6] Conversely, when Prev is *high*, threat detection systems often exhibit poor values for *N*PV, even while exhibiting excellent values for $P_d$, $P_{fa}$, and PPV. Such cases are not discussed in this article, since fewer scenarios in the DoD and DHS involve a *high* prevalence of threat among clutter.

[7] PMs must decide whether the 10 FNs from the first system are acceptable with respect to the tiered system's capability requirements, since the first system's FNs will not have the opportunity to pass through the second system and be found. Setting capability requirements is beyond the scope of this article.

[8] PMs must determine what constitutes a "good" performance when setting the capability requirements for the tiered system.

[9] Statistical tests can show which differences are statistically significant (Fleiss et al., 2013), while subject matter expertise can determine which differences are operationally significant.

[10] Once again, PMs must determine what constitutes a "good" performance when setting the capability requirements for the tiered system.

[11] Once again, statistical tests can show which differences are statistically significant (Fleiss et al., 2013), while subject matter expertise can determine which differences are operationally significant.

[12] All four of these metrics are correlated, since all four metrics depend on the system's threshold for alarm. For example, tuning a system to lower its alarm threshold will increase its $P_d$ at the cost of also increasing its $P_{fa}$. Thus, $P_d$ cannot be considered in the absence of $P_{fa}$ and vice versa. To examine this correlation, $P_d$ and $P_{fa}$ are often plotted against each other while the system's alarm threshold is systematically varied, creating a Receiver-Operating Characteristic curve (Urkowitz, 1967). Similarly, lowering the system's alarm threshold will also affect its PPV. To explore the correlation between $P_d$ and PPV, these metrics can also be plotted against each other while the system's alarm threshold is systematically varied in order to form a Precision-Recall curve (Powers, 2011). (Note that PPV and $P_d$ are often referred to as Precision and Recall, respectively, in the information retrieval community [Powers, 2011]. Also, $P_d$ and $P_{fa}$ are often referred to as Sensitivity and One Minus Specificity, respectively, in the medical community [Altman & Bland, 1994a].) Furthermore, although $P_d$ and $P_{fa}$ do not depend upon Prev, PPV and NPV do. Therefore, PMs must take Prev into account when setting and testing system requirements based on PPV and NPV. Such considerations can be done in a cost-effective way by designing the test to have an artificial prevalence of 0.5 and then calculating PPV and NPV from the $P_d$ and $P_{fa}$ values calculated during the test and the more realistic Prev value estimated for operational settings (Altman & Bland, 1994b).

## Biography

**Dr. Shelley M. Cazares** is a research staff member at the Institute for Defense Analyses (IDA). Her research involves machine learning and physiology to reduce collateral damage in the military theater. Before IDA, she was a principal research scientist at Boston Scientific Corporation, where she designed algorithms to diagnose and treat cardiac dysfunction with implantable medical devices. She earned her BS from MIT in EECS and PhD from Oxford in Engineering Science.

*(E-mail address: scazares@ida.org)*