Could better tests have predicted the rare circumstances of the Germanwings crash? Probably not

May 28 2015, by Norman A Paradis

Credit: AI-generated image (disclaimer)

When people do terrible things, it seems reasonable to believe we should have taken steps to identify them beforehand. If we can do that, then surely we can prevent them from doing harm.

The crash of Germanwings Flight 9525 in March, which appears to have been an intentional act, is an example. It shocks us (and understandably so) when a trusted professional harms those who have entrusted their lives to him or her.

So why not identify pilots at risk and take steps to prevent similar events from ever occurring again?

Because it is likely impossible, and maybe even counterproductive.

And that's not just my opinion. The limits of what can be achieved in predicting an event represent a dilemma we face all the time in biomedical testing.

Let me take you through such an analysis, and show you how futile such programs would likely be in preventing events like the air crash in Europe.

Medical test can be sensitive or specific, but rarely both

Any interview or written survey instrument intended to identify individuals at risk of perpetrating rare and horrific acts is essentially a medical test. And the performance of such tests is described by its sensitivity and specificity. Simply put, sensitivity is the ability of the test to detect the disease, and specificity is the accuracy of its result.

For most tests, you make trade-offs between one or the other: sensitivity versus specificity. For instance, highly sensitive tests generally have many false positives – they call patients sick when the patient does not have the disease. And highly specific tests often have many false negatives – they miss many patients with the disease.

Generally, you can have have a sensitive test or a specific test, but you can't have a sensitive *and* specific test. Using a simple metaphor, this can be called the "no free lunch law" of medical testing.

This limitation becomes overwhelming when biomedical tests are used in populations with a very low incidence of the disease tested for.

An absurd example can help to understand this. Modern pregnancy tests are very accurate, over 99%. However, let's say you apply a pregnancy test in a population of 10,000 men. You will get a handful of positive tests, 100% of which will be false positives.

For this reason, standard blood tests cannot generally be used to screen for very rare diseases without being paired with a second specific confirmatory test.

Turning our attention back to Germanwings Flight 9525, the incidence of an event like this is so uncommon that it is within a rounding error of male pregnancy.

There have been 660 million commercial airline departures since 1959, with only a handful of crashes believed to have been intentional acts by the pilot. Even if we assume there may have been crashes intentionally caused by pilots but not attributed to them, it is still a very rare event. Maybe not the rarest of events (at least one person among the approximately 100 billion people who have ever lived claims to have been both struck by lightning and bitten by a shark), but for our purposes it's particularly unusual.

So, even if we could develop a test or a screening process to find a pilot who would intentionally crash a plane, and that system was very, very good – both specific and sensitive – virtually all positives would be false positives.

Psycho-social medical tests aren't very accurate

And there is a hierarchy for test performance that makes all of this more complicated. Tests in which you cut the patient open and examine tissue under a microscope have the best performance, with nearly perfect sensitivity and specificity. Imaging tests, such as CAT scans and MRIs, provide millions of visual data points and also have very good performance. But by the time we get down to measuring the concentration of molecules in blood, problems develop. Such tests should not be used without a thorough understanding of the incidence of the disease.

At the very bottom of the hierarchy of performance are psycho-social survey instruments – tests in which a series of questions are asked with the intention of making psychological diagnosis. Some experts have asserted that once publication bias (the tendency to publish only positive results) is removed, most if not all such instruments will be found to lack any predictive performance whatsoever.

A large systematic review published in the *British Medical Journal* studied the performance of assessment tools for the prediction of violence in people at risk and found that two people would need to be detained, or somehow otherwise prevented from acting, to prevent one violent act. They concluded "even after 30 years of development, the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence based."

Prediction can lead to false positive results

Even precisely diagnosing a disease is more difficult than most people realize. There is also a hierarchy when it comes to disease diagnostics. Well-understood and immediately life-threatening illnesses such as

advanced cancer or heart disease can often be easily diagnosed. On the other end of the spectrum, nonspecific aches and pains, or diseases in their very early stages, challenge even the best clinicians.

Don't be misled by the vast psychiatric and psychological literature; the underlying pathophysiology and molecular biology of these disorders are not really understood. It comes as no surprise that our ability to definitively predict their risk is minimal.

So what would happen if we used some interview-based diagnostic instrument to predict the risk that a pilot might intentionally crash a plane? For the purposes of argument, let's assume that such an event might occur in the range of one in a few hundred million take-offs.

Since we're dealing with poorly performing diagnostic tools, in the setting of a poorly understood behavioral disease, it is likely that we will get tens of thousands of positive tests. And because we are trying to predict an extraordinarily rare complication of that disease, all, or almost all, positives will be false positives.

Even worse, these false positives may not be benign. There are at least two additional dimensions inherent to this exercise that make it worrisome:

1. The airlines and regulatory organizations may overreact to the recent crash by revoking the flying credentials of pilots who "fail" such a testing.
2. Because their job is at risk, pilots will attempt to hide dark thoughts and concerns that are normal to all human beings.

It is possible – even likely – that such a program might cause pilots with symptoms of depression to hide their disease and possibly avoid treatment for a treatable and not altogether uncommon condition –

increasing the overall risk to passengers, since diseases like depression may be associated with cognitive and performance impairment when untreated.

False positives can have major consequences

These concepts, by the way, are applicable in settings less rare than plane crashes. They come into play whenever a test – or even a test equivalent – is used to refine our estimation that something exists or may happen. Medical testing is the classic example, but the detection of defective jet turbine blades would be equally valid.

The extreme rarity of a pilot intentionally crashing an airliner, and the poor performance of psychological tests, make it easy to conclude that such "testing" would be futile. It is much more difficult to figure out what to do with things like screening for breast cancer or predicting risk of Alzheimer's dementia.

But it is also much more important.

The underlying mathematics informs us that one needs to know the performance of the test and the incidence of the outcome of interest. What the math doesn't teach us is that our response to the result is also very important.

If the use of a test only causes us to non-invasively recheck more frequently or more carefully, that is one thing. It is a whole other thing to respond by cutting open a patient or exposing them to X-rays.

When the consequences of a false-positive test are large, we must be much more careful if we are to avoid harm.

One of my favorite examples is the drug testing of athletes. The

organizations responsible act like their programs perform to a high degree of certainty. But unless they are using laboratory tests with performance unavailable to clinical medicine, and the incidence of drug use among athletes is very high, their false-positive rate is likely greater than people realize.

It may be possible to prevent rare events such as this one – "smart" cockpit doors or some such technological solution. But predicting their occurrence by looking more closely at the individuals involved is doomed to fail. It is an extreme version of a problem we all confront daily, mostly without realizing it.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation