Modern Health Science; Vol. 2, No. 1; 2019
ISSN 2576-7291 E-ISSN 2576-7305
https://doi.org/10.30560/mhs.v2n1p60

Screening, Sensitivity, Specificity, and So Forth: A Second, Somewhat Skeptical, Sequel

Robert Trevethan[1]

[1] Independent academic researcher and author, Albury, NSW, Australia

Correspondence: Robert Trevethan. E-mail: robertrevethan@gmail.com

Received: May 7, 2019; Accepted: June 6, 2019; Published: June 13, 2019

Abstract

This article is concerned with the sensitivity, specificity, predictive values, and other metrics associated with screening tests. It has direct origins in two previous articles. In this third article, the author of the first article writes about topics and issues that were addressed only minimally in his previous article and expands on topics raised by authors of the second article. In particular, attention is turned to wording and terminology that can be idiosyncratic and confusing with regard to screening versus diagnosis as well as to issues associated with reference ("gold") standards and screening tests, and to the importance of cutpoints and prevalence in relation to metrics associated with screening tests. The primary aims are to help readers attain clarity about topics that they might have felt unsure about; gain reassurance about conceptual difficulties in the field that, once recognized for what they are, can become less problematic because it is possible to be confident about not being confident; and, where appropriate, adopt a skeptical attitude about screening tests and their associated metrics. Examples are drawn from the use of ankle–brachial and toe–brachial indices for identifying peripheral artery disease, although wider applicability is intended.

Keywords: screening, diagnosis, sensitivity, specificity, predictive values, likelihood ratios

1. Introduction

In November 2017, *Frontiers in Public Health - Epidemiology* published an article that I authored concerning sensitivity, specificity, and predictive values (Trevethan, 2017b). Approximately 10 months later, Grunau and Linn published an article in the same journal (Grunau & Linn, 2018), indicating that my original article did not contain sufficient detail about some topics and suggesting that additional topics should be considered. In preparing this "second sequel", my intention is to present material not included in my first article because of its journal's word limit but also to expand on, and address, topics presented by the authors of the second article (the first sequel).

In my original article, I attempted to provide clear and helpful descriptions of sensitivity, specificity, and predictive values as well as to identify and resolve what I believed to be common misconceptions concerning those metrics. I also dealt with ways in which those metrics could inform decision making in health contexts. The impetus for this was my sense that researchers and clinicians seem to have difficulty understanding those metrics—a lack of understanding that has been borne out empirically (Manrai et al., 2014; Puhan et al., 2005; Steurer et al., 2002; Whiting, Davenport, et al., 2015). In this current article, I retain an intention to write clearly. This intention is driven by a sense that much writing about screening tests could be described as hermetic—somehow sealed from ready comprehension (Note 1)—and that I might be able to present material that would resonate for a readership that sought clarity and reassurance in contrast to material that often appears to be confusing, unclear, or even misleading.

In this article, a major parallel aim is to help readers realize why some confusions exist, but I also pursue the recommendation made at the end of my original article where I wrote that healthy skepticism is advisable in the context of screening tests. I first provide information concerning complexities associated with the words *screening* and *diagnosis*. I then raise issues concerning the nature of reference standards and screening tests, and follow that by demonstrating how cutpoints and sample prevalence can, but in some respects may not, substantially influence sensitivity, specificity, predictive values, and other metrics. In the concluding sections I provide additional information and insights.

In order to support these topics, I use examples relating to the ankle–brachial index (ABI) and toe–brachial index (TBI), both of which are noninvasive indicators of peripheral artery disease (PAD). They are commonly regarded as screening rather than diagnostic tests and are ideal for current purposes because they have features that permit

revelation of complexities related to screening tests. Both indices are described in the next subsection to provide an appropriate foundation for the material that follows. Although ABIs and TBIs inevitably have some features that do not translate readily to other contexts, they are used because, without concrete examples, information and issues can fail to become apparent or to consolidate.

2. The Ankle–Brachial and Toe–Brachial Indices

Peripheral artery disease refers to narrowing of arteries caused by atherosclerotic (i.e., fatty) deposits in arteries throughout any part of the body other than the heart and brain (the word *peripheral* can be deceiving because it refers not only to limbs but also to other, more "internal", parts of the body, including kidneys and stomach). Estimates of the prevalence of PAD vary widely from near zero to > 60% depending on how the disease is identified and on variables within subpopulations that include age, ethnicity, gender, hypertension, diabetes, and smoking history (Caro et al., 2005; Criqui & Aboyans, 2015). Atherosclerosis reduces effective flow of blood, leading to a variety of health problems. It is often first evidenced as stenoses or occlusions in the arteries of the legs and, because of that, when PAD is present there is likely to be less blood supply to the ankles and feet which therefore produces a difference in blood pressure between the upper parts of the body and the lower extremities, with lower pressure in the latter. In order to detect whether that difference is present, systolic blood pressures are often obtained from the arm (known medically as the brachium) and compared with blood pressure at either the ankle or a toe (usually the big toe, or hallux). These hemodynamically logical, and relatively simple, procedures are used because PAD is often asymptomatic.

When ankle blood pressure is divided by brachial blood pressure, the single-figure outcome is known as the ankle–brachial index, or ABI. For a healthy person, ankle pressure is typically higher than brachial pressure (Goldstein et al., 2014; Gong et al., 2015; Xu et al., 2010), (Note 2) and therefore the ABI is likely to be > 1.0, often in the range 1.1 to 1.3. In a person with PAD, the ABI is, at least theoretically, likely to be noticeably < 1.0, and an ABI ≤ 0.90 is often regarded as indicative of PAD (Bundó et al., 2013; Xu et al., 2010).

When toe blood pressure is divided by brachial pressure, the outcome is known as the toe–brachial index, or TBI. For a healthy person, toe pressure is often, although not always, lower than brachial pressure, and therefore TBIs for healthy people are often < 1.0 (Quong, 2016; Watanabe et al., 2015). Høyer et al. (2013) have estimated that a TBI of 0.71 is the lowest threshold for a normal TBI, and TBIs ≥ 0.75 have been regarded as unlikely in the presence of PAD (Hinchliffe et al., 2016).

3. Defining Screening and Diagnostic Tests

In the opening paragraph of my former article (Trevethan, 2017b), I defined screening tests as typically having "advantages over diagnostic tests such as placing fewer demands on the healthcare system and being more accessible as well as less invasive, less dangerous, less expensive, less time-consuming, and less physically and psychologically discomforting for clients". In addition, screening tests are often relatively simple and are not intended to be definitive (Wilson & Jungner, 1968). By way of contrast, I characterized diagnostic tests as not usually possessing the typical features of screening tests (they often possess *opposite* features) and as being defined primarily in terms of providing definitive evidence about the presence or absence of a target disease or condition.

Screening tests are usually, therefore, acknowledged to have some degree of imprecision or ambiguity. Because of that, their sensitivity and specificity might both be only moderate, or one might be satisfactory but the other unsatisfactory. This is evident in a number of studies examined by Xu et al. (2010) where, for the ABI, nine pairs of sensitivity and specificity results were {.69 and .83}, {.71 and .89}, {.76 and .90}, {.79 and .96}, {.76 and .95}, {.68 and .99}, {.63 and .97}, {.20 and .99}, and {.15 and .99}. In research by Sonter, Tehan, and Chuter (2017), both sensitivity and specificity for the TBI were consistently moderate to unsatisfactory, the paired results being {.71 and .77}, {.78 and .61}, {.74 and .67}, and {.70 and .62}.

Another important, and quite different, feature of screening tests is that they come in different forms, and with dissimilar purposes. Within a seminal document, Wilson and Jungner (1988) described seven kinds of screening tests or activities. Among these, it is possible to conceive of a continuum, at one end of which is mass screening of large population groups as well as widespread screening of people identified as belonging to high-risk groups, both of which might be seen as falling primarily within the remit of conventionally conceived epidemiology with its focus on identifying the determinants, prevalence, and incidence of diseases within a population, as well as controlling those diseases. At the other end of the Wilson and Jungner continuum are *case findings*. These might be seen as falling within the remit of clinical epidemiology with its focus—often medically oriented—on the prevention, identification, diagnosis, prognosis, and treatment of disease, as well as on the accuracy of tests, *for individual people*. Wilson and Jungner point out that differences in the kinds of screening activities need to be kept

in mind—and it is the latter kind that is the focus within this present article. Other authors might approach the topic of screening with a different focus, which, ideally, should be made explicit in order to avoid confusion.

Just as the word *screening* can have different meanings, so, also, can the word *diagnosis*. Initially, an acceptable definition seems simple: diagnostic tests provide clear evidence concerning the presence or absence of a condition (Trevethan, 2017b), and these tests can therefore be expected to have high sensitivity and specificity and, consequently, to yield very few false positive or false negative results.

In practice, however, drawing a distinction between screening and diagnostic tests is not clearcut. Sometimes one diagnostic test might be compared with another diagnostic test, with the former regarded as highly trustworthy and therefore able to serve as a valid yardstick for assessing the latter's trustworthiness. In this situation, diagnostic tests might be conceived of as hierarchically related to each other. Furthermore, what some people regard as a diagnostic test can be regarded by others as being, less definitively, at the level of indication, or initial detection, that is associated with screening. This difference in perspectives is evident in the titles of the following journal articles about the ABI and TBI, where words relating to diagnosis and screening (the latter being referred to in terms of detection) have been bolded:

- *Sensitivity and specificity of the ankle–brachial index to **diagnose** peripheral artery disease* (Xu et al., 2010)
- *The toe-brachial index in the **diagnosis** of peripheral arterial disease* (Høyer et al., 2013)
- *Toe brachial index measured by automated device compared to duplex ultrasonography for **detecting** peripheral arterial disease* (Sonter, Tehan, & Chuter, 2017)
- *A systematic review of the sensitivity and specificity of the toe–brachial index for **detecting** peripheral artery disease* (Tehan, Santos, & Chuter, 2016).

At times, the distinction between screening and diagnosis can even be blurred within the space of a few words, as in: "This is essential to determine the clinical efficacy of using this technique for early **diagnosis** of PAD and ongoing peripheral vascular **screening** and monitoring" (Formosa et al., 2018, p. 281, bolding added).

Although it may be helpful to make a distinction between screening tests as providing (merely) an indication that a condition is present, and diagnostic tests as providing more definitive evidence, there are two important caveats. These are that a screening test might have such high and well-founded sensitivity that a negative test result can be regarded as ruling the condition out (commonly referred to as *snout*), and that a screening test might have such high and well-founded specificity that a positive test result can be regarded as ruling the condition in (commonly referred to as *spin*), concerning which I provided extended information in my previous article (Trevethan, 2017b).

Under some circumstances, therefore, aspects of screening tests might validly acquire the status of diagnostic tests. Nevertheless, and while keeping the snout and spin principles in mind, the main purpose of screening tests might be regarded as providing a "ballpark" indication concerning the presence or absence of a condition so that appropriate decisions can be made—perhaps that nothing at all needs be done, that evasive or remedial action should be recommended or implemented immediately, that regular follow-up monitoring should be put in place, or that further exploration would be needed to ascertain whether or not a condition is likely to be, or is actually, present.

In order to assess the validity, or trustworthiness, of a screening test, that test is assessed against a diagnostic test, or reference standard. In an ideal situation, there would be reliable and valid procedures within both the reference standard and the screening test as well as appropriate criteria for determining whether a condition is present or not. An example of this situation is provided in Figure 1. There, for current purposes, it is assumed that the ABI is an excellent screening test for PAD when it has a cutpoint set at 0.90 and therefore that it has been able to accurately identify 120 people as either having, or as not having, PAD—and, on both the reference standard and screening test, 30 of those 120 people were identified as having PAD, and 90 as not having PAD. Therefore, all of the people tested could be categorized as either true positives or true negatives. This would result in sensitivity and specificity both being 100%, and, if the people assessed by the reference standard were representative of a specified subpopulation, the predictive values for other people belonging to that subpopulation would also equal 100% (Note 3).

When situations are less ideal than depicted in Figure 1, there are deficiencies in either the reference standard or the screening test, or in both. Examples of the kind of deficiencies that can exist will be provided below, first with

regard to reference standards, and then with regard to screening tests. Several of the issues that are raised pertain to both reference standards and screening tests, so some duplication of content is inevitable but will be avoided as much as possible. There are additional issues, and also more extensive issues within those considered below, but the issues raised serve to indicate that the validity of neither reference standards nor screening tests should be assumed unquestioningly and that, where possible, steps should be taken to improve the sensitivity, specificity, predictive values, and other metrics associated with screening tests by addressing problems related to both reference standards and screening tests.

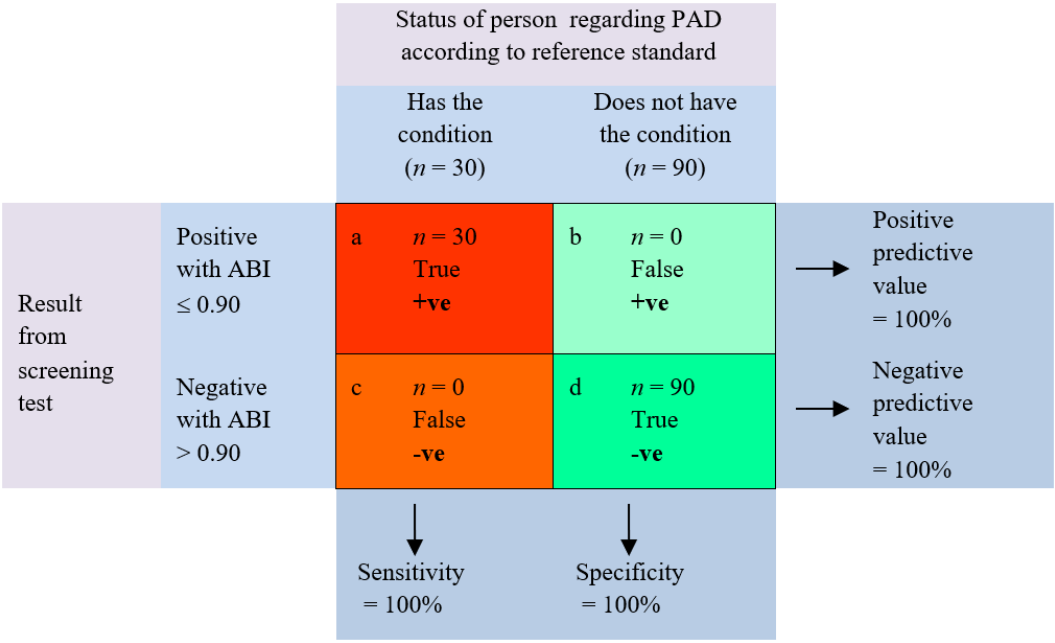

Figure 1. Diagram depicting the ideal situation in which 120 people were tested and sensitivity, specificity, and predictive values are all 100%

4. Issues Pertaining to Reference Standards

A number of issues pertain to reference standards—the diagnostic tests that are assumed to definitively establish the presence or absence of a condition. Three will be raised here. As indicated above, they are not intended to be exhaustive but rather to indicate a need for caution, even skepticism, when considering the adequacy of reference standards.

A major issue is that criteria are not always unequivocal when people are initially categorized as either having or as not having a particular condition. For example, when determining whether or not people have PAD, which is often based on the presence of atherosclerosis within arteries of the legs, a number of different criteria are used. These criteria include the number and size of any stenoses or occlusions in the legs, where those stenoses or occlusions are located, and the number of arteries affected. These differences in the reference standard might not be trivial. For example, Okamoto et al. (2006) conducted one set of analyses in which lesions below the knee (infragenicular) were regarded as indicative of PAD if there was complete obstruction with blocking of contrast media, and another set of analyses in which lesions above the knee (supragenicular) were regarded as indicative of PAD if there was more than 75% stenosis of the vascular lumen. These two reference standards, when applied to four different screening tests, produced strikingly different outcomes. For example, the respective sensitivity and specificity of the TBI with the first method (infragenicular) were 45 and 100%, and with the second method (supragenicular) were 92 and 87%. (Note 4)

Other examples of problems concerning reference standard criteria are not difficult to find. In the final stage of a refined literature search concerning the ABI, Xu et al. (2010) excluded 10 of 33 studies for not possessing an appropriate reference standard. From a different angle, within their systematic review focusing on the TBI in screening for PAD, Tehan, Santos, and Chuter (2016) examined seven studies across which there was little consistency in the criteria used to define the presence of PAD according to the reference standards used.

Clearly, finding an appropriate reference standard for defining and measuring PAD is not straightforward, and careful thought must therefore be given to identify the most valid indicator of that target condition. It might also be desirable to use a combination of two or more indicators, rather than a single indicator, as the reference standard to determine whether or not people have a particular condition (Bossuyt et al., 2003; Parikh et al., 2008).

When different methodological procedures are used within reference standards to establish whether or not people have a condition, another issue arises because there is no guarantee that the different procedures will produce similar or optimal results. According to the articles examined in their systematic review concerning identification of PAD, Tehan, Santos, and Chuter (2016) found that procedures included unspecified angiography, digital subtraction angiography, angiography in conjunction with multidetector computed tomography, and color duplex sonography/ultrasound—each of which could generate an outcome that differed from other outcomes and might therefore produce differing results, some possibly more valid than others, when the adequacy of a screening test is being assessed.

A third issue arises when different operators within a particular method might have different levels of competence or different decision-making practices and could therefore produce differing results even when using the same procedure(s) and attempting to apply the same criteria. Often, researchers provide no indication about the reliability of the operators in their studies. For example, five of the seven studies reviewed by Tehan, Santos, and Chuter. (2016) contained no information about rater reliability. Even when that information has been provided, it is not always reassuring. For example, Tehan, Bray, and Chuter (2016) assessed interrater reliability across three operators who used color duplex sonography to identify the presence of PAD, and reported a kappa of .78—a level of agreement that McHugh (2012) has categorized as being only moderate, albeit at the high end of moderate. Under these circumstances, it is difficult to regard a reference standard as having satisfactory reliability.

A fourth issue relates to unsound investigative logic sometimes being employed. For example, Chongthawonsatid and Dutsadeevettakul (2017) assessed sensitivity, specificity, predictive values, and accuracy (Note 5) of the ABI by comparing results from one method of obtaining ABIs with another, the second of which they referred to as a "gold standard". In this situation, two highly similar screening tests were compared with each other and no diagnostic-level measure of PAD was involved. In the penultimate paragraph of my previous article (Trevethan, 2017b), I referred to a different logical flaw. There, Jönelid et al. (2016) assessed the ABI as a screening test but also used the same ABI results from the screening test as part of their reference standard, thus inevitably obtaining an inflated similarity between the two measures because one was embedded within the other.

Given the above, it is obvious that any assessment of a screening test should include an evaluation of the adequacy of the reference standard against which that screening test's results are being compared. So much is this necessary that Parikh et al. (2008) considered the anomalous, even amusing, possibility that a screening test might be more indicative of a condition's presence than is the reference standard used to assess the screening test's adequacy. Because of circumstances such as these, where the validity of reference standards can be called into question, it is understandable that many methodologists now prefer the term *reference standard* to *gold standard*. The term gold standard could well refer to a myth that creates an inappropriate sense of validity and certainty.

5. Issues Pertaining to Screening Tests

Screening tests, also, cannot be assumed to produce appropriate or accurate results. In general, the relevant issues are more extensive than are the issues associated with reference standards, but many also pertain to reference standards. Although an exhaustive treatment will not be attempted here, what follows should be sufficient to indicate that the results from screening tests should not be accepted at face value and, again, that a degree of skepticism could well be warranted.

As an introductory springboard, it might be noted that, in their systematic review concerning the TBI as a screening test for PAD, Tehan, Santos, and Chuter (2016) selected 17 studies as being relevant, excluded 10 of those studies for a variety of reasons (including five for not having an acceptable reference standard), assessed the remaining seven studies using the QUADAS-2 (see Whiting, Rutjes, et al., 2011), and adjudged none of those seven studies to be without risk of bias regarding measurement of TBIs.

An initial issue associated with screening tests is that measurement methods might be unreliable because of inherent inaccuracy related to device characteristics or poor calibration at the time measurements were taken. Rich (2015), for example, has indicated that toe pressures obtained manually with an aneroid sphygmomanometer and Doppler device might be less accurate than are other methods. Independent of that, manufacturers of the SysToe device evidently believe that, in order to obtain accurate toe pressure measurements, the toe pulp should be exsanguinated prior to occlusion occurring, and therefore both exsanguination and occlusion cuffs are provided

with the SysToe (Pérez-Martin et al., 2010), but other manufacturers of automated devices for measuring toe pressure, by providing only an occlusion cuff, evidently do not regard initial exsanguination to be necessary.

Second, different procedures might produce different outcomes. Aboyans et al. (2012) and others (Al-Qaisi et al., 2009; Caruana et al., 2005; Jelinek & Austin, 2006; Xu et al., 2010) refer to an array of procedures in relation to measurement of the ABI—including the number of readings taken at the arm and ankle, the arteries from which readings are taken at the ankle, and whether averaged readings or the higher/highest ankle and brachial readings are used in the index—an array that suggests some measurements are likely to be different from, and preferable to, others. A parallel range of procedures exists for the measurement of TBIs, including which arms and toes blood pressures are taken from, the size and location of toe cuffs, and whether or not pressure readings are averaged (McAra & Trevethan, 2018). Some of these procedures have been demonstrated to yield substantially different outcomes (Bhamidipaty et al., 2015; Påhlsson et al., 2007).

Third, the people who conduct the tests might not be reliable. It would appear that in many situations associated with measurement of the ABI there is high intra- and inter-rater reliability if operators are skilled or have had sufficient practice with the assessment equipment (Mätzke et al., 2003). However, rater reliability is often not assessed and therefore can be indeterminate. When evidence is provided, the results are not always reassuring. For example, Tehan, Bray, Keech, et al. (2015) reported interrater ICCs of 0.80 for toe pressures and 0.66 for brachial pressures, with both ICCs falling well short of the desired 0.90 that is regarded as desirable for clinical situations (Nunnally & Bernstein, 1994; Portney & Watkins, 2009; Trevethan, 2017a). It should be conceded, however, that the inconsistency in readings might have occurred because those readings were taken up to a week apart, so endogenous and exogenous variables relevant to the participants, rather than rater disagreement, could have been the predominant driver of inconsistency. In other research, the interrater reliability ICCs for toe pressures taken within 15 minutes of each other ranged from .82 to .91 (Romanos et al., 2010; Sonter, Chuter, & Casey, 2015), but for brachial pressures were as low as .49 and did not exceed .82 (Sonter, Chuter, & Casey, 2015; Sonter, Sadler, & Chuter, 2015). These results are far from reassuring given the desirability of ICCs > .90 in clinical settings.

Fourth, the people being assessed might have fluctuant characteristics that are not sufficiently taken into account. For example, people who have diabetes can have highly variable brachial blood pressures (Parati et al., 2013; Trevethan, 2019), even within the span of a single clinical visit (Okada et al., 2015), so procedures such as averaging two or more blood pressure readings when obtaining ABIs and TBIs might be advisable. Furthermore, confounders such as the white coat phenomenon, environmental temperature, and time of day might need to be controlled for or overcome, or at least taken into account.

Fifth, the people being assessed might not present themselves or be prepared satisfactorily. For the assessment of ABIs and TBIs, such preconditions as avoidance of recent medication, food, tobacco, caffeine, and vigorous physical activity; a pretest resting period; and heating of limbs can all be of importance (Bonham, 2011; Campbell et al., 1990; Høyer et al., 2013; Sawka & Carter, 1992).

Given the range of problems that might beset screening tests, it is understandable that two or more of those tests have sometimes been used in conjunction with each other (Cadogan, McNair, Laslett, & Hing, 2013; Cadogan, McNair, Laslett, Hing, & Taylor, 2013; Lewis et al., 2016).

5. Issues Associated with Cutpoints

Many variables of interest to health professionals are continuous, not binary, in nature and therefore cutpoints need to be established either side of which people can be regarded as having, or as not having, the condition of interest. With both the ABI and TBI, all readings below a chosen cutpoint are assumed to be indicative of PAD, and, conversely, all readings above that cutpoint are assumed to indicate absence of PAD.

This kind of distinction appears to be uncomplicated. Cutpoint values *can* be complicated, however, primarily because of uncertainty, even disagreement, about where they should be placed. For the ABI, a cutpoint of 0.80, rather than the conventional 0.90 has sometimes been referred to (European Stroke Organisation et al., 2011; McAra et al., 2017). Higher cutpoints have also been proposed—at 1.00 (Chen et al., 2016; Criqui, McClelland, et al., 2010) and even at 1.10 (Gornik, 2009). These higher cutpoints have been recommended because medial arterial calcification (MAC), a condition usually associated with increasing age and diabetes, prevents ankle arteries from being compressed and, as a result, both ankle blood pressure readings and ABIs become spuriously elevated. Therefore, an apparently normal ABI might mask PAD, so the level for "normal" has sometimes been raised in order to avoid false negative screening results.

Cutpoints > 0.90 for the ABI are associated with two kinds of problems, however. On the one hand, they could generate too many false positive results, thus reducing the specificity of the ABI and lead to unnecessary alarm,

distress, overreferral, and overtreatment. On the other hand, the higher ABIs could be indicative of MAC, which is considered by some researchers to be of similar concern as PAD with regard to cardiovascular events (Nishimura et al., 2016; Suominen et al., 2008). Although MAC is sometimes regarded to be likely only if ABIs are > 1.40 (Bundó et al., 2013; Chen et al., 2016), cutpoints of 1.30, 1.40, and 1.50 have been investigated as prospective cutpoints (Suominen et al., 2008), and even lower ABI cutpoints of 1.20 (McAra et al., 2017) and 1.10 (Cardenas et al., 2018) have been regarded as indicating MAC. Under these circumstances, it is difficult to interpret mid-range ABI values with confidence. They could be normal, could mask the presence of PAD by being spuriously inflated as a result of atherosclerosis, or could be indicative of MAC as an advanced, clinically significant, and possibly separate vascular condition that should not be ignored (see Ix, Miller, Criqui, & Orchard, 2012).

Because MAC within blood vessels in the ankle could be present in people who have PAD, and because calcification is less likely to occur in the toes (Aboyans et al., 2012; Young et al., 1993), TBIs, rather than ABIs, are sometimes recommended over ABIs when screening for PAD in order to avoid ambiguity (Hinchliffe et al., 2016; Mills et al., 2014; Rooke et al., 2011; Trevethan, 2018). For the remainder of this article, therefore, examples will be based on TBIs.

Despite their prospective advantages, TBIs are not without issues concerning cutpoints, however. These cutpoints have been set at 0.50 (Suzuki, 2007), 0.60 (Bundó et al., 2013; Suominen et al., 2008), 0.65 (McAra et al., 2017), and 0.75 (Hinchliffe et al., 2016), and in a review concerning the TBI, Høyer et al. (2013) estimated that 0.71 was the lowest threshold for a normal TBI for limbs that had been heated, but they also referred to studies in which the cutpoint for PAD ranged from 0.54 to 0.75.

These differences are not trivial. On a screening test that has continuous scores, the cutpoints used when determining whether people should be regarded as having, or as not having, a condition can have a considerable effect on sensitivity, specificity, and predictive values. This can be demonstrated in Figures 2 and 3. The data set used in creating entries for these figures, although fictitious, was informed by data obtained in other research (McAra, 2015; McAra & Trevethan, 2018). The full set of these data, with supporting annotations, is provided in Appendix A. These data represent TBIs that might be found in a sample of 100 people considered at high risk of PAD, namely people over 75 years of age, some of whom are healthy but many of whom have combinations of diabetes, hypertension, hyperlipidemia, obesity, a history of smoking, chronic kidney disease, symptoms of intermittent claudication, and leg pain at rest. In this dataset, the mean TBI = 0.57, the standard deviation = 0.24, and TBIs range from 0.13 to 1.05.

As pointed out earlier in this article, prevalence of PAD varies between subpopulations. With that in mind, many people in the previous paragraph might not only be regarded as at high risk of PAD but also as providing a reasonable indication of prevalence for that group. Therefore, based on results from the reference standard, prevalence of PAD for that subpopulation can be calculated from entries in Figures 2 and 3 by the following formula:

$$[\ a + c\ /\ (a + b + c + d)\ ]\ x\ 100$$

which, for both figures yields an outcome of 60%—a percentage that is, for current purposes, intentionally high for PAD.

In order to demonstrate how cutpoint values can influence sensitivity, specificity, and predictive values, in Figures 2 and 3, those metrics have been calculated for TBI cutpoints at 0.55 and 0.85, respectively, while holding PAD prevalence constant. (Note 6) The stringent (i.e., conservative, or low) cutpoint of 0.55 (Figure 2) might be based on a need to identify people for whom immediate referrals should be made, whereas the more precautionary (i.e., tentative, or high) cutpoint of 0.85 (Figure 3) might be set to identify a greater number of people who, less urgently, might undergo further investigative testing, receive advice about lifestyle changes, or be monitored more frequently.

Four conclusions can be drawn from the contents of Figures 2 and 3. First, the cutpoints chosen on a screening test can have noticeable effects on the sensitivity and specificity values associated with that test: sensitivity is 70% in Figure 2, and a much higher 98% in Figure 3; specificity is 90% in Figure 2, but only 38% in Figure 3.

Second, when screening tests involve cutpoints, it is not sufficient to speak of those tests' sensitivity and specificity without also indicating the associated cutpoint. Furthermore, all three variables (sensitivity, specificity, and

cutpoint) should not only be taken into account, but taken into account *simultaneously*, when assessing and reporting the adequacy of a screening test.

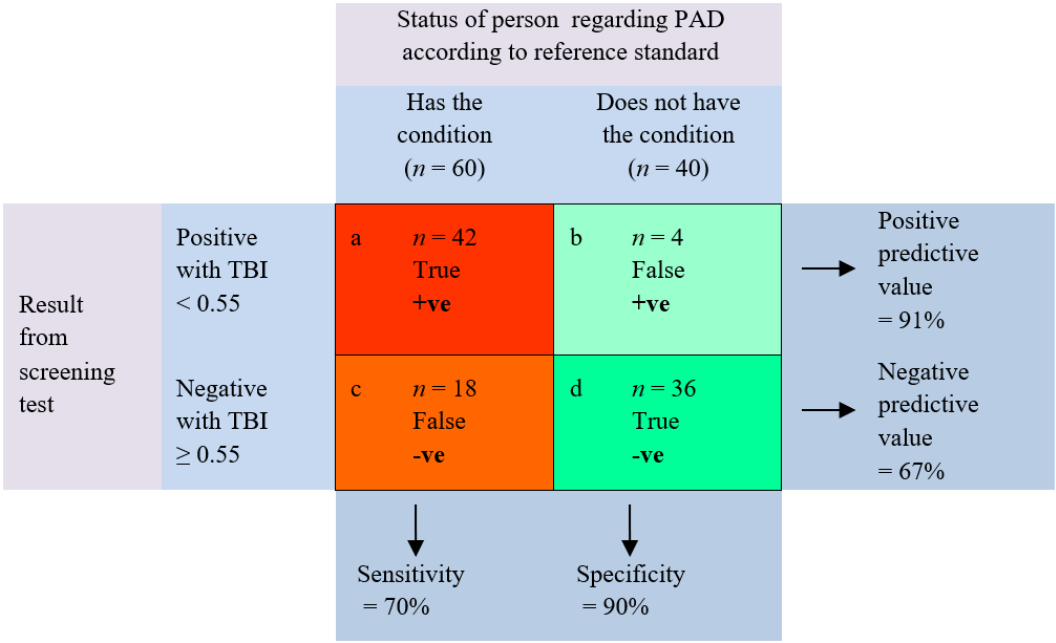

Figure 2. Diagram showing sensitivity, specificity, predictive values, and accuracy based on a fictitious dataset of 100 people, 60 of whom had been diagnosed on a reference standard as having PAD, and 40 as not having PAD. The TBI cutpoint on the screening test is set at 0.55. This depicts a high prevalence (60%), stringent cutpoint situation.

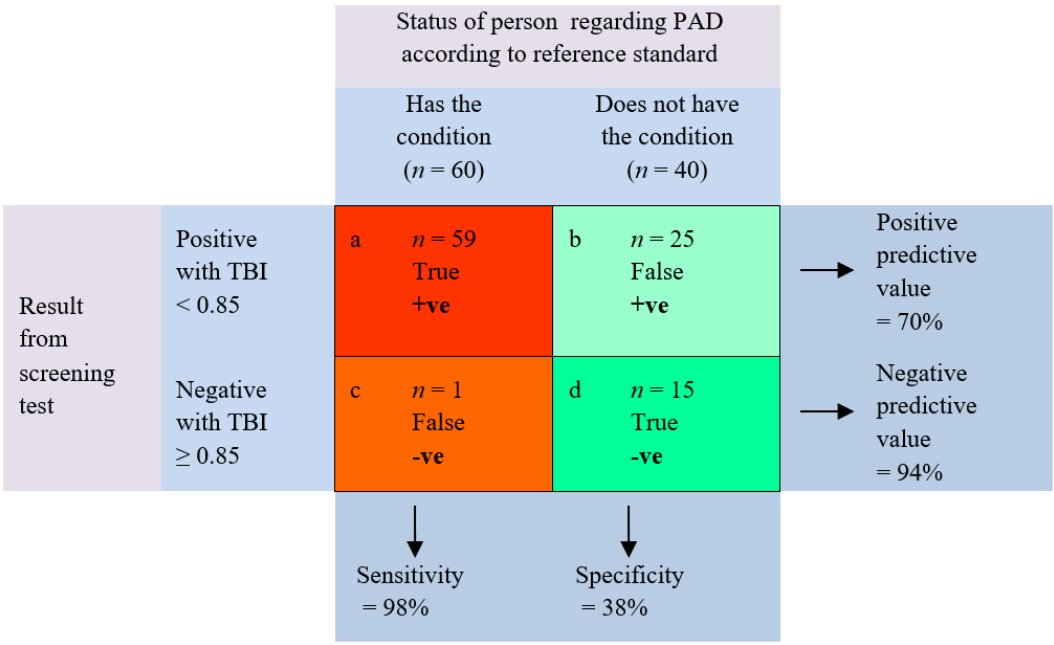

Figure 3. Diagram showing sensitivity, specificity, predictive values, and accuracy based on a fictitious dataset of 100 people, 60 of whom had been diagnosed on a reference standard as having PAD, and 40 as not having PAD. The TBI cutpoint on the screening test is set at 0.85. This depicts a high prevalence (60%), precautionary cutpoint situation.

Third, cutpoints can have noticeable effects on predictive values. The positive predictive value (PPV) is 88% in Figure 2, and 70% in Figure 3; in contrast, the negative predictive value (NPV) is 67% in Figure 2, and 94% in Figure 3.

Fourth, therefore, because clinical decisions are based on predictive values, when cutpoints are involved it is advisable to be aware of the cutpoint that had been used when determining those predictive values.

6. Issues Associated with Prevalence

In addition to screening tests' cutpoints influencing sensitivity, specificity, and predictive values, the sample of people who are recruited for initial analysis has a bearing on those metrics, sometimes to a noticeable extent. (Note 7) There are two distinctly different situations under which the initial samples can be recruited, and those situations have substantial implications concerning how prevalence is taken into account when determining predictive values.

In the first of these situations, a sample is chosen solely to determine the sensitivity and specificity of a screening test (Grunau & Linn, 2018). Here, the only requirement is that some people within the sample can be regarded as definitely having the condition of interest, and the others can be regarded as definitely not having that condition—and that determination can occur regardless of the prevalence of the condition within either the general population or any specific subpopulation of interest. Under these circumstances, calculation of predictive values is almost certainly inappropriate because predictive values are influenced by prevalence, but prevalence might well not have been taken into account and, in fact, was probably *not* taken into account (Molinaro, 2015). In order to calculate predictive values, an appropriate prevalence level needs to be identified and subsequently used in conjunction with Bayes' theorem. In order to obtain the PPV under these circumstances, the formula is:

$$\frac{\text{Sensitivity x prevalence}}{\text{(Sensitivity x prevalence)} + \text{[(1 – specificity) x (1 – prevalence)]}}$$

In order to obtain the NPV, the formula is:

$$\frac{\text{Specificity x (1 – prevalence)}}{\text{[(1 – sensitivity) x prevalence]} + \text{[specificity x (1 – prevalence)]}}$$

In the second situation, a sample is deliberately chosen to correspond, as much as possible, to a particular subpopulation of interest and therefore prevalence will be built into the analyses (Molinaro, 2015) and Bayes' theorem does not need to be applied in order to obtain predictive values. It is this second situation that pertains when creating the data for the entries within Figures 1 to 5 in this article, and therefore calculations of predictive values in those figures—calculations based solely on entries within those figures—can be regarded as valid for illustrative purposes.

The importance of sample prevalence can begin to be appreciated by demonstrating that screening tests based on samples from subpopulations in which a target condition is prevalent are likely to have higher sensitivity and lower specificity than do tests based on samples from subpopulations in which a condition has low prevalence—and vice versa. These have been labeled sensitivity and specificity biases, respectively (Kesson, 2009). They are revealed by comparing entries in Figure 2 with those in Figure 4. In both of those figures, the screening test cutpoint for TBIs is constant at 0.55, but the samples are based on different sample prevalences.

As already noted, the data in Figure 2 are based on a fictitious sample of 100 people at high risk of PAD, in which prevalence was demonstrably high. In order to create a contrast in sample prevalences, another fictitious sample of 100 people was created for Figure 4 (again with recourse to actual data used in previous research to be assured of verisimilitude). This second sample was constructed such that few people were expected to have PAD. These people might, for illustrative purposes, be 100 ostensibly healthy adults between 50 and 70 years of age. The full set of data, with supporting annotations, is provided in Appendix B. Compared with the data in Figure 2, where prevalence was calculated at 60%, prevalence for the data in Figure 4 can be calculated to be only 6%. For this data set, the mean TBI = 0.80, the standard deviation = 0.21, and TBIs range from 0.40 to 1.10.

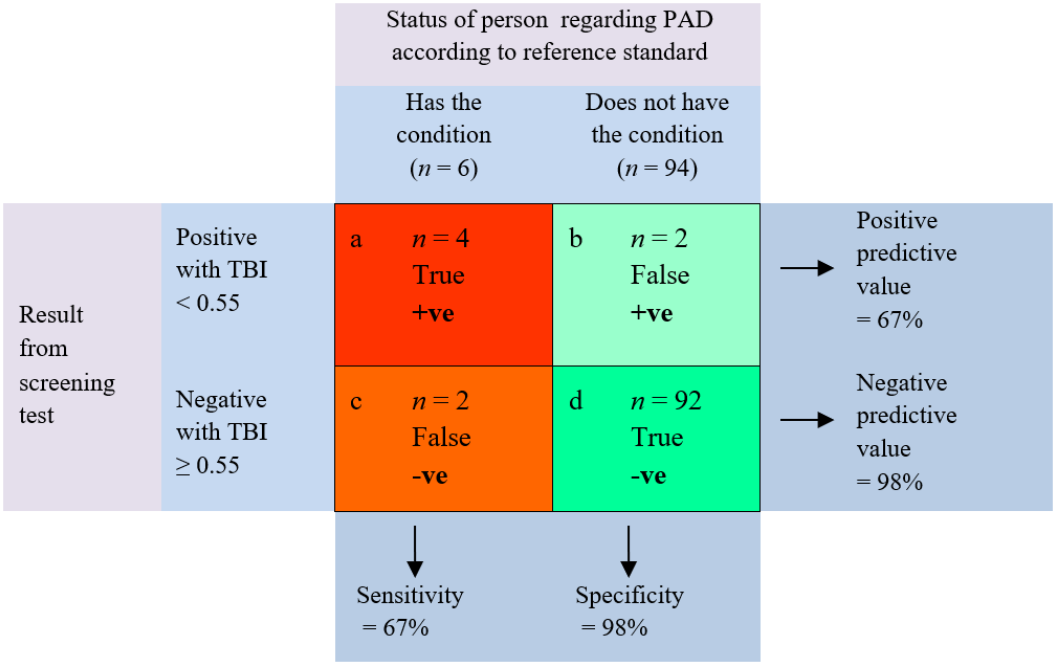

Figure 4. Diagram showing sensitivity, specificity, predictive values, and accuracy based on a fictitious dataset of 100 people, 6 of whom had been diagnosed on a reference standard as having PAD, and 94 as not having PAD. The TBI cutpoint on the screening test is set at 0.55. This depicts a low prevalence (6%), stringent cutpoint situation.

A comparison of entries in Figures 2 and 4 reveals that, at least with the stringent cutpoint of 0.55 held constant, the sensitivity and specificity biases hold true. The sensitivity of 70% associated with high prevalence (Figure 2) is higher than sensitivity of 67% associated with low prevalence (Figure 4), and, conversely, the specificity of 90% associated with high prevalence (Figure 2) is lower than specificity of 98% associated with low prevalence (Figure 4). The differences between the two sensitivity values, and between the two specificity values, are not large, however.

Prevalence differences affect predictive values to a much greater extent. With regard to PPVs, the sample with high prevalence (Figure 2) yielded a considerably higher PPV of 91% relative to the sample with low prevalence (Figure 4) with its PPV of 67%. This difference in PPVs can be explained in two ways. On the one hand, there is more opportunity for true positive results to occur when prevalence is high given that most results fall into cells a and c. Because of that, if a screening test is reasonably effective, cell a will almost inevitably have greater numbers than will cell b, resulting in less opportunity for false positives to occur in cell b. Therefore, the PPV, being calculated from cells a and b, is likely to be quite high. Expressed differently, high prevalence simply means that a person being tested is likely to have the condition of interest and therefore, based on this fact alone, a positive test result is likely to be correct (Loong, 2003). On the other hand, low prevalence of a condition is likely to be associated with smaller PPVs. To begin with, most observations will be located in cells b and d. However, although most of those observations will probably be in cell d if a screening test is reasonably effective, there could well be a proportionally greater number of observations in cell b relative to cell a than when prevalence is high.

The converse occurs with NPVs. The sample with higher prevalence (Figure 2) yielded a considerably lower NPV of 67% than did the sample that had low prevalence (Figure 4) with its NPV of 98%. With appropriate explanatory reversals, the reasons for this difference parallel the reasons provided in the paragraph immediately above for prevalence-influenced differences in PPVs. On the one hand, if a condition is prevalent but the cutpoint for detecting it is stringent (i.e., a condition must be quite pronounced before it is considered to exist), there is likely to be a disproportionate number of false negative results (cell c) because many people with the condition will not be identified by the screening test. That will drive the NPV down. On the other hand, when prevalence is low (and a cutpoint is stringent, as before), there are likely to be relatively few observations in cell c if a screening test is reasonably effective, so an NPV is likely to be higher. Low prevalence simply means that a person being tested is

unlikely to have the condition and therefore, based on this fact alone, a negative test result is likely to be correct (Loong, 2003).

In summary, as prevalence increases, sensitivity and PPVs also increase, but specificity and NPVs decrease—and vice versa. This means that when screening tests are being described and evaluated, the characteristics of the samples on which those tests are predicated should be fully disclosed. Furthermore, it has been argued that when those tests are used in clinical situations, the characteristics of comparison samples should be taken into account because there should be a match between those samples and the people subsequently being screened (Akobeng, 2006; Coulthard, 2007; Kesson, 2009; Mulherin & Miller, 2002; Pewsner et al., 2004; Ray et al., 2010; Vetter et al., 2018).

7. Interactive Effects and Additional Findings Concerning Cutpoints and Prevalence

In the previous two subsections, the analyses were first simplified in that contrasting cutpoints were examined when prevalence was held constant (Figures 2 and 3), then the influence of prevalence was examined while holding a cutpoint constant (Figures 2 and 4). Although those analyses revealed some important insights, additional insights are possible and become evident when cutpoints and prevalences are examined in combination with each other.

Some analysis of the interaction between cutpoints and prevalence is possible by comparing the entries in Figures 3 and 4 because they have dissimilar cutpoints as well as dissimilar prevalence levels (precautionary cutpoint, high prevalence in Figure 3; stringent cutpoint, low prevalence in Figure 4)—and some analysis of entries in those two figures will be made below. However, the prospect of exploring interactive effects, as well as additional unidimensional effects, can be extended by creating Figure 5. In it, the same data that had been generated for Figure 4 were used, but the TBI cutpoint was raised to 0.85 to correspond with the precautionary cutpoint used in Figure 3.

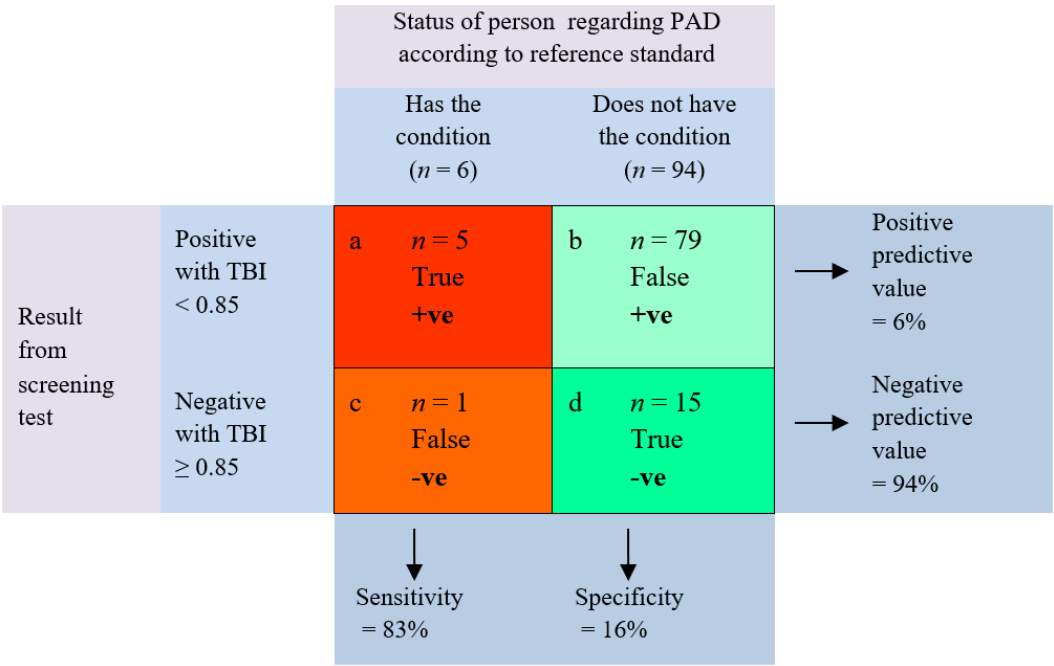

Figure 5. Diagram showing sensitivity, specificity, predictive values, and accuracy based on a fictitious dataset of 100 people, 6 of whom had been diagnosed on a reference standard as having PAD, and 94 as not having PAD. The TBI cutpoint on the screening test is set at 0.85. This depicts a low prevalence (6%), precautionary cutpoint situation.

New sets of comparisons are therefore possible. Figures 2 and 5 have different cutpoints *and* prevalences (stringent cutpoint, high prevalence in Figure 2; precautionary cutpoint, low prevalence in Figure 5), Figures 4 and 5 have different cutpoints but the same prevalence (both low), and Figures 3 and 5 have the same cutpoint (precautionary) but different prevalences. The six sets of comparisons, including those made possible by the creation of Figure 5, are depicted in Figure 6.

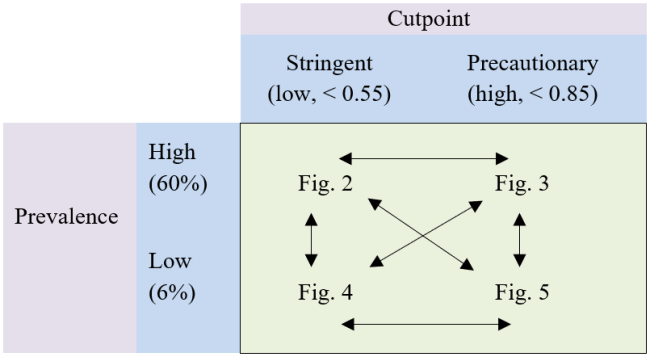

Figure 6. Six sets of comparisons between Figures 2 to 5

The number and variety of comparisons is large when sensitivity, specificity, and predictive values from Figures 2 to 5 are taken into account. A summary is provided in Table 1 where, for subsequent examination and comment, likelihood ratios, and three global/summary metrics comprising test accuracy, the Youden index (Grunau & Linn, 2018; Youden, 1950), and the predictive summary index, or PSI (Grunau & Linn, 2018) have also been included. (Note 8)

Table 1. Sensitivity, specificity, predictive values, and other metrics for fictitious samples that differ on screening test cutpoints and population prevalence[a]

							Global/summary metrics		
Sample	Sensitivity	Specificity	PPV[b]	NPV[c]	PLR[d]	NLR[e]	Accuracy[f]	Youden index[g]	PSI[h]
Stringent cutpoint (0.55)	.70	.90	.91	.67	7.0	0.33	.78	.60	.58
High prevalence (**60%,** Fig. 2)									
Low prevalence (**6%,** Fig. 4)	.67	.98	.67	.98	31.3	0.34	.96	.65	.65
Precautionary cutpoint (0.85)	.98	.38	.70	.94	1.6	0.05	.74	.36	.64
High prevalence (**60%,** Fig. 3)									
Low prevalence (**6%,** Fig. 5)	.83	.16	.06	.94	.99	1.06	.20	-.01	.00

[a] Metrics associated with sensitivity, specificity, and predictive values have been converted from percentages to decimal format in order to conform with the representation of other metrics.
[b] Positive predictive value (PPV) = a / (a + b)
[c] Negative predictive value (NPV) = d / (c + d)
[d] Positive likelihood ratio (PLR) = sensitivity / (1 – specificity)
[e] Negative likelihood ratio (NLR) = (1 – sensitivity) / specificity
[f] Accuracy = (a + d) / (a + b + c + d)
[g] Youden index = sensitivity + (specificity – 1)
[h] Predictive summary index **(**PSI) = PPV + NPV – 1

The entries in Table 1 would inevitably change if different screening test cutpoints and sample prevalences had been used when producing Figures 2 to 5. Because of that, and to avoid overanalysis, most comparisons that follow will focus on entries in Figure 2 versus Figure 5, and on entries in Figure 3 versus Figure 4, because each of those paired comparisons contains differences in *both* cutpoints and prevalences. Although the specific details are inevitably fluid, these comparisons yield general insights and principles.

Entries for Figure 2 indicate that a stringent cutpoint in the presence of high prevalence is likely to produce moderate sensitivity and high specificity as well as a high PPV and a moderate NPV. In contrast, entries for Figure 5 indicate that a precautionary cutpoint in the presence of low prevalence is likely to produce moderate sensitivity and high specificity as well as an extremely low PPV and an extremely high NPV.

Between Figures 2 and 5, there is no consistency regarding either size or pattern of the outcomes for sensitivity, specificity, or predictive values—a point that I made elsewhere (Trevethan, 2017b) but is worth reiterating lest researchers and clinicians make unwarranted assumptions about values of some metrics on the basis of knowledge of values on only one or two other metrics.

Entries for Figure 3 indicate that a precautionary cutpoint in the presence of high prevalence is likely to produce high sensitivity and NPVs, low specificity, and moderate PPVs. In contrast, the entries for Figure 4 reveal that a stringent cutpoint in the presence of low prevalence is likely to produce moderate sensitivity and PPVs in conjunction with extremely high specificity and NPVs. Again, there is no consistency regarding either size or pattern of the outcomes for sensitivity, specificity, or predictive values between the two figures.

The sensitivity and specificity biases that were confirmed when comparing entries for Figures 2 and 4 (i.e., high prevalence is associated with higher sensitivity and lower specificity than is low prevalence, and vice versa) are only partially apparent when comparing entries for Figures 3 and 5. In the latter pair of figures, the sensitivity bias was again evident (sensitivity is higher in Figure 3 than in Figure 5), but the specificity bias was reversed (specificity is higher in Figure 3 than in Figure 5). These results run counter to the sometimes-argued notion that sensitivity and specificity are unaffected by prevalence (see Lalkhen & McCluskey, 2008; Wong & Lim, 2011), a notion that has been disproven elsewhere (Brenner & Gefeller, 1997; Li & Fine, 2011). There is an added revelation, however, in that it is difficult to predict with confidence how prevalence levels will influence sensitivity and specificity.

8. Additional Considerations and Insights

Cutpoints and prevalence are not the only variables that can influence sensitivity, specificity, and predictive values. An additional influence, referred to as spectrum bias or sometimes as spectrum effect (Mulherin & Miller, 2002), can occur if either extremely unwell or extremely healthy people are included in samples when the metrics of a screening test are being established (Al Fattani & Aljoudi, 2015; Goehring et al., 2004; Leeflang et al., 2009; Parikh et al., 2008; Pewsner et al., 2004; Schmidt & Factor, 2013; Šimundić, 2009; Willis, 2008). These situations could be seen as creating extensions of the sensitivity and specificity biases. Within subpopulations that have greater prevalence of a condition, there are likely to be more people who exhibit severe forms of that condition and who are therefore more easily identified, thus creating greater sensitivity because of fewer false negative results. Conversely, extreme healthiness in a subpopulation is likely to lessen the probability of false positive results, and will therefore enhance specificity. Essentially, screening tests are likely to perform, or *appear* to perform, more effectively when used with samples in which the conditions of interest are manifestly expressed as being present or absent, which could be related to (sub)population prevalence.

Likelihood ratios are provided in Table 1 primarily for completeness in case readers are interested in how those ratios relate to entries in Figures 2 to 5. Because of evidence that likelihood ratios are seldom used by health professionals (Puhan et al., 2005; Reid et al., 1998; Whiting, Davenport, et al., 2015), these ratios are accorded minimal attention here. However, it is probably helpful to acknowledge that, as positive likelihood ratios rise above 1.0, the more they are likely to indicate presence of a condition, and as negative likelihood ratios fall below 1.0, the more they are likely to indicate absence of a condition—but also that being aware of these trends might not be particularly informative. More helpful is knowing that positive likelihood ratios convincingly indicate presence of a condition only when > 10.0, and negative likelihood ratios convincingly indicate absence of a condition only when < 0.10 (Ray et al., 2010; Šimundić, 2009; Vetter et al., 2018). Rules of thumb for interpreting likelihood ratios are provided in Table 2 because guidelines of that kind seem to be elusive.

Table 2. Rules of thumb for interpreting likelihood ratios in terms of their predictive strength[a]

Predictive strength	Positive likelihood ratio	Negative likelihood ratio
None	1	1
Poor	1 – 5	1 – 0.2
Good	5 – 10	0.2 – 0.1
Excellent	> 10	< 0.1

[a.] Adapted from Table 3 of Ray et al. (2010)

An interesting relationship of likelihood ratios with sensitivity and specificity is observable in Table 1 where the only positive likelihood ratio that exceeds 10 is associated with high specificity, providing support for the spin principle, and the only negative likelihood ratio that is < 0.10 is associated with high sensitivity, providing support for the snout principle. This indicates that the useful information provided by likelihood ratios is already available from sensitivity and specificity values.

Entries in Table 1 for the three global/summary metrics comprising accuracy, the Youden index, and the PSI indicate that accuracy is always higher than the other two. Apart from that, however, there is almost no systematic pattern evident in the relationships among these metrics.

These discrepancies suggest that global metrics might not be particularly informative in clinical situations because each can be influenced in different ways by cutpoints and prevalence and therefore can mask unique features of, and important differences between, sensitivity, specificity, predictive values, and likelihood ratios. By masking these unique features, global/summary metrics might not be merely uninformative. They might be deceptive.

Reservations about these global/summary metrics have been raised elsewhere. Accuracy has been referred to by a number of authors as needing to be viewed with caution (Ray et al., 2010; Šimundić, 2009; Whiting, Davenport, et al., 2015; Zhu et al., 2010). Added to that, Youden's index has been characterized as insensitive to differences in sensitivity and specificity (Šimundić, 2009), and the PSI has been described as most effective when sensitivity and specificity are both high and prevalence is 50% (Irving & Holden, 2013), thus severely limiting its applicability because screening tests are typically unsatisfactory with regard to either sensitivity or specificity, or to both of those metrics, and prevalence of a condition is seldom 50%. Furthermore, there seem to be no guidelines concerning when any of these three global/summary metrics can be regarded as indicating that a test is satisfactory, thus further depriving them of usefulness. Entries in Table 1 demonstrate that it would be unwise to make assumptions about any one of sensitivity, specificity, predictive values, or likelihood ratios associated with a test based merely on knowledge about any others among those metrics, let alone based on global/summary metrics such as accuracy, the Youden index, and the PSI.

9. Concluding Remarks

Although different metrics would inevitably be obtained with other data than were used in this article, a number of overarching points can be made. Predominant among these is that the foundations of screening tests, and the metrics associated with those tests, are by no means simple and that some degree of cautious sophistication needs to be exercised when evaluating and using them. In particular, the metrics associated with screening tests can be substantially influenced not only by where cutpoints are placed but also by, and in interaction with, the prevalence of a condition in the samples chosen for analysis. Therefore, when the characteristics of a screening test are being described, information should be provided about the cutpoints that had been applied *and* the prevalence levels of the target condition in the sample, or subpopulation, with which the test's sensitivity, specificity, and predictive values were determined.

The contents of this article should also indicate that an informed skeptical stance is reasonable with regard to most metrics associated with screening tests. For those metrics to be relied on, the validity of both reference standards and the screening tests themselves needs to be guaranteed as much as possible, or at least taken into account. Reference standards can fail to be as "golden" as is often assumed, and screening tests, ipso facto, are likely to have inbuilt uncertainty and inaccuracy.

Furthermore, in the Wilson and Jungner (1968) context of case findings, which is the focus of this article, Willis (2008) has pointed out that it might be difficult to attain a satisfactory degree of correspondence between a particular patient and the group(s) on which a test's PPV and NPV were determined. For example, relevant prevalence data might not be available in relation to a particular female patient of a specific age and ethnic background who had never smoked but currently has hypertension and symptoms of prediabetes. Therefore perhaps, in reality, it is only *best approximations* of prevalence that can be taken into account in decision making, and consequently predictive values should not be expected to attract high levels of certainty.

It is also important to recognize that overreliance on, or preoccupation with, screening test metrics could result in losing sight of the reality that many conditions are not dichotomous in nature, but rather lie on a continuum (see Trevethan, 2019). Appropriate decisions are therefore unlikely to be based on two clearly demarcated sets of considerations. For example, a different range of possibilities will be considered, and decisions made, when readings are extreme than when readings are only moderately disquieting, or when readings lie only a little above or a little below a specified cutpoint.

In light of these considerations, it should be acknowledged that effective clinicians are likely to use metrics associated with screening tests in order to complement, rather than dominate, decision making; not feel overly concerned if they disregard metrics beyond sensitivity, specificity, and predictive values; not rely on a single screening test, particularly if results from it are ambiguous; and engage in careful history taking and observation of signs and symptoms as part of a comprehensive assessment of each patient. There is, after all, an art as well as a science to good practice (Saunders, 2000).

Acknowledgements

Sylvia McAra and Rod Pope provided valuable feedback on earlier drafts of this manuscript.

Compliance with Ethical Standards

Conflicts of Interest

None.

Human and Animal Rights

This article does not contain any studies with human participants performed by the author.

Notes

1. It is revealing that, at the time of writing, the Wikipedia site for sensitivity and specificity is prefaced with a request that the site be written in a more comprehensible way, and almost all subsections within that site are indicated as being in need of editing.
2. The higher blood pressure at the ankle than at the brachium occurs because the vascular bed of the foot creates a reflection that increases blood pressure at the ankle (Singh et al., 2013).
3. Calculations to obtain these metrics are provided in the footnotes for Table 1.
4. All percentages within this article have been rounded to the nearest integer.
5. In this context, and elsewhere in this article, test accuracy is based specifically on the following formula in the standard 2 x 2 contingency table: (a + d) / (a + b + c + d).
6. Confidence intervals are not provided in this article in order to avoid complexity that might be counterproductive. Confidence intervals for the main metrics associated with screening tests are easily obtained from the following website: https://www.medcalc.org/calc/diagnostic_test.php
7. There is a common belief that sensitivity and specificity are not affected by prevalence, but that is not always the case. This will be considered in more detail, and demonstrated, later in this article.
8. A number of additional metrics can be derived from entries in the basic 2 x 2 contingency table. They are not dealt with in this article because they appear to attract little attention in clinical and research contexts.

References

Aboyans, V., Criqui, M. H., Abraham, P., Allison, M. A., Creager, M. A., Diehm, C., … Treat-Jacobson, D. (2012). Measurement and interpretation of the ankle-brachial index: A scientific statement from the American Heart Association. *Circulation, 126*, 2890–2909. https://doi.org/10.1161/CIR.0b013e318276fbcb

Akobeng, A. K. (2006). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Paediatrica, 96*, 338–341. https://doi.org/10.1111/j.1651-2227.2006.00180

Al Fattani, A. A., & Aljoudi, A. (2015). Sources of bias in diagnostic accuracy studies. *Journal of Applied Hematology, 6*, 178–180. https://doi.org/10.4103/1658-5127.171991

Al-Qaisi, M., Nott, D. M., King, D. H., & Kaddoura, S. (2009). Ankle brachial pressure index (ABPI): An update for practitioners. *Vascular Health and Risk Managment, 5*, 833–841. https://doi.org/10.2147/VHRM.S6759

Bhamidipaty, V., Dean, A., Yap, S. L., Firth, J., Barron, M., Allard, B., & Chan, S. T. F. (2015). Second toe systolic pressure measurements are valid substitutes for first toe systolic pressure measurements in diabetic patients: A prospective study. *European Journal of Vascular and Endovascular Surgery, 49*, 77–82. https://doi.org/10.1016/j.ejvs.2014.09.011

Bonham, P. A. (2011). Measuring toe pressures using a portable photoplethysmograph to detect arterial disease in high risk patients: An overview of the literature. *Ostomy Wound Management, 57*, 36–44. https://www.o-wm.com/files/owm/pdfs/OWM_November2011_Bonham.pdf

Bossuyt, P. M., Reitsma, J. B., Bruns D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Clinical Chemistry, 49*, 1–6. https://doi.org/10.1373/49.1.1

Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine, 16*, 981–991. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N

Bundó, M., Urrea, M., Muñoz, L., Llussà, J., Forés, R., & Torán, P. (2013). [Correlation between toe-brachial index and ankle-brachial index in patients with diabetes mellitus type 2]. *Medicina Clinica (Barc), 140*, 390–394. Spanish. https://doi.org/10.1016/j.medcli.2012.03.012

Cadogan, A., McNair, P., Laslett, M., & Hing, W. (2013). Shoulder pain in primary care: Diagnostic accuracy of clinical examination tests for non-traumatic acromioclavicular joint pain. *BMC Musculoskeletal Disorders, 14*, 156–166. https://doi.org/10.1186/1471-2474-14-156

Cadogan, A., McNair, P., Laslett, M., Hing, W., & Taylor, S. (2013). Diagnostic accuracy of clinical examination features for identifying large rotator cuff tears in primary health care. *Journal of Manual and Manipulative Therapy, 21*, 148–159. https://doi.org/10.1179/2042618612Y.0000000020

Campbell, N., Chockalingam, A., Fodor, J. G., & McKay, D. W. (1990). Accurate, reproducible measurement of blood pressure. *Canadian Medical Association Journal, 143*, 19–24. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1452062/

Cardenas, V., Seo, K., Sheth, S., & Meyr, A. J. (2018). Prevalence of lower-extremity arterial calcification in patients with diabetes mellitus complicated by foot disease at an urban US tertiary-care center. *Journal of the American Podiatric Medical Association, 108*, 267–271. https://doi.org/10.7547/16-075

Caro, J., Migliaccio-Walle, K., Ishak, K., & Proskorovsky, I. (2005). The morbidity and mortality following a diagnosis of peripheral arterial disease: Long-term follow-up of a large database. *BMC Cardiovascular Disorders, 5*, 14. https://www.ncbi.nlm.nih.gov/pubmed/15972099

Caruana, M. F., Bradbury, A. W., & Adam, D. (2005). The validity, reliability, reproducibility and extended utility of ankle to brachial pressure index in current vascular surgical practice. *European Journal of Vascular and Endovascular Surgery, 29*, 443–451. https://doi.org/10.1016/j.ejvs.2005.01.015

Chen, J., Mohler, E. R. 3rd, Garimella, P. S., Hamm, L. L., Xie, D., Kimmel, S., Townsend, R. R., … He, J. (2016). Ankle brachial index and subsequent cardiovascular disease in patients with chronic kidney disease. *Journal of the American Heart Association, 5*(6), pii:e003339. https://doi.org/10.1161/JAHA.116.003339

Chongthawonsatid, S., & Dutsadeevettakul, S. (2017). Validity and reliability of the ankle-brachial index by oscillometric blood pressure and automated ankle-brachial index. *Journal of Research in Medical Science, 22*, 44. https://doi.org/10.4103/jrms.JRMS_728_16

Coulthard, M. G. (2007). Quantifying how tests reduce diagnostic uncertainty. *Archives of Disease in Childhood, 92*, 404–408. https://doi.org/10.1136/adc.2006.111633

Criqui, M. H., & Aboyans, V. (2015). Epidemiology of peripheral artery disease. *Circulation Research, 116*, 1509–1526. https://www.ahajournals.org/doi/abs/10.1161/CIRCRESAHA.116.303849?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed&

Criqui, M. H., McClelland, R. L., McDermott, M. M., Allison, M. A., Blumenthal, R. S., Aboyans V., … Shea S. (2010). The ankle-brachial index and incident cardiovascular events in the MESA (Multi-Ethnic Study of Atherosclerosis). *Journal of the American College of Cardiology, 56*, 1506–1512. https://doi.org/10.1016/j.jacc.2010.04.060

European Stroke Organisation, Tendera, M., Aboyans, V., Bartelink, M. L., Baumgartner, I., Clément, D., … Van Damme, H. (2011). ESC guidelines on the diagnosis and treatment of peripheral artery diseases: Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteries: The Task Force on the Diagnosis and Treatment of Peripheral Artery Diseases of the European Society of Cardiology (ESC). *European Heart Journal, 32*, 2851–2906. https://doi.org/10.1093/eurheartj/ehr211

Formosa, C., Ellul, C., Mizzi, A., Mizzi, S., & Gatt, A. (2018). Interrater reliability of spectral Doppler waveform analysis among podiatric clinicians. *Journal of the American Podiatric Medical Association, 108*, 280–284. https://doi.org/10.7547/16-026

Goehring, C., Perrier, A., & Morabia, A. (2004). Spectrum bias: A quantitative and graphical analysis of the variability of medical diagnostic test performance. *Statistics in Medicine, 23*, 125–135. https://www.ncbi.nlm.nih.gov/pubmed/14695644

Goldstein, L. N., Wells, M., & Sliwa, K. (2014). Blood pressure measurements in the ankle are not equivalent to blood pressure measurements in the arm. *South African Medical Journal, 104*, 869–873. https://doi.org/10.7196/SAMJ.8102

Gong, Y., Cao, K. W., Xu, J. S., Li, J. X., Hong, K., Cheng, X. S., & Su, H. (2015). Valuation of normal range of ankle systolic blood pressure in subjects with normal arm systolic blood pressure. *PLoS One., 10*(6), e0122248. https://doi.org/10.1371/journal.pone.0122248

Gornik, H. L. (2009). Rethinking the morbidity of peripheral arterial disease and the "normal" ankle-brachial index. *Journal of the American College of Cardiology, 53*, 1063–1064. https://doi.org/10.1016/j.jacc.2008.12.019

Grunau, G. L., & Linn, S. (2018). Commentary: Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 6,* Article 256. https://doi.org/10.3389/fpubh.2018.00256

Hinchliffe, R. J., Brownrigg, J. R. W., Apelqvist, J., Boyko, E. J., Fitridge, R., Mills, J. L., … Schaper, N. C. (2016). IWGDF guidance on the diagnosis, prognosis and management of peripheral artery disease in patients with foot ulcers in diabetes. *Diabetes Metabolism Research Reviews, 32*(Suppl 1), 37–44. https://doi.org/10.1002/dmrr.2698

Høyer, C., Sandermann, J., & Petersen, L. J. (2013). The toe-brachial index in the diagnosis of peripheral arterial disease. *Journal of Vascular Surgery, 58*, 231–238. https://doi.org/10.1016/j.jvs.2013.03.044

Irving, G., & Holden, J. (2013). The time-efficiency principle: Time as the key diagnostic strategy in primary care. *Family Practice, 30*, 386–389. https://doi.org/10.1093/fampra/cmt007

Ix, J. H., Miller, R. G., Criqui, M. H., & Orchard, T. J. (2012). Test characteristics of the ankle-brachial index and ankle-brachial difference for medial arterial calcification on X-ray in type 1 diabetes. *Journal of Vascular Surgery, 56*, 721–727. https://doi.org/10.1016/j.jvs.2012.02.042

Jelinek, H. F., & Austin, M. (2006). The ankle–brachial index in clinical decision making. *Foot (Edinb), 16*, 153–157. https://doi.org/10.1016/j.foot.2006.04.003

Jönelid, B., Johnston, N., Berglund, L., Andrén, B., Kragsterman, B., & Christersson, C. (2016). Ankle brachial index most important to identify polyvascular disease in patients with non-ST elevation or ST-elevation myocardial infarction. *European Journal of Internal Medicine, 30*, 55–60. https://doi.org/10.1016/j.ejim.2015.12.016

Kesson, A. M. (2009). Predictive values, sensitivity and specificity in clinical virology. Retrieved June 8, 2019 from http://www.virologyresearch.unsw.edu.au/virology/wp-content/uploads/2013/08/VIM09_AlisonKesson_ClinicalVirology.pdf

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain, 8*, 221–223. https://doi.org/10.1093/bjaceaccp/mkn041

Leeflang, M. M., Bossuyt, P. M., & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: Implications for evidence-based diagnosis. *Journal of Clinical Epidemiology, 62*, 5–12. https://doi.org/10.1016/j.jclinepi.2008.04.007

Lewis, J. E. A., Williams, P., & Davies, J. H. (2016). Non-invasive assessment of peripheral arterial disease: Automated ankle brachial index measurement and pulse volume analysis compared to duplex scan. *SAGE Open Medicine, 4*. https://doi.org/10.1177/2050312116659088

Li, J., & Fine, J. P. (2011). Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics*, *12*, 710–722. https://doi.org/10.1093/biostatistics/kxr008

Loong, T. W. *(*2003). Understanding sensitivity and specificity with the right side of the brain. *BMJ., 327*, 716–719. https://doi.org/10.1136/bmj.327.7417.716

Manrai, A. K., Bhatia, G., Strymish, J., Kohane, I. S., & Jain, S. H. (2014). Medicine's uncomfortable relationship with math: Calculating positive predictive value. *JAMA Internal Medicine, 174*, 991–993. https://doi.org/10.1001/jamainternmed.2014.1059

Mätzke, S., Franckena, M., Albäck, A., Railo, M., & Lepäntalo, M. (2003). Ankle brachial index measurements in critical leg ischaemia – The influence of experience on reproducibility. *Scandinavian Journal of Surgery, 92*, 144–147. https://doi.org/10.1177/145749690309200206

McAra, S. (2015). *Glyceryl trinitrate and toe–brachial indexes in pedal ischaemia*. (Unpublished doctoral dissertation). Charles Sturt University. Albury, NSW, Australia. Retrieved from https://researchoutput.csu.edu.au/en/publications/glyceryl-trinitrate-and-toe-brachial-indexes-in-pedal-ischaemia-3

McAra, S., & Trevethan, R. (2018). Measurement of toe-brachial indices in people with subnormal toe pressures: Complexities and revelations. *Journal of the American Podiatric Medical Association, 108*, 115–125. https://doi.org/10.7547/16-036

McAra, S., Trevethan, R., Wang, L., & Tinley, P. (2017). Vascular screening of the foot: For life and limb. *Diabetes and Primary Care Australia*, *2*(1), 16–24. http://pcdsa.com.au/wp-content/uploads/2016/12/DPCA2-1_16%E2%80%9324_wm.pdf

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica (Zagreb), 22*, 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Mills, J. L. Sr, Conte, M. S., Armstrong, D. G., Pomposelli, F. B., Schanzer, A., Sidawy, A. N., & Andros, G. (2014). The Society for Vascular Surgery Lower Extremity Threatened Limb Classification System: Risk stratification based on wound, ischemia, and foot infection (WIfI). *Journal of Vascular Surgery, 59*, 220–234. e1-2. https://doi.org/10.1016/j.jvs.2013.08.003

Molinaro, A. M. *(2015).* Diagnostic tests: How to estimate the positive predictive value. *Neuro-Oncology Practice, 2*, 162–166. https://doi.org/10.1093/nop/npv030

Mulherin, S. A., & Miller W. C. (2002). Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine, 137*, 598–602. https://www.ncbi.nlm.nih.gov/pubmed/12353947

Nishimura, H., Miura, T., Minamisawa, M., Ueki, Y., Abe, N., Hashizume, N., … Kuwahara, K. (2016). Clinical characteristics and outcomes of patients with high ankle-brachial index from the IMPACT-ABI study. *PLoS ONE, 11*(11), e0167150. https://doi.org/10.1371/journal.pone.0167150

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.

Okada, R., Yasuda, Y., Tsushita, K., Wakai, K., Hamajima, N., & Matsuo, S. (2015). Within-visit blood pressure variability is associated with prediabetes and diabetes. *Scientific Reports, 5*, Article number 7964. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4295087/

Okamoto, K., Oka, J., Maesato, K., Ikee, R., Mano, T., Moriya, H., … Kobayashi, S. (2006). Peripheral arterial occlusive disease is more prevalent in patients with hemodialysis: Comparison with findings of multidetector-row computed tomography. *American Journal of Kidney Diseases, 48*, 269–276. https://doi.org/10.1053/j.ajkd.2006.04.075

Påhlsson, H. I., Laskar, C., Stark, K., Andersson, A., Jogestrand, T., & Wahlberg, E. (2007). The optimal cuff width for measuring toe blood pressure. *Angiology, 58*, 472–476. https://doi.org/10.1177/0003319706294606

Parati, G., Ochoa, J. E., Salvi, P., Lombardi, C., & Bilo, G. (2013). Prognostic value of blood pressure variability and average blood pressure levels in patients with hypertension and diabetes. *Diabetes Care, 36*(Supplement 2), S312-S324. https://doi.org/10.2337/dcS13-2043

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*, 45–50. https://doi.org/10.4103/0301-4738.37595

Pérez-Martin, A., Meyer, G., Demattei, C., Böge, G., Laroche, J-P., Quéré, I., & Dauzat, M. (2010). Validation of a fully automatic photoplethysmographic device for toe blood pressure measurement. *European Journal of Vascular and Endovascular Surgery, 40*, 515–520. https://doi.org/10.1016/j.ejvs.2010.06.008

Pewsner, D., Battaglia, M., Minder, C., Marx, A., Bucher, H. C., & Egger, M. (2004). Ruling a diagnosis in or out with "SpPIn" and "SnNOut": A note of caution. *BMJ, 329*, 209–213. https://doi.org/10.1136/bmj.329.7459.209

Portney, L. G., & Watkins, M. P. (2009). Foundations of clinical research: Applications to practice (3rd. ed). Upper Saddle River, NJ: Pearson Education.

Puhan, M. A., Steurer, J., Bachmann, L. M., & ter Riet, G. (2005). A randomized trial of ways to describe test accuracy: The effect on physicians' post-test probability estimates. *Annals of Internal Medicine, 143*, 184–189. https://doi.org/10.7326/0003-4819-143-3-200508020-00004

Quong, W. L., Fung, A. T., Yu, R. Y., & Hsaing, Y. N. H. (2016). Reassessing the normal toe-brachial index in young healthy adults. *Journal of Vascular Surgery, 63*, 652–656. https://doi.org/10.1016/j.jvs.2015.09.019

Ray, P., Le Manach, Y., Riou, B., & Houle, T. T. (2010). Statistical evaluation of a biomarker. *Anesthesiology, 112*, 1023–1040. https://doi.org/10.1097/ALN.0b013e3181d47604

Reid, M. C., Lane, D. A., & Feinstein, A. R. (1998). Academic calculations versus clinical judgments, practicing physicians' use of quantitative measures of test accuracy. *American Journal of Medicine, 104*, 374–380. https://doi.org/10.1016/S0002-9343(98)00054-0

Rich, K. (2015). Toe blood pressure and toe-brachial index. *Journal of Vascular Nursing., 33*, 164–166. https://doi.org/10.1016/j.jvn.2015.09.002

Romanos, M. T., Raspovic, A., & Perrin, B. M. (2010). The reliability of toe systolic pressure and the toe brachial index in patients with diabetes. *Journal of Foot and Ankle Research, 3*, 31. https://doi.org/10.1186/1757-1146-3-31

Rooke, T. W., Hirsch, A. T., Misra, S., Sidawy, A. N., Beckman, J. A., Findeiss L. K., … Zierler, R. E. (2011). ACCF/AHA Focused update of the guideline for the management of patients with peripheral artery disease. *Circulation, 124*, 2020–2045. https://doi.org/10.1161/CIR.0b013e31822e80c3

Saunders, J. (2000). The practice of clinical medicine as an art and as a science. *Medical Humanities, 26*(1). https://doi.org/10.1136/mh.26.1.18

Sawka, A., & Carter, S. A. (1992). Effect of temperature on digital systolic pressures in lower limb in arterial disease. *Circulation, 85*, 1097–1101. https://doi.org/10.1161/01.CIR.85.3.1097

Schmidt, R. L., & Factor, R. E. (2013). Understanding sources of bias in diagnostic accuracy studies. *Archives of Pathology and Laboratory Medicine, 137*, 558–565. https://doi.org/10.5858/arpa.2012-0198-RA

Šimundić, A-M. *(*2009). Measures of diagnostic accuracy: Basic definitions. *Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine, 19*, 203–211. https://www.ncbi.nlm.nih.gov/pubmed/27683318

Singh, S., Bailey, K. R., & Kullo, I. J. (2013). Ethnic differences in ankle brachial index are present in middle-aged individuals without peripheral arterial disease. *International Journal of Cardiology, 162*, 228–233. https://doi.org/10.1016/j.ijcard.2011.05.068

Sonter, J., Chuter, V., & Casey, S. (2015). Intratester and intertester reliability of toe pressure measurements in people with and without diabetes performed by podiatric physicians. *Journal of the American Podiatric Medical Association., 105*, 201–208. https://doi.org/10.7547/0003-0538-105.3.201

Sonter, J., Sadler, S., & Chuter, V. (2015). Inter-rater reliability of automated devices for measurement of toe systolic blood pressure and the toe brachial index. *Blood Pressure Monitoring, 20*, 47–51. https://doi.org/10.1097/MBP.0000000000000083

Sonter, J., Tehan, P., & Chuter, V. (2017). Toe brachial index measured by automated device compared to duplex ultrasonography for detecting peripheral arterial disease in older people. *Vascular, 25*, 612–617. https://doi.org/10.1177/1708538117705293

Steurer, J., Fischer, JE., Bachmann, LM., Koller, M., *&* ter Riet, G. (2002). Communicating accuracy of tests to general practitioners: A controlled study. *BMJ, 324*(7341), 824–826. https://doi.org/10.1136/bmj.324.7341.824

Suominen, V., Rantanen, T., Vanermo, M., Saarinen, J., & Salenius J. (2008). Prevalence and risk factors of PAD among patients with elevated ABI. *European Journal of Vascular and Endovascular Surgery, 35*, 709–714. https://doi.org/10.1016/j.ejvs.2008.01.013

Suzuki, K. (2007). How to diagnose peripheral artery disease. *Podiatry Today, 20*(4), 54–65. http://www.podiatrytoday.com/article/6952

Tehan, P. E., Bray, A., & Chuter, V. H. (2016). Non-invasive vascular assessment in the foot with diabetes: Sensitivity and specificity of the ankle brachial index, toe brachial index and continuous wave Doppler for detecting peripheral arterial disease. *Journal of Diabetes and its Complications, 30*, 155–160. https://doi.org/10.1016/j.jdiacomp.2015.07.019Tehan, P. E., Santos, D., & Chuter, V. H. (2016). A systematic review of the sensitivity and specificity of the toe-brachial index for detecting peripheral artery disease. *Vascular Medicine, 21*, 382–389. https://doi.org/10.1177/1358863X16645854

Tehan, P., Bray, A., Keech, R., Rounsley, R., Carruthers, A., & Chuter, V. H. (2015). Sensitivity and specificity of the toe brachial index for detecting peripheral arterial disease: Initial findings. *Journal of Ultrasound in Medicine, 34*, 1737–1743. https://doi.org/10.7863/ultra.15.14.09071

Trevethan, R. (2017a). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services Outcomes Research Methodology, 17*, 127–143. https://doi.org/10.1007/s10742-016-0156-6

Trevethan, R. (2017b). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*, Article 307. https://doi.org/10.3389/fpubh.2017.00307

Trevethan, R. (2018). Subjecting the ankle-brachial index to timely scrutiny: Is it time to say goodbye to the ABI? *Scandinavian Journal of Clinical and Laboratory Investigation, 78*, 94–101. https://doi.org/10.1080/00365513.2017.1416665

Trevethan, R. (2019). Consistency of toe systolic pressures, brachial systolic pressures, and toe-brachial indices in people with and without diabetes. *Current Diabetes Reviews, 15*, 85–92. https://doi.org/10.2174/1573399814666180123113619

Trevethan, R. (2019). Toe systolic pressures and toe-brachial indices: Uses, abuses, and shades of gray. *Blood Pressure Monitorin*g, *24*, 45–51. https://doi.org/10.1097/MBP.0000000000000372

Vetter, T. R., Schober P., & Mascha, E. J. (2018). Diagnostic testing and decision-making: Beauty is not just in the eye of the beholder. *Anesthesia and Analgesia, 127*, 1085–1091. https://doi.org/10.1213/ANE.0000000000003698

Watanabe, Y., Masaki, H., Yunoki, Y., Tabuchi, A., Morita, I., Mohri, S., & Tanemoto, K. (2015). Ankle-brachial index, toe-brachial index, and pulse volume recording in healthy young adults. *Annals of Vascular Diseases, 8*, 227–235. https://doi.org/10.3400/avd.oa.15-00056

Whiting, P. F., Davenport, C., Jameson, C., Burke, M., Sterne, J. A. C., Hyde, C., & Ben-Shlomo, Y. (2015). How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open, 5*(7), e008155. https://doi.org/10.1136/bmjopen-2015-008155

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., … QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*, 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Willis, B. H. (2008). Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. *Family Practice*, *25*, 390–396. https://doi.org/10.1093/fampra/cmn051

Wilson, J. M. G., & Jungner, G. (1968). *Principles and practice of screening for disease*. Geneva: World Health Organization. Retrieved from http://www.who.int/iris/handle/10665/37650

Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. *Proceedings of Singapore Healthcare, 40*, 316–318. https://doi.org/10.1177/201010581102000411

Xu, D., Li, J., Zou, L., Xu, Y., Hu, D., Pagoto, S., & Ma, Y. (2010). Sensitivity and specificity of the ankle–brachial index to diagnose peripheral artery disease: A structured review. *Vascular Medicine, 15*, 361–369. https://doi.org/10.1177/1358863x10378376

Youden, W. J. *(1950). Index for rating diagnostic tests. Cancer, 3,* 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

Young, M. J., Adams, J. E., Anderson, G. F., Boulton, A. J. M., & Cavanagh, P. R. (1993). Medial arterial calcification in the feet of diabetic patients and matched non-diabetic control subjects. *Diabetologia, 36*, 615–621. https://doi.org/10.1007/BF00404070

Zhu, W., Zeng, N., & Wang, N. (2010). *Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with Practical SAS® implementations*. Retrieved June 7 2019, from http://www.lexjansen.com/nesug/nesug10/hl/hl07.pdf

Appendix A

Toe–brachial indices for data in Figures 2 and 3

As indicated in the text, the toe–brachial index (TBI) values in the following table are based on a hypothetical sample of 100 people who were considered to be at high risk of PAD, namely people over 75 years of age, some of whom are healthy but many of whom have combinations of diabetes, hypertension, hyperlipidemia, obesity, a history of smoking, chronic kidney disease, symptoms of intermittent claudication, and leg pain at rest. This data set is the basis for entries in Figures 2 and 3. Among these people, differing numbers had TBIs indicative of PAD depending on where the TBI cutoff is placed –at either 0.55 or at 0.85.

	TBI value	Reference standard: Has condition = 1 Does not have cond. = 2	Notes
1.	.14	1	
2.	.15	1	
3.	.15	1	
4.	.16	1	
5.	.17	1	
6.	.20	1	
7.	.21	1	
8.	.22	1	
9.	.22	1	
10.	.23	1	
11.	.24	1	
12.	.25	1	
13.	.26	1	
14.	.27	1	
15.	.28	1	
16.	.29	1	
17.	.31	1	
18.	.32	1	
19.	.33	1	
20.	.34	1	
21.	.35	1	
22.	.36	1	
23.	.36	1	
24.	.37	1	
25.	.37	1	
26.	.39	1	
27.	.40	1	
28.	.41	1	
29.	.42	1	
30.	.42	1	
31.	.42	1	
32.	.43	1	
33.	.44	1	
34.	.45	2	
35.	.45	1	
36.	.46	1	
37.	.46	1	
38.	.47	1	
39.	.49	1	
40.	.51	2	
41.	.51	1	
42.	.51	1	
43.	.51	2	

44.	.52	1	
45.	.53	1	
46.	.54	2	*The following notes refer to Figure 2.*
47.	.55	1	Above here in the table there are 42 x 1s and 4 x 2s ($n = 46$). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.55, there will be 42 true positives (the 1s above here in the table) and 4 false positives (the 2s above here in the table). At or below here in the table there are 18 x 1s and 36 x 2s ($n = 54$). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.55, there will be 18 false negatives (the 1s below here in the table) and 36 true negatives (the 2s below here in the table).
48.	.56	1	
49.	.57	1	
50.	.58	1	
51.	.58	1	
52.	.59	1	
53.	.59	1	
54.	.60	2	
55.	.61	1	
56.	.62	1	
57.	.63	2	
58.	.64	1	
59.	.65	1	
60.	.65	2	
61.	.66	1	
62.	.66	1	
63.	.67	2	
64.	.68	1	
65.	.69	2	
66.	.59	2	
67.	.70	2	
68.	.70	2	
69.	.71	1	
70.	.71	2	
71.	.72	1	
72.	.72	2	
73.	.73	2	
74.	.74	2	
75.	.75	2	
76.	.76	2	
77.	.77	1	
78.	.78	2	
79.	.79	2	
80.	.80	2	
81.	.81	2	
82.	.82	2	
83.	.83	2	
84.	.84	2	*The following notes refer to Figure 3.*
85.	.85	1	Above here in the table there are 59 x 1s and 25 x 2s ($n = 84$). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.85, there will be 59 true positives (the 1s above here) and 25 false positives (the 2s above here). At or below here there is only one 1 and there are 15 x 2s ($n = 16$). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.85, there will be 1 false negative (the 1 at or below here in the table) and there are 15 true negatives (the 2s below here in the table).

86.	.86	2
87.	.87	2
88.	.88	2
89.	.89	2
90.	.90	2
91.	.91	2
92.	.92	2
93.	.93	2
94.	.94	2
95.	.97	2
96.	1.00	2
97.	1.01	2
98.	1.02	2
99.	1.03	2
100.	1.05	2

Appendix B

Toe–brachial indices for data in Figures 4 and 5

As indicated in the text, the TBI values in the following table are based on a hypothetical sample of 100 ostensibly healthy adults between 50 and 70 years of age who are presumably not at risk of PAD, so almost all of them had quite high TBIs, and among them differing numbers had TBIs indicative of PAD depending on where the TBI cutoff is placed – in this case at 0.55 or at 0.85. This data set is the basis for entries in Figures 4 and 5. Among these people, as for Figures 2 and 3, differing numbers had TBIs indicative of PAD depending on where the TBI cutoff is placed – again at either 0.55 or 0.85

	TBI value	Reference standard: Has condition = 1 Does not have cond. = 2	Notes
1.	.40	1	
2.	.41	1	
3.	.42	1	
4.	.47	2	
5.	.48	1	
6.	.54	2	*The following notes refer to Figure 4.*
7.	.55	2	Above here in the table there are 4 x 1s and 2 x 2s ($n = 6$). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.55, there will be 4 true positives (the 1s above here) and 2 false positives (the 2s above here in the table). At or below here in the table there are 2 x 1s and 92 x 2s ($n = 94$). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.55, there will be only 2 false negatives (the 1s at or below here) and 92 true negatives (the 2s below here in the table).
8.	.56	1	
9.	.56	2	
10.	.84	2	
11.	.84	2	
12.	.84	2	
13.	.83	2	
14.	.82	2	
15.	.80	2	
16.	.81	2	
17.	.82	2	
18.	.83	2	
19.	.84	2	

20.	.82	2
21.	.81	2
22.	.81	2
23.	.82	2
24.	.83	2
25.	.83	2
26.	.83	2
27.	.84	2
28.	.80	2
29.	.84	2
30.	.81	2
31.	.82	2
32.	.83	2
33.	.81	2
34.	.82	2
35.	.83	2
36.	.80	2
37.	.79	2
38.	.80	2
39.	.79	2
40.	.78	2
41.	.80	2
42.	.82	2
43.	.83	2
44.	.84	2
45.	.79	2
46.	.82	2
47.	.83	2
48.	.80	2
49.	.84	2
50.	.84	2
51.	.84	2
52.	.82	2
53.	.83	2
54.	.81	2
55.	.79	2
56.	.80	2
57.	.83	2
58.	.84	2
59.	.78	2
60.	.84	2
61.	.79	2
62.	.78	2
63.	.77	2
64.	.78	2
65.	.81	2
66.	.84	2
67.	.80	2
68.	.80	2
69.	.81	2
70.	.81	2
71.	.82	2
72.	.82	2
73.	.83	2
74.	.84	2
75.	.82	2

76.	.82	2	
77.	.83	2	
78.	.83	2	
79.	.84	2	
80.	.80	2	
81.	.81	2	
82.	.82	2	
83.	.83	2	
84.	.84	2	
85.	.85	2	*The following notes refer to Figure 5.* Above here in the table there are 5 x 1s and 79 x 2s (n = 84). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.85, there will be 5 true positives (the 1s above here) and 79 false positives (the 2s above here in the table). At or below here in the table there is one 1, and 15 x 2s (n = 16). So, if PAD is regarded as unlikely when the TBI cutpoint is drawn at ≥ 0.85, there will be one false negative (the only 1 at or below here) and there are 15 true negatives (the 2s below here in the table).
86.	.86	1	
87.	.91	2	
88.	. 92	2	
89.	. 92	2	
90.	. 99	2	
91.	. 92	2	
92.	.94	2	
93.	.93	2	
94.	. 99	2	
95.	.95	2	
96.	.96	2	
97.	.98	2	
98.	1.05	2	
99.	1.08	2	
100.	1.10	2	