The Journal of Pain
OFFICIAL JOURNAL OF THE AMERICAN PAIN SOCIETY
American Pain Society
www.jpain.org
www.sciencedirect.com
CELEBRATING OUR 10th ANNIVERSARY

Using Screening Tests to Predict Aberrant Use of Opioids in Chronic Pain Patients: *Caveat Emptor*

Robert W. Bailey, Kevin E. Vowles

PII: S1526-5900(17)30627-2

DOI: 10.1016/j.jpain.2017.06.004

Reference: YJPAI 3434

To appear in: *Journal of Pain*

Received Date: 9 December 2016

Revised Date: 10 June 2017

Accepted Date: 16 June 2017

Please cite this article as: Bailey RW, Vowles KE, Using Screening Tests to Predict Aberrant Use of Opioids in Chronic Pain Patients: *Caveat Emptor*, *Journal of Pain* (2017), doi: 10.1016/j.jpain.2017.06.004.

Using Screening Tests to Predict Aberrant Use of Opioids in Chronic Pain Patients: *Caveat Emptor*

Robert W. Bailey[1]* and Kevin E. Vowles[2]

University of New Mexico

Disclosures: No conflicts of interest to report. This work was partially supported by a grant from the National Center for Complementary and Integrative Health (R34AT008398).

[1] **Corresponding author*. University of New Mexico; 1 University of New Mexico, MSC03 2220, Albuquerque, NM, 87131; ph. 617-953-1584; Fax: 505.277.1394; email: rwbailey4@gmail.com

[2] University of New Mexico; 1 University of New Mexico, MSC03 2220, Albuquerque, NM, 87131; ph. 505-974-5343; email: k.e.vowles@gmail.com

Abstract

Screening tests represent a critical tool in chronic pain treatment for predicting aberrant opioid use, which has emerged as a significant public health issue. Nevertheless, there remains a significant potential for the misapplication of screeners in this context. The potential difficulties in evaluating the diagnostic efficiency of screeners have been well established, particularly with regard to the impact that the prevalence of a disorder has on predictive value. The wide range in the reported prevalence of aberrant opioid use behaviors makes it difficult to interpret data obtained from popular screeners for assessing the potential for the aberrant use of opioids. Given the prevalence of opioid problems, however, formulating clear clinical guidelines on such screeners appears highly important. The aims of the present paper include (1) providing a review of the salient issues necessary for interpreting diagnostic efficiency statistics of screening tests, (2) identifying the critical differences between sensitivity, specificity and predictive value, and (3) discussing the characteristic effects that disease prevalence has on statistical prediction. The paper also reviews key processes in screening measure development and highlights several key considerations relevant to their appropriate use in clinical decision-making.

Perspective: This article highlights common metrics for evaluating the clinical utility of screening tests in predicting aberrant opioid use. In addition, it explores a series of considerations key to developing clinical guidelines for interpreting the results of screeners in this context.

Keywords: screening tests; decision-making; opioid therapy; chronic pain

Introduction

Chronic pain is a substantial healthcare concern that affects up to 25% of the adult population in developed countries [6, 21, 28]. Among the many treatment options, perhaps the most controversial is opioid pharmacotherapy [2, 5]. Notwithstanding evidence that opioids may represent a valuable treatment option in pain management, a public health crisis has emerged with regard to their aberrant use. Prevalence estimates for the recreational use of opioids, for example, have risen along with the increased number of prescription sales [20]. Opioids also appear to be a driving force behind an alarming increase in the number of fatal and nonfatal drug poisonings in the United States over the past several years [35].

There are certainly risks involved in prescribing opioids for chronic pain, yet a subset of patients do experience pain reduction or increased functioning, or both [10, 14]. Consequently, one potential solution to this public health dilemma has involved developing systematic screening protocols to predict the potential for aberrant use prior to initiating opioid therapy [27, 34]. Information derived from valid psychometric instruments could help identify those at high risk for aberrant use and ensure that they receive increased monitoring or alternative non-opioid pain treatments. For those whom a psychometric instrument indicates low risk for aberrant use, and for whom the benefits of opioids are perceived to be appreciable, opioid pharmacotherapy can remain a viable treatment option. These instruments offer an important and impartial alternative to methods that do not reliably classify individuals according to risk, including clinical interviews, provider observation of problematic behaviors, and urine toxicology screening [27, 44].

Unfortunately, while screening tests are available for assessing the potential for future aberrant opioid use, there are a number of considerations that must be addressed to prevent misapplication of their findings. The purpose of the present paper is to provide an overview of key areas for potential misapplication, synthesize the literature on appropriate screener development, highlight the particular requirements for accurate prediction of future behavior, and evaluate a widely used aberrant opioid use screening measure in relation to these issues. The paper is organized around five key points highlighting important considerations related to the use of screening measures in the prediction of future behavior, which include:

1. Statistical methods are more accurate than clinical judgment.
2. Sensitivity and Specificity are not Predictive Value.
3. Population base rates affect a measure's diagnostic efficiency.
4. Sensitivity and Specificity are not fixed properties.
5. Benchmark diagnostic tests are necessary.

In addition to providing an overview of these key issues, several recommendations for future work in this highly important area are offered.

Considerations Related to Predicting the Aberrant Use of Opioids

1. Statistical Methods Are More Accurate Than Clinical Judgment

The benefits of using statistical methods to predict uncertain future events have long been understood in psychological science. Notably, Meehl[32] was responsible for an important publication that outlined the superior accuracy of statistical prediction over clinical prediction, even when clinicians were given additional information not included in the statistical model[13, 19]. In the context of predicting opioid use patterns in those with chronic pain, the primary means of making statistical-based decisions has involved the use of self-report screening measures. In order to best understand the specific requirements for predicting future behavior, it is necessary to provide a brief overview of procedures used to develop and initially validate screening measures. The development and initial validation process will then be discussed in relation to one of the most frequently used screening measures for predicting aberrant opioid use, the Revised Screener and Opioid Assessment for Patients with Pain (SOAPP-R)[7].

Screening measure development procedures. Screening tests have a single primary purpose: to detect the potential presence or absence of a particular attribute in people [41]. Screening tests, or screeners, have demonstrated utility in multiple contexts, from predicting the likelihood of academic or professional success through the use of aptitude tests to detecting the presence of tuberculosis based on the results of a purified protein derivative (PPD) test in medical settings. Screening measures are intended for use among large populations to detect the presence of an attribute, particularly when individuals are either unaware of the attribute or unwilling to admit having it.[41, 42]

The results of any screening test include four possibilities, which can be captured in a 2 x 2 cross-tabulation table (see Table 1). Only two of the four possible outcomes are correct classifications, in that the screener accurately classifies those who have the attribute (true positives) and those who do not (true negatives). Because the variables included in screeners are only tendencies, in other words, people with the attribute tend to behave in particular ways or tend to have other, related characteristics, the results of screeners will include error [30, 42]. Thus, the two remaining outcomes comprise the misclassifications (false positives and false negatives). The primary task when developing screening tests is to maximize the number of individuals who are accurately classified, known as the *hit rate*. During the screening test development process, the screener is being evaluated relative to an established benchmark test to determine how well the screener can detect the presence or absence of the attribute at the present time in relation to that benchmark. Thus, this analysis is cross-sectional in nature.

Screening measure development metrics. The primary objective while developing a screening measure is to determine four indices of accuracy or *diagnostic efficiency*, including the more commonly known sensitivity and specificity, but also the less well known positive and negative likelihood ratios [41]. *Sensitivity* is defined as the proportion of people who have the attribute and are correctly classified by the screener as possessing that attribute. It is calculated by dividing the number of *true positives* by the sum of *true positives* and *false negatives.* Conversely, *specificity* is the proportion of people who do not have the attribute and are correctly classified by the screener as not possessing that attribute. It is calculated by dividing the number of *true negatives* by the sum of *true negatives* and *false positives*. Thus, in a sample of 100 individuals, each of whom possessed the attribute of interest (e.g., current aberrant opioid use), a measure with a sensitivity of 0.70 would accurately classify 70 out of every 100 people as possessing an attribute, while the remaining 30 individuals would be incorrectly classified as not possessing the attribute. Sensitivity and specificity are typically referred to as *column-based indexes* because they are calculated using the information in the columns of Table 1.

The *positive likelihood ratio* (LR^+) is calculated as the proportion of true positives to all who actually have the attribute (i.e., sensitivity) divided by the proportion of false positives to all who are diagnosed as not having attribute (i.e., 1 – specificity). The LR^+ indicates the likelihood a positive screening result comes from someone who actually has the attribute. In terms of interpretation, an LR^+ of 1, for example, indicates a random or useless test with no diagnostic value [31], where a positive screener result is equally as likely to have come from someone who has the attribute as it is from someone who does not, as diagnosed by the benchmark test. Positive likelihood ratios greater than 1 indicate increased accuracy in the screener’s ability to correctly identify attribute presence. An LR^+ of 2, for instance, indicates that a positive screener is twice as likely to come from someone who actually possesses the attribute than from someone who does not (i.e., a false positive).

The *negative likelihood ratio* (LR$^-$) is the false negative rate (i.e., 1 – sensitivity) divided by the rate of true negatives (i.e., specificity). The LR$^-$ indicates the likelihood a negative screening result comes from someone who does not possess the attribute, and as the LR$^-$ tends toward 0 there is a decreased likelihood that the attribute is present.

The SOAPP-R – a representative example for predicting aberrant opioid use. The SOAPP-R[7] was developed by integrating an item pool of risky behaviors relevant to opioid misuse and abuse into an instrument with decision rules that could aid in opioid prescription decisions. The final version of the SOAPP-R included 24 self-report items intended to predict aberrant use behaviors in patients using opioids to manage long-term pain. The results of the validation process, which used a sample of 283 pain patients, indicated a cut score of 18 had adequate sensitivity (81%) and specificity (68%). Overall, the development process for the SOAPP-R was exemplary as it involved a thorough process of items selection and refinement, evaluation of content validity, and thorough reporting of all diagnostic efficiency statistics, as well as predictive validity. The SOAPP-R will therefore be used for illustration purposes in the following sections.

Considerations related to sensitivity and specificity. Sensitivity and specificity are affected by the *cut score*, which is determined by the instrument developers to delineate between those who do and do not have the attribute according to the screener [33]. In a screening test like the SOAPP-R, a very low cut line results in most respondents being classified as aberrant users. Although a screener with a low cut line will correctly classify most individuals with aberrant opioid use behaviors, it will also incorrectly classify many individuals as aberrant opioid users, when in fact opioid use is non-aberrant. Thus, in practice, only a small percentage of those who do not have the attribute will be correctly identified as being non-aberrant opioid users. Too low a cut score is not very efficient in practice because there will be many false positives. Conversely, if specificity is maximized using a high cut score, the screener will correctly classify most individuals who are not currently aberrant users, but with a high false negative rate.

Deciding whether to emphasize sensitivity or specificity comes at the discretion of the instrument developers, who must explore the relative merits of various cut scores and provide a defensible justification for the final cut score. For the SOAPP-R, Butler et al.[7] reported on sensitivity and specificity for all possible cut scores and then chose a cut score of 18 in order to emphasize sensitivity over specificity (.81 and .68, respectively). As noted above, sensitivity is related to the LR^+, which was 3.8 for the SOAPP-R and indicated a positive result was 3.8 times as likely to have come from a person currently using opioids aberrantly than from a person using opioids non-aberrantly. The authors also reported an LR^- of .29, indicating that a negative screener result was about 3.5 times as likely to come from a person who did not in fact have the attribute than from someone who did possess it. In determining the final cut score for the SOAPP-R, Butler et al. stated that the priority was to identify those at high risk and minimize false negatives, even if that meant there were a number of false positives identified.

Sensitivity, specificity, and likelihood ratios are examples of diagnostic efficiency statistics that are commonly evaluated during screener development. As noted, these metrics evaluate the performance of the screening test against an established diagnostic benchmark, often referred to as a "gold standard". The SOAPP-R[7] was tested against a benchmark system called the "aberrant behavior drug index," which combined physician and self-report along with the results of a urine drug screen. Essentially, this step in the development process evaluates how well the screening measure performs at detecting the presence or absence of an attribute *at present*. Once a screening measure like the SOAPP-R is put into clinical practice, however, the primary concern shifts from how well it classifies individuals against a benchmark to a focus on the screener's ability to distinguish between those who will go on to have and not have the attribute in the future.[41] This latter focus pertains to *predictive value*, which involves additional considerations that are detailed in the following sections.

2. Sensitivity and Specificity Are Not Predictive Value

Predictive accuracy is indicated by the positive predictive value (PPV) and negative predictive value (NPV) of the screening test.[29] In contrast to the column-based indexes of sensitivity and specificity, these indices are referred to as *row-based indexes* because their values are calculated using the information in the rows (see Table 1). Although we will illustrate that there are expected relationships between the column- and row-based indexes, sensitivity and specificity can provide very little information in relation to predictive values. An examination of the cells in Table 1 illustrates that the calculations for sensitivity and specificity do not account for false positives and false negatives, respectively, thus the row-based indices provide additional information in comparison to the indices of the columns.

Sensitivity and specificity versus predictive values. Conditional probability statements are particularly helpful for distinguishing sensitivity and specificity from positive and negative predictive values. A conditional probability is simply the probability of some event occurring given another event is true, notated as: *P*(event occurs | some other event is true). Sensitivity, for example, is simply the probability of scoring above the cut line, *given the individual is an aberrant user*, or *P*(positive result | aberrant user). Specificity, on the other hand, is the probability of scoring below the cut line, *given the individual is a non-aberrant user*, or

P(negative result | non-aberrant user).

Notice that the above conditional statements comport with the notion of accurate classification and screener development, *where attribute status is known*. When attribute status is unknown, such as in clinical practice, the conditional probabilities of interest are the exact inverse of those that pertain to sensitivity and specificity. For instance, *given an individual scores above the cut line*, the primary interest will be the probability that the individual goes on to be an aberrant user, or *P*(aberrant user | positive result). As much as *P*(positive result | aberrant user) and *P*(aberrant user | positive result), which refer to sensitivity and PPV, respectively, appear similar, the actual probabilities can differ greatly [11, 12]. Cohen[12] demonstrated that what affects the degree of difference between a conditional probability and its inverse is the *prior probability*, which refers to the base rate of aberrant opioid use in the present example.

The PPV can be calculated from its inverse, sensitivity, by using the prior probability and Bayes' Theorem (see Cohen[12] or Streiner[41] for more information). An equivalent and simpler method for calculating PPV involves using the row-based information in Table 2 and dividing the number of *true* positives detected by the SOAPP-R by the total number of individuals who scored above the cut line, or 62/109 = .57. Butler et al. reported that the sensitivity of the SOAPP-R indicated that it was able to "accurately identify" 81% of users who "turn out to be at high risk," yet the calculated PPV of 57% can be interpreted as the probability that someone will be an aberrant user given that the individual scores above 18 on the SOAPP-R. Therefore a clinician using this test alone to predict aberrant use could expect to be wrong over 40% of the time. While it is important to note that a screening test should not be used diagnostically, a follow-up diagnostic test cannot be applied when the SOAPP-R is used to predict future use patterns. Therefore a PPV in the range of 57% indicates that the positive result (i.e., that the individuals will go on to be an aberrant opioid user) must be interpreted with caution and not in isolation. Additional screening is warranted, such as a diagnostic interview or urine drug screen. The apparent discrepancy between sensitivity and PPV in this situation underscores the critical need to understand the limitations of a screening test in this context with regard to predictive value. Note that Butler et al.'s choice of a cut score that prioritized identification of true positives at the expense of over-identifying false-positives implies maximizing sensitivity and NPV. In other words, emphasizing sensitivity means more false positives, which do not factor into the calculation of the column-based sensitivity statistic. On the other hand, the proportion of false positives is accounted for in PPV. In summary, the SOAPP-R performs well at correctly classifying individuals who are diagnosed with aberrant behaviors (81% sensitivity), but many of the individuals who are not diagnosed with aberrant use are misclassified, resulting in a PPV of 57%. Thus, the calculation of predictive values provides a more complete picture of the limitations of the SOAPP-R.

Clinical implications. The results from screening tests for any attribute will always involve "noise," or error, and predictive values offer information that is more clinically relevant compared to sensitivity and specificity [24, 30, 39]. With regard to the SOAPP-R and prediction, the results do indicate that the screener has strong *negative* predictive value, in that it performs well at predicting individuals who use opioids non-aberrantly (NPV = .87; see Table 2), which seems a significant increase over the probability of non-aberrant users assumed using prevalence (1-.345 = .655 prevalence of non-aberrant users; see Table 2). In other words, the SOAPP-R is likely to be very accurate in identifying those that will *not* go on to use opioids aberrantly.

Based on the guidelines of Meehl and Rosen[33] and Streiner[41], although the high false positive rate indicates it is impractical to use the SOAPP-R to "rule in" future aberrant users, the screener predicts safe opioid use patterns with acceptable accuracy. Remembering that the primary concern in screener use is to maximize the utility in clinical settings, sensible clinical guidelines for the use of the SOAPP-R might specify that it is most appropriate for use in identifying those who are *unlikely* to go on to develop aberrant use behaviors.

This issue of correctly identifying those who will go on to be non-aberrant opioid users is also relevant to our next consideration, which pertains to the substantial effects that base rates have on predictive value. As is true with any screening test, if the base rates change, the predictive validity will also shift accordingly [12, 33, 41].

3. Population Base Rates Effect Diagnostic Efficiency

According to Meehl and Rosen,[33] the base rate, or prevalence of a condition in the population, is an essential component to quantify when seeking to evaluate the predictive value of a psychometric device. This guidance is germane to the topic of aberrant opioid use behaviors because the base rates are reported to vary considerably. For instance, the rates of abuse and addiction among chronic pain patients who were prescribed opioids have been reported to be as low as 3.27% [15], yet other reports have indicated “addiction problems” in as many as 50% of patients [23]. In a more recent review, with means that were adjusted based on sample size and study quality, misuse and addiction rates were approximately 25% and 10%, respectively [45]. The variability in base rates can be attributed to the many ways in which aberrant use behaviors were defined and the various locations in which these data were collected [36].

In the present subsection, the primary aim will be to demonstrate how substantial variability in base rates will affect the predictive power of any screener, including the SOAPP-R[7], and therefore impact the screener’s utility as a clinical decision-making aid. To illustrate this point, let us assume three separate “true” base rates of aberrant opioid use behaviors that fall into the general range reported in the literature: 3%, 25%, and 50%. Using these base rates, as well as the sensitivity and specificity reported by Butler et al. for the SOAPP-R, we have calculated three 2 x 2 tables (Tables 3a-c) using Butler et al.'s original sample size (*N* = 223). To simplify the demonstration and discussion of predictive value, sensitivity and specificity are held constant in accordance with the SOAPP-R validation study.

In the following examples, diagnostic accuracy for the three different prevalence rates is based on PPV, NPV, and overall efficiency (the number of correct decisions divided by all decisions). We also evaluated the *incremental validity* provided by the screener above and beyond simply using the base rates alone to determine the percentage of incorrect classifications (i.e. assuming all individuals will use opioids without problems and thus the misclassification rate equals the base rate for aberrant use). For this calculation, we used Kraemer's[29] calibrated positive and negative predictive value (CPPV and CNPV, respectively), which is essentially a conservative measure of predictive validity that corrects for chance agreement, similar to Cohen's kappa.[41]. (For a more detailed discussion of the importance of calibration in conjunction with diagnostic efficiency statistics, see Kraemer[29].)

Diagnostic efficiency assuming 3% base rate. When the base rates for aberrant opioid use are assumed to be 3% in the population, PPV is also extremely low (Table 3a), such that of the 75 individuals who score above the cut line on the SOAPP-R, only 8% are predicted to aberrantly use opioids. However, NPV is excellent: over 99% of individuals who score below the cut line will go on to use opioids non-aberrantly. At a 3% base rate, a similar relationship exists between CPPV and CNPV; whereas CPPV indicates that the test increases diagnostic value by only 4% for predicting aberrant use, CNPV suggests that diagnostic value is increased by 71% for predicting non-aberrant use of opioids. The overall efficiency of the test is 69%. In short, if a

3% base rate of aberrant opioid use is assumed to be true in the population of opioid users, then the SOAPP-R is best used to identify those who will go on to use opioids in a non-aberrant manner and should not alone be used to identify those who will go on to use opioids aberrantly.

Diagnostic efficiency assuming 25% base rate. When the base rate for aberrant use is set at 25% (Table 3b), one begins to see the resulting characteristic directions in which the efficiency statistics shift. The PPV has significantly increased to 46% and NPV decreased slightly to 91%, which indicates the SOAPP-R still accurately predicts non-aberrant opioid use, but is essentially no better than a coin flip in the prediction of aberrant use. Similarly, CPPV has increased and CNPV has decreased, such that the test now increases the diagnostic value by 28% for a positive diagnosis (CPPV), and the increase in diagnostic value for a negative diagnosis (CNPV) has only decreased slightly to 66%. Overall efficiency has increased slightly to 71%, but, as with a 3% base rate, if the true base rate of aberrant use is 25%, then the SOAPP-R is best used to identify those who will go on to use opioids in a non-aberrant manner.

Diagnostic efficiency assuming 50% base rate. At 50% prevalence for aberrant use behaviors, the overall prediction efficiency of the SOAPP-R improves markedly. The PPV and NPV are both high at 71% and 78%, respectively. Similarly, CPPV has increased to 43% and CNPV remains high at 56%. Lastly, overall efficiency is optimized at 75%. One could argue that this base rate is when the SOAPP-R may be appropriately used to classify both those who will and who will not go on to develop aberrant opioid use.

According to Meehl and Rosen[33] and Streiner,[41] the positive and negative predictive values are optimized at 50% prevalence in the population. As prevalence in the population increases above 50%, PPV will continue to increase while NPV declines. In other words, as population prevalence decreases below 50%, PPV decreases and NPV increases; as population prevalence increases above 50%, PPV increases and NPV decreases. At this point, it is important to emphasize that the limited utility of the SOAPP-R's PPV at low base rates is characteristic of all screeners, rather than a limitation unique to the SOAPP-R alone.

Rules of thumb regarding screeners and base rates. There are several heuristics that can be derived from the pattern observed in the tables examining the influence of base rates (Tables 3a-c). First, when prevalence is low, as demonstrated in the present example at 25% or 3% prevalence for opioid aberrant use behaviors, the majority of positive predictions are incorrect, and therefore the screener should be used only to rule out the condition. It would therefore be inadvisable to interpret scores above the cut line for the SOAPP-R in this situation (too many false positives). Additionally, as prevalence continues to increase past 50%, NPV will decrease to the point that the majority of negative predictions would be incorrect. In this scenario, the screener should be used only to rule in the condition, and one should not interpret scores below the cut line.

Perhaps the most important heuristic offered by Meehl and Rosen[33] underscores the importance of comparing the predictive validity of the screening test against the base rates – the *incremental validity* of the screener. Table 3a illustrates this point quite well. At a prevalence of 3%, the NPV appears excellent, such that about 99% of people will be correctly predicted to use their opioid medications non-aberrantly. Thus, the screener will be wrong about 1% of the time. Forgoing the screener entirely, however, and assuming all individuals will be non-aberrant users will result in an error rate equal to the prevalence of 3%. This scenario would require examining the marginal costs of administering the screener, as the marginal benefit appears small (a 1% versus 3% rate of incorrect decisions). This example highlights the difficulty in interpreting the absolute percentages for predictive value without a reference point such as prevalence. Even tests with seemingly low predictive value can offer an important incremental benefit beyond relying on the base rates, such as in the prediction of adolescent suicide [25]. In terms of the 2% incremental predictive benefit by using the SOAPP-R, the marginal cost of test administration is probably negligible, as the test takes up few clinical resources. Making any kind of case for how to use the SOAPP-R, or any other screener, in the clinical context, however, involves two critical assumptions: 1) the medical community agrees on a universal definition for "aberrant use" and 2) the base rates for the agreed upon definition are known. In the domain of opioid use behaviors, there is reason to doubt that there is a universal definition of aberrant use, a topic relevant to our final consideration regarding the importance of benchmarks. Before we discuss benchmarks, however, it is important to discuss one more aspect of sensitivity and specificity – they are not fixed properties of the measure.

4. Sensitivity and Specificity Are Not Fixed Properties

Sensitivity and specificity were held constant in the preceding examples examining the impact of base rates on PPV and NPV. In order for sensitivity and specificity to remain constant in practice, however, both patients who do and do not have the attribute must respond with "absolute homogeneity" to the test across clinical populations[29]. In other words, sensitivity and specificity of any measure can be only expected to remain stable as long as the sample size is large enough to be sufficiently representative of the entire population of interest.

Given the range of different populations of chronic pain patients one could potentially sample when evaluating a screener like the SOAPP-R[7], it is improbable that the calculated sensitivity and specificity are representative of the entire population of opioid-using individuals with chronic pain. The number of individuals with chronic pain totals into the millions, and the initial SOAPP-R validation study included a sample of 223 chronic pain patients who do not appear to have been randomly selected. Just as the base rates will change to some degree depending on the clinical context, the response tendencies, and therefore the screener's psychometric properties, will also vary depending on the context of each administration. Therefore, until such a time that samples are large enough to be assumed to be representative of the population, it is recommended that sensitivity and specificity, as well as PPV and NPV, are calculated with each new sample analyzed.

5. Benchmark Diagnostic Tests are Necessary

The challenges in using screeners to predict aberrant use behaviors among patients who are prescribed opioids for chronic pain are not isolated to this specific context, as prevalence will always affect the predictive value of a screening test. For example, in HIV screening among the general population, the probability of being infected by the virus is approximately .01%, and using Bayes' Theorem will indicate that the posterior probability of actually having the virus following a positive screener result is only 50% [18]. In this case, however, practitioners working with infectious diseases have several advantages over those working in clinical psychology or examining human behavior. For one, infectious disease approaches generally use screening tests to detect current disease presence, whereas a screener like the SOAPP-R predicts future behavior. After a positive result on the SOAPP-R, the only way to confirm the predictive validity in any particular case is to follow future opioid use patterns, as there is no diagnostic follow-up test that can be administered. Secondly, an HIV infection involves an identifiable pathogen, and, after the initial screening test, which only detects anti-bodies, a more involved diagnostic follow-up test can identify the virus itself [8]. The lack of an identifiable pathogen in aberrant opioid use behaviors presents additional challenges to evaluating screening tests because a gold standard diagnostic test is so essential to screener validation.

At present, both the Diagnostic and Statistical Manual Of Mental Disorders (DSM-5)[1] and the ICD-10 Classification Of Mental And Behavioural Disorders (ICD-10)[46] are commonly used classification systems for diagnosing opioid use disorder. Upon examining the criteria from each system, however, the lack of consensus about which behaviors are most salient becomes clear. For example, the DSM-5 no longer delineates between abuse and dependence, yet the ICD-10 makes a distinction between dependence and harmful use, which is similar to DSM-IV-defined abuse, and considered less severe of the two [38]. The ICD-10 and DSM criteria also include some key differences in the specific criteria, which may result in different prevalence rates for opioid use disorder between the two systems [23, 43]. Furthermore, the complexities of using opioids to manage chronic pain can present as a confound to the criteria for diagnosing aberrant use[5, 17, 37, 44]. For example, taking opioids in larger amounts over time may denote poorly controlled pain or the intractable nature of chronic pain, or both, and not necessarily indicate medication tolerance. Taken together, it would appear that addressing the base rates for aberrant use would first require coming to a consensus on how opioid use disorder is defined in those who are using opioids in the treatment of chronic pain, and include only those criteria for aberrant use behaviors that are most salient from a public health standpoint. As Fordyce[16] recognized long ago, it is important to examine whether analgesic use is causing problems or impeding the process of chronic pain management. Sage advice, though four decades later, we are still trying to determine exactly *how* to define clinically significant opioid use problems.

Discussion

The goal of any psychometric measure is to support the process of clinical decision-making. In order to accomplish this goal, the instrument in question must demonstrate adequate predictive validity and offer a measureable incremental benefit over other methods of making clinical decisions, such as relying on base rates or making use of clinical judgment. Further, following sound measurement development practices, including reporting on key psychometric properties, are not alone sufficient for establishing predictive validity. Sensitivity and specificity, by themselves, become relatively unimportant following screener development, and the base rates for the attribute of interest significantly affect how a screener should be used to make predictions for future behavior [33, 41]. The base rates for aberrant opioid use are particularly difficult to pinpoint, especially given the lack of consensus on diagnostic criteria, which creates challenges for creating clinical guidelines for any screening test, such as the SOAPP-R.

The consequences for using a screener lacking robust positive predictive validity may include inappropriate diagnosis, stigma, lack of access to effective pain control for low-risk individuals who score above the cut line, and, for a screener lacking robust negative predictive validity, adverse outcomes related to aberrant use for high-risk individuals who score below the cut line [40]. The issue of stigma may be particularly salient for African-American patients with chronic pain, who may be less likely than White patients to receive opioid prescriptions for chronic pain[9] and who may also be subjected to closer monitoring and more likely to be sent for substance abuse assessments[3, 22]. Screeners like the SOAPP-R, when used properly, may lead to more equitable treatment for all chronic pain patients under consideration for opioids.

As indicated by Kraemer[29], sensitivity and specificity as well as positive and negative predictive values should be expected to vary across clinical populations. Therefore, just as with the psychometric properties of any other instrument, the efficiency statistics of a screener will change to some degree with every test administration. The reliability of diagnostic efficiency statistics like sensitivity and specificity can only begin to be assumed to reflect the true population parameters after many different administrations in various settings. Yet even population parameters may be of little use when using a screening test at a specific medical setting working with a specific clinical population. Diagnostic efficiency statistics are best understood in the context in which a screener is used: a screener like the SOAPP-R may, for example, be used to effectively rule out future aberrant use in a general clinic where the base rate of opioid use disorder is low, yet the SOAPP-R should be used to rule in aberrant use in a clinic with a high prevalence of pre-existing substance use disorders, a risk factor for opioid use disorders. Furthermore, one must be careful about generalizing the diagnostic efficiency of a screener beyond the population for which it was intended. The known risk factors and SOAPP-R validation study are specific to adults, for instance, and it cannot be assumed that any of the diagnostic efficiency statistics will generalize to adolescent populations. Therefore understanding base rates at the level of the local clinic may be the best tool for understanding how to use screening tests in context.

General Guidelines for Using Screeners

The statistician George Box is famously quoted as stating that all of our statistical models are wrong, but some are useful [4]. Given the wide range in base rates for aberrant opioid use, the clinical utility of our models to predict the aberrant use of opioids is indeed dubious, *caveat emptor*, though several guidelines can be offered that would help prevent misapplication, thus fulfilling Box's aspiration of usefulness. This paper is intended to provide a primer on understanding diagnostic efficiency and apply the findings of seminal works in this area of predicting aberrant opioid use through screening tests. An understanding of the following key points will facilitate more effective use of screening measures in relation to this need:

Screening tests are useful tools for predicting risk. When used properly, screeners represent a more standardized and objective assessment compared to clinical judgment, such as clinical interviews and looking for problematic use indicators. Screeners have demonstrated their value in both medicine and clinical psychology. Key diagnostic efficiency statistics, including sensitivity, specificity, and predictive values, can help inform interpretation of the results and the decision-making process. Still, it is important to understand that a screener like the SOAPP-R only evaluates one aspect of the decision of whether to prescribe an opioid – risk potential. Another key evaluation area for those deemed to be low risk is the expected benefit from opioids, which is beyond the scope of a screener like the SOAPP-R.

Sensitivity and specificity do not imply predictive value. While not inaccurate to say that a screener like the SOAPP-R has demonstrated 81% accuracy in identifying individuals with aberrant use behaviors, this phrasing can be misconstrued as prediction. In this example, the 81% refers only to the sensitivity of the screener, which is a statistic of classification accuracy used to develop a screener. The differences in language between sensitivity and specificity compared to PPV and NPV are quite subtle, but each can lead one to different conclusions about the clinical utility of a screener. Sensitivity and specificity are determined at the discretion of the instrument developer, but predictive values are primarily affected by base rates.

Base rates affect predictive values in expected ways. As demonstrated by Cohen [11, 12], Meehl and Rosen [33], and Streiner[41], as well as in Tables 3a-c, the base rates for an attribute affect the predictive value of a screener. When prevalence is high, NPV will be low, so only those who score above the cut line can be clearly interpreted. Alternatively, when prevalence is low, PPV will be low, and only those who score below the cut line can be clearly interpreted. The PPV and NPV are optimized at a prevalence of 50%. Based on evidence suggesting that the base rates for aberrant opioid use falls below 50%, the SOAPP-R cannot be expected to accurately predict those who will go on to aberrantly use opioids. Therefore positive results should be interpreted with caution and should not be the primary driver behind a recommendation not to prescribe. A more cautious approach might involve additional testing, closer monitoring (including ongoing diagnostic testing), or starting out at a lower dosage. Screeners like the SOAPP-R can, however, accurately predict those who will go on to use opioids in a non-aberrant manner when base rates are low. Further, the specific predictive values are best determined by understanding base rates at the level of the clinic itself. For example, base rates in a primary care setting may be below 50%, while they may be higher in other services, such as so-called “Co-Occurring Disorder” clinics that specialize in chronic pain treatment for those with problematic opioid use patterns.

Diagnostic efficiency statistics are not fixed properties of a screener. Just as one should report on statistics like internal consistency for a psychometric instrument based on the sample in a given study, basic diagnostic efficiency statistics should be offered in a similar manner when using screeners. At a minimum, studies should report on sensitivity, specificity, PPV and NPV for the sample under investigation whenever a screener is used, rather than reporting on the sensitivity and specificity from the original measure development study, as is common practice.

The diagnostic system implemented affects diagnostic efficiency. Presently, there is no universally-accepted “gold standard” against which to compare the results of a screener for opioid use disorder [7]. The validity of the screener, however, is significantly tied to the accuracy of the benchmark test [41]. As explored in the previous section, the lack of a universal benchmark test is perhaps the greatest barrier to improving the predictive validity of screeners for opioid addiction, particularly because of the aforementioned issues regarding classification criteria in the DSM-5[1]. Using different classification systems to confirm the result of the screener may lead to discrepant diagnostic efficiency statistics. Until a universally agreed upon diagnostic system is established, it is recommended that consistent diagnostic systems are used at the level of specific studies or within specific clinics. Furthermore, if a specific setting uses a different diagnostic system than that which was used during screener development, diagnostic efficiency statistics, including sensitivity, specificity, PPV, and NPV, must be recalculated.

Conclusions

It seems clear that predicting the aberrant use of opioids in the context of chronic pain management using screening tests is a complex process. Therefore, it is incumbent upon clinicians involved in these decisions to enact good monitoring practices that are fair to patients and recognize the potential for high false positive rate for those who score above the cut line on screeners. For those who are deemed a higher risk, there is support for providing close monitoring and substance use counseling in increasing opioid use compliance [26], though we must also keep in mind the stigmatizing effect of assuming at-risk status when the rate of false positives is high. Screening tests themselves are not diagnostic, so it is important to develop monitoring protocols that recognize the potential for the false positives that arise from screeners. Ultimately, the only way to confirm the results of a screener is to follow all patients during long-term opioid therapy. This provides an additional benefit of enhancing a clinic's data on the base rates for aberrant use, which can then help inform clinic-specific guidelines for interpreting screening tests for predicting aberrant use of opioids in chronic pain.

References

1. American Psychiatric Association: Diagnostic and statistical manual of mental disorders: DSM-5, American Psychiatric Association, Washington, DC, 2013

2. Bailey RW, Vowles KE. Chronic noncancer pain and opioids: Risks, benefits, and the public health debate. *Professional Psychology: Research and Practice.* 46:340-347, 2015

3. Becker WC, Starrels JL, Heo M, Li X, Weiner MG, Turner BJ. Racial differences in primary care opioid risk reduction strategies. *Annals of Family Medicine.* 9:219-225, 2011

4. Box GE. Science and statistics. *Journal of the American Statistical Association.* 71:791-799, 1976

5. Brady KT, McCauley JL, Back SE. Prescription opioid misuse, abuse, and teatment in the United States: An update. *American Journal of Psychiatry.* 173:18-26, 2015

6. Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *European Journal of Pain.* 10:287-287, 2006

7. Butler SF, Fernandez K, Benoit C, Budman SH, Jamison RN. Validation of the revised screener and opioid assessment for patients with pain (SOAPP-R). *Journal of Pain.* 9:360-372, 2008

8. Centers for Disease Control and Prevention: HIV/AIDS: HIV testing. Available at: http://www.cdc.gov/hiv/testing/index.html Accessed March 31, 2017.

9. Chen I, Kurz J, Pasanen M, Faselis C, Panda M, Staton LJ, O'Rorke J, Menon M, Genao I, Wood J. Racial differences in opioid use for chronic nonmalignant pain. *Journal of General Internal Medicine.* 20:593-598, 2005

10. Chou R, Fanciullo GJ, Fine PG, Miaskowski C, Passik SD, Portenoy RK. Opioids for chronic noncancer pain: prediction and identification of aberrant drug-related behaviors: a review of the evidence for an American Pain Society and American Academy of Pain Medicine clinical practice guideline. *Journal of Pain.* 10:131-146, 2009

11. Cohen J: Things I have learned (so far). In: Methodological issues & strategies in clinical research, American Psychological Association; US, Washington, DC, 1990, pp. 407-424

12. Cohen J. The earth is round (p < .05). *American Psychologist.* 49:997-1003, 1994

13. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. *Science.* 243:1668-1674, 1989

14. Dowell D, Haegerich TM, Chou R: CDC Guideline for Prescribing Opioids for Chronic Pain — United States, 2016, Centers for Disease Control, 2016

15. Fishbain DA, Cole B, Lewis J, Rosomoff HL, Rosomoff RS. What Percentage of Chronic Nonmalignant Pain Patients Exposed to Chronic Opioid Analgesic Therapy Develop Abuse/Addiction and/or Aberrant Drug-Related Behaviors? A Structured Evidence-Based Review. *Pain Medicine.* 9:444-459, 2008

16. Fordyce WE: Behavioral methods for chronic pain and illness, Mosby, St. Louis, 1976

17. Garland EL, Froeliger B, Zeidan F, Partin K, Howard MO. The downward spiral of chronic pain, prescription opioid misuse, and addiction: cognitive, affective, and neuropsychopharmacologic pathways. *Neuroscience and Biobehavioral Reviews.* 37:2597-2607, 2013

18. Gigerenzer G, Krauss S, Vitouch O: The null ritual. In: The sage handbook of quantitative methodology for the social sciences.(Kaplan, D., Ed.), Sage, Thousand Oaks, CA: Sage, 2004, pp. 391

19. Grove WM, Lloyd M. Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology*. 115:192-194, 2006

20. Hall AJ, Logan JE, Toblin RL, Kaplan JA, Kraner JC, Bixler D, Crosby AE, Paulozzi LJ. Patterns of abuse among unintentional pharmaceutical overdose fatalities. *JAMA.* 300:2613-2620, 2008

21. Hardt J, Jacobsen C, Goldberg J, Nickel R, Buchwald D. Prevalence of chronic pain in a representative sample in the United States. *Pain Medicine.* 9:803-812, 2008

22. Hausmann LR, Gao S, Lee ES, Kwoh CK. Racial disparities in the monitoring of patients on chronic opioid therapy. *Pain.* 154:46-52, 2013

23. Højsted J, Sjøgren P. Addiction to opioids in chronic pain patients: A literature review. *European Journal of Pain.* 11:490-518, 2007

24. Howell DC: Statistical methods for psychology. 7th edition, Wadsworth, Belmont, CA, 2010

25. Hsu LM. Diagnostic validity statistics and the MCMI-III. *Psychological Assessment.* 14:410-422, 2002

26. Jamison RN, Ross EL, Michna E, Chen LQ, Holcomb C, Wasan AD. Substance misuse treatment for high-risk chronic pain patients on opioid therapy: A randomized trial. *Pain.* 150:390-400, 2010

27. Jamison RN, Serraillier J, Michna E. Assessment and treatment of abuse risk in opioid prescribing for chronic pain. *Pain Research and Treatment.* 2011:1-12, 2011

28. Johannes CB, Le TK, Zhou X, Johnston JA, Dworkin RH. The prevalence of chronic pain in United States adults: Results of an Internet-based survey. *Journal of Pain.* 11:1230-1239, 2010

29. Kraemer HC: Evaluating medical tests: Objective and quantitative guidelines, Sage Newbury Park, CA, 1992, pp. 91-99

30. Kruschke J: Introduction: Credibility, models, and parameters. In: Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, Academic Press, 2014, pp. 15-32

31. McGee S. Simplifying likelihood ratios. *Journal of General Internal Medicine.* 17:647-650, 2002

32. Meehl PE: Clinical versus statistical prediction: A theoretical analysis and a review of the evidence, University of Minnesota Press, Minneapolis, 1954

33. Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin.* 52:194-216, 1955

34. Michna E, Ross EL, Hynes WL, Nedeljkovic SS, Soumekh S, Janfaza D, Palombi D, Jamison RN. Predicting aberrant drug behavior in patients treated for chronic pain: importance of abuse history. *Journal of Pain and Symptom Management.* 28:250-258, 2004

35. Park H, Bloch M: How the epidemic of drug overdose deaths ripples across America. In: New York Times, New York, NY, 2016

36. Reid MC, Engles-Horton LL, Weber MB, Kerns RD, Rogers EL, O'Connor PG. Use of opioid medications for chronic noncancer pain syndromes in primary care. *Journal of General Internal Medicine.* 17:173-179, 2002

37. Savage SR. Assessment for addiction in pain-treatment settings. *Clinical Journal of Pain.* 18:S28-38, 2001

38. Schuckit MA, Hesselbrock V, Tipp J, Anthenelli R, Bucholz K, Radziminski S. A comparison of DSM-III-R, DSM-IV and ICD-10 substance use disorders diagnoses in 1922 men and women subjects in the COGA study. *Addiction.* 89:1629-1638, 1994

39. Silver N: The signal and the noise: Why so many predictions fail - but some don't, Penguin Group, New York, NY, 2012, pp. 244-250

40. Smith S: Prominent pain doctor investigated by DEA after patient deaths. Available at: http://www.cnn.com/2013/12/20/health/pain-pillar/index.html Accessed 3/31/17.

41. Streiner DL. Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment.* 81:209-219, 2003

42. Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest.* 1:1-26, 2000

43. Tarrahi MJ, Rahimi-Movaghar A, Zeraati H, Amin-Esmaeili M, Motevalian A, Hajebi A, Sharifi V, Radgoodarzi R, Hefazi M, Fotouhi A. Agreement between DSM-IV and ICD-10 criteria for opioid use disorders in two Iranian samples. *Addictive Behaviors.* 39:553-557, 2014

44. Turk DC, Swanson KS, Gatchel RJ. Predicting opioid misuse by chronic pain patients: a systematic review and literature synthesis. *Clinical Journal of Pain.* 24:497-508, 2008

45. Vowles KE, McEntee ML, Julnes PS, Frohe T, Ney J, van der Goes D. Rates of opioid misuse, abuse, and addiction in chronic pain: A systematic review and data synthesis. *Pain.* 156:569-576, 2015

46. World Health Organization: The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines, Geneva: World Health Organization, 1992

List of Tables

Table 1. *Possible Screening Test Results*

Table 2. *Results of Screener Validation for the SOAPP-R*

Table 3a. *Assuming 3% Prevalence - Hypothetical Results for the SOAPP-R*

Table 3b. *Assuming 25% Prevalence - Hypothetical Results for the SOAPP-R*

Table 3c. *Assuming 50% Prevalence - Hypothetical Results for the SOAPP-R*

Tables

Table 1. *Possible Screening Test Results*

Screener Result	Attribute Status		
	Present	Absent	
Present	True Positive	False Positive	*Positive Predictive Value*
Absent	False Negative	True Negative	*Negative Predictive Value*
	Sensitivity	*Specificity*	

Note. Sensitivity and specificity are calculated from the columns and are known as column-based indices; positive and negative predictive values are calculated from the rows and are knows as row-based indices.

Table 2. *Results of Screener Validation for the SOAPP-R*

SOAPP-R Result	Aberrant Drug Behavior Index		
	Aberrant User	Normal User	Row Total
Positive	62	47	109
Negative	15	99	114
Column Total	77	146	223

Note. Data calculated from Butler et al. (2008). A positive result indicates a score above 18 on the SOAPP-R. Prevalence = 77/223 = .345; test level = 109/223 = .489; Sensitivity = 62/77 = .81; Specificity = 99/146 = .68; Positive Predictive Value (PPV) = 62/109 = .57; Negative Predictive Value (NPV) = 99/114 = .87. PPV and NPV can also be calculated using Bayes' Theorem.

Table 3a. *Assuming 3% Prevalence - Hypothetical Results for the SOAPP-R*

SOAPP-R Result	Aberrant Drug Behavior Index		
	Aberrant User	Normal User	Row Total
Positive	6	69	75
Negative	1	147	148
Column Total	7	216	223

Note. Prevalence = 7/223 = .03; test level = 75/223 = .336; PPV = 6/75 = .080; NPV = 147/148 = .991; Overall efficiency = (6+147)/223 = .686; CPPV = (.08-.03)/(1-.03) = .043; CNPV = (.991-.97)/(1-.97) = .714. Meehl and Rosen's ratio = .087.

Table 3b. *Assuming 25% Prevalence - Hypothetical Results for the SOAPP-R*

SOAPP-R Result	Aberrant Drug Behavior Index		
	Aberrant User	Normal User	Row Total
Positive	45	53	98
Negative	11	114	125
Column Total	56	167	223

Note. Prevalence = 56/223 = .25; test level = 98/223 = .439; PPV = 45/98 = .461; NPV = 114/125 = .910; Overall efficiency = (45+114)/223 = .713; CPPV = (.46-.25)/(1-.25) = .277; CNPV = (.910-.75)/(1-.75) = .659; Meehl and Rosen's ratio = .843.

Table 3c. *Assuming 50% Prevalence - Hypothetical Results for the SOAPP-R*

SOAPP-R Result	Aberrant Drug Behavior Index		
	Aberrant User	Normal User	Row Total
Positive	90	36	126
Negative	21	76	97
Column Total	112*	112*	223

Note. Prevalence = 112.5/223 = .5; test level = 126/223 = .565; PPV = 90/126 = .714; NPV = 76/97 = .784; Overall efficiency = (90+76)/223 = .745; CPPV = (.72-.5)/(1-.5) = .434; CNPV = (.78-.5)/(1-.5) = .563; Meehl and Rosen's ratio = 2.53.
*Rounded up to nearest whole number

Highlights:

- Screening tests are an important risk evaluation tool for aberrant opioid use
- Screeners also have the potential to be misused in clinical decision-making
- Proper interpretation recognizes that sensitivity does not imply predictive value
- The variation in base rates for aberrant opioid use complicates predictive validity
- When base rates are low, positive results on screeners must be interpreted with caution