Evaluating the Screening Ability of Patient-Reported Outcome Instruments

Cheryl D Coon,[1] Lori D McLeod,[1] Lesley M Arnold,[2] Arthi B Chandran,[3] Susan A Martin[4],

[1] RTI Health Solutions, Research Triangle Park, NC, United States; [2] University of Cincinnati College of Medicine, Cincinnati, OH, United States; [3] Pfizer, Inc., New York, NY, United States; [4] RTI Health Solutions, Ann Arbor, MI, United States

BACKGROUND

- Assessments composed of patient-reported outcome (PRO) measures can be used in health care settings as screeners for medical conditions to achieve various objectives:
  - Quickly identify patients who are likely to benefit from a formal diagnostic evaluation
  - Avoid unnecessary diagnostic procedures, particularly when these are time- or resource-intensive or invasive in nature
  - Simply rule out the existence of a particular condition.
- The evaluation of a PRO screening assessment ideally occurs through analyses using a "gold standard" diagnosis of the condition of interest.

OBJECTIVE

- To provide an overview of a set of statistics often used to evaluate PRO screening measures, including definitions and interpretations.
- To illustrate the use of these statistics using a screening tool for fibromyalgia.

METHODS

Statistics

- The statistics assume that the condition of interest is binary (i.e., a person has or does not have the underlying condition), and the screener provides a binary result (i.e., the screener determines that the person is likely or unlikely to have the condition).
- Table 1 shows the four possible classifications that result from the use of a PRO measure to screen for a medical condition: true positive, false positive, false negative, and true negative.

Table 1. Example of Condition Versus Screener Results

		Condition	
		Positive	Negative
Screener	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

N = TP+FP+FN+TN.

- The goal of any PRO screening instrument is to maximize the true positives and true negatives while minimizing the false positives and false negatives. A number of statistics can be used to evaluate the performance of a screener to do this. (Please see yellow box on statistical measures for assessing screeners.)

Example

- The gold standard diagnosis of fibromyalgia was established by the American College of Rheumatology (ACR) criteria[1] and includes the clinical evaluation of pain at 18 tender points throughout the body. These criteria have been reported to be used by most rheumatologists (~ 65%) but are believed to be much less commonly used in usual practice by primary care physicians.
- The Arnold Fibromyalgia Diagnostic Screen (AFDS)[2] was developed for use by primary care physicians to easily screen for the likely presence of fibromyalgia using a PRO instrument coupled with an abbreviated clinician examination.
- Two scoring models for the AFDS are considered here: the AFDS Primary, which includes only selected PRO responses, and the AFDS Alternative, which includes the selected PRO responses coupled with selected clinician examinations.
- The AFDS was completed by 141 patients, 73 of whom had an ACR diagnosis of fibromyalgia and 68 of whom did not have an ACR diagnosis of fibromyalgia (Table 2).

Table 2. Comparison of AFDS Screen Results to the ACR Criteria Diagnosis of Fibromyalgia

		ACR Criteria	
		Fibromyalgia	No Fibromyalgia
AFDS Primary	Positive screen	50	21
	Negative screen	23	47
AFDS Alternative	Positive screen	50	12
	Negative screen	23	56

STATISTICAL MEASURES FOR ASSESSING SCREENERS

Sensitivity $= P(\text{Positive Screen}|\text{Positive Condition}) = \dfrac{TP}{TP+FN}$

- Provides the probability of a positive screen in people with the condition—i.e., the percentage of patients with the condition who are correctly classified.
- Helps rule out a disease—if sensitivity is high, then people with the condition are likely to have a positive screen, so the probability is low of having the condition when the screen is negative.
- Ranges from 0 to 1. Higher values are better.
- Evaluates the screening ability of a PRO but limited interpretation in clinical practice because it gives the rate of positive screens for people known to have underlying condition (which would be unknown at the time of the screener).

Specificity $= P(\text{Negative Screen}|\text{Negative Condition}) = \dfrac{TN}{TN+FP}$

- Provides the probability of a negative screen in people without the condition—i.e., the percentage of patients without the condition who are correctly classified
- Helps rule in a disease—if specificity is high, then people without the condition are likely to have a negative screen, so the probability is low of not having the condition when the screen is positive.
- Ranges from 0 to 1. Higher values are better.
- Same limitation as sensitivity.

Youden's Index = Sensitivity+Specificity−1

- Communicates in a simple way the sensitivity and specificity as one number.
- Ranges from 0 to 1—higher values are better.
- Does not allow for the consideration of tradeoffs of high sensitivity for low specificity and vice versa.

Positive predictive value (PPV) $= P(\text{Positive Condition}|\text{Positive Screen}) =$

$$\frac{TP}{TP+FP} \;\langle\text{OR}\rangle\; \frac{\text{Sensitivity} \times \text{Prevalence}}{(\text{Sensitivity}\times\text{Prevalence})+\left((1-\text{Specificity})\times(1-\text{Prevalence})\right)}$$

- Provides the probability of having the condition in people with a positive screen—i.e., the percentage of patients with a positive screen who are correctly classified.
- Ranges from 0 to 1. Higher values are better.
- Greatly depends on the prevalence of the condition. Low prevalence conditions will naturally have low PPV. In rare conditions, there is more uncertainty that a positive screen indicates the presence of the condition.

Negative predictive value (NPV) $= P(\text{Negative Condition}|\text{Negative Screen}) =$

$$\frac{TN}{TN+FN} \;\langle\text{OR}\rangle\; \frac{\text{Specificity}\times(1-\text{Prevalence})}{\left((1-\text{Sensitivity})\times\text{Prevalence}\right)+\left(\text{Specificity}\times(1-\text{Prevalence})\right)}$$

- Provides the probability of not having the condition in people with a negative screen—i.e., the percentage of patients with a negative screen who are correctly classified.
- Ranges from 0 to 1. Higher values are better.
- Greatly depends on the prevalence of the condition. High prevalence conditions will naturally have low NPV. In common conditions, there is more uncertainty that a negative screen indicates the absence of the condition.

Positive likelihood ratio (LR+) $= \dfrac{\text{Sensitivity}}{1-\text{Specificity}}$

- Compares the probability of a positive screen in people with the condition to the probability of a positive screen in people without the condition.
- Ranges from 0 to infinity. Values > 1 suggest that a positive screen is associated with a higher probability of having the condition.
- Is useful for making head-to-head comparisons for screeners but does not aid in interpreting the result of a screener for a particular individual.

Negative likelihood ratio (LR–) $= \dfrac{1-\text{Sensitivity}}{\text{Specificity}}$

- Compares the probability of a negative screen in people with the condition to the probability of a negative screen in people without the condition.
- Ranges from 0 to infinity. Values < 1 suggest that a negative screen is associated with a higher probability of not having the condition.
- Same limitation as LR+.

$$\textbf{Kappa} = \frac{\left[\frac{TP}{N}+\frac{TN}{N}\right]-\left[\left(\frac{TP}{N}+\frac{FP}{N}\right)\times\left(\frac{TP}{N}+\frac{FN}{N}\right)+\left(\frac{TN}{N}+\frac{FN}{N}\right)\times\left(\frac{TN}{N}+\frac{FP}{N}\right)\right]}{1-\left[\left(\frac{TP}{N}+\frac{FP}{N}\right)\times\left(\frac{TP}{N}+\frac{FN}{N}\right)+\left(\frac{TN}{N}+\frac{FN}{N}\right)\times\left(\frac{TN}{N}+\frac{FP}{N}\right)\right]}$$

- Measures agreement between the condition and the screener.
- Ranges from 0 to 1. Higher values are better.
- Is higher when the probability of having the condition is roughly the same as the probability of not having the condition, so kappa might be deflated when prevalence is low (or high).

Accuracy $= \dfrac{TP+TN}{N}$

- Measures proportion of people correctly classified by the screener.
- Ranges from 0 to 1. Higher values are better.
- Can be improved if the screener were to classify everyone as positive (or everyone as negative), so it is possible for an improvement in accuracy to result in a less useful screener.

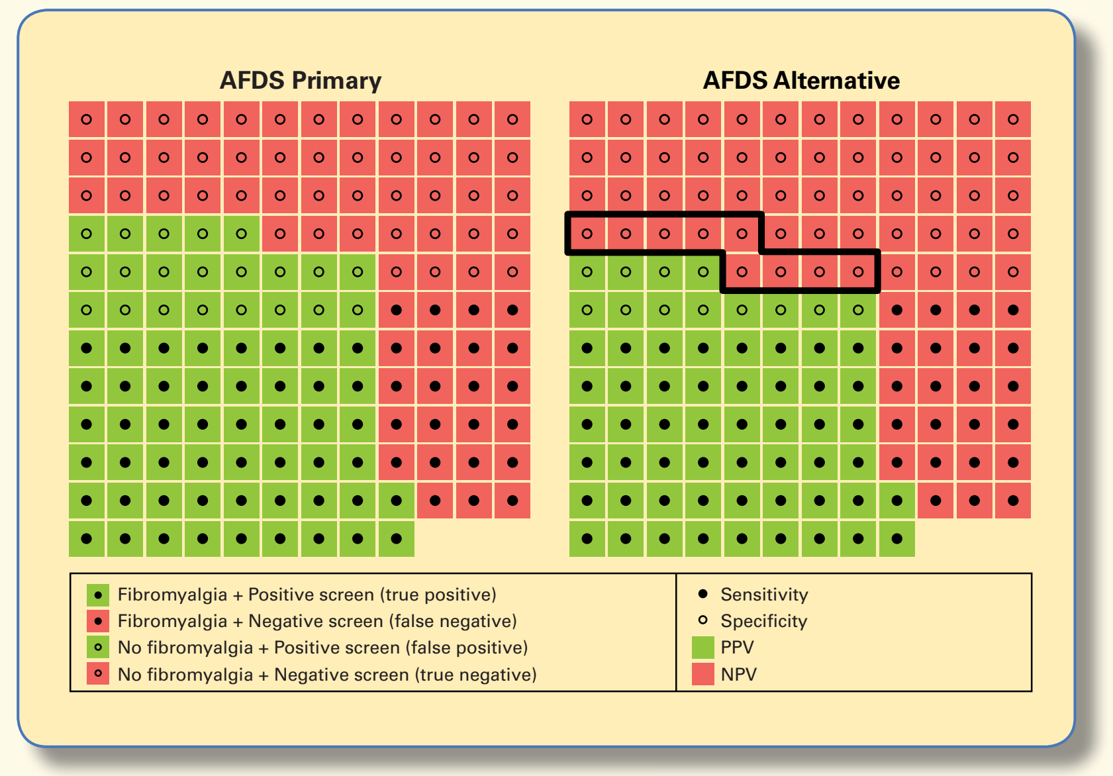
Odds ratio $= \dfrac{TP\times TN}{FP\times FN}$

- Measures ratio of correct classifications to incorrect classifications.
- Ranges from 0 to infinity. Values > 1 indicate that a positive screen is associated with a higher probability of having the condition and a negative screen is associated with a higher probability of not having the condition.
- Makes head-to-head comparisons for screeners but does not aid in interpreting the result of a screener for a particular individual.

RESULTS

- Figure 1 shows a graphical demonstration of sensitivity, specificity, PPV, and NPV based on Loong's work.[3]

Figure 1. Diagram of AFDS Results



- Sensitivity of AFDS Primary $= \dfrac{50}{50+23} = 0.68$
- Sensitivity of AFDS Alternative $= \dfrac{50}{50+23} = 0.68$
  - Of patients with an ACR diagnosis of fibromyalgia, the AFDS Primary and AFDS Alternative screener both correctly identified 68% of patients.
- Specificity of AFDS Primary $= \dfrac{47}{47+21} = 0.69$
- Specificity of AFDS Alternative $= \dfrac{56}{56+12} = 0.82$
  - Of patients without an ACR diagnosis of fibromyalgia, the AFDS Primary screener correctly identified 69% of patients, whereas the AFDS Alternative screener correctly identified 82% of patients.
- PPV of AFDS Primary =

$$\frac{50}{50+21} \;\langle\text{OR}\rangle\; \frac{0.68\times0.02}{(0.68\times0.02)+\left((1-0.69)\times(1-0.02)\right)} = 0.70 \;\langle\text{OR}\rangle\; 0.04$$

- PPV of AFDS Alternative =

$$\frac{50}{50+12} \;\langle\text{OR}\rangle\; \frac{0.68\times0.02}{(0.68\times0.02)+\left((1-0.82)\times(1-0.02)\right)} = 0.81 \;\langle\text{OR}\rangle\; 0.07$$

  - Of patients from the population recruited for this study with an AFDS Primary positive screen, 70% of them are likely to have an ACR diagnosis of fibromyalgia. However, of patients in the general population with an AFDS Primary positive screen, only 4% are likely to have an ACR diagnosis of fibromyalgia.
  - Of patients from the population recruited for this study with an AFDS Alternative positive screen, 81% are likely to have an ACR diagnosis of fibromyalgia. However, of patients in the general population with an AFDS Alternative positive screen, 7% are likely to have an ACR diagnosis of fibromyalgia.
  - The prevalence of fibromyalgia in the population recruited for this study was 52%; therefore, both AFDS models offered an improvement in identifying patients with fibromyalgia. However, the rates of 70% and 81% applied only to clinical practice settings with similar criteria as those patients screened in this study (i.e., fibromyalgia or chronic pain for at least 3 months prior).
  - The more generalizable PPV was the one that incorporated the real-world prevalence, which was estimated to be 2%. These rates also indicated that either AFDS model offers some value in identifying patients with fibromyalgia. The AFDS Primary did twice as well as chance alone, whereas the AFDS Alternative offered value more than three-fold over chance.
- NPV of AFDS Primary =

$$\frac{47}{47+23} \;\langle\text{OR}\rangle\; \frac{0.69\times(1-0.02)}{\left((1-0.68)\times0.02\right)+\left(0.69\times(1-0.02)\right)} = 0.67 \;\langle\text{OR}\rangle\; 0.99$$

- NPV of AFDS Alternative =

$$\frac{56}{56+23} \;\langle\text{OR}\rangle\; \frac{0.82\times(1-0.02)}{\left((1-0.68)\times0.02\right)+\left(0.82\times(1-0.02)\right)} = 0.71 \;\langle\text{OR}\rangle\; 0.99$$

  - Of patients from the population recruited for this study with an AFDS Primary negative screen, only 67% are likely to not have an ACR diagnosis of fibromyalgia. However, of patients in the general population with an AFDS Primary negative screen, 99% are likely to not have an ACR diagnosis of fibromyalgia.
  - Of patients from the population recruited for this study with an AFDS Alternative negative screen, 71% are likely to not have an ACR diagnosis of fibromyalgia. However, of patients in the general population with an AFDS Alternative negative screen, 99% are likely to not have an ACR diagnosis of fibromyalgia.
  - The percentage of people without fibromyalgia in the population recruited for this study was 48%; therefore, both AFDS models offered an improvement in identifying patients without fibromyalgia. However, the rates of 67% and 71% applied only to clinical practice settings with similar criteria as the patients screened in this study.
  - The more generalizable PPV was the one that incorporated the real-world prevalence, which estimated that 98% of the general population did not have fibromyalgia. The AFDS screeners offered some value for identifying people without fibromyalgia.
- LR+ for AFDS Primary $= \dfrac{0.68}{1-0.69} = 2.19$
- LR+ for AFDS Alternative $= \dfrac{0.68}{1-0.82} = 3.78$
  - The probability of an AFDS Primary positive screen in patients with an ACR diagnosis of fibromyalgia was 2.19 times larger than the probability of an AFDS Primary positive screen in patients without an ACR diagnosis of fibromyalgia.
  - The probability of an AFDS Alternative positive screen in patients with an ACR diagnosis of fibromyalgia was 3.88 times larger than the probability of an AFDS Alternative positive screen in patients without an ACR diagnosis of fibromyalgia.
- LR– for AFDS Primary $= \dfrac{1-0.68}{0.69} = 0.46$
- LR– for AFDS Alternative $= \dfrac{1-0.68}{0.82} = 0.39$
  - The probability of an AFDS Primary negative screen in patients without an ACR diagnosis of fibromyalgia was 2.17 (i.e., 1/0.46) times larger than the probability of an AFDS Primary negative screen in patients with an ACR diagnosis of fibromyalgia.
  - The probability of an AFDS Alternative negative screen in patients without an ACR diagnosis of fibromyalgia was 2.56 (i.e., 1/0.39) times larger than the probability of an AFDS Alternative negative screen in patients with an ACR diagnosis of fibromyalgia.
- Kappa for AFDS Primary =

$$\frac{\left[\frac{50}{141}+\frac{47}{141}\right]-\left[\left(\frac{50}{141}+\frac{21}{141}\right)\times\left(\frac{50}{141}+\frac{23}{141}\right)+\left(\frac{47}{141}+\frac{23}{141}\right)\times\left(\frac{47}{141}+\frac{21}{141}\right)\right]}{1-\left[\left(\frac{50}{141}+\frac{21}{141}\right)\times\left(\frac{50}{141}+\frac{23}{141}\right)+\left(\frac{47}{141}+\frac{23}{141}\right)\times\left(\frac{47}{141}+\frac{21}{141}\right)\right]} = 0.38$$

- Kappa for AFDS Alternative =

$$\frac{\left[\frac{50}{141}+\frac{56}{141}\right]-\left[\left(\frac{50}{141}+\frac{12}{141}\right)\times\left(\frac{50}{141}+\frac{23}{141}\right)+\left(\frac{56}{141}+\frac{23}{141}\right)\times\left(\frac{56}{141}+\frac{12}{141}\right)\right]}{1-\left[\left(\frac{50}{141}+\frac{12}{141}\right)\times\left(\frac{50}{141}+\frac{23}{141}\right)+\left(\frac{56}{141}+\frac{23}{141}\right)\times\left(\frac{56}{141}+\frac{12}{141}\right)\right]} = 0.51$$

  - Kappa can be interpreted as an intraclass correlation coefficient.
  - There was fair agreement between the AFDS Primary screener and the ACR fibromyalgia diagnosis.
  - There was moderate agreement between the AFDS Alternative screener and the ACR fibromyalgia diagnosis.
- Accuracy for AFDS Primary $= \dfrac{50+47}{141} = 0.69$
- Accuracy for AFDS Alternative $= \dfrac{50+56}{141} = 0.75$
  - The AFDS Primary screener correctly classified 69% of patients who were screened for having fibromyalgia, or the AFDS Primary screener produced incorrect fibromyalgia screens 31% of the time.
  - The AFDS Alternative screener correctly classified 75% of patients who were screened for having fibromyalgia, or the AFDS Alternative screener produced incorrect fibromyalgia screens 25% of the time.
- Odds ratio for AFDS Primary $= \dfrac{50\times47}{21\times23} = 4.87$
- Odds ratio for AFDS Alternative $= \dfrac{50\times56}{12\times23} = 10.14$
  - There is almost a fivefold greater odds of an ACR diagnosis of fibromyalgia when the AFDS Primary measure produced a positive screen versus a negative screen. There is almost a fivefold greater odds of not having an ACR diagnosis of fibromyalgia when the AFDS Primary measure produced a negative screen versus a positive screen.
  - There is a tenfold greater odds of an ACR diagnosis of fibromyalgia when the AFDS Alternative measure produced a positive screen versus a negative screen. There is a tenfold greater odds of not having an ACR diagnosis of fibromyalgia when the AFDS Alternative measure produced a negative screen versus a positive screen.

CONCLUSIONS

- The AFDS Alternative model is superior to the AFDS Primary model on specificity, PPV, LR+ and LR-, kappa, accuracy, and odds ratio.
  - The AFDS Alternative is preferable for ruling in the presence of fibromyalgia because people without fibromyalgia are likely to have a negative screen, so the probability is low of not having fibromyalgia when the AFDS Alternative screen is positive.
- There are a number of agreement statistics to consider when evaluating screeners, each with pros and cons.
  - Researchers must understand what conclusions can and cannot be made with each agreement statistic.
  - No one agreement statistic paints the full picture of the value of a screener, so researchers should consider a battery of statistics.
  - It is important to consider the prevalence rate of the sample recruited for the screening study and to compare it with the prevalence in the population for which the screener is intended.
- Ultimately, it is desirable to minimize false positives and false negatives. However, it is difficult to minimize both simultaneously, so tradeoffs must be considered.
  - If the formal diagnostic procedures are particularly invasive or time- or resource-intensive (e.g., lumbar puncture), then screeners should minimize false positives so that patients are not unnecessarily subjected to these procedures.
  - If the disease has exceptional risks when left undiagnosed (e.g., breast cancer), then screeners should minimize false negatives so that patients in need of treatment do not go without.

REFERENCES

1. Wolfe F, Smythe HA, Yunnus MB, Bennett RM, Bombardier C, Goldenberg DL, et al. The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. Arthritis Rheum. 1990 Feb;33(2):160-72.
2. Arnold L, Stanford S, Welge J, Crofford L. Development and testing of the fibromyalgia diagnostic screen for primary care. J Pain 2011;12(4);Suppl 1:P8.
3. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ. 2003 Sep 27;327(7417):716-9.

KEY SOURCES IN DIAGNOSTIC TESTING

Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. BMJ. 1994 Jun 11;308(6943):1552.

Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ. 1994 Jul 9;309(6947):102.

Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004 Jul 17;329(7458):168-9.

Grimes DA, Schultz KF. Uses and abuses of screening tests. Lancet. 2002 Mar 9;359(9309):881-4.

CONTACT INFORMATION

Cheryl D Coon
Director, Psychometrics

RTI Health Solutions
200 Park Offices Drive
Research Triangle Park, NC 27709

Phone: +1.702.818.4249
Fax: +1.702.818.4249
E-mail: ccoon@rti.org

Presented at: ISPOR 16th Annual International Meeting
May 21-25, 2011
Baltimore, MD, United States