Screening for Obstructive Sleep Apnea: Bayes Weighs in

Matt T. Bianchi*

Sleep Division, Department of Neurology, Massachusetts General Hospital, Wang 720, Boston, MA 02114, USA

Abstract: A fundamental challenge associated with screening tests is recognition of the impact of disease prevalence upon the predictive value of the result. For example, in the common circumstance of screening for low prevalence diseases, even good tests may have unacceptably high false positive rates. The converse situation, screening in high prevalence populations, is less common but occurs with obstructive sleep apnea (OSA): even ostensibly good screening tests may have unacceptably high false negative rates. The challenge of recognizing false negative OSA screening results has important implications as screens are increasingly implemented in high risk populations. This raises two clinically important questions: 1) How sensitive and specific should a screening test be to minimize false negative results across a spectrum of baseline prevalence; and 2) Given a screening test with known sensitivity and specificity, in what range of disease prevalence may the test be reasonably applied? Simple graphics are presented that incorporate acceptable risk thresholds and illustrate combinations of prevalence, sensitivity, and specificity in which disease probability remains high despite a negative test result. Adopting a Bayesian approach, together with acceptable risk thresholds, may help to avoid potential pitfalls of false negative screening results.

Although the baseline prevalence in the United States of obstructive sleep apnea (OSA) is estimated in the range of 3-28% [1], in certain populations the prevalence is much higher. For example, higher prevalence has been reported for patients with refractory epilepsy (33%) [2], recent stroke (58%) [3], refractory hypertension (63%) [4], heart failure (35%) [5], polycystic ovary syndrome (65%) [6], anterior ischemic optic neuropathy (89%) [7], children with Down's Syndrome (63%) [8], and those undergoing bariatric surgery (80%) [9-10]. Treatment of OSA may be associated with improved outcomes in some settings [11]. One particular area of interest involves OSA in the peri-operative setting, which may be associated with complications and/or longer hospital stay [12-13]. Given the limited resources for laboratory polysomnogram (PSG) as a screening tool, screening questionnaires may assist in OSA risk stratification.

A screening tool with the acronym "STOP-Bang" was recently developed to provide dichotomous risk stratification (low versus high OSA risk) in general surgical populations [12-13]. In this population, adverse outcomes involving respiratory compromise were increased in the peri-operative period in those with OSA, defined as an apnea-hypopnea index (AHI) >5 by standard PSG. The STOP-Bang screen involves four yes/no patient questions (snoring, daytime tiredness, observed nocturnal apnea, high blood pressure), combined with four yes/no clinical features (BMI >35, age >50, neck circumference >40 cm, male gender). This simple tool was validated in a surgical population (177 patients, including general, gynecologic, orthopedic, urologic, plastic, ophthalmologic, and neurosurgery cases) [13], with a positive result defined as ≥3 "yes" responses. Within the study limitations (including high refusal and no-show rates for PSG), the sensitivity for OSA varied with OSA severity (as expected), defined by AHI: >5, 83.6% (CI 75.8-89.7%); >15, 92.9% (CI 84.1-97.6%); >30, 100% (CI 91.0-100%). Specificity was much lower, and also varied (as expected) with the AHI: ≥5, 56.4% (CI 42.3-69.7%); ≥15, 43.0% (CI 33.5-52.9%); ≥30, 37.0% (CI 28.9-45.6%). Although there were no serious complications or deaths associated with OSA, respiratory complications were more than doubled for AHI ≥5. The STOP-Bang screen was superior to the STOP, Berlin, and American Society of Anesthesiology (ASA) questionnaires in terms of sensitivity and specificity for OSA.

How should a clinician approach OSA evaluation and treatment in high risk populations such as this? When pretest probability is high, a positive screen adds little, while a negative screen in principle should allow assignment of risk stratification low enough (using risk thresholds) to forego further testing. Understanding the impact of prevalence (as well as sensitivity and specificity) on the negative predictive value (NPV) of OSA screens is critical for avoiding the potential for false reassurance from a negative screen. In other words, one should recognize and estimate the residual OSA risk after a negative screen. Formal decision analysis may eventually facilitate risk threshold determinations by balancing the morbidity and cost of peri-operative complications against the cost and time delays of systematic PSG testing versus screening-based stratification. In the STOP-Bang study, the prevalence of OSA was quite high (69%). While the screen detected moderate and severe OSA with higher negative predictive value (NPV of 90 and 100%, respectively), the screen had only 60% NPV for mild OSA with AHI >5. The complication rates were similar at the three levels of OSA severity, suggesting that the sensitivity and specificity for patients with AHI >5 is a clinically relevant cutoff. Unexpected results, such as a negative screen in a high risk patient, represent a challenge to diagnostic interpretation, especially when the pre-test probability (pre-TP) is uncertain [14]. OSA screening faces the potentially false

*Address correspondence to this author at the Sleep Division, Department of Neurology, Massachusetts General Hospital, Wang 720, Boston, MA 02114, USA; E-mail: mtbianchi@partners.org

1874-6209/09

reassurance of a negative screen: although the negative result lowers the probability of OSA, the remaining risk may be unacceptably high.

Introducing a threshold of OSA probability, above which formal PSG testing would be clinically warranted, could guide clinical interpretation and decrease the risks associated with false negative screening results. For example, if decision analysis (or consensus estimation) concluded that 25% risk of OSA was "acceptable" in a given population, then the NPV of a screening test result should be higher than 75% (100% minus the acceptable risk threshold). In the study by Chung *et al.* the NPV for the screen (to detect AHI >5) was only 60.8%, which means a nearly 40% chance of having OSA despite the negative screen. The chance of moderately severe OSA (AHI 15-30) despite a negative screening test was 10% (90% NPV). Having a threshold framework would be helpful to guide subsequent decisions, specifically by addressing the question: How much residual risk of OSA, after a negative screen, is acceptable? The answer may depend on the clinical population (a 10% risk of OSA remaining after a negative screen might be unacceptable in a professional driver, but acceptable for an asymptomatic adult).

Another approach could involve portable screening devices in high-risk populations. For example, patients felt to have unacceptably high risk despite a negative STOP-Bang screen could be tested at home using one of multiple home testing options in order to solidify their risk assessment [15-16]. As these devices are less expensive and more convenient than PSG, they are well-poised as an intermediate step in OSA evaluation, with relative savings of cost and time. Although serial testing involves using the post-test probability (post-TP) after the first screen as the pre-TP for the second step of screening, this requires that the two tests are independent. Nevertheless, as a first approximation, one could follow a negative screening survey with a home study, with the goal of improving the overall NPV without resorting to formal PSG.

Although the idea of acceptable risk involves many factors, not all of which are quantitative, clinical decision making may be guided by the relationship between prevalence, sensitivity and specificity. A simple graphic is presented (Fig. **1**) to address the following questions: 1) Given a known population prevalence, how sensitive and specific should a screening test be to achieve a clinically relevant NPV; and 2) Given a screening test with known sensitivity and specificity, in what population (in terms of prevalence) can the appropriate NPV goal be achieved.

The NPV for a spectrum of theoretical tests, spanning combinations of sensitivity and specificity from 5-95%, is shown in Fig. (**1**). The NPV of the same spectrum of tests varies with baseline prevalence, ranging from 2% to 70% in panels A through F. Thus, each panel represents an easily referenced "landscape" highlighting the dependence of NPV on sensitivity, specificity, and prevalence. Note that NPV depends more strongly on sensitivity than specificity, but is also profoundly dependent on prevalence. Potential risk thresholds are highlighted as follows: the gray shading indicates >50% NPV, the dotted line border indicates >90% NPV, and the solid line border indicates >95% NPV. Note that only combinations of sensitivity and specificity percentages whose sum exceeds 100% are shown, since failure to meet this criteria renders a test meaningless (by producing positive likelihood ratios <1, and negative likelihood ratios >1, which are paradoxical in Bayesian terms, as positive tests would reduce disease probability and negative tests would increase disease probability).

When prevalence is low at baseline, for example, 2% (Fig. **1A**), all tests in the landscape have a NPV >98%, and so are shaded gray and bounded by a solid line. As prevalence increases, the NPV landscape shifts to reflect not only the higher baseline disease probability, but also the extent to which the probability is adjusted downward for each theoretical test (sensitivity-specificity pair). Note that the impact of prevalence on the predictive value is non-linear. The baseline prevalence can thus be considered a "boundary" condition, setting the minimum NPV of any negative screen (regardless of sensitivity or specificity), from which the negative result further modifies the probability. At 10% prevalence (Fig. **1B**), all tests provide a NPV of at least 90% and are thus all shaded and bounded by a dotted line. Those tests providing a NPV >95%, approximately half of the landscape, are indicated by the solid line. At any prevalence below 50%, the NPV will be at least 50%, and thus all tests in panels C and D are also shaded gray. As prevalence increases, for example, to 30% or 50%, only the most powerful tests provide NPV greater than 90 or 95% (Fig. **1C** and **1D**). At 60 and 70% prevalence, such as occurs in some surgical populations as above, no tests in this range of sensitivity and specificity (max 95% each) can provide NPV >90%, and only a subset of tests can even provide NPV>50% (see non-shaded tests in Fig. **1E** and **1F**). It is for these high prevalence conditions, when a negative screen might be followed by a home screening device, to improve risk stratification and, if negative, reduce the residual risk below the physician-defined acceptable threshold.

In the population studied by Chung *et al.* with baseline prevalence of ~70% for OSA with at least AHI >5 severity, the STOP-Bang questionnaire yielded NPV of 60% (this can be estimated by finding in Fig. (**1F**) the sensitivity/specificity pair closest to 83.6% and 56.4%). The figure provides a simple reference to consider the NPV of a test across a spectrum of prevalence conditions, especially if thresholds of acceptable risk are to be utilized, and thus facilitate estimation of residual risk despite a negative screen.

CONCLUSION

Inaccurate estimation of (or failure to consider) the pre-TP of disease represents an important (and avoidable) pitfall in diagnostic test interpretation [17-21]. Screening high prevalence populations requires particular caution in interpreting negative results, which may falsely reassure despite substantial "residual" OSA risk. A Bayesian approach, incorporating OSA prevalence into decision making, provides a useful quantitative framework, particularly when combined with acceptable risk thresholds. The main clinical pitfall in screening high risk populations is that a screening test alone may not have sufficient discriminative power for a negative result to effectively lower disease probability below an acceptable OSA risk threshold. Although acceptable risk may be an elusive target (depending on patient-specific clinical features), it is notable that even the best of the four screens described above (STOP-Bang) provided a NPV of ~60% in a high prevalence population, leaving a substantial residual

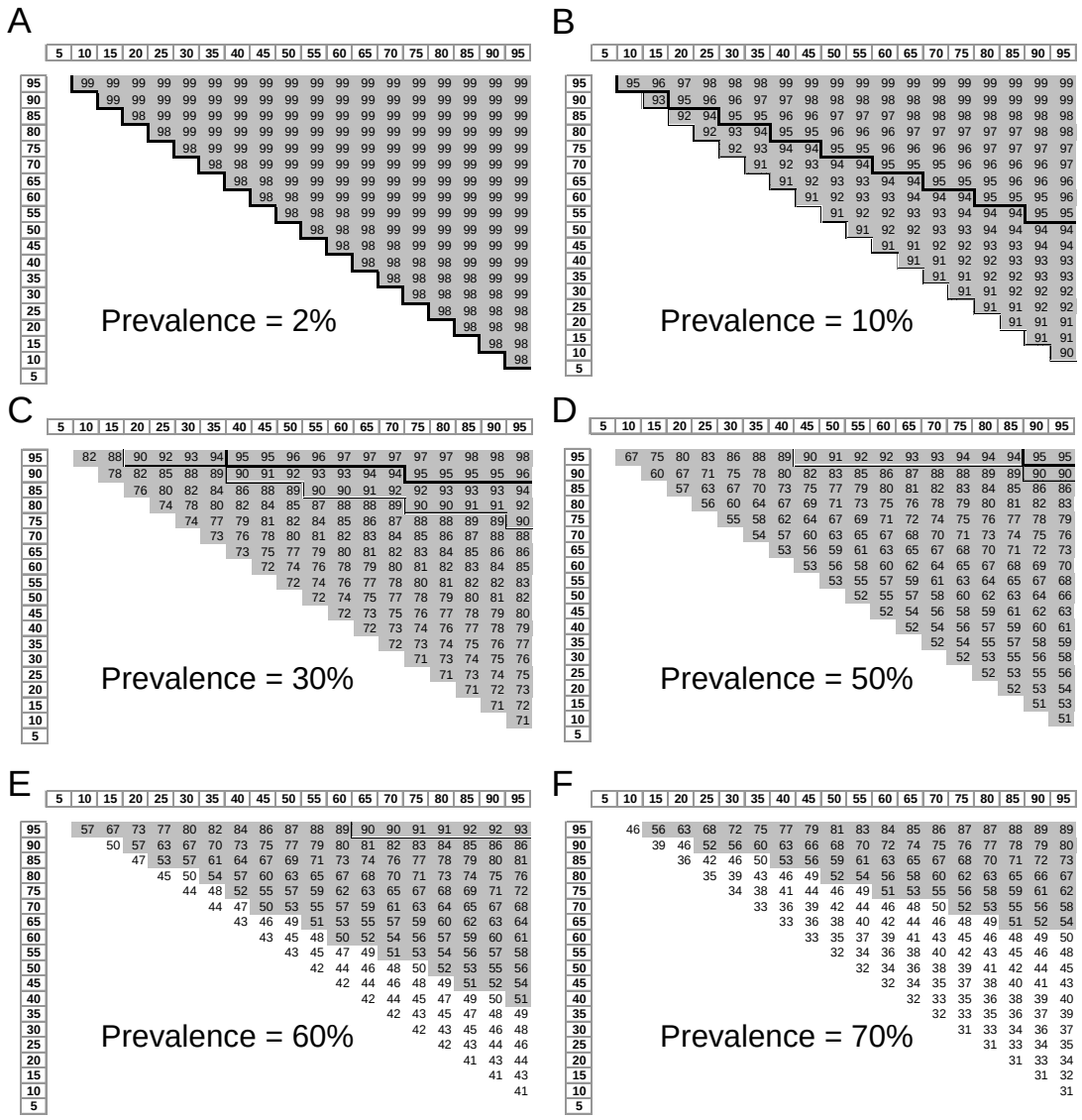

Fig. (1). Negative predictive value depends on sensitivity, specificity, and prevalence.

NPV is shown for combinations of sensitivity (Y-axis) and specificity (X-axis) ranging from 5-95%. Each panel (**A-F**) represents a different baseline disease prevalence, ranging from 2 to 70% as indicated. Gray shading indicates tests providing a NPV of at least 50%, while tests providing a NPV of at least 90% are indicated by a dotted line border, and tests providing a NPV of at least 95% are indicated by a solid line border. Note that the NPV for each test (sensitivity-specificity pair) varies across baseline prevalence. NPV was calculated (true negatives divided by all negatives) *via* the standard 2x2 "box" method of dichotomized disease absence versus presence, and positive versus negative test results.

risk despite a negative screening result. Since it may not be feasible to pursue PSG testing in all patients of high-prevalence populations, utilizing screening tests for risk stratification is a reasonable first step. Further data linking medical and surgical risks to OSA severity (AHI-based) will add important information to this evolving clinical challenge. Home screening may also be incorporated into the decision making process as an intermediate step between screening surveys and dedicated PSG testing. Clinical decision making can be further enhanced by incorporating (qualitative or quantitative) thresholds for testing or treating. Finally, recognizing the relationship of NPV with baseline prevalence may help to avoid the potential pitfalls of false negative screening results.

ACKNOWLEDGEMENTS

This work was not funded. The author thanks Dr. Geoff Gilmartin for valuable comments.

REFERENCES

[1] Young T, Peppard PE, Gottlieb DJ. Epidemiology of obstructive sleep apnea: a population health perspective. Am J Respir Crit Care Med 2002; 165: 1217-39.

[2] Malow BA, Levy K, Maturen K, Bowes R. Obstructive sleep apnea is common in medically refractory epilepsy patients. Neurology 2000; 55: 1002-7.

[3] Bassetti CL, Milanova M, Gugger M. Sleep-disordered breathing and acute ischemic stroke: diagnosis, risk factors, treatment, evolution, and long-term clinical outcome. Stroke 2006; 37: 967-72.

[4] Logan AG, Perlikowski SM, Mente A, *et al*. High prevalence of unrecognized sleep apnoea in drug-resistant hypertension. J Hypertens 2001; 19: 2271-7.

[5] Sin DD, Fitzgerald F, Parker JD, Newton G, Floras JS, Bradley TD. Risk factors for central and obstructive sleep apnea in 450 men and women with congestive heart failure. Am J Respir Crit Care Med 1999; 160: 1101-6.

[6] Fogel RB, Malhotra A, Pillar G, Pittman SD, Dunaif A, White DP. Increased prevalence of obstructive sleep apnea syndrome in obese women with polycystic ovary syndrome. J Clin Endocrinol Metab 2001; 86: 1175-80.

[7] Palombi K, Renard E, Levy P, *et al*. Non-arteritic anterior ischaemic optic neuropathy is nearly systematically associated with obstructive sleep apnoea. Br J Ophthalmol 2006; 90: 879-82.

[8] Marcus CL, Keens TG, Bautista DB, von Pechmann WS, Ward SL. Obstructive sleep apnea in children with Down syndrome. Pediatrics 1991; 88: 132-9.

[9] O'Keeffe T, Patterson EJ. Evidence supporting routine polysomnography before bariatric surgery. Obes Surg 2004; 14: 23-6.

[10] Lopez PP, Stefan B, Schulman CI, Byers PM. Prevalence of sleep apnea in morbidly obese patients who presented for weight loss surgery evaluation: more evidence for routine screening for obstructive sleep apnea before weight loss surgery. Am Surg 2008; 74: 834-8.

[11] Bradley TD, Floras JS. Obstructive sleep apnoea and its cardiovascular consequences. Lancet 2009; 373: 82-93.

[12] Chung F, Yegneswaran B, Liao P, *et al*. STOP questionnaire: a tool to screen patients for obstructive sleep apnea. Anesthesiology 2008; 108: 812-21.

[13] Chung F, Yegneswaran B, Liao P, *et al*. Validation of the Berlin questionnaire and American Society of Anesthesiologists checklist as screening tools for obstructive sleep apnea in surgical patients. Anesthesiology 2008; 108: 822-30.

[14] Bianchi MT, Alexander BM, Cash SS. Incorporating Uncertainty Into Medical Decision Making: An Approach to Unexpected Test Results. Med Decis Making 2009; 29: 116-24.

[15] Flemons WW, Littner MR, Rowley JA, *et al*. Home diagnosis of sleep apnea: a systematic review of the literature. An evidence review cosponsored by the American Academy of Sleep Medicine, the American College of Chest Physicians, and the American Thoracic Society. Chest 2003; 124: 1543-79.

[16] Ward Flemons W, McNicholas WT. Clinical prediction of the sleep apnea syndrome. Sleep Med Rev 1997; 1: 19-32.

[17] Bianchi MT, Alexander BM. Evidence based diagnosis: does the language reflect the theory? BMJ 2006; 333: 442-5.

[18] Attia JR, Nair BR, Sibbritt DW, *et al*. Generating pre-test probabilities: a neglected area in clinical decision making. Med J Aust 2004; 180: 449-54.

[19] Lyman GH, Balducci L. The effect of changing disease risk on clinical reasoning. J Gen Intern Med 1994; 9: 488-95.

[20] Lyman GH, Balducci L. Overestimation of test effects in clinical judgment. J Cancer Educ 1993; 8: 297-307.

[21] Phelps MA, Levitt MA. Pretest probability estimates: a pitfall to the clinical utility of evidence-based medicine? Acad Emerg Med 2004; 11: 692-4.

Received: May 08, 2009 Revised: September 02, 2009 Accepted: October 26, 2009