The Reliability and Diagnostic Utility of the Orthopaedic Clinical Examination

1

The health sciences and medical professions continue to focus on evidence-based practice defined as the integration of the best available research evidence and clinical expertise with the patient's values.[1,2] Evidence should be incorporated into all aspects of physical therapy patient and client management, including the examination, evaluation, diagnosis, prognosis, and intervention. Perhaps the most crucial component is a careful, succinct clinical examination that can lead to an accurate diagnosis, the selection of appropriate interventions, and the determination of a prognosis. Thus, it is of utmost importance to incorporate evidence of how well clinical tests and measures can distinguish between patients who present with specific musculoskeletal disorders and patients who do not.[1,2]

The diagnostic process entails obtaining a patient history, developing a working hypothesis, and selecting specific tests and measures to confirm or refute the formulated hypothesis. The clinician must determine the pretest (before the evaluation) probability that the patient has a particular disorder. Based on this information the clinician selects appropriate tests and measures that will help determine the posttest (after the evaluation) probability of the patient having the disorder, until a degree of certainty has been reached such that patient management can begin (the *treatment threshold*). The purpose of clinical tests is not to obtain diagnostic certainty but rather to reduce the level of uncertainty until the treatment threshold is reached.[2] The concepts of pretest and posttest probability and treatment threshold are elaborated later in this chapter.

As the number of reported clinical tests and measures continues to grow, it is essential to thoroughly evaluate a test's diagnostic properties before incorporating the test into clinical practice.[3] Integrating the best evidence available for the diagnostic utility of each clinical test is essential in determining an accurate diagnosis and implementing effective, efficient treatment. It seems only sensible for clinicians and students to be aware of the diagnostic properties of tests and measures and to know which have clinical utility. This text assists clinicians and students in selecting tests and measures to ensure the appropriate classification of patients and to allow for quick implementation of effective management strategies.

The assessment of diagnostic tests involves examining several properties, including reliability and diagnostic accuracy. A test is considered *reliable* if it produces precise and reproducible information. A test is considered to have *diagnostic accuracy* if it can discriminate between patients who have a specific disorder and patients who do not have it.[4] Scientific evaluation of the clinical utility of physical therapy tests and measures involves comparing the examination results with reference standards such as radiographic studies (which represent the closest measure of the truth). Using statistical methods from the field of epidemiology, the diagnostic accuracy of the test, that is, its ability to determine which patients have a disorder and which do not, is then calculated. This chapter focuses on the characteristics that define the reliability and diagnostic accuracy of specific tests and measures. The chapter concludes with a discussion of the quality assessment of studies investigating reliability and diagnostic utility.

Reliability

For a clinical test to provide information that can be used to guide clinical decision making, it must have acceptable reliability. *Reliability* is the degree of consistency with which an instrument or rater measures a particular attribute.[5] When we investigate the reliability of a measurement, we are determining the proportion of that measurement that is a true representation and the proportion that is the result of measurement error.[6]

When discussing the clinical examination process, it is important to consider two forms of reliability: intraexaminer and interexaminer reliability. *Intraexaminer reliability* is the ability of a single rater to obtain identical measurements during separate performances of the same test. *Interexaminer reliability* is a measure of the ability of two or more raters to obtain identical results with the same test.

The kappa coefficient (κ) is a measure of the proportion of potential agreement after chance is removed[1,5,7]; it is the reliability coefficient most often used for categorical data (positive or negative).[5] The correlation coefficient commonly used to determine the reliability of data that are continuous in nature (e.g., range-of-motion data) is the intraclass correlation coefficient (ICC).[7]

Although interpretations of reliability vary, coefficients are often evaluated by the criteria described by Shrout,[8] with values less than 0.10 indicating no reliability, values between 0.11 and 0.40 indicating slight reliability, values between 0.41 and 0.60 indicating fair reliability, values between 0.61 and 0.80 indicating moderate reliability, and values greater than 0.81 indicating substantial reliability. "Acceptable reliability" must be decided by the clinician using the specific test or measure[9] and should be based on the variable being tested, the reason a particular test is important, and the patient on whom the test will be used.[6] For example, a 5% measurement error may be very acceptable when measuring joint range of motion but is not nearly as acceptable when measuring pediatric core body temperature.

Diagnostic Accuracy

Clinical tests and measures can never absolutely confirm or exclude the presence of a specific disease.[10] However, clinical tests can be used to alter the clinician's estimate of the probability that a patient has a specific musculoskeletal disorder. The accuracy of a test is determined by the measure of agreement between the clinical test and a reference standard.[11,12] A reference standard is the criterion considered the closest representation of the truth of a disorder being present.[1] The results obtained with the reference standard are compared with the results obtained with the test under investigation to determine the percentage of people correctly diagnosed or the diagnostic accuracy.[13] Because the diagnostic utility statistics are completely dependent on both the reference standard used and the population studied, we have specifically listed these within this text to provide information to consider when selecting the tests and measures reported. Diagnostic accuracy is often expressed in terms of positive and negative predictive values (PPVs and NPVs), sensitivity and specificity, and likelihood ratios (LRs).[1,14]

2×2 Contingency Table

To determine the clinical utility of a test or measure, the results of the reference standard are compared with the results of the test under investigation in a 2×2 contingency table, which provides a direct comparison between the reference standard and the test under investigation.[15] It allows for the calculation of the values associated with diagnostic accuracy to assist with determining the utility of the clinical test under investigation (Table 1-1).

The 2×2 contingency table is divided into four cells (a, b, c, d) for the determination of the test's ability to correctly identify true positives (cell a) and rule out true negatives (cell d). Cell b represents the false-positive findings wherein the diagnostic test was found to be positive yet the reference standard obtained a negative result. Cell c represents the false-negative findings wherein the diagnostic test was found to be negative yet the reference standard obtained a positive result.

Once a study investigating the diagnostic utility of a clinical test has been completed and the comparison with the reference standard has been performed in the 2×2 contingency table, determination of the clinical utility in terms of overall accuracy, PPVs and NPVs, sensitivity and specificity, and LRs can be calculated. These statistics are useful in determining whether a diagnostic test is useful for either ruling in or ruling out a disorder.

Table 1-1 2×2 Contingency Table Used to Compare the Results of the Reference Standard with Those of the Test under Investigation

	Reference Standard Positive	Reference Standard Negative
Clinical Test Positive	True-positive results a	False-positive results b
Clinical Test Negative	False-negative results c	True-negative results d

Table 1-2 2×2 Contingency Table Showing the Calculation of Positive Predictive Values (PPVs) and Negative Predictive Values (NPVs) Horizontally and Sensitivity and Specificity Vertically

	Reference Standard Positive	Reference Standard Negative	
Clinical Test Positive	True positives a	False positives b	$PPV = a/(a+b)$
Clinical Test Negative	c False negatives	d True negatives	$NPV = d/(c+d)$
	$Sensitivity = a/(a+c)$	$Specificity = d/(b+d)$	

Overall Accuracy

The overall accuracy of a diagnostic test is determined by dividing the correct responses (true positives and true negatives) by the total number of patients.[16] Using the 2×2 contingency table, the overall accuracy is determined by the following equation:

$$\text{Overall accuracy} = 100\% \times (a+d)/(a+b+c+d) \quad (1\text{-}1)$$

A perfect test would exhibit an overall accuracy of 100%. This is most likely unobtainable in that no clinical test is perfect, and each will always exhibit at least a small degree of uncertainty. The accuracy of a diagnostic test should not be used to determine the clinical utility of the test, because the overall accuracy can be a bit misleading. The accuracy of a test can be significantly influenced by the prevalence of a disease, or the total instances of the disease in the population at a given time.[5,6]

Positive and Negative Predictive Values

PPVs estimate the likelihood that a patient with a positive test actually has a disease.[5,6,17] PPVs are calculated horizontally in the 2×2 contingency table (Table 1-2) and indicate the percentage of patients accurately identified as having the disorder (true positive) divided by all the positive results of the test under investigation. A high PPV indicates that a positive result is a strong predictor that the patient has the disorder.[5,6] The formula for the PPV is:

$$PPV = 100\% \times a/(a+b) \quad (1\text{-}2)$$

NPVs estimate the likelihood that a patient with a negative test does not have the disorder.[5,6] NPVs are also calculated horizontally in the 2×2 contingency table (see Table 1-2) and indicate the percentage of patients accurately identified as not having the disorder (true negative) divided by all the negative results of the test under investigation.[11] The formula for the NPV is as follows:

$$NPV = 100\% \times d/(c+d) \quad (1\text{-}3)$$

The predictive values are significantly influenced by the prevalence of the condition.[11] Hence, we have not specifically reported these in this text.

Sensitivity

The *sensitivity* of a diagnostic test indicates the test's ability to detect those patients who actually have a disorder as indicated by the reference standard. This is also referred to as the *true-positive rate*.[1] Tests with high sensitivity are good for ruling out a particular disorder. The acronym *SnNout* can be used to remember that a test with high *Sen*sitivity and a *N*egative result is good for ruling *out* the disorder.[1]

Consider, for example, a clinical test that, compared with the reference standard, exhibits a high sensitivity for detecting lumbar spinal stenosis. Considering the rule above, if the test is negative it assists with ruling out lumbar spinal stenosis. If the test is positive, it is likely to accurately identify a high percentage of patients presenting with stenosis. However, it also may identify as positive

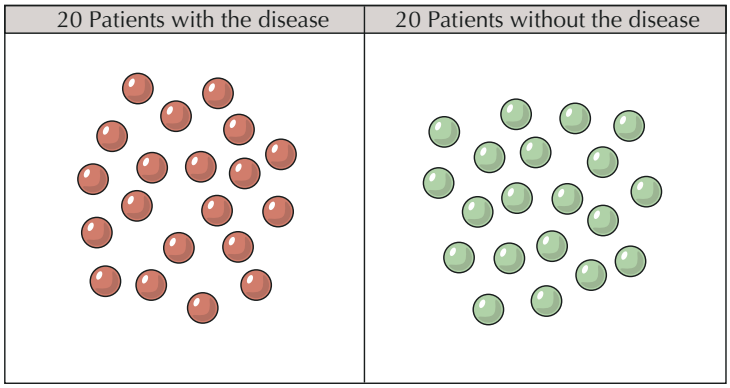

Figure 1-1
Sensitivity and specificity example. Twenty patients with and 20 patients without the disorder.

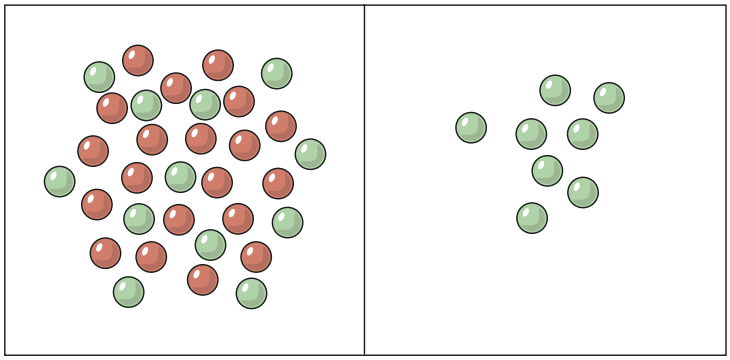

Figure 1-2
100% Sensitivity. One hundred percent sensitivity infers that if the test is positive, all those with the disease will be captured. However, although this test captured all those with the disease, it also captured many without it. Yet if the test result is negative, we are confident that the disorder can be ruled out (SnNout).

many of those without the disorder (false positives). Thus, although a negative result can be relied on, a positive test result does not allow us to draw any conclusions (Figs. 1-1 and 1-2).

The sensitivity of a test also can be calculated from the 2×2 contingency tables. However, it is calculated vertically (see Table 1-2). The formula for calculating a test's sensitivity is as follows:

$$\text{Sensitivity} = 100\,\% \times a/(a+c) \tag{1-4}$$

Specificity

The *specificity* of a diagnostic test simply indicates the test's ability to detect those patients who actually do not have the disorder as indicated by the reference standard. This is also referred to as the *true-negative rate*.[1] Tests with high specificity are good for ruling in a disorder. The acronym *SpPin* can be used to remember that a test with high *Sp*ecificity and a *P*ositive result is good for ruling *in* the disorder.[16,18,19]

Consider a test with high specificity. It would demonstrate a strong ability to accurately identify all patients who do not have a disorder. If a highly specific clinical test is negative, it is likely to identify a high percentage of those patients who do not have the disorder. However, it is also possible that the highly specific test with a negative result will identify a number of patients who actually have the disease as being negative (false negative). Therefore, we can be fairly confident that a highly specific test with a positive finding indicates that the disorder is present (Fig. 1-3).

The formula for calculating test specificity is as follows:

$$\text{Specificity} = 100\,\% \times d/(b+d) \tag{1-5}$$

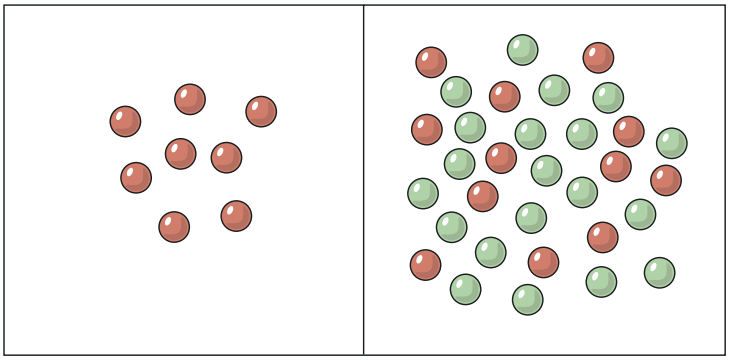

Figure 1-3
100% Specificity. One hundred percent specificity infers that if the test is negative, all those without the disease will be captured. However, although this test captured all those without the disease, it also captured many with it. Yet if the test is positive, we are confident that the patient has the disorder (SpPin).

Sensitivity and specificity have been used for decades to determine a test's diagnostic utility; however, they possess a few clinical limitations.[11] Although sensitivity and specificity can be useful in assisting clinicians in selecting tests that are good for ruling in or out a particular disorder, few clinical tests demonstrate both high sensitivity and high specificity.[11] Also the sensitivity and specificity do not provide information regarding a change in the probability of a patient having a disorder if the test results are positive or negative.[18,20] Instead, LRs have been advocated as the optimal statistics for determining a shift in pretest probability that a patient has a specific disorder.

Likelihood Ratios

A test's result is valuable only if it alters the pretest probability of a patient having a disorder.[21] LRs combine a test's sensitivity and specificity to develop an indication in the shift of probability given the specific test result and are valuable in guiding clinical decision making.[20] LRs are a powerful measure that can significantly increase or reduce the probability of a patient having a disease.[22]

LRs can be either positive or negative. A positive LR indicates a shift in probability favoring the existence of a disorder, whereas a negative LR indicates a shift in probability favoring the absence of a disorder. Although LRs are often not reported in studies investigating the diagnostic utility of the clinical examination, they can be calculated easily if a test's sensitivity and specificity are available. Throughout this text, for studies that did not report LRs but did document a test's sensitivity and specificity, the LRs were calculated by the authors.

The formula used to determine a positive LR is as follows:

$$\text{LR} = \text{Sensitivity} / (1 - \text{Specificity}) \tag{1-6}$$

The formula used to determine a negative LR is as follows:

$$\text{LR} = (1 - \text{Sensitivity}) / \text{Specificity} \tag{1-7}$$

A guide to interpreting test results can be found in Table 1-3. Positive LRs higher than 1 increase the odds of the disorder given a positive test, and negative LRs less than 1 decrease the odds of the disorder given a negative test.[22] However, it is the magnitude of the shifts in probability that determines the usefulness of a clinical test. Positive LRs higher than 10 and negative LRs close to zero often represent large and conclusive shifts in probability. An LR of 1 (either positive or negative) does not alter the probability that the patient does or does not have the particular disorder and is of little clinical value.[22] Once the LRs have been calculated, they can be applied to the nomogram (Fig. 1-4)[23] or a mathematical equation[24] can be used to determine more precisely the shifts in probability given a specific test result. Both methods are described in further detail later in the chapter.

Table 1-3 Interpretation of Likelihood Ratios

Positive Likelihood Ratio	Negative Likelihood Ratio	Interpretation
>10	<0.1	Generate large and often conclusive shifts in probability
5 to 10	0.1 to 0.2	Generate moderate shifts in probability
2 to 5	0.2 to 0.5	Generate small but sometimes important shifts in probability
1 to 2	0.5 to 1.0	Alter probability to a small and rarely important degree

Adapted from Jaeschke R, Guyatt GH, Sackett DL III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA*. 1994;271:703-707.

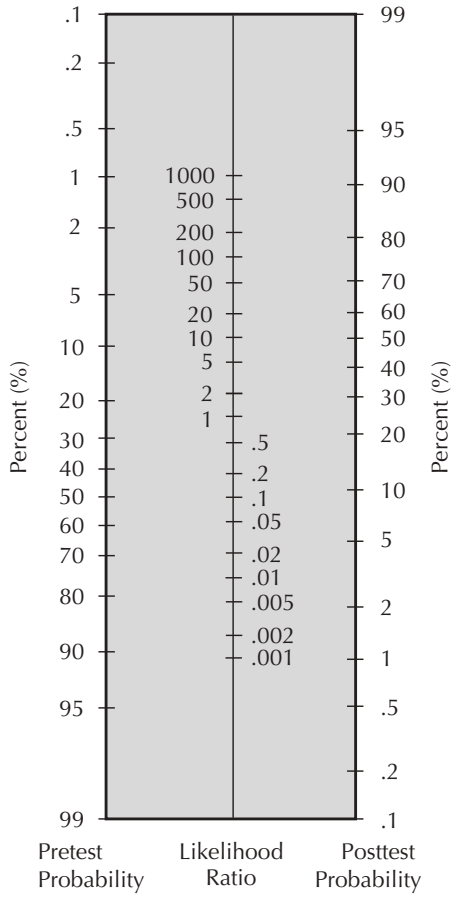

Figure 1-4

Fagan's nomogram. (Adapted with permission from Fagan TJ. Letter: nomogram for Bayes theorem. *N Engl J Med.* 1975;293:257. Copyright 2005, Massachusetts Medical Society. All rights reserved.)

If a diagnostic test exhibits a specificity of 1, the positive LR cannot be calculated because the equation will result in a zero for the denominator. In these circumstances, a suggestion has been made to modify the 2×2 contingency table by adding 0.5 to each cell in the table to allow for the calculation of LRs.[25]

Consider, for example, the diagnostic utility of the Crank test[5,26] in detecting labral tears in the shoulder compared with arthroscopic examination, the reference standard. This is revealed in a 2×2 contingency table (Table 1-4). The inability to calculate a positive LR becomes obvious in the following:

$$\text{Positive LR} = \text{Sensitivity} / (1 - \text{Specificity}) = 1/(1-1) = 1/0 \quad (1\text{-}8)$$

Table 1-4 Results of the Crank Test in Detecting Labral Tears When Compared with the Reference Standard of Arthroscopic Examination

	Arthroscopic Examination Positive (n = 12)	Arthroscopic Examination Negative (n = 3)	
Crank Test Positive	10 a	0 b	PPV = 100 × 10 / 10 = 100 %
Crank Test Negative	c 2	d 3	NPV = 100 × 3 / 5 = 60 %
	Sensitivity = 100 % × 10 / 12 = 83 %	Specificity = 100 % × 3 / 3 = 100 %	

Because zero cannot be the denominator in a fraction, the 2×2 contingency table is modified by adding 0.5 to each cell.

Although the addition of 0.5 to each cell is the only reported method of modifying the contingency table to prevent zero in the denominator of an LR calculation, considering the changes that occur with the diagnostic properties of sensitivity, specificity, and predictive values, this technique has not been used in this text. In circumstances in which the specificity is zero and the positive LR cannot be calculated, it is documented as "undefined" (UD). In these cases, although we are not calculating the positive LR, the test is indicative of a large shift in probability.

Confidence Intervals

Calculations of sensitivity, specificity, and LRs are known as *point estimates*. That is, they are the single best estimates of the population values.[5] However, because point estimates are based on small subsets of people (samples), it is unlikely that they are a perfect representation of the larger population. It is more accurate, therefore, to include a range of values (*interval estimate*) in which the population value is likely to fall. A *confidence interval* (CI) is a range of scores around the point estimate that likely contains the population value.[27] Commonly, the 95% CI is calculated for studies investigating the diagnostic utility of the clinical examination. A 95% CI indicates the spread of scores in which we can be 95% confident that they contain the population value.[5] In this text, the 95% CI is reported for all studies that provided this information.

Pretest and Posttest Probability

Pretest probability is the likelihood that a patient exhibits a specific disorder before the clinical examination. Often prevalence rates are used as an indication of pretest probability, but when prevalence rates are unknown, the pretest probability is based on a combination of the patient's medical history, the results of previous tests, and the clinician's experience.[16] Determining the pretest probability is the first step in the decision-making process for clinicians. Pretest probability is an estimate by the clinician and can be expressed as a percentage (e.g., 75%, 80%) or as a qualitative measure (e.g., somewhat likely, very likely).[11,16] Once the pretest probability of a patient having a disorder is identified, tests and measures that have the potential to alter the probability should be selected for the physical examination. Posttest probability is the likelihood that a patient has a specific disorder after the clinical examination procedures have been performed.

Calculating Posttest Probability

As previously mentioned, LRs can assist with determining the shifts in probability that would occur following a given test result and depend on the respective LR ratios of that given test. The quickest method to determine the shifts in probability once an LR is known for a specific test is the nomogram (Fig. 1-5).[23] The nomogram is a diagram that illustrates the pretest probability on the left and the posttest probability on the right, with the LRs in the middle. To determine the shift in probability, a mark is placed on the nomogram representing the pretest probability. Then a mark is made on the nomogram at the level of the LR (either negative or positive). The two lines are

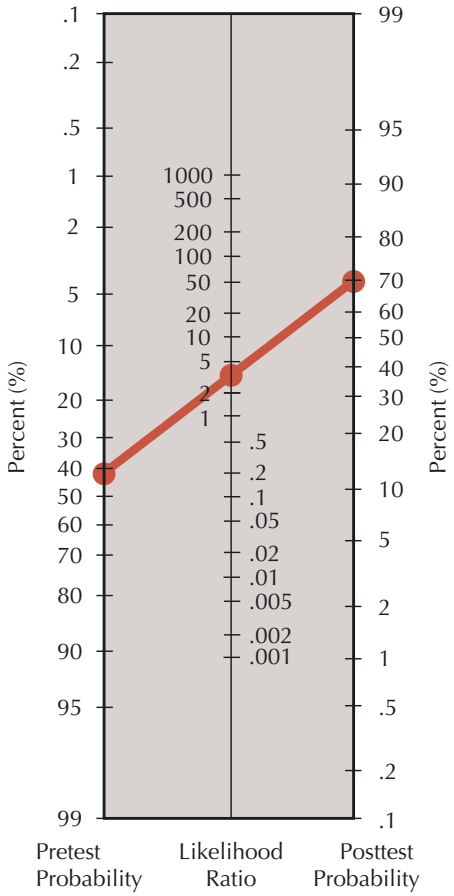

Figure 1-5
Nomogram representing the change in pretest probability from 42% if the test was positive (positive likelihood ratio = 4.2) to a posttest probability of 71%. (Adapted with permission from Fagan TJ. Letter: nomogram for Bayes theorem. *N Engl J Med.* 1975;293:257. Copyright 2005, Massachusetts Medical Society. All rights reserved.)

connected with a straight line and the line is carried through the right of the diagram. The point at which the line crosses the posttest probability scale indicates the shift in probability.

A more precise determination of the shift in probability can be calculated algebraically with the following formula[16]:

Step 1. Pretest odds = Pretest probability / 1 − Pretest probability (1-9)

Step 2. Pretest odds × LR = Posttest odds (1-10)

Step 3. Posttest odds / Posttest odds + 1 = Posttest probability (1-11)

The clinician must make a determination of when the posttest probability is either low enough to rule out the presence of a certain disease or when the posttest probability is high enough that the clinician feels confident in having established the presence of a disorder. The level at which evaluation ceases and treatment begins is known as the *treatment threshold* (Fig. 1-6).[16]

Assessment of Study Quality

Once relevant articles are retrieved, the next step is critical analysis of their content for adequate methodologic rigor. It has been reported that the methodologic quality of studies investigating the diagnostic utility of the clinical examination is generally inferior to that of studies investigating the effectiveness of therapies.[28,29] Unfortunately, studies with significant methodologic flaws

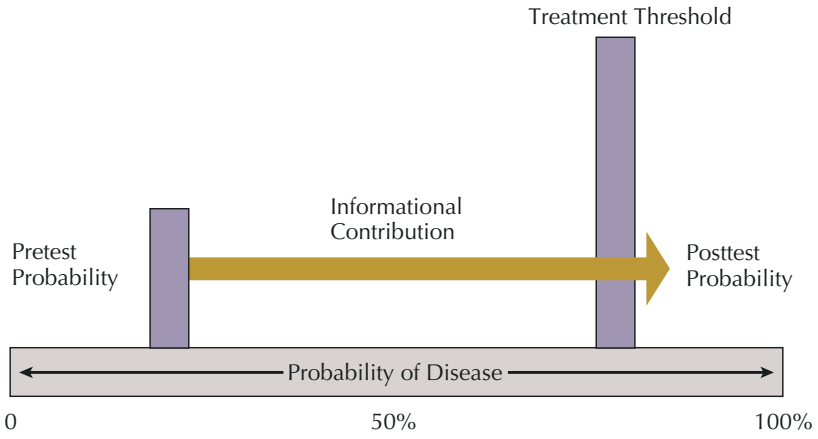

Figure 1-6

Treatment threshold. Clinicians must use the pretest probability and likelihood ratios to determine the treatment threshold as indicated in this illustration.

reporting the usefulness of specific tests and measures can lead to premature incorporation of ineffective tests. This can result in inaccurate diagnoses and poor patient management. Alternatively, identification and use of rigorously appraised clinical tests can improve patient care and outcomes.[29]

The Quality Appraisal for Reliability Studies (QAREL) was developed to assess the quality of diagnostic reliability studies.[30] The QAREL is an 11-item checklist developed in consultation with a reference group of experts in diagnostic research and quality appraisal that is used to assess a study's methodologic quality. Each item is scored as "yes," "no," "unclear," or "N/A." The QAREL has been found to be a reliable assessment tool when reviewers are given the opportunity to discuss the criteria by which to interpret each item.[31] Reliability of 9 of the 11 items was identified as good reliability, whereas reliability of only 2 of the 11 items was identified as fair reliability.[31] We have used the QAREL to evaluate each study related to reliability referenced in this text. For the purpose of this study we considered good quality to be ≥75% and fair quality to be between 74% and 50%. The percentages were calculated by dividing the number of "yes" responses by 11, minus the number of "N/A" responses. Green symbols indicate a high level of methodologic quality and imply that readers can be confident in study results. Yellow symbols indicate fair methodologic quality and imply that readers should interpret such study results with caution. Studies deemed to be of poor methodologic quality have not been included in the diagnostic utility tables throughout the chapters.

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) was developed to assess the quality of diagnostic accuracy studies.[32] A four-round Delphi panel identified 14 criteria that are used to assess a study's methodologic quality. Each item is scored as "yes," "no," or "unclear." The QUADAS is not intended to quantify a score for each study but rather provides a qualitative assessment of the study with the identification of weaknesses.[32] The QUADAS has demonstrated adequate agreement for the individual items in the checklist.[33] We have used the QUADAS to evaluate each study referenced in this text. For the purpose of this text we considered good quality to be ≥75% and fair quality to be between 74% and 50%. This was calculated by dividing the number of "yes" responses by 14 (the total number of criteria). Green symbols indicate a high level of methodologic quality and imply that readers can be confident in study results. Yellow symbols indicate fair methodologic quality and imply that readers should interpret such study results with caution. Studies deemed to be of poor methodologic quality have not been included in the diagnostic utility tables throughout the chapters.

Summary

It is important to consider the reliability and diagnostic utility of tests and measures before including them as components of the clinical examination. Tests and measures should demonstrate adequate reliability before they are used to guide clinical decision making. Throughout this text, the reliability of many tests and measures is reported. It is essential that clinicians consider these reported levels of reliability in the context of their own practice.

Before implementing tests and measures into the orthopaedic examination, it is first essential to consider each test's diagnostic utility. Table 1-5 summarizes the statistics related to diagnostic accuracy as well as the mathematical equations and operational definitions for each. The usefulness of a test or measure is most commonly considered in terms of the respective test's diagnostic properties. These can be described in terms of sensitivity, specificity, PPVs, and NPVs. However, perhaps the most useful diagnostic property is the LR, which can assist in altering the probability that a patient has a specific disorder.

No clinical test or measure provides absolute certainty as to the presence or absence of disease. However, clinicians can determine when enough data have been collected to alter the probability beyond the treatment threshold where the evaluation can cease, and therapeutic management can begin. Furthermore, careful methodologic assessment provides greater insight into the scientific rigor of each study and its performance, applicability, reliability, and reproducibility within a given clinical practice.

Table 1-5 2×2 Contingency Table and Statistics Used to Determine the Diagnostic Utility of a Test or Measure

	Reference Standard Positive	Reference Standard Negative
Diagnostic Test Positive	True-positive results a	False-positive results b
Diagnostic Test Negative	c False-negative results	d True-negative results

Statistic	Formula	Description
Overall accuracy	$(a+d)/(a+b+c+d)$	The percentage of individuals who are correctly diagnosed
Sensitivity	$a/(a+c)$	The proportion of patients with the condition who have a positive test result
Specificity	$d/(b+d)$	The proportion of patients without the condition who have a negative test result
Positive predictive value	$a/(a+b)$	The proportion of individuals with a positive test result who have the condition
Negative predictive value	$d/(c+d)$	The proportion of individuals with a negative test result who do not have the condition
Positive likelihood ratio	Sensitivity / (1 – Specificity)	If the test is positive, the increase in odds favoring the condition
Negative likelihood ratio	(1 – Sensitivity) / Specificity	If the test is positive, the decrease in odds favoring the condition

References

1. Sackett DL, Straws SE, Richardson WS, et al. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. London: Harcourt Publishers Limited; 2000.
2. Kassirer JP. Our stubborn quest for diagnostic certainty: a cause of excessive testing. *N Engl J Med*. 1989;320:1489–1491.
3. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061–1066.
4. Schwartz JS. Evaluating diagnostic tests: what is done–what needs to be done. *J G Intern Med*. 1986;1:266–267.
5. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Upper Saddle River, NJ: Prentice Hall Health; 2000.
6. Rothstein JM, Echternach JL. *Primer on Measurement: An Introductory Guide to Measurement Issues*. Alexandria, VA: American Physical Therapy Association; 1999.
7. Domholdt E. *Physical Therapy Research*. 2nd ed. Philadelphia: WB Saunders; 2000.
8. Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res*. 1998;7:301–317.
9. Van Genderen F, De Bie R, Helders P, Van Meeteren N. Reliability research: towards a more clinically relevant approach. *Phys Ther Rev*. 2003;8:169–176.
10. Bossuyt PMM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem*. 2003;49:1–6.
11. Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. *Phys Ther*. 2001;81:1546–1564.
12. Jaeschke R, Guyatt GH, Sackett III DL. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA*. 1994;271:389–391.
13. Bossuyt PMM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49:7–18.
14. McGinn T, Guyatt G, Wyer P, et al. Users' guides to the medical literature XXII: how to use articles about clinical decision rules. *JAMA*. 2000;284:79–84.
15. Greenhalgh T. Papers that report diagnostic or screening tests. *BMJ*. 1997;315:540–543.
16. Bernstein J. Decision analysis (current concepts review). *J Bone Joint Surg*. 1997;79:1404–1414.
17. Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. *Phys Ther*. 1985;65:1671–1675.
18. Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn? *Med Decis Making*. 1994;14:175–180.
19. Riddle DL, Stratford PW. Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. *Phys Ther*. 1999;79:939–948.
20. Hayden SR, Brown MD. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decision making. *Ann Emerg Med*. 1999;33:575–580.
21. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44:763–770.
22. Jaeschke R, Guyatt GH, Sackett DL. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA*. 1994;271:703–707.
23. Fagan TJ. Letter: nomogram for Bayes theorem. *N Engl J Med*. 1975;293:257.
24. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston: Little, Brown; 1991.
25. Wainner RS, Fritz JM, Irrgang JJ, et al. Reliability and diagnostic accuracy of the clinical examination and patient self-report measures for cervical radiculopathy. *Spine*. 2003;28:52–62.
26. Mimori K, Muneta T, Nakagawa T, Shinomiya K. A new pain provocation test for superior labral tears of the shoulder. *Am J Sports Med*. 1999;27:137–142.
27. Fidler F, Thomason N, Cumming G, et al. Editors can lead researchers to confidence intervals, but can't make them think. *Psychol Sci*. 2004;15:119–126.
28. Moons KGM, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem*. 2004;50:473–476.
29. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. *JAMA*. 1995;274:645–651.
30. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*. 2010;63(8):854–861.
31. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Med Res Methodol*. 2013;13:111.
32. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol*. 2005;5:19.
33. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006;6:9.

Temporomandibular Joint 2

Patient History	
Questions	• Screening instruments have been shown to be very good at identifying temporomandibular disorder (TMD) pain (+LR [likelihood ratio] of 33). • A subject complaint of "periodic restriction" (the inability to open the mouth as wide as was previously possible) has been found to be the best single history item to identify anterior disc displacement, both in patients with reducing discs and in those with nonreducing discs.
Physical Examination	
Palpation	• Reproducing pain during palpation of the temporomandibular joint (TMJ) and related muscles has been found to be moderately reliable and appears to demonstrate good diagnostic utility for identifying TMJ effusion confirmed by magnetic resonance imaging (MRI) and TMD when compared with a comprehensive physical examination. We recommend that palpation at least include the TMJ (+LR = 4.87 to 5.67), the temporalis muscle (+LR = 2.73 to 4.12), and the masseter muscle (+LR = 3.65 to 4.87). • If clinically feasible, pressure pain threshold (PPT) testing is helpful because it demonstrates superior diagnostic utility in identifying TMD when compared with a comprehensive physical examination.
Joint Sounds	• Detecting joint sounds (clicking and crepitus) during jaw motion is a generally unreliable sign demonstrating moderate diagnostic utility except in attempts to detect moderate to severe osteoarthritis (+LR = 4.79) and nonreducing anterior disc displacement (+LR = 2.6 to 15.2).
Range-of-Motion and Dynamic Movement Measurements	• Measuring mouth range of motion appears to be a highly reliable test, and when the range of motion is restricted or deviated from the midline, the measurement has moderate diagnostic utility in identifying nonreducing anterior disc displacement. • Detecting pain during motion is a less reliable sign, but it also demonstrates moderate to good diagnostic utility in identifying nonreducing anterior disc displacement and self-reported TMJ pain. • The combination of *motion restriction* and *pain during assisted opening* has been found to be the best combination for identifying nonreducing anterior disc displacement (+LR = 7.71). • Consistent with assessment of other body regions, assessment of "joint play" and "end feel" is highly unreliable and has unknown diagnostic utility.
Combination of Tests	• A combination of clinical examination findings has been shown to be beneficial in identifying disc displacement without reduction (5 positive tests +LR = 7.9)
Interventions	• Patients with TMD who report (1) *symptoms* ≥4/10 (*10* being severe pain) and (2) pain for 10 months' duration or less may benefit from nightly wearing of an occlusal stabilization splint, especially if they have (3) *nonreducing anterior disc displacement* and (4) show *improvement after 2 months* (+LR = 10.8 if all four factors are present).

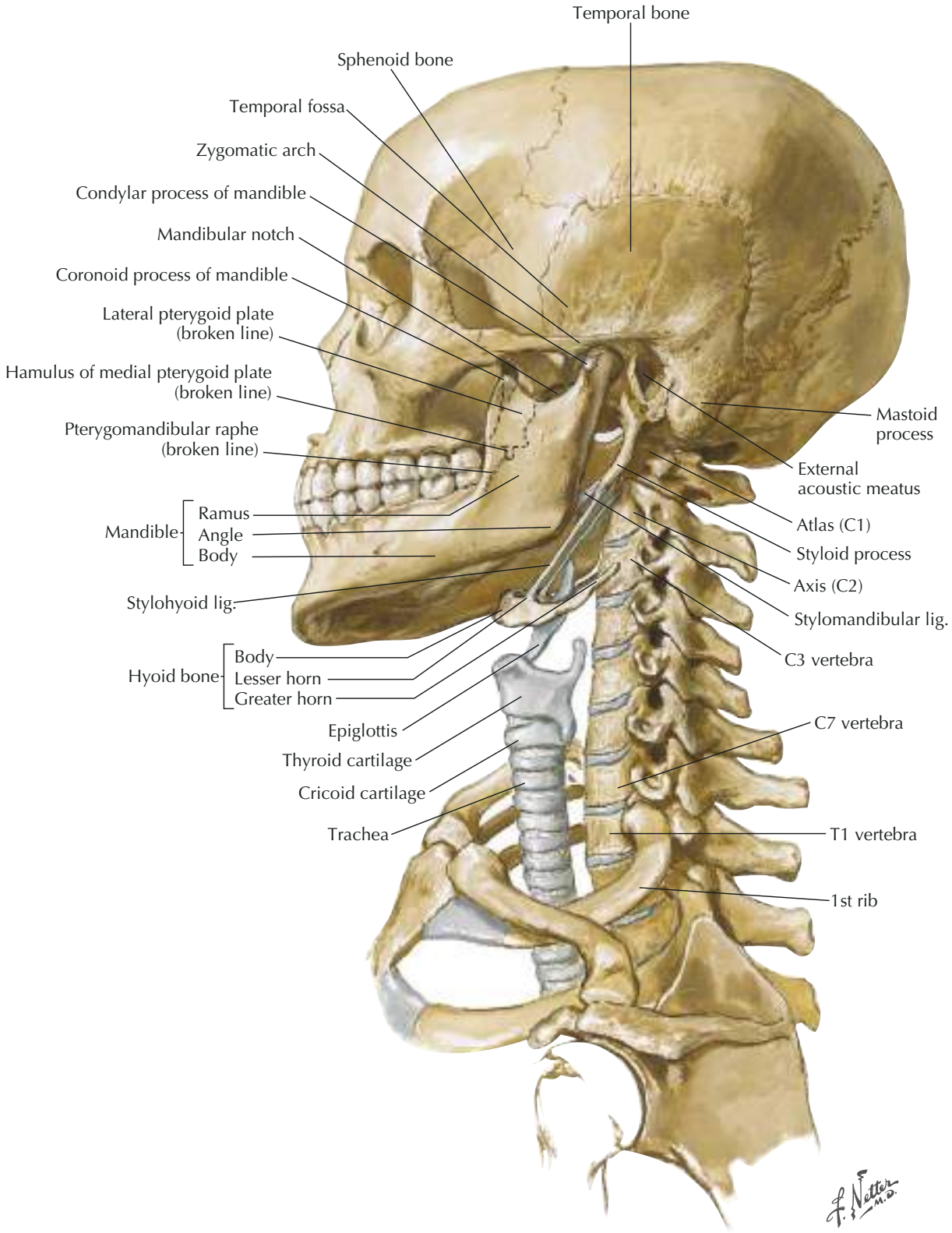

Figure 2-1
Bony framework of head and neck.

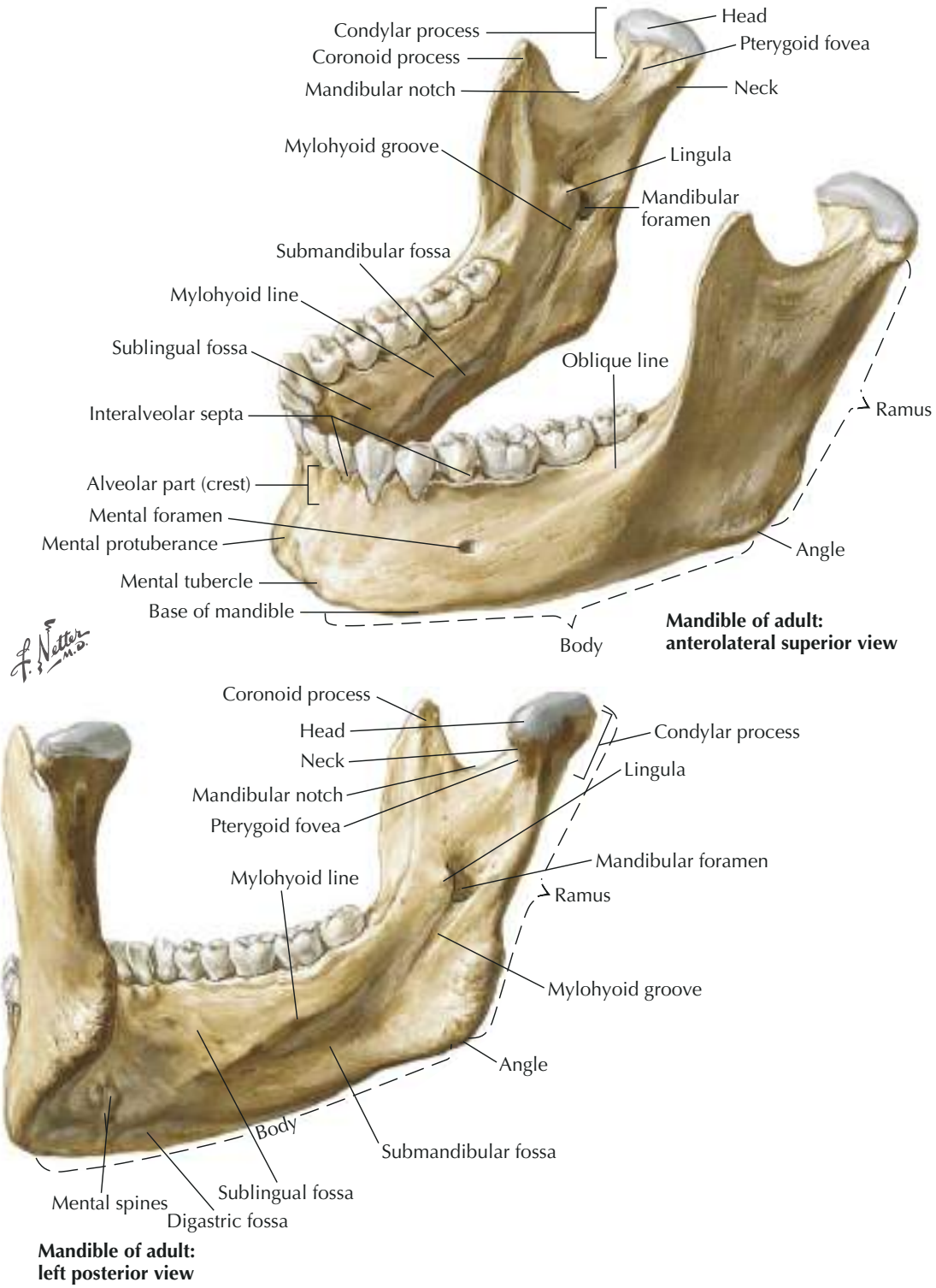

Figure 2-2
Mandible.

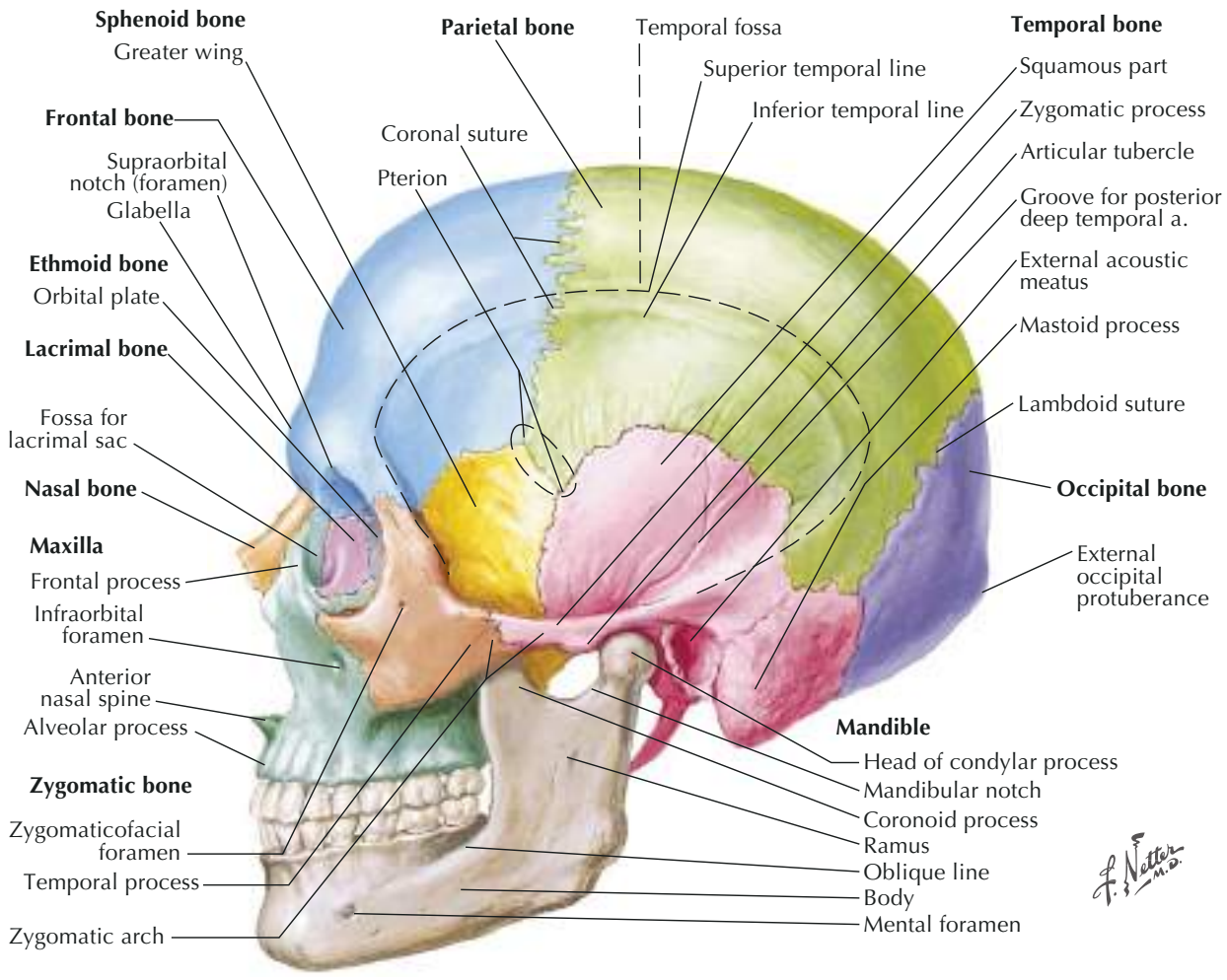

Figure 2-3
Lateral skull.

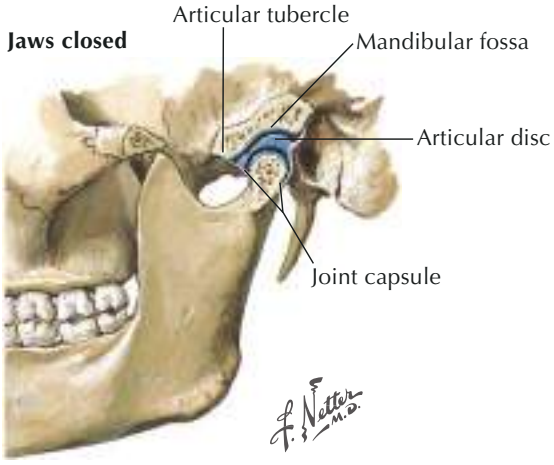

Figure 2-4
Temporomandibular joint.

The temporomandibular joint (TMJ) is divided by an intraarticular biconcave disc that separates the joint cavity into two distinct functional components. The upper joint is a plane, or gliding, joint that permits translation of the mandibular condyles. The lower joint is a hinge joint that permits rotation of the condyles. The closed pack position of the TMJ is full occlusion. A unilateral restriction pattern primarily limits contralateral excursion but also affects mouth opening and protrusion.

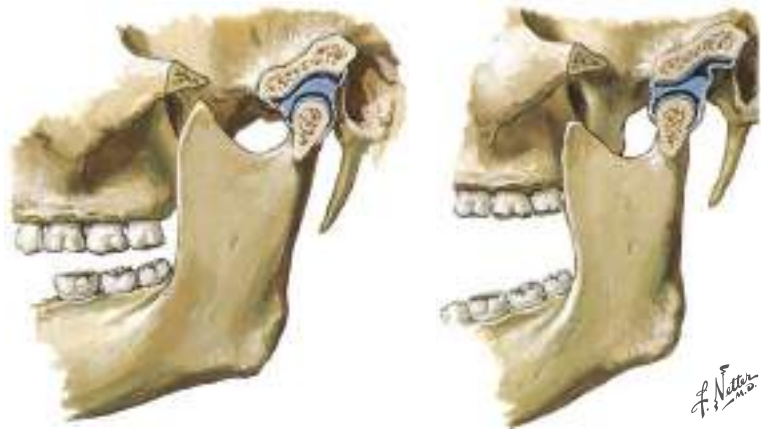

Figure 2-5
Temporomandibular joint mechanics.

During mandibular depression from a closed mouth position, the initial movement occurs at the lower joint as the condyles pivot on the intraarticular disc. This motion continues to approximately 11 mm of depression. With further mandibular depression, motion begins to occur at the upper joint and causes anterior translation of the disc on the articular eminence. Normal mandibular depression is between 40 and 50 mm.

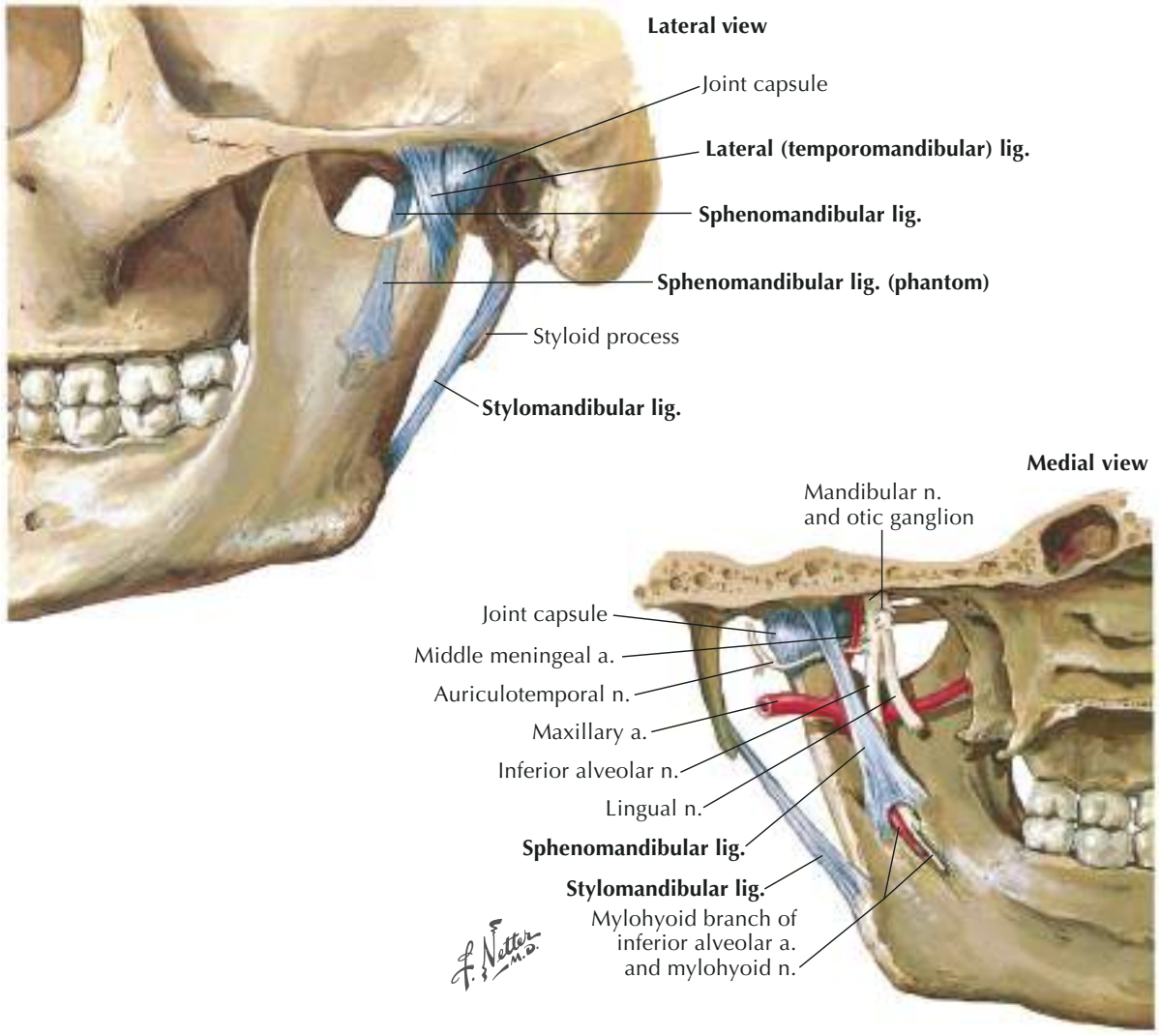

Figure 2-6
Temporomandibular joint ligaments.

Ligaments	Attachments	Function
Temporomandibular	Thickening of anterior joint capsule extending from neck of mandible to zygomatic arch	Strengthen the TMJ laterally
Sphenomandibular	Sphenoid bone to mandible	Serve as a fulcrum for and reinforcer of TMJ motion
Stylomandibular	Styloid process to angle of mandible	Provide minimal support for joint

Muscles Involved in Mastication

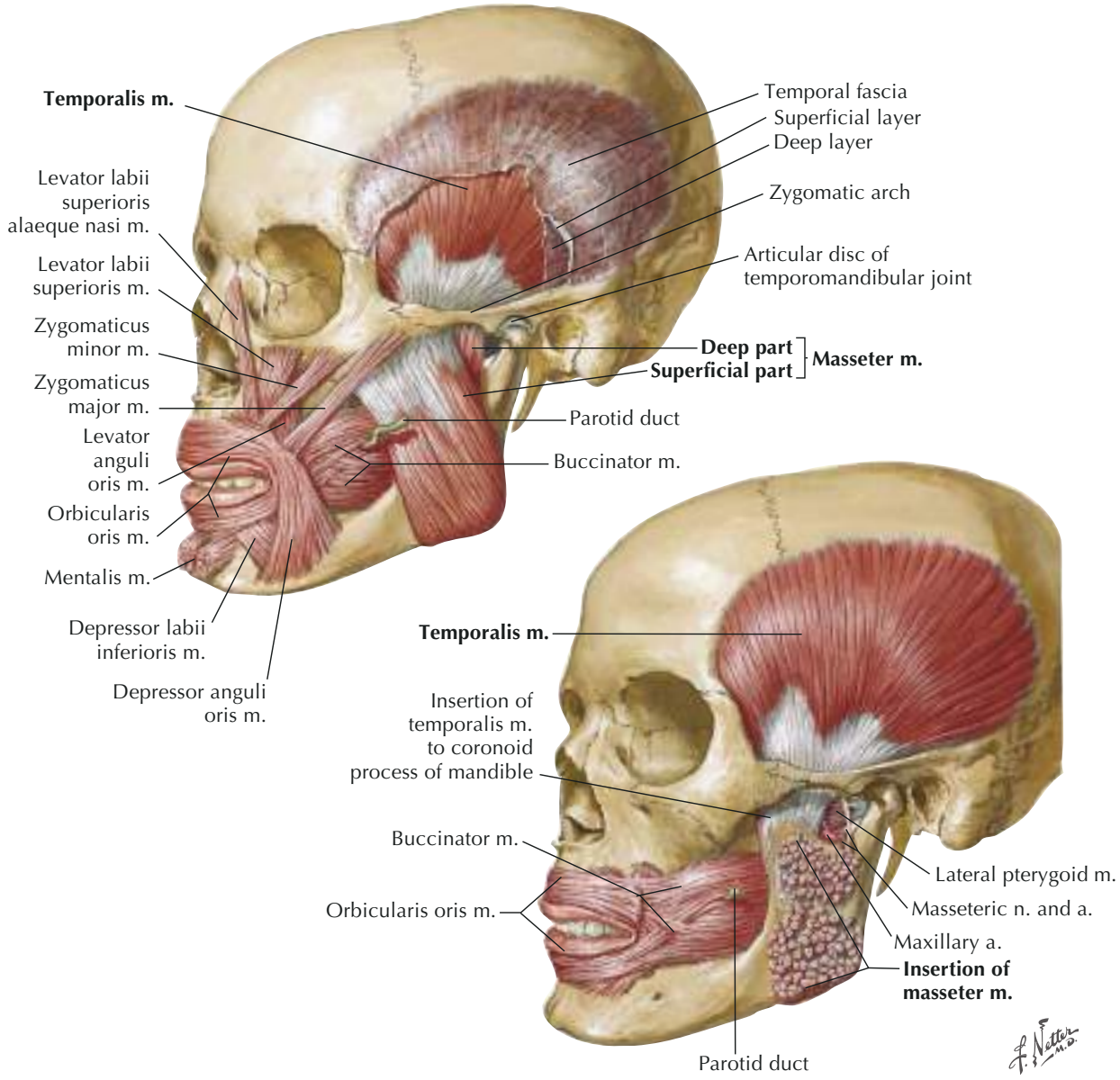

Figure 2-7
Muscles involved in mastication, lateral views.

Muscle	Proximal Attachment	Distal Attachment	Nerve and Segmental Level	Action
Temporalis	Temporal fossa	Coronoid process and anterior ramus of mandible	Deep temporal branches of mandibular nerve	Elevate mandible
Masseter	Inferior and medial aspects of zygomatic arch	Coronoid process and lateral ramus of mandible	Mandibular nerve via masseteric nerve	Elevate and protrude mandible

Muscles Involved in Mastication—cont'd

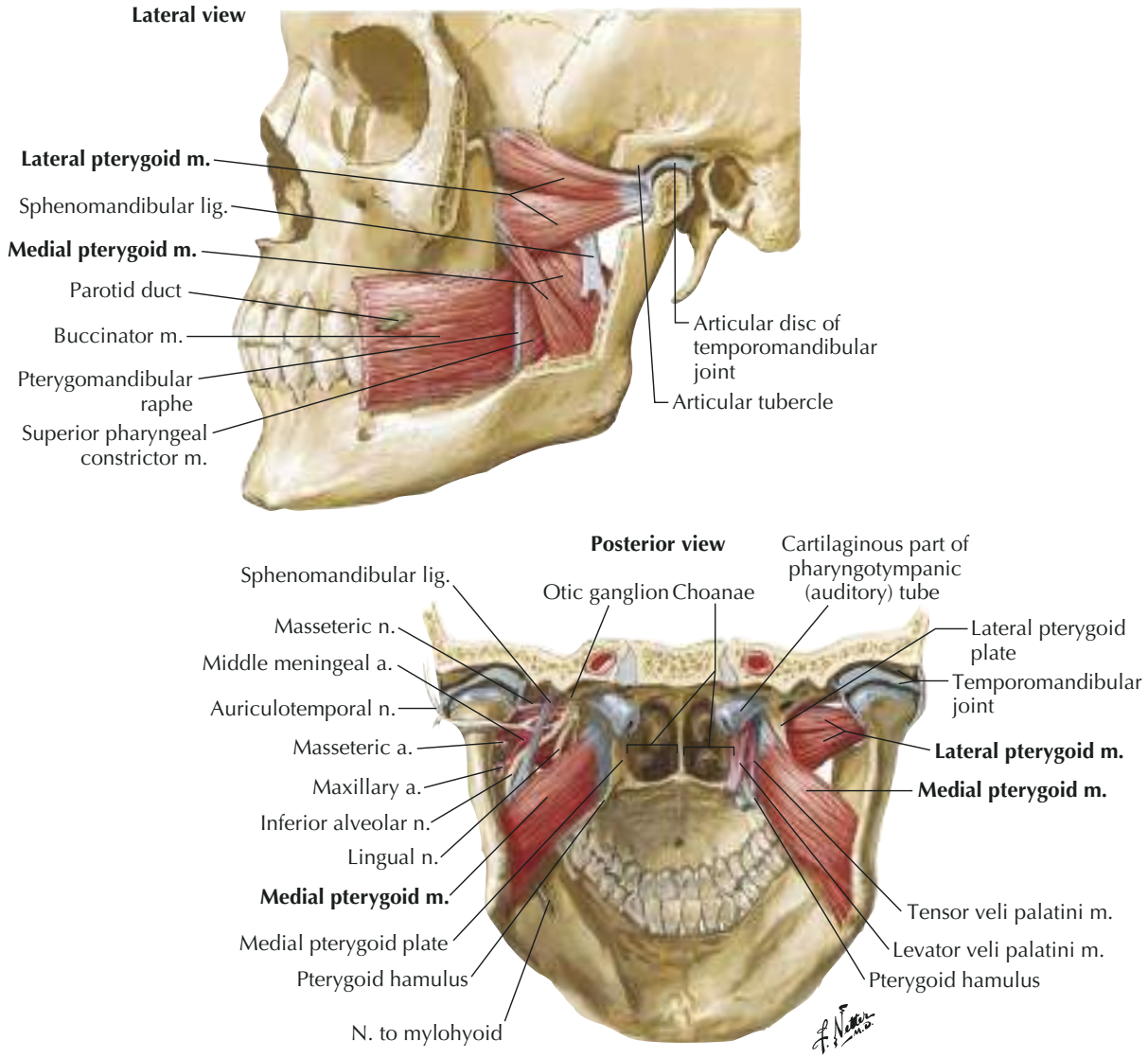

Figure 2-8
Muscles involved in mastication, lateral and posterior views.

Muscle	Proximal Attachment	Distal Attachment	Nerve and Segmental Level	Action
Medial pterygoid	Medial surface of lateral pterygoid plate, pyramidal process of palatine bone, and tuberosity of maxilla	Medial aspect of mandibular ramus	Mandibular nerve via medial pterygoid nerve	Elevate and protrude mandible
Lateral pterygoid (superior head)	Lateral surface of greater wing of sphenoid bone	Neck of mandible, articular disc, and TMJ capsule	Mandibular nerve via lateral pterygoid nerve	Acting bilaterally: protrude and depress mandible
Lateral pterygoid (inferior head)	Lateral surface of lateral pterygoid plate			Acting unilaterally: laterally deviate mandible

Muscles of the Floor of the Mouth

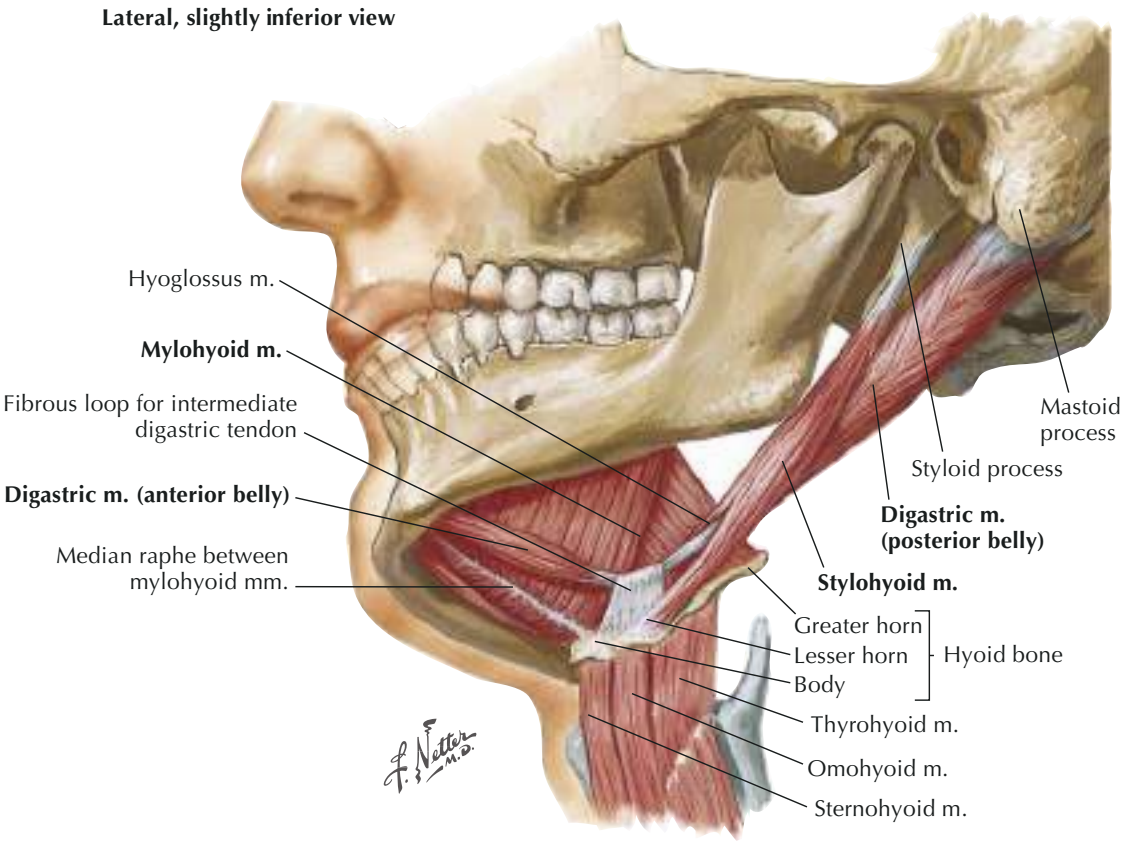

Figure 2-9
Floor of mouth, inferior view.

Muscle	Proximal Attachment	Distal Attachment	Nerve and Segmental Level	Action
Mylohyoid	Mylohyoid line of mandible	Hyoid bone	Mylohyoid nerve (branch of cranial nerve [CN] V_3)	Elevates hyoid bone
Stylohyoid	Styloid process of temporal bone	Hyoid bone	Cervical branch of facial nerve	Elevates and retracts hyoid bone
Geniohyoid	Inferior mental spine of mandible	Hyoid bone	C1 via hypoglossal nerve	Elevates hyoid bone anterosuperiorly
Digastric (anterior belly)	Digastric fossa of mandible	Intermediate tendon to hyoid bone	Mylohyoid nerve	Depresses mandible; raises and stabilizes hyoid bone
Digastric (posterior belly)	Mastoid notch of temporal bone		Facial nerve	

Muscles of the Floor of the Mouth—cont'd

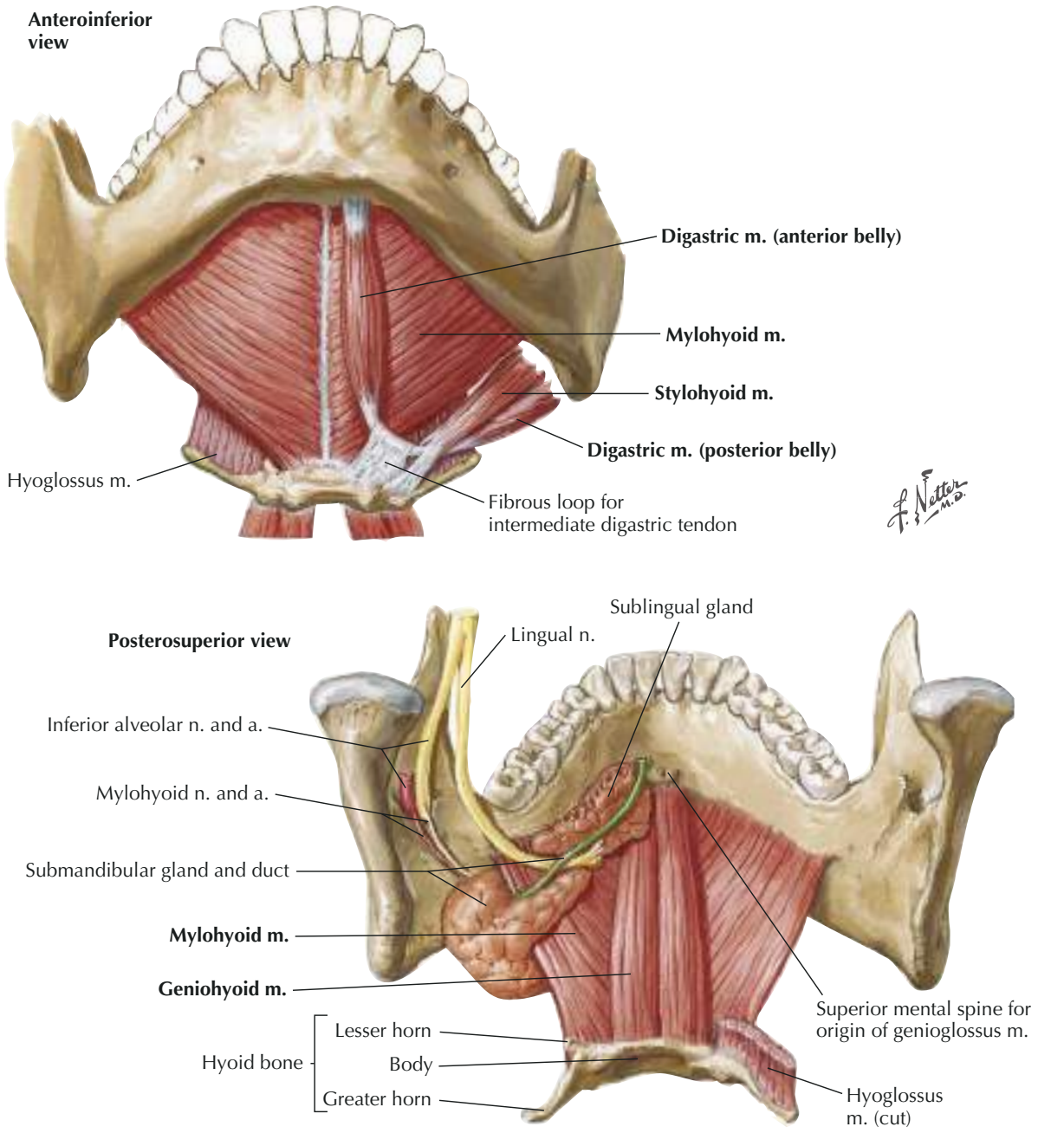

Figure 2-10
Floor of mouth, anteroinferior and posterosuperior views.

Mandibular Nerve

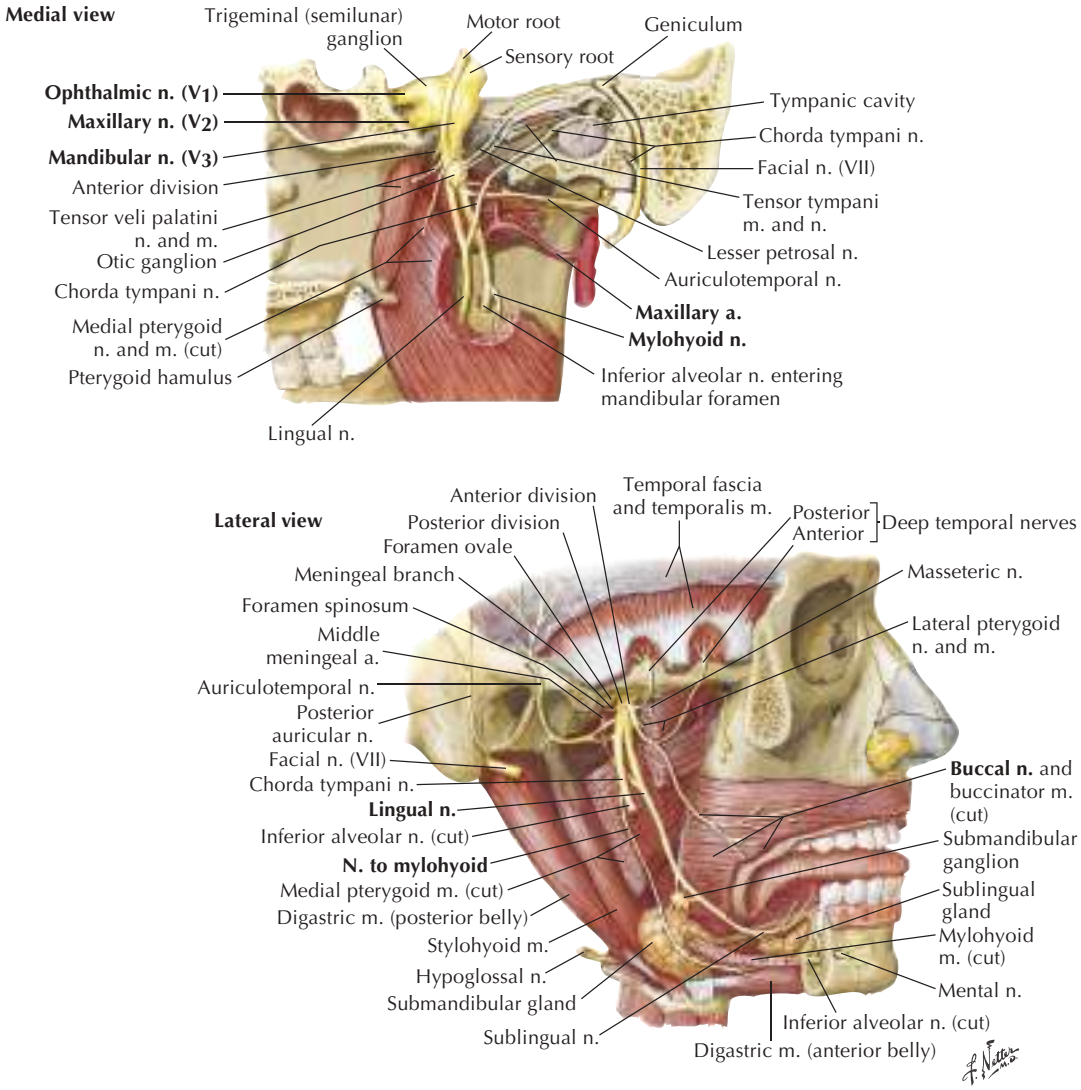

Figure 2-11
Mandibular nerve, medial and lateral views.

Nerves	Segmental Levels	Sensory	Motor
Mandibular	CN V_3	Skin of inferior third of face	Temporalis, masseter, lateral pterygoid, medial pterygoid, digastric, mylohyoid
Nerve to mylohyoid	CN V_3	No sensory	Mylohyoid
Buccal	CN V_3	Cheek lining and gingiva	No motor
Lingual	CN V_3	Anterior tongue and floor of mouth	No motor
Maxillary	CN V_2	Skin of middle third of face	No motor
Ophthalmic	CN V_1	Skin of superior third of face	No motor

CN V, trigeminal nerve.

Patient Reports	Initial Hypothesis
Patient reports jaw crepitus and pain during mouth opening and closing. Might also report limited opening with translation of the jaw to the affected side at the end range of opening	Possible osteoarthrosis Possible capsulitis Possible internal derangement consisting of an anterior disc displacement that does not reduce[1-3]
Patient reports jaw clicking and pain during opening and closing of the mouth	Possible internal derangement consisting of anterior disc displacement with reduction[1,4,5]
Patient reports limited motion to about 20 mm with no joint noise	Possible capsulitis Possible internal derangement consisting of an anterior disc displacement that does not reduce[1]

The Association of Oral Habits with Temporomandibular Disorders

Figure 2-12
Frequent leaning of head on the palm.

Gavish and colleagues[6] investigated the association of oral habits with signs and symptoms of TMDs in 248 randomly selected female high school students. Although sensitivity and specificity were not reported, the results demonstrated that chewing gum, jaw play (nonfunctional jaw movements), chewing ice, and frequent leaning of the head on the palm were associated with the presence of TMJ disorders.

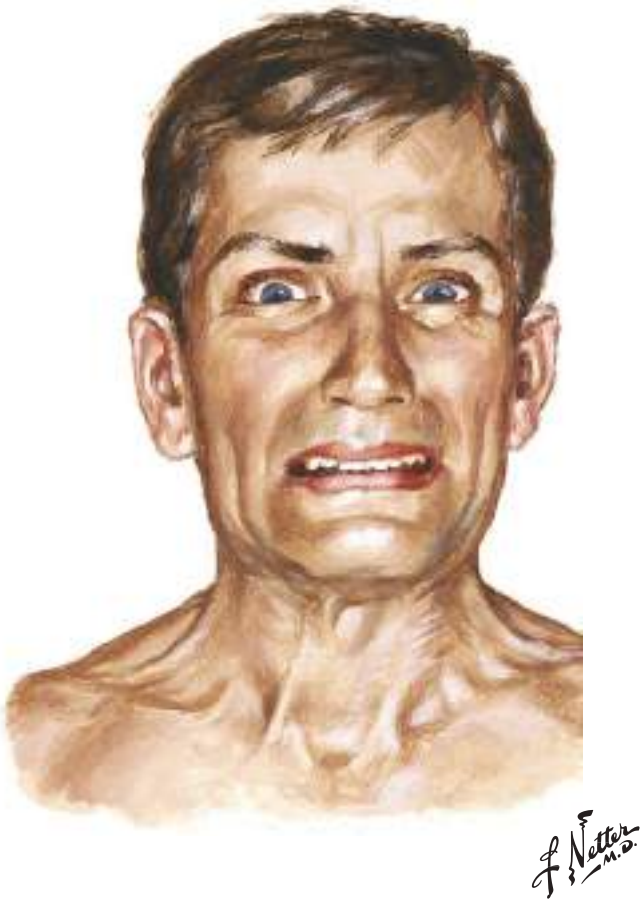

Figure 2-13
Temporomandibular joint pain.

Historical Finding and Study Quality	Description and Positive Findings	Population	Test-Retest Reliability
Visual analog scale (VAS)[7] ●	A 100-mm line, with ends defined as "no pain" and "worst pain imaginable"	38 consecutive patients referred with TMD	$\kappa = .38$
Numerical scale[7] ●	An 11-point scale, with 0 indicating "no pain" and 10 representing "worst pain"		$\kappa = .36$
Behavior rating scale[7] ●	A 6-point scale ranging from "minor discomfort" to "very strong discomfort"		$\kappa = .68$
Verbal scale[7] ●	A 5-point scale ranging from "no pain" to "very severe pain"		$\kappa = .44$

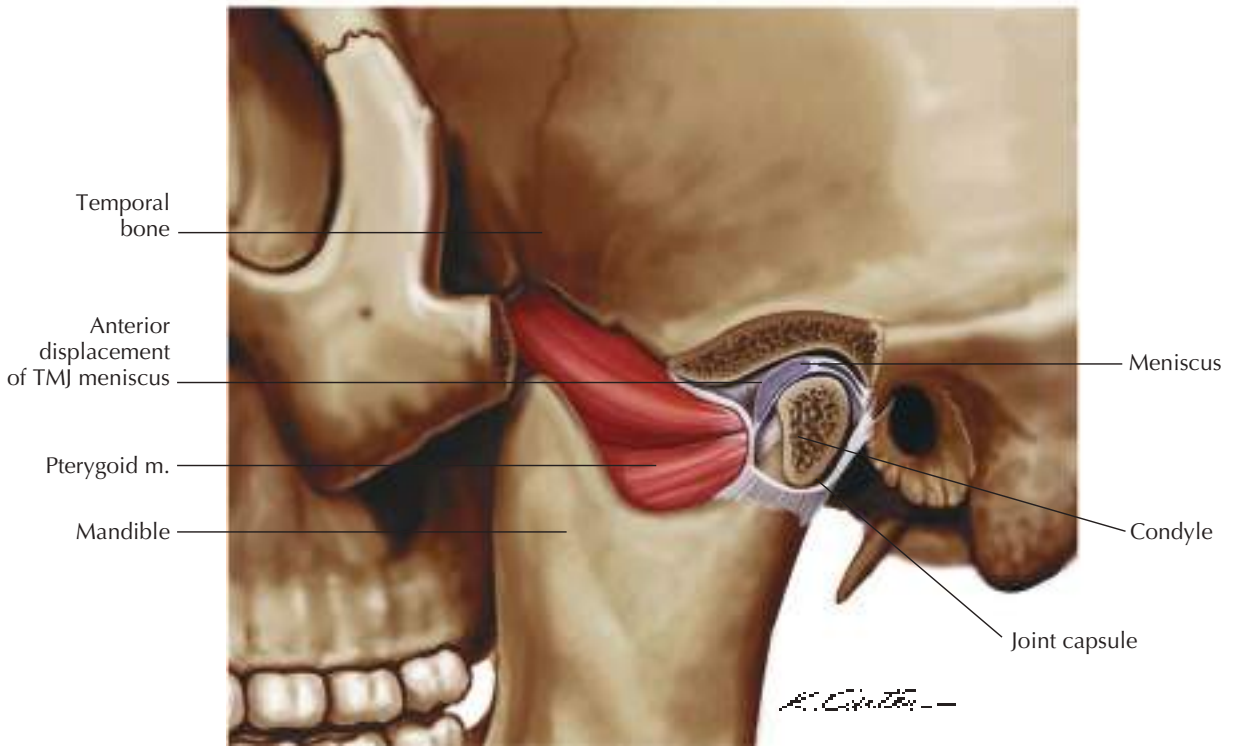

Figure 2-14
Anterior disc displacement.

Historical Finding and Study Quality	Description and Positive Findings	Population	Reference Standard	Sens	Spec	+LR	–LR
Clicking[8] ●	Momentary snapping sound during opening or functioning	70 patients (90 TMJs) referred with complaints of craniomandibular pain	Anterior disc displacement via MRI	In presence of reducing disc			
				.82	.19	1.01	.95
				In presence of nonreducing disc			
				.86	.24	1.13	.58
Locking[8] ●	Sudden onset of restricted movement during opening or closing			In presence of reducing disc			
				.53	.22	.68	2.14
				In presence of nonreducing disc			
				.86	.52	1.79	.27
Restriction after clicking[8] ●	Inability to open as wide as was previously possible after clicking			In presence of reducing disc			
				.26	.40	.43	1.85
				In presence of nonreducing disc			
				.66	.74	2.54	.46

Continued

Historical Finding and Study Quality	Description and Positive Findings	Population	Reference Standard	Sens	Spec	+LR	–LR
Periodic restriction[8] ●	Periodic inability to open as wide as was previously possible			In presence of reducing disc			
				.60	.90	6.0	.44
				In presence of nonreducing disc			
				.12	.95	2.4	.93
Continuous restriction[8] ●	Continuous inability to open as wide as was previously possible			In presence of reducing disc			
				.35	.26	.47	2.5
				In presence of nonreducing disc			
				.78	.62	2.05	.35
Function related to joint pain[8] ●	Not reported			In presence of reducing disc			
				.82	.10	.91	1.8
				In presence of nonreducing disc			
				.96	.24	1.26	.17
Complaint of clicking[8] ●				In presence of reducing disc			
				.28	.24	.37	3.00
				In presence of nonreducing disc			
				.82	.69	2.65	.26
Complaint of movement-related pain[8] ●				In presence of reducing disc			
				.71	.31	1.03	.94
				In presence of nonreducing disc			
				.74	.36	1.16	.72
Complaint of severe restriction[8] ●				In presence of reducing disc			
				.60	.65	1.71	.62
				In presence of nonreducing disc			
				.38	.93	5.43	.67

Reliability of Self-Reported Temporomandibular Pain

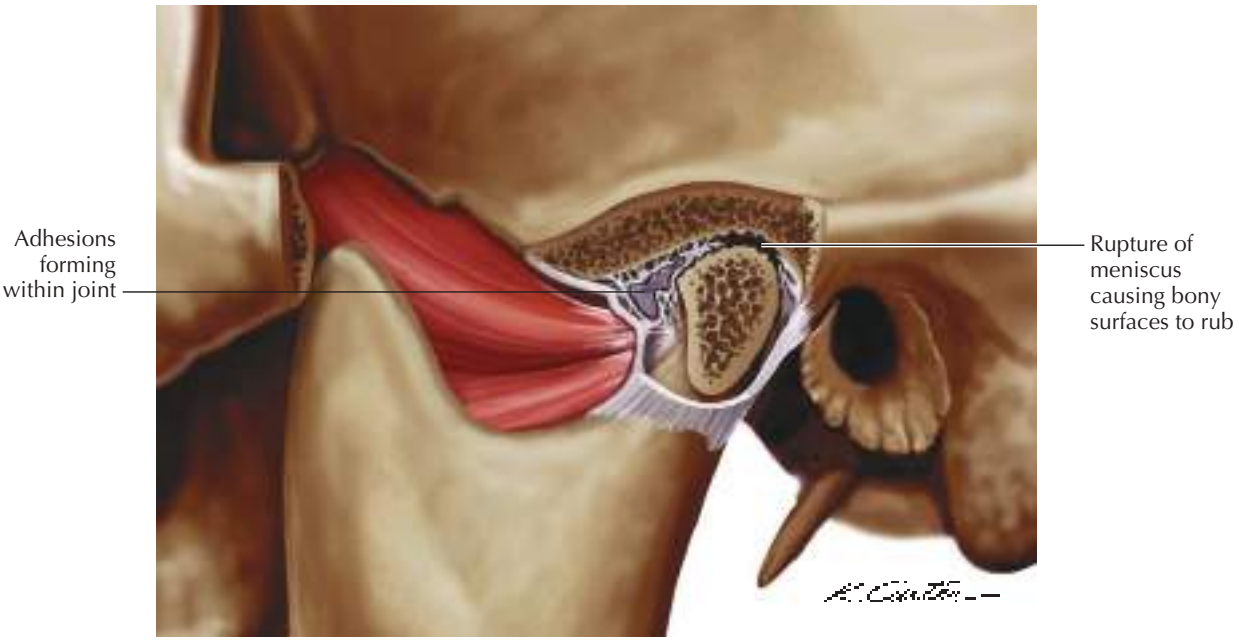

Figure 2-15
Temporomandibular arthrosis.

Historical Finding and Study Quality	Description and Positive Findings	Population	Reliability
TMD pain screening questionnaire[10] ●	See diagnostic table on following page. Participants were asked same questions 2 to 7 days apart	549 participants: 212 with pain-related TMD, 116 with TMD, 80 with odontalgia, 45 with headache without TMD pain, and 96 healthy controls	ICC = .83
Self-report of TMJ pain[9] ●	See diagnostic table on following page. Participants were asked same questions 2 weeks apart	120 adolescents: 60 with self-reported TMJ pain and 60 age- and sex-matched controls	Test-retest κ = .83 (.74, .93)

Diagnostic Utility of Self-Reported Temporomandibular Pain

Historical Finding and Study Quality	Description and Positive Findings	Population	Reference Standard	Sens	Spec	+LR	–LR
TMD pain screening questionnaire[10] ◆	Participants were asked: (1) "In the last 30 days, on average, how long did any pain in your jaw or temple area on either side last?" (a) There was no pain (b) Pain lasted from a very brief time to more than a week, but it did stop (c) Pain was continuous (2) "In the last 30 days, have you had pain or stiffness in your jaw on awakening?" (a) No (b) Yes (3) "In the last 30 days, did […] chewing hard or tough food […] change any pain (i.e., make it better or make it worse) in your jaw or temple area on either side?" (a) No (b) Yes An (a) response received 0 points, a (b) response received 1 point, and a (c) response received 2 points. The test was positive for scores of 2 or higher	549 participants: 212 with pain-related TMD, 116 with TMJ disorder, 80 with odontalgia, 45 with headache without TMD pain, and 96 healthy controls	RDC/TMD assessment protocol	.99	.97	33.0	.01
Self-report of TMJ pain[9] ●	Participants were asked: (1) "Do you have pain in your temple, face, TMJ, or jaw once a week or more?" (2) "Do you have pain when you open your mouth wide or chew once a week or more?" If answer was "yes" to either question, test was positive	120 adolescents: 60 with self-reported TMJ pain and 60 age- and sex-matched controls	RDC/TMD diagnosis of myofascial pain or arthralgia, arthritis, and arthrosis	.98	.90	9.8 (4.8, 20.0)	.02 (.00, .16)

RDC/TMD, Research Diagnostic Criteria for Temporomandibular Disorders
diagnostic accuracy statistics reported for participants with pain-related TMD versus healthy controls.

The Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) provides evidence-based criteria for assessing patients with TMD. It superseded the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) as of 2014 and is intended for immediate implementation in both clinical and research settings.[11] All tools required for clinical implementation are available at the International RDC-TMD Consortium website (www.rdc-tmdinternational.org/, accessed February 2015). A summary of the DC/TMD is presented here along with the associated reliability and diagnostic utility statistics. However, because the sources of the statistical estimates were not always clear, we were unable to assess the quality of the studies that provided the reliability and diagnostic utility values. The previous version of RDC/TMD showed fair to moderate agreement for most diagnoses and no to slight agreement for some diagnoses.

Diagnosis	History	Examination	Interexaminer Reliability	Sens	Spec	+LR	–LR
Myalgia	Positive for both: 1. Pain in jaw, temple, ear, front of ear 2. Pain modified with jaw movement, function, or parafunction	Positive for both: 1. Confirmation of pain in temporalis or masseter muscle 2. Report of familiar pain with one or more of following: (a) Palpation of temporalis muscle; (b) Palpation of masseter muscle; (c) Maximum unassisted or assisted opening movement	κ = .94 (.83, 1.00)	.90	.99	90.0	.10
Local myalgia	Positive for both: 1. Pain in jaw, temple, ear, or front of ear 2. Pain modified with jaw movement, function, or parafunction	Positive for all: 1. Confirmation of pain in temporalis or masseter muscle 2. Report of familiar pain with palpation of temporalis or masseter muscle 3. Report of pain localized to site of palpation	Not reported	Not established	Not established	Not established	Not established
Myofascial pain	Positive for both: 1. Pain in jaw, temple, ear, or front of ear 2. Pain modified with jaw movement, function, or parafunction	Positive for all: 1. Confirmation of pain in temporalis or masseter muscle 2. Report of familiar pain with palpation of temporalis or masseter muscle 3. Report of pain spreading beyond site of palpation but within boundary of muscle	Not reported	Not established	Not established	Not established	Not established

Continued

Diagnosis	History	Examination	Interexaminer Reliability	Sens	Spec	+LR	–LR
Myofascial pain with referral	Positive for both: 1. Pain in jaw, temple, ear, or front of ear 2. Pain modified with jaw movement, function, or parafunction	Positive for all: 1. Confirmation of pain in temporalis or masseter muscle 2. Report of familiar pain with palpation of temporalis or masseter muscle 3. Report of pain at site beyond boundary of muscle palpated	κ = .85 (.55, 1.00)	.86	.98	43.0	.14
Arthralgia	Positive for both: 1. Pain in jaw, temple, ear, or front of ear 2. Pain modified with jaw movement, function, or parafunction	Positive for both: 1. Confirmation of pain in area of TMJ 2. Report of familiar pain in TMJ with at least one of the following provocation tests: (a) Palpation of lateral pole or around lateral pole (b) Maximum unassisted or assisted opening, right or left lateral, or protrusive movement	κ = .86 (.75, .97)	.89	.98	44.5	.11
Headache attributed to TMD	Positive for both: 1. Headache of any type in temple 2. Headache modified with jaw movement, function, or parafunction	Positive for both: 1. Confirmation of headache in area of temporalis muscle 2. Report of familiar headache in temple with at least one of the following provocation tests: (a) Palpation of temporalis muscle (b) Maximum unassisted or assisted opening, right or left lateral, or protrusive movement	Not reported	.89	.87	6.85	.13

Note: Reliability and validity are derived from the datasets of the Validation Project and TMJ Impact Project Finalization of DC/TMD.[11]

Diagnosis	History	Examination	Interexaminer Reliability	Sens	Spec	+LR	–LR
Disc displacement with reduction	Positive for at least one: 1. In last 30 days, any TMJ noise present with jaw movement or function 2. Patient reports any noise present during examination	Positive for at least one: 1. Clicking, popping, and/or snapping noise during both opening and closing movements, detected with palpation during at least one of three repetitions of jaw opening and closing movements 2. Clicking, popping, and/or snapping noise detected with palpation during at least one of three repetitions of opening or closing movements AND right or left lateral or protrusive movement(s)	κ = .58 (.33, .84)	.34	.92	4.25	.72
Disc displacement with reduction with intermittent locking	Positive for both: 1. In last 30 days, any TMJ noise with jaw movement or function or patient reports any noise present during examination 2. In last 30 days, jaw locks with limited mouth opening and then unlocks	Positive for at least one: 1. Clicking, popping, and/or snapping noise during both opening and closing movements, detected with palpation during at least one of three repetitions of jaw opening and closing movements 2. Clicking, popping, and/or snapping noise detected with palpation during at least one of three repetitions of opening or closing movements AND right or left lateral or protrusive movement	Not reported	.38	.98	19.0	.63

Continued

Diagnosis	History	Examination	Interexaminer Reliability	Sens	Spec	+LR	–LR
Disc displacement without reduction with limited opening	Positive for both: 1. Jaw locked so that mouth would not open all the way 2. Limitation in jaw opening severe enough to limit jaw opening and interfere with ability to eat	Positive for the following: 1. Maximum assisted opening (passive stretch) movement, including vertical incisal overlap less than 40 mm	Not reported	.80	.97	26.7	.21
Disc displacement without reduction without limited opening	Positive for both of the following in the past: 1. Jaw locked so that mouth would not open all the way 2. Limitation in jaw opening severe enough to limit jaw opening and interfere with ability to eat	Positive for the following: 1. Maximum assisted opening (passive stretch) movement, including vertical incisal overlap of 40 mm or more	κ = .84 (.38, 1.00)	.54	.79	2.57	.58
Degenerative joint disease	Positive for at least one: 1. In last 30 days, any TMJ noise present with jaw movement or function 2. Patient reports any noise present during examination	Positive for the following: 1. Crepitus detected with palpation during at least one of the following: opening, closing, right or left lateral movement, or protrusive movement	κ = .33 (.01, .65)	.55	.61	1.41	.74
Subluxation	Positive for both: 1. In last 30 days, jaw locking or catching in a wide-open mouth position so could not close from wide-open position 2. Inability to close mouth from wide-open position without a self-maneuver	No examination findings required	Not reported	.98	1.00	Undefined	.02

Note: Reliability and validity are derived from the datasets of the Validation Project and TMJ Impact Project Finalization of DC/TMD.[11]