The Hypothesis-Driven Physical Examination

Brian T. Garibaldi, MD\textsuperscript{a,\,*}, Andrew P.J. Olson, MD\textsuperscript{b}

KEYWORDS
• Evidence-based physical diagnosis • Hypothesis-driven physical examination
• Likelihood ratios • Accuracy • Reliability

KEY POINTS
• The physical examination remains a vital part of the clinical encounter.
• Many physical examination maneuvers are just as reliable as diagnostic gold standard tests.
• A hypothesis-driven approach to the physical examination emphasizes the performance of specific physical examination maneuvers that are able to alter the likelihood of disease in a given patient.
• The physical examination should be taught to trainees in a context-specific manner as opposed to the traditional head-to-toe approach.
• Likelihood ratios are diagnostic weights that facilitate interpretation of physical examination findings at the bedside.

INTRODUCTION

For centuries, clinicians have used bedside observation to make diagnostic decisions. Over time, additional modalities have been added to aid in the diagnostic process. Perhaps the greatest example is the introduction of the stethoscope by Laennec in the early nineteenth century.\textsuperscript{1} As technology has advanced beyond the stethoscope, diagnosis has moved further away from the bedside in the form of laboratory testing and diagnostic imaging.\textsuperscript{2} However, the key to the accurate diagnosis of many conditions still lies in the bedside observations of an astute clinician. In some patients, these observations are the only way to determine the presence or absence of disease (eg, herpes zoster, Parkinson disease, cellulitis, and so forth).\textsuperscript{3} In other conditions, additional tests are needed for a definitive diagnosis (eg, myocardial infarction); but the physical examination plays a key role in substantially revising the probability of
disease in order to effectively guide further evaluation. Even after a diagnosis is made, the physical examination is important in following the disease’s trajectory and severity. For example, the presence of an S3 gallop in patients with heart failure predicts mortality and might prompt more aggressive intervention beyond the information found on an echocardiogram.4

### Several Factors Have Led to a Decline in Physical Examination Skills

Despite its enduring importance, several factors have led to a decline in physical examination skills in recent years.5–7 In the modern hospital, graduate medical trainees spend as little as 12% of their time in direct contact with patients and their families.8 This lack of time at the bedside has decreased opportunities for deliberate practice and reduced the number of practitioners who are confident in their ability to teach examination skills.6,9

There are also many practitioners and learners who question the value and relevance of the physical examination in the age of technology.10 Some fail to recognize that many physical examination maneuvers are just as reliable as gold standard technology-based tests. Reliability is commonly measured either through simple agreement or by calculating a kappa score. A kappa score of 0 means that agreement between two observers happens by chance alone. A kappa score of 1 indicates perfect agreement. In general, a kappa score greater than 0.4 is considered reasonable for a diagnostic test.3,11 Many physical examination maneuvers have kappa scores between 0.4 and 0.75, which indicates intermediate to good reliability (Table 1). Many diagnostic gold standards have kappa scores that are in that same range. Technology-based tests are not inherently more reliable than the physical examination.

Another reason that some clinicians, particularly those who trained more recently, hold a nihilistic view about the utility of the physical examination is that the examination is often taught as a list of maneuvers to be performed, regardless of the clinical context. Students learn this head-to-toe approach instead of tailoring their examination to each individual patient.12 They are then assessed on their ability to perform

### Table 1

<table>
<thead>
<tr>
<th>Physical Finding</th>
<th>Kappa</th>
<th>Diagnostic Standard</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver span &gt;9 cm by percussion</td>
<td>0.11</td>
<td>Classification of coronary artery lesions (by catheterization)</td>
<td>0.33</td>
</tr>
<tr>
<td>Delayed carotid upstroke</td>
<td>0.26</td>
<td>Pulmonary infiltrate (by chest radiograph)</td>
<td>0.38</td>
</tr>
<tr>
<td>Diminished cardiac dullness</td>
<td>0.49</td>
<td>Cardiomegaly (by chest radiograph)</td>
<td>0.48</td>
</tr>
<tr>
<td>Facial palsy (present or absent)</td>
<td>0.57</td>
<td>Severity of valvular regurgitation (by echo)</td>
<td>0.32–0.55</td>
</tr>
<tr>
<td>Clubbing (Schamroth sign)</td>
<td>0.64</td>
<td>Cirrhosis (by liver biopsy)</td>
<td>0.59</td>
</tr>
<tr>
<td>Systolic hypertension (SBP &gt;160 mm Hg)</td>
<td>0.75</td>
<td>Calf DVT (by ultrasound)</td>
<td>0.69</td>
</tr>
<tr>
<td>Tachycardia (pulse &gt;100 bpm)</td>
<td>0.85</td>
<td>Diagnosis of narrow complex tachycardia (by ECG)</td>
<td>0.70</td>
</tr>
<tr>
<td>Abdominal jugular test</td>
<td>0.92</td>
<td>Intertitial edema (by chest radiograph)</td>
<td>0.83</td>
</tr>
</tbody>
</table>

*Abbreviations: bpm, beats per minute; DVT, deep venous thrombosis; ECG, electrocardiogram; echo, echocardiogram; SBP, systolic blood pressure.*

the maneuvers on this extensive list. This approach is in stark contrast to how other diagnostic tests are taught and obtained and is an approach that is increasingly frowned on in the age of precision medicine. Imagine if medical students were encouraged to obtain a complete blood count, comprehensive metabolic panel, and coagulations studies in every patient in order to be thorough rather than tailoring their diagnostic evaluation. It is not at all surprising, then, that some learners find the physical examination to be less useful than other diagnostic tests.

Clinicians Tailor the Physical Examination to Each Individual Patient in a Hypothesis-Driven Fashion

Although the physical examination is taught and learned in a head-to-toe manner, this is rarely the way in which the examination is actually performed at the bedside. Clinicians in practice perform selected maneuvers in a sequence that is choreographed for each patient. Take, for example, the difference between the clinic room examination for a patient presenting with shortness of breath versus the examination for a patient with knee pain. Clinicians tailor their examination to the individual likelihood of various diseases in each patient and perform maneuvers that are likely to revise these probabilities. This approach to the physical examination is referred to as the hypothesis-driven physical examination (HDPE). Such a framework that encourages the application of the physical examination in a directed, tailored approach instead of a rote, acontextual manner can make the physical examination more useful and decrease unnecessary clinical testing.

We can all think of examples where an unexpected finding on a thorough physical examination has led to a correct and previously unconsidered diagnosis. This approach is particularly valuable in patients who are unable to provide a clear history (eg, patients with altered mental status, critically ill patients in the intensive care unit, and so forth) and in patients with a complex disease or unexplained symptoms in which there is a high level of diagnostic uncertainty and when myriad diagnoses are realistically being considered. In addition to uncovering important and unsuspected findings, performing a thorough physical examination in such patients often helps providers shift from system 1 to system 2 thinking as discussed in Bennett W. Clark and colleagues’ article, “Diagnostic Errors and the Bedside Clinical Examination,” in this issue. But this complete physical examination is actually a hypothesis-driven process where maneuvers are performed with anticipation for their results, rather than just going through the motions. The HDPE can resemble a complete physical examination under the right circumstances.

There is tremendous value in the physical examination for both patients and providers that goes beyond its ability to diagnose disease. There is an immeasurable sacredness and intimate vulnerability that occurs during a physical examination. As described in Cari Costanzo and Abraham Verghese’s article, “The Physical Exam as Ritual: Social Sciences and Embodiment in the Context of the Physical Exam,” in this issue, the physical examination is a ritual that in and of itself may have therapeutic properties for both patients and physicians alike. A well-performed physical examination may be beneficial, but a hastily or poorly performed examination has the potential to cause harm. A hypothesis-based approach to the physical examination requires that clinicians abandon physical examination maneuvers that have been shown to have little, if any, diagnostic utility and may, in fact, be harmful (eg, the screening pelvic examination in asymptomatic patients). By eliminating maneuvers that lack diagnostic value, the examination will be more useful and efficient, while still preserving the important social and cultural aspects of the physical examination as ritual.
**BASICS OF THE HYPOTHESIS-DRIVEN PHYSICAL EXAMINATION**

The performance of the HDPE can be separated into 3 tasks that can be then integrated into the whole task of a patient encounter (or simulated encounter for learning). First, accurate pretest probabilities for the likelihood of possible conditions in each patient must be considered. Second, physical examination maneuvers that have adequate operating characteristics to revise the probability of disease should be selected and performed. Lastly, the results of these findings must be combined with the respective pretest probability in order to arrive at a posttest probability. This will allow the creation of a prioritized differential diagnosis to guide the remainder of the evaluation. The authors discuss this process in more detail later.

**Determination of Pretest Probabilities**

One of the foundational aspects to making accurate, timely diagnoses is the determination of pretest probabilities for specific conditions; these are the probabilities that are revised during history taking and the physical examination. Overestimation or underestimation of pretest probability during diagnostic evaluation may lead to expenditure of cognitive and clinical resources on improbable diagnostic hypotheses. This issue is called base-rate neglect and is one of the most important factors in diagnostic error. Pretest probability can be determined using published data on the epidemiology of a disease, but it can also be modified based on a clinician's personal experience with a given population or in a certain practice. For example, a 65-year-old man with a 3-month history of progressive cough who is referred to an interstitial lung disease clinic at a tertiary care center is much more likely to have idiopathic pulmonary fibrosis than postnasal drip. This prevalence estimate is critical when evaluating the significance of test results. The pretest probability of some diagnostic considerations are orders of magnitude higher than others and should be given more initial diagnostic consideration.

**Using Likelihood Ratios to Select Appropriate Physical Examination Maneuvers**

In order for the HDPE to be maximally effective, clinicians must focus on performing maneuvers that have the best operating characteristics to revise the probability of a particular diagnostic hypothesis. Cognitive load is limited by this narrowing of focus, and more attention may be paid to the test’s findings rather than its performance alone.

How can we best understand which physical examination maneuvers have the most discriminatory/diagnostic value? In addition to understanding the reliability of a test (ie, its reproducibility from one examiner to another), a compelling, and pragmatic, answer lies in the concept of likelihood ratios (LRs). Some physical examination findings when present change the likelihood that patients have a disease. These findings are called positive LRs (LR+). For example, the finding of delayed carotid pulses increases the likelihood of aortic stenosis in patients with a systolic ejection murmur. Other physical examination findings when absent change the likelihood that patients have a disease. These factors are called negative LRs (LR−). This terminology can be confusing because both LR+ and LR− can increase or decrease the probability of disease. For example, the presence of an early peaking murmur decreases the likelihood of severe aortic stenosis in patients with a systolic ejection murmur. Similarly, the absence of radiation to the neck, decreases the likelihood of aortic stenosis.

LRs are derived from the sensitivity and specificity of the examination maneuver. The sensitivity of a physical examination maneuver refers to the proportion of patients who have the disease who have the physical finding. Specificity refers to the proportion of
patients without the disease who lack the physical examination finding. In general, a highly sensitive test when absent, decreases the likelihood of a disease. A highly specific test when present, increases the likelihood of disease. When visually displayed, one can see why sensitivity and specificity are used for decreasing and increasing likelihood, respectively (Fig. 1). A more sensitive test has fewer false negatives (and, thus, a negative test is likely to be a true negative), whereas a more specific test has fewer false positives (and, thus, a positive test is likely to be a true positive).

LRs combine these two performance characteristics. A positive LR is the number of patients with the disease who have the finding divided by the number of patients without the disease who have the finding (ie, the sensitivity divided by [1 – specificity]). A negative LR is the number of patients with the disease who do not have the finding, divided by the number of patients without disease who do not have the finding (ie, [1 – sensitivity] divided by specificity). For an excellent explanation of LRs, please refer to Stephen McGee’s text, *Evidence-Based Physical Diagnosis.*

![Fig. 1](image) Changes in sensitivity and specificity change false-positive and false-negative rates. (A) In tests with higher specificity, the false-positive rate is lower (compare light gray boxes in [A] and [B]). A positive test makes the diagnosis more likely. (B) In tests with higher sensitivity, the false-negative rate is lower (compare dark gray boxes in [A] and [B]). A negative test makes the diagnosis less likely.
Arriving at a Posttest Probability

If you have already estimated the pretest probability of a disease, you can apply a specific examination maneuver and, based on the presence of absence of a finding, you can determine the posttest probability of disease. LRs greater than 1 increase the posttest probability, whereas LRs less than 1 decrease the posttest probability. Calculating actual posttest probability is beyond the scope of this article, but there are a few helpful tips that allow the use of LRs at the bedside without complex calculations. A LR of 1 does not change the pretest probability, so a physical examination finding with an LR of 1 does not help in making a diagnosis. For example, Homans sign for detecting calf vein deep venous thrombosis has an LR of approximately 1 and is not helpful in ruling in or ruling out that diagnosis.\(^{21}\) LRs of 2, 5, and 10 increase the pretest probability by 15%, 30%, and 45%, respectively. LRs of 0.5, 0.2, and 0.1 decrease the pretest probability by 15%, 30%, and 45%, respectively. Physical examination LRs are essentially diagnostic weights that either increase or decrease the probability of disease (Fig. 2).\(^{3,20}\)

There are several useful resources that provide the prevalence of common diseases as well as the LRs for common physical examination findings.\(^{3,22}\)

Using Hypothesis-Driven Physical Examination in Practice

The use of HDPE in clinical practice is best illustrated by a case. Imagine that you are seeing a 64-year-old former smoker for a routine health visit. He asks you if he might have chronic obstructive pulmonary disease (COPD). The prevalence of COPD in adults in the United States who ever smoked is estimated to be 22%.\(^{23}\) On chest examination, he has absence of superficial cardiac dullness at the left lower sternal border. The LR for the finding of absent cardiac dullness at the left lower sternal border indicating COPD is approximately 10.\(^{3}\) The posttest probability of COPD in this patient based on that finding would be 67% (pretest probability of 22% \* 45% [the increase in probability associated with a LR of 10]). This posttest probability is high enough to warrant further diagnostic evaluation and may be an effective piece of information to positively inform smoking cessation counseling and to select disease-specific therapies.

The finding of absent cardiac dullness on anterior chest percussion is an underutilized physical examination maneuver. In addition to having a high LR, it is also a test with moderate reliability (kappa = 0.49).\(^{24}\) Many clinicians do not routinely perform anterior chest percussion. However, in the appropriate context it can provide significant diagnostic value. In contrast, percussing diaphragmatic excursion is a maneuver that is commonly taught to medical students that provides no significant diagnostic value when trying to determine if patients have COPD. In the case of suspected COPD, percussion of diaphragmatic excursion could be confidently omitted, much to the chagrin of the clinical skills course preceptor.

IMPORTANT CAVEATS FOR THE USE OF LIKELIHOOD RATIOS
Not All Examination Maneuvers Have Associated Likelihood Ratios

It is important to note that not all examination maneuvers have reported LRs. For example, if the presence of a physical examination finding defines the disease (i.e., it is the gold standard test), there will not be an associated LR. Think of a patient with a herpes zoster rash. One does not need to calculate pretest or posttest probabilities of disease if the rash is characteristic; the patient has zoster. There are also examination maneuvers for which robust LR data are not available. This lack of data does not mean that the examination maneuver has no value and should not be performed; that maneuver needs to be interpreted in the context of the individual patient and the findings considered based on experience and the best available evidence.

Using Multiple Examination Findings to Calculate Posttest Probability

Clinicians usually perform several different physical examination maneuvers when considering a diagnosis. When multiple examination maneuvers are performed, the LRs of each individual maneuver can probably be added together, as long as it is reasonable to assume that the findings are physiologically independent. This is oftentimes not the case, and this approach will likely overestimate the posttest probability of disease. Consider a 77-year-old patient who is referred to you because of a cardiac murmur that was heard during a recent hospitalization for urosepsis. You are concerned about possible aortic stenosis. On physical examination, you hear a mid to late peaking systolic ejection murmur that radiates to the carotids and note that he has delayed carotid upstrokes. The findings of delayed carotid upstrokes and a late peaking murmur each have LRs favoring the diagnosis of aortic stenosis. However, both findings are caused by obstruction to blood flow at the level of a stenotic aortic valve. It would not be fair to add their LRs together to determine the final posttest probability of disease because the findings have the same pathophysiologic basis. It would probably be more appropriate to take the larger of the two LRs and use that to calculate posttest probability.

In contrast, when examination findings are physiologically independent, they can likely be combined to modify posttest probability. For example, jaundice and splenomegaly are physiologically independent (the former arising from hepatic synthetic dysfunction and the latter arising from portal hypertension). These findings could be reasonably combined to revise the probability that a patient has cirrhosis.

One way to confidently use multiple examination maneuvers is to look for studies that have identified independent findings and combined them into prediction rules or “stop rules.” A recent and highly successful version of a stop rule is the HINTS (head-impulse test, direction-changing nystagmus, and test of skew) test for patients who present with acute dizziness. If a patient with acute dizziness has an abnormal head impulse test (implying a peripheral cause of vertigo), and no direction-changing nystagmus or skew deviation (both findings that imply a central cause of vertigo), the LR for an acute stroke is 0.02. This LR is better than the LR of a normal diffusion-weighted MRI in ruling out acute stroke. Other well-known examples are the Wells score for venous thromboembolism and the Centor criteria for streptococcal pharyngitis. Clinical prediction rules have the advantage of being more predictive than the individual tests they contain.

TEACHING THE HYPOTHESIS-DRIVEN PHYSICAL EXAMINATION
The Concept of Coselection

In addition to routinely performing more than one examination maneuver during a patient encounter, clinicians usually consider more than one diagnosis at a time. The
process by which multiple potential diagnoses are considered and evaluated is called coselection. Instead of creating diagnostic hypotheses after the history and physical examination (as is done in many morning reports), the HDPE encourages clinicians to consider diagnostic hypotheses before and during the application of history and physical examination maneuvers. This practice makes intuitive sense. Information obtained from the history guides additional questions and relevant physical examination maneuvers. The information obtained on the physical examination prompts new avenues of questioning as well as additional physical examination maneuvers. Although this is a process that many physicians conduct intuitively, there is concern that the cognitive load of coselection may be too great for trainees, who are learning the maneuvers at the same time they are learning the clinical context. Some of the cognitive load can be decreased by limiting the number of diagnostic hypotheses during simulation and early clinical experiences. This strategy is likely good for practicing clinicians as well: focusing on a few key potential diagnoses is a more doable task than considering many diagnoses without focused attention. Focusing physical examination training on key findings that discriminate between different diagnoses can also reduce the cognitive load for learners.

Medical Education Training Programs and Hypothesis-Driven Physical Examination

Medical education training programs should encourage mastery of physical examination maneuvers with proven or potential diagnostic utility and should de-emphasize maneuvers that lack diagnostic value. Examples of tests that should not be taught or routinely performed include Homans sign for diagnosing deep vein thrombosis and Tinel’s sign for diagnosing carpel tunnel syndrome. These tests lack discriminative value and are not helpful in changing the likelihood of disease. In addition, when another diagnostic modality (such as point-of-care ultrasound) is clearly superior to an existing physical diagnosis maneuver for the question at hand, clinicians should adapt their practice, when possible, to include that additional testing modality. Training programs at the undergraduate and graduate level should also incorporate newer diagnostic modalities. The physical examination is not in competition with technology; in many cases, they provide the most valuable information when used together.

A reflective approach that emphasizes discriminatory features improves learners’ diagnostic performance in simulated cases. This approach asks learners to identify what features for a given diagnostic hypothesis are present and expected, which features are expected but absent, and which features are present and unexpected. Early research shows that the concerns about students’ inability to learn and perform the physical examination in a hypothesis-driven manner are likely unfounded. First-year medical students naturally perform the physical examination in a hypothesis-driven, patient-tailored manner when allowed to do so during a clinical skills examination. This finding is striking given that first-year students have little, if any, contextual clinical knowledge.

SUMMARY

The physical examination remains a vital part of the clinical encounter. For some diagnoses, the physical examination remains the gold standard diagnostic test. For other diagnoses, the physical examination provides prognostic information above and beyond technologically based tests. Although some physicians think that technology is more reliable than bedside observation, this is simply not true for several important physical examination maneuvers. When faced with a clinical question, it is most often not practical to perform every head-to-toe physical examination maneuver from
medical school in the hopes of finding something relevant. Physicians select particular aspects of the examination that will likely be important to the issue at hand. This approach is often done intuitively based on experience and prior training. However, a more intentional focus on examination maneuvers that are accurate and reliable can greatly increase the efficiency and diagnostic yield of the physical examination. The HDPE is a valuable tool to practice and teach a high-yield approach to bedside clinical examination and is part of a growing movement to return practitioners and trainees to the bedside.6,7,30

REFERENCES