Diagnosing Acute Heart Failure in the Emergency Department: A Systematic Review and Meta-analysis

Jennifer L. Martindale, MD, Abel Wakai, MD, Sean P. Collins, MD, MSc, Phillip D. Levy, MD, MPH, Deborah Diercks, MD, Brian C. Hiestand, MD, Gregory J. Fermann, MD, Ian deSouza, MD, and Richard Sinert, DO

Abstract

**Background:** Acute heart failure (AHF) is one of the most common diagnoses assigned to emergency department (ED) patients who are hospitalized. Despite its high prevalence in the emergency setting, the diagnosis of AHF in ED patients with undifferentiated dyspnea can be challenging.

**Objectives:** The primary objective of this study was to perform a systematic review and meta-analysis of the operating characteristics of diagnostic elements available to the emergency physician for diagnosing AHF. Secondary objectives were to develop a test–treatment threshold model and to calculate interval likelihood ratios (LRs) for natriuretic peptides (NPs) by pooling patient-level results.

**Methods:** PubMed, EMBASE, and selected bibliographies were searched from January 1965 to March 2015 using MeSH terms to address the ability of the following index tests to predict AHF as a cause of dyspnea in adult patients in the ED: history and physical examination, electrocardiogram, chest radiograph (CXR), B-type natriuretic peptide (BNP), N-terminal proB-type natriuretic peptide (NT-proBNP), lung ultrasound (US), bedside echocardiography, and bioimpedance. A diagnosis of AHF based on clinical data combined with objective test results served as the criterion standard diagnosis. Data were analyzed using Meta-DiSc software. Authors of all NP studies were contacted to obtain patient-level data. The Quality Assessment Tool for Diagnostic Accuracy Studies-2 (QUADAS-2) for systematic reviews was utilized to evaluate the quality and applicability of the studies included.

**Results:** Based on the included studies, the prevalence of AHF ranged from 29% to 79%. Index tests with pooled positive LRs ≥ 4 were the auscultation of S3 on physical examination (4.0, 95% confidence interval [CI] = 2.7 to 5.9), pulmonary edema on both CXR (4.8, 95% CI = 3.6 to 6.4) and lung US (7.4, 95% CI = 4.2 to 12.8), and reduced ejection fraction observed on bedside echocardiogram (4.1, 95% CI = 2.4 to 7.2). Tests with low negative LRs were BNP < 100 pg/mL (0.11, 95% CI = 0.07 to 0.16), NT-proBNP < 300 pg/mL (0.09, 95% CI = 0.03 to 0.34), and B-line pattern on lung US LR (0.16, 95% CI = 0.05 to 0.51). Interval LRs of BNP concentrations at the low end of “positive” results as defined by a cutoff of 100 pg/mL were substantially lower (100 to 200 pg/mL: 0.29, 95% CI = 0.23 to 0.38) than those associated with higher BNP concentrations (1000 to 1500 pg/mL: 7.12, 95% CI = 4.53 to 11.18). The interval LR of NT-proBNP concentrations even at very high values (30,000 to 200,000 pg/mL) was 3.30 (95% CI = 2.05 to 5.31).

**Conclusions:** Bedside lung US and echocardiography appear to the most useful tests for affirming the presence of AHF while NPs are valuable in excluding the diagnosis.
Acute heart failure (AHF) is defined as a gradual or rapid deterioration in heart failure signs and symptoms in need of urgent treatment. Dyspnea is the most common presenting complaint prompting AHF patients to seek acute care. A primary diagnosis of AHF accounts for approximately one million emergency department (ED) visits in the United States. Despite its high prevalence in the ED setting, diagnosing AHF in ED patients with undifferentiated dyspnea can be challenging, especially in patients with advanced age and comorbid disease. There is no single historical, physical examination, electrocardiographic (ECG) or radiographic finding that can on its own reliably diagnose or rule out AHF as the cause of dyspnea. An ED diagnosis of AHF based on history, physical examination, chest radiograph (CXR), and ECG is qualified as “uncertain” in 44% of cases and is discordant with the final discharge diagnosis in nearly one out of every four cases. Natriuretic peptide (NP) testing improves diagnostic uncertainty for acutely dyspneic patients and is now a routine component of the workup of patients with possible AHF. However, even when NP testing is incorporated into the clinical workup of acute dyspnea, the misclassification rate remains 14% to 29%. Other diagnostic modalities such as bedside echocardiography, lung ultrasound (US), and bioimpedance have been shown to help discriminate between AHF and other primary causes of dyspnea, but their added clinical utility has yet to be fully characterized.

Reviews and meta-analyses aimed at helping clinicians sort through the array of available resources for evaluation of dyspneic patients have been previously published, but the most comprehensive of these is now more than a decade old. As new diagnostic modalities have been integrated into clinical practice and new data have been published in the ensuing period, the primary objective of this systematic review was to cohere the current best evidence concerning the diagnostic accuracy of index tests that might help discriminate AHF from other clinical conditions in patients presenting to the ED with dyspnea. The index tests evaluated in this review include: clinical history, symptoms, physical examination findings, ECG, CXR, B-type natriuretic peptide (BNP), N-terminal proB-type natriuretic peptide (NT-proBNP), lung US, ED-based bedside echocardiography, and bioimpedance.

METHODS

Study Design
We conducted a systematic review of studies that examined the operating test characteristics of the modalities used by emergency physicians (EPs) for the diagnosis of AHF among patients presenting to the ED with dyspnea. The systematic review was conducted using the Preferred Reporting Items for Systematic Review and Meta-analyses (PRISMA) guidelines.

Search Strategy
The design and manuscript structure of this systematic review conform to the recommendations from the Meta-analysis of Observational Studies in Epidemiology (MOOSE) statement. The medical literature was searched using PubMed and EMBASE from their inception through March 2015. With the assistance of an experienced health sciences librarian, the selected Medical Subject Headings (MeSH) terms heart failure and dyspnea were individually combined with the MeSH terms sensitivity and specificity, predictive value of tests, history taking, physical examination, electrocardiography, natriuretic peptide, ultrasonography, ultrasonics, echocardiography, and bioelectrical impedance. References from review articles identified by these searches were searched for relevant studies. Two authors independently screened the search results of each diagnostic modality of interest.

Studies eligible for inclusion in the systematic review were those that focused on the diagnosis of AHF in the ED population. Studies that focused on diagnostic tests that were not available shortly after ED presentation were excluded, as were studies that focused on patients with compensated, chronic heart failure and those that focused on prognosis or therapeutics. Case studies and reports or studies published solely in abstract form were excluded. The search was also restricted to human studies published in English.

Shortly after this systematic review project began, a comprehensive systematic review by Hill et al. on NPs for the diagnosis of AHF in the ED was published. All references identified by Hill et al. were screened for possible inclusion in this review. To avoid duplicating the PubMed and EMBASE NP searches performed by Hill et al., we limited our search for NP articles using these databases to a time frame beyond the stop point of their review (June 2012).

A review author (JLM) was tasked with ensuring that reported diagnostic data outside of the scope of each individual search was included in the analysis of the other relevant diagnostic elements in this review (e.g., a study identified only by the bioimpedance search strategy that reported BNP data was referred to reviewers also assigned to BNP/NT-proBNP).

Criteria for Considering Studies for this Review

Types of Participants. We included studies that recruited adult patients presenting to the ED with dyspnea as a primary complaint. The authors chose to exclude studies that recruited patients who presented to an urgent care setting, as we felt these patients might undergo an abbreviated diagnostic workup and likely represent a different spectrum of AHF than those who present to the ED with true, emergent dyspnea. Patients were not excluded based on comorbidities, etiology of AHF, or the presence of arrhythmia.

Types of Index Tests. We included studies that used history, symptoms, and physical examination findings as index tests for the diagnosis of AHF. Our search for NP testing was limited to BNP and NT-proBNP since these are the two most commonly used peptides for the ED diagnosis of AHF. We did not limit our search to point-of-care testing, as results from standard NP tests conducted in a hospital laboratory are typically made available to EPs during a patient’s ED course. We also did not limit our search and inclusion to any specific
type of assay. Our search for lung US studies was not limited to a specific protocol. We narrowed our inclusion of studies investigating lung US and echocardiography to those that had EPs both performing and interpreting these tests. Inclusion of bioimpedance studies was not limited to any particular type of protocol, bioimpedance metric, or specific device.

**Types of Reference Standard.** Acute heart failure is a clinical diagnosis, and there are currently no universally accepted diagnostic criteria to serve as the criterion standard for AHF in the acute care setting. Previously published diagnostic criteria (Framingham, Boston, National Health and Nutrition Examination Survey [NHANES]) may be used as clinical guides but on their own lack sensitivity for the diagnosis of AHF. We therefore chose to include studies that used as a reference standard a final diagnosis of AHF based on adjudication of clinical data by independent reviewers who were blinded to the study’s primary index test. We imposed no restrictions on the timing of the final diagnosis, the type of physician making the final diagnosis, or the type of clinical data upon which the final diagnosis was based.

**Data Abstraction.** Two or more authors for each index test independently selected articles from the combined PubMed/EMBASE search for full text review (history and physical, JM and AW; CXR, BH and GF; ECG, JM and AW; NT-proBNP and BNP, SC, DD, and PL; echocardiography, RS and Ids; lung US, RS and Ids; bioimpedance, BH and GF). Each reviewer independently selected potentially eligible studies before both authors agreed on the list of studies for full text review. Differences in study selection were resolved by consensus. Having read the methods sections of the full-text version of the studies potentially eligible for inclusion, each author then applied the stated inclusion and exclusion criteria to determine which studies to include in our systematic review. Differences were resolved by consensus after discussion and adjudication.

A standardized data collection form (see Data Supplement S1, available as supporting information in the online version of this paper) was used to abstract data pertaining to study funding, study location and setting, patient selection, patient demographics, prevalence of heart failure, manufacturer of the index test, definition of a positive diagnostic test, specialist performing and interpreting the index test, blinding with respect to the index test and criterion standard, and data incorporated into the criterion standard diagnosis of AHF. Study authors were contacted when study methodology or results required clarification.

**Data Analysis**

Sensitivities, specificities, and likelihood ratios (LRs) were calculated based on constructed two-by-two tables for each included study. To compute meta-analysis summary estimates when more than one study assessed the same index test, we combined test characteristic data using a random-effects model with MetaDiSc software. Interstudy heterogeneity was assessed for pooled estimates of sensitivity and specificity using the DerSimonian-Laird random-effects model. Publication bias, or overrepresentation of studies yielding positive results in published literature, was not assessed because a consensus approach toward funnel plot analysis is lacking for meta-analyses of diagnostic studies.

**BNP and NT-proBNP Analysis**

Summary analyses for dichotomous NP results were performed separately for each type of peptide, assay manufacturer, and common cutoff value. Cutoff values within 5 pg/mL for BNP and 50 pg/mL for NT-proBNP were considered common. Because they are ubiquitous in clinical practice, and have various decision-making cut-points, we also used patient-level data to compute interval LRs for BNP/NT-proBNP. The interval LR represents the probability of a test result within a user-defined interval in patients with a criterion standard diagnosis of AHF divided by the probability of a result in the same interval in patients with an alternative (non-AHF) diagnosis. To derive patient-level data, we contacted the authors of included studies that examined the test characteristics of BNP or NT-proBNP and requested actual patient-level NP results and their associated final diagnoses (AHF/not AHF).

**Quality Assessment**

Two authors independently used the revised Quality Assessment Tool for Diagnostic Accuracy Studies-2 (QUADAS-2) to evaluate the overall quality of evidence for included studies relating to each index test. The QUADAS-2 tool assists review authors in making explicit judgments for the risk of bias in four domains of study methodology: 1) patient selection, 2) index test, 3) reference standard, and 4) flow and timing. The tool also allows authors to rate how well a study’s 1) patient selection, 2) index test, and 3) reference standard apply to the specific research question posed by the systematic review. There are currently no criteria for assessing risk of bias in studies that compare multiple index tests. Studies that evaluated more than one index test were excluded from our QUADAS-2 analysis. The following signaling questions and statements were used to tailor the QUADAS-2 tool to this systematic review and guide quality judgments:

**Patient Selection.** The risk of spectrum bias was considered high if 1) the study was a case-control design or 2) the study made inappropriate exclusions that eliminated diagnoses with overlapping features of AHF or reduced the burden of disease in the non-AHF population. Examples of unacceptable exclusions included renal insufficiency, cirrhosis, morbid obesity, and other causes of dyspnea such as pneumonia, pulmonary embolism, and chronic obstructive pulmonary disease (COPD). Excluding these conditions would inflate the specificity of diagnostic tests. Acceptable exclusions included trauma patients, patients with an acute coronary syndrome unless dyspnea was the predominant complaint, and obvious noncardiac diagnoses such as pneumothorax. Loss of applicability was rated as “high” if the study...
failed to consider all patients presenting to the ED with a chief complaint of dyspnea for which AHF was in the differential. Studies that excluded patients based on age demographic (other than > 18 years) or dyspnea severity or had significant gender imbalance were also at high risk for losing applicability.

**Index Test.** If the result of the index test was interpreted with the foreknowledge of the initial diagnosis or other clinical data, the study was at high risk of over-estimating the accuracy of the diagnostic test. This applied to all index tests except history and examination, bioimpedance, and NP. If the study failed to use a predetermined cut-off value or definition of a positive test, the study was also considered to be at high risk for bias in this domain. Loss of applicability was rated as “high” if the diagnostic test was performed or interpreted by specialists other than EPs. Sonographic studies performed by fellowship-trained EPs were considered to be at high risk for losing applicability to the majority of EPs without formal US training.

**Reference Standard.** If the results from the index test were incorporated into the final diagnosis of AHF, this domain was considered at high risk for incorporation bias and inflated estimates of both sensitivity and specificity. Exception for history and examination elements was made, as these were considered essential for the reference standard. If the approach to the criterion standard diagnosis of heart failure was not explicit or if it failed to incorporate previously established clinical criteria (for example, Framingham, Boston, European Society of Cardiology, NHANES (32)), the risk of bias was considered high. If the study’s criterion standard did not match our review’s clinical definition (new or worsening dyspnea in addition to objective evidence supporting the diagnosis of AHF), this domain was considered less applicable to our review. For example, a study that diagnosed AHF based on echocardiographic evidence alone would be less applicable.

**Flow and Timing.** If there was a significant delay between ED presentation and performance of the index test, this domain was considered to be at high risk for bias. We used 6 hours as the cutoff after which a patient would likely have received therapeutic interventions and change clinical course enough to affect diagnostic testing. Sonographic and bioimpedance studies that excluded patients because of lack of feasibility (for example, body habitus, poor acoustic windows) were also considered to be at high risk for bias.

Each pair of authors independently piloted the QUADAS-2 tool with a sample of one to three included studies. Signaling questions that led to incongruent piloted responses were refined as deemed necessary. Kappa analysis using SPSS Statistics version 17.0 was used to calculate statistical agreement regarding blinded QUADAS-2 answers when more than three studies per index test could be evaluated by these criteria. When consensus agreement could not be reached between the two authors, a third reviewer (JLM) adjudicated.

Risk of bias and applicability for the 35 NP studies derived from the systematic review by Hill et al. was already examined using the QUADAS-2 tool. Evaluation of bias and applicability was therefore limited to the more recent studies identified by our search and included in this review.

**Test–Treatment Threshold**

The Pauker and Kassirer decision threshold model, which incorporates diagnostic test characteristics as well as estimated risks and benefits of treatment of AHF, was used to determine testing and treatment thresholds. Briefly, the Pauker and Kassirer decision threshold model is based on five variables: sensitivity, specificity, risk of a diagnostic test, risk of treatment, and anticipated benefit of treatment.

**RESULTS**

From the 9,405 citations identified by the PubMed searches, 9,317 citations identified by the EMBASE searches, and two studies identified from the bibliographies of review articles, a total of 57 studies including 52 unique patient cohorts were chosen for inclusion in this systematic review. A summary of the selection process for the systematic review is presented in Figure 1. Study selection diagrams illustrating the separate searches applied to each index test are shown in Data Supplement S2 (available as supporting information in the online version of this paper). Most studies were prospective and cross-sectional in design. Only two of the included studies published after 2003 were conducted using the Standards for Reporting of Diagnostic Accuracy (STARD) criteria.

**Prevalence**

The combined population from the 52 unique cohorts included in this review was 17,893 patients. The summary prevalence of AHF in these studies was 45.6% (95% CI = 44.9 to 46.4) with a range from 29% to 79%.

**History and Physical Examination**

A total of 31 studies were selected for inclusion and are fully described in Data Supplement S3 (available as supporting information in the online version of this paper). Studies that duplicated the cohort of an included study reported insufficient information to calculate test characteristics, or took place in the prehospital setting rather than the ED were excluded after full text review (Data Supplement S2). Two of the included studies focused on history and physical examination as the primary index tests of interest.

Point estimates for variables relating to clinical history, symptoms, and physical examination findings that were reported in at least four separate patient cohorts are included in Table 1. Additional point estimates are included in Data Supplement S3. Statistical heterogeneity was very high for all pooled estimates of variables reported in more than two studies except for diabetes (I² = 21%). There was no single historical variable, symptom, or physical examination finding that could significantly reduce the likelihood of AHF. A history of prosthetic valve, aortic or mitral valve disease, and...
hypertensive crisis were found to increase the likelihood of AHF, but these point estimates were derived from a single study rather than pooled studies (Data Supplement S3). The physical examination finding with the highest positive LR was an S3 gallop. For the 14 studies that reported S3 data, positive LRs ranged from 1.658 to 13.0.59

Based on the QUADAS-2 tool and due to the patient exclusion criteria used, the risk of bias in patient selection was high for the majority of the 17 studies subjected to this analysis (Data Supplement S2). Exclusion of patients with significant renal dysfunction, acute coronary syndrome, and other comorbidities in these studies likely resulted in a spectrum of patients that were less severely ill, leading to an overestimation of specificity for the diagnostic tests evaluated in these studies. Risk of bias for the criterion standard was rated low in 65% of the included studies. The reliability for the authors’ QUADAS-2 assessments was moderate (κ = 0.59, 95% CI = 0.43 to 0.75).

ECG
Searches from PubMed, EMBASE, and references of review articles identified 2,258 citations, 13 of which were selected for full text review. Of these, three were selected for inclusion in the systematic review. Searches dedicated to other index tests in this review identified an additional eight studies. ECG variables were grouped with baseline patient characteristic data in each of the included studies; electrocardiogram was not the index test of primary interest in any of these studies. All but two studies excluded patients with acute coronary ischemia or myocardial infarction. None of the abnormal electrographic findings substantially increased or decreased the likelihood of AHF (Table 2). The QUADAS-2 criteria could be applied to seven of the included studies (Data Supplement S2). Agreement between the two investigators was moderate (κ = 0.55, 95% CI = 0.29 to 0.81).

Chest Radiography
From the 1,845 citations selected from PubMed, EMBASE, and bibliographic review of key review articles, 18 were selected for full text review. However, only one study from full-text review was ultimately included; the other 17 studies included in the systematic review were identified from searches designed for the other index tests (Data Supplement S2). Chest radiography was not the primary index test of interest in any of the included
Table 1
Pooled Test Performance Characteristics for History and Physical Examination Findings

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>No. of Studies</th>
<th>No. of Patients</th>
<th>% AHF (95% CI)</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>LR− (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthopnea</td>
<td>15</td>
<td>5,430</td>
<td>45.5 (44.2–46.9)</td>
<td>52.1 (50.1–54.0)</td>
<td>70.5 (68.8–72.1)</td>
<td>1.9 (1.4–2.5)</td>
<td>0.74 (0.64–0.85)</td>
</tr>
<tr>
<td>PND</td>
<td>9</td>
<td>2,216</td>
<td>44.8 (42.8–46.9)</td>
<td>46.2 (43.7–48.6)</td>
<td>73.9 (71.9–75.9)</td>
<td>1.6 (1.2–2.1)</td>
<td>0.79 (0.71–0.88)</td>
</tr>
<tr>
<td>Dyspnea at rest</td>
<td>4</td>
<td>2,038</td>
<td>37.9 (35.9–40.0)</td>
<td>54.6 (51.2–58.0)</td>
<td>49.6 (46.9–52.3)</td>
<td>1.1 (0.9–1.4)</td>
<td>0.88 (0.74–1.04)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>History</th>
<th>No. of Patients</th>
<th>% (95% CI)</th>
<th>(95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>3</td>
<td>3,009</td>
<td>42.8 (40.1–44.6)</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>5</td>
<td>3,489</td>
<td>40.2 (38.6–41.9)</td>
</tr>
<tr>
<td>Renal failure</td>
<td>5</td>
<td>2,840</td>
<td>40.9 (39.1–42.7)</td>
</tr>
<tr>
<td>ML, history of</td>
<td>4</td>
<td>4,208</td>
<td>40.5 (39.1–42.0)</td>
</tr>
<tr>
<td>ARI</td>
<td>6</td>
<td>1,935</td>
<td>51.9 (49.8–54.2)</td>
</tr>
<tr>
<td>CAD</td>
<td>14</td>
<td>4,963</td>
<td>42.9 (41.5–44.3)</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>5</td>
<td>2,923</td>
<td>39.8 (38.1–41.6)</td>
</tr>
<tr>
<td>DM</td>
<td>19</td>
<td>7,707</td>
<td>47.3 (46.2–48.4)</td>
</tr>
<tr>
<td>HTN</td>
<td>25</td>
<td>10,137</td>
<td>45.6 (44.6–46.6)</td>
</tr>
</tbody>
</table>

| No history of COPD | 18 | 8,053 | 42.8 (41.7–43.9) | 78.9 (77.4–80.3) | 34.1 (32.6–35.6) | 1.22 (1.11–1.36) | 0.7 (0.6–0.8) |

<table>
<thead>
<tr>
<th>Examination findings</th>
<th>No. of Studies</th>
<th>No. of Patients</th>
<th>% AHF (95% CI)</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>LR− (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>14</td>
<td>5,900</td>
<td>45.2 (44.0–46.5)</td>
<td>12.7 (11.5–14.0)</td>
<td>97.7 (97.2–98.2)</td>
<td>4.0 (2.7–5.9)</td>
<td>0.91 (0.88–0.95)</td>
</tr>
<tr>
<td>JVD</td>
<td>23</td>
<td>8,012</td>
<td>47.8 (46.7–48.9)</td>
<td>37.2 (35.7–38.7)</td>
<td>87.0 (85.9–88.0)</td>
<td>2.8 (1.7–4.5)</td>
<td>0.76 (0.69–0.84)</td>
</tr>
<tr>
<td>Hepatogastroduodenal reflex</td>
<td>4</td>
<td>1,209</td>
<td>60.4 (57.6–63.1)</td>
<td>14.1 (11.9–16.6)</td>
<td>93.4 (91.2–95.5)</td>
<td>2.2 (1.3–3.7)</td>
<td>0.91 (0.88–0.94)</td>
</tr>
<tr>
<td>Leg edema</td>
<td>26</td>
<td>9,626</td>
<td>47.2 (46.2–48.2)</td>
<td>51.9 (50.5–53.4)</td>
<td>75.2 (74.0–76.4)</td>
<td>1.9 (1.6–2.3)</td>
<td>0.98 (0.81–0.76)</td>
</tr>
<tr>
<td>Murmurs</td>
<td>8</td>
<td>4,004</td>
<td>45.3 (43.8–46.8)</td>
<td>27.8 (25.8–29.9)</td>
<td>93.2 (91.6–94.8)</td>
<td>1.9 (0.9–3.9)</td>
<td>0.93 (0.79–1.08)</td>
</tr>
<tr>
<td>Rales</td>
<td>22</td>
<td>8,775</td>
<td>48.2 (47.1–49.2)</td>
<td>62.3 (60.8–63.7)</td>
<td>88.1 (86.7–89.4)</td>
<td>1.8 (1.5–2.1)</td>
<td>0.60 (0.51–0.69)</td>
</tr>
<tr>
<td>Wheezing</td>
<td>13</td>
<td>6,970</td>
<td>44.2 (43.0–45.3)</td>
<td>22.3 (20.9–23.8)</td>
<td>64.0 (62.5–65.4)</td>
<td>0.6 (0.5–0.8)</td>
<td>1.19 (1.09–1.30)</td>
</tr>
<tr>
<td>Absent fever</td>
<td>7</td>
<td>3,157</td>
<td>43.6 (41.9–45.3)</td>
<td>92.4 (90.9–93.8)</td>
<td>20.6 (18.6–22.5)</td>
<td>1.14 (1.02–1.27)</td>
<td>0.4 (0.3–0.6)</td>
</tr>
</tbody>
</table>

AFIB = atrial fibrillation; CAD = coronary artery disease; CHF = congestive heart failure; COPD = chronic obstructive pulmonary disease; CRI = chronic renal insufficiency; DM = diabetes mellitus; HTN = hypertension; JVD = jugular venous distension; LR+ = positive likelihood ratio; LR− = negative likelihood ratio; MI = myocardial infarction; PND = paroxysmal nocturnal dyspnea.
Table 2

Pooled Test Performance Characteristics for Chest Radiograph and Electrocardiogram Findings

<table>
<thead>
<tr>
<th>No. of Studies</th>
<th>No. of Patients</th>
<th>Electrocardiogram Findings</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>LR+</th>
<th>LR−</th>
<th>95% CI</th>
<th>p (LR+)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ischemic changes15,51</td>
<td>2 / 1,138</td>
<td>55.8 (53.7–63.8)</td>
<td>5.6 (3.9–7.7)</td>
<td>96.5 (94.2–98.1)</td>
<td>2.0 (1.0–1.00)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Atrial fibrillation19,20,36,58,60,65</td>
<td>6 / 2,242</td>
<td>55.8 (53.7–63.8)</td>
<td>5.6 (3.9–7.7)</td>
<td>96.5 (94.2–98.1)</td>
<td>2.0 (1.0–1.00)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ST-depression58,65</td>
<td>2 / 1,024</td>
<td>60.8 (57.8–63.8)</td>
<td>5.5 (4.9–6.1)</td>
<td>91.7 (89.9–92.7)</td>
<td>2.2 (1.0–1.00)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Normal sinus rhythm8,12,62</td>
<td>3 / 1,207</td>
<td>39.6 (36.9–42.4)</td>
<td>5.2 (2.1–10.5)</td>
<td>91.8 (88.4–95.1)</td>
<td>5.2 (1.7–16.2)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kerley B-lines36,72</td>
<td>2 / 814</td>
<td>46.8 (43.4–50.5)</td>
<td>3.1 (1.2–8.1)</td>
<td>99.3 (98.4–99.9)</td>
<td>9.2 (6.5–12.5)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interstitial edema</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>98.8 (97.3–99.6)</td>
<td>9.8 (6.5–12.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alveolar edema</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>97.1 (95.2–98.7)</td>
<td>9.4 (5.9–13.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pulmonary edema*7,8,12,14,16,18</td>
<td>15 / 3,501</td>
<td>48.3 (46.2–50.5)</td>
<td>5.7 (4.7–6.8)</td>
<td>96.9 (95.6–98.4)</td>
<td>4.8 (3.6–6.4)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enlarged cardiac silhouette8,12,15,18,20,21,54,58,60,64</td>
<td>15 / 3,501</td>
<td>48.3 (46.2–50.5)</td>
<td>5.7 (4.7–6.8)</td>
<td>96.9 (95.6–98.4)</td>
<td>4.8 (3.6–6.4)</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pulmonary venous hypertension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96.8 (95.1–98.0)</td>
<td>9.8 (6.5–13.0)</td>
</tr>
</tbody>
</table>

LR+ = positive likelihood ratio; LR− = negative likelihood ratio; CI = confidence interval. *Refers to generalized pulmonary edema in studies that did not report specifically on both “alveolar edema” and “interstitial edema.”

**Note:** The radiographic variable most commonly reported by the included studies was pulmonary edema. For the purpose of pooling results, the terms “condensation,” “pulmonary congestion,” “edema,” pulmonary venous hypertension,” “interstitial and/or alveolar edema,” and “interstitial edema” (when alveolar edema was not an included study variable) were equated with “pulmonary edema.” The positive LR associated with pulmonary edema was 4.8 (95% CI = 3.6 to 6.4; Table 2; forest plot, Data Supplement S2). Pulmonary edema, however, was an insensitive finding (sensitivity = 56.9%; 95% CI = 54.7 to 59.1%). The presence of enlarged cardiac silhouette and pleural effusions were less helpful findings.

QUADAS-2 analysis could be performed on 114,15,16,18,23,54,58,60,62,64,67,70,73,80 of the included studies (Data Supplement S2). Agreement between the two investigators applying the QUADAS-2 criteria was fair (k = 0.40, 95% CI 0.24 to 0.56). Patient selection was at risk for decreased applicability in 73% of the studies because they excluded subgroups of dyspneic patients based on comorbidities. Risk of bias for the criterion standard diagnosis was rated high in each of the included studies because radiographic results were consistently incorporated into final diagnoses.

**BNP/NT-proBNP**

A total of 417,8,10,11,14,16–21,25,35,46,50,51,54,55,58,60–62,64,67,70,73–88 unique studies were included in this review; 35 were identified from the systematic review by Hill et al.,29 four25,49,79,83 from our 2012–2015 PubMed and EMBASE searches and two18,78 from searches performed for other index tests as part of this systematic review (Data Supplement S2). Of the 41 included studies, 20 evaluated BNP,11,14,16,18,19,21,25,35,50,51,54,55,58,60,74,76,78,84–86 and 14 evaluated NT-proBNP,7,8,10,17,20,49,61,62,64,70,79,81,83,88 and seven evaluated both BNP and NT-proBNP.34,66,74,75,77,80,82

With the exception of three studies,10,60,81 NPs were the primary index tests of interest. Three of the included studies were funded by companies that manufactured the NP assay under investigation.7,35,83

The main inclusion criterion for most studies was a chief complaint of dyspnea, although a few studies broadened this criterion to include patients with lower extremity edema19 or signs or symptoms of AHF (dyspnea, edema, fatigue).4,45 Data Supplement S3. A common exclusion criterion was renal insufficiency,4,16,25,35,55,60,62,64,70,74,78,88 or renal failure,55,61,79 The criterion standard diagnosis of AHF was based on
clinical judgment alone or fulfillment of clinical criteria after independent chart review of hospitalization records. Blinding to all NP results was reported in 29 (69%) of the included studies.

Test characteristics were pooled according to assay manufacturers, given the systematic differences that have been shown among commercially available NP assays. At the most commonly reported cutoff value of 100 pg/mL, BNP demonstrated high sensitivity and a negative LR less than 0.2 (forest plot, Data Supplement S2). The pooled sensitivity and specificity among the 19 studies of the Elecsys Roche immunoassay for NT-proBNP provided test characteristic data for the cutoff value of 300 pg/mL. The pooled negative LR at this cutoff was 0.09 (95% CI = 0.03 to 0.34; forest plot, Data Supplement S2). At a significantly higher cutoff value of 1500 pg/mL, specificity improved, but only to 72.9% (95% CI = 70.6 to 75.0%). Diagnostic performance data for other NP assays used in the included studies of this systematic review are shown in Table 3.

We were able to obtain patient-level data from the authors of five studies that evaluated BNP only,14,16,51,76,78 of two studies that evaluated NT-proBNP only,86,88 and of two studies that evaluated both BNP and NT-proBNP.75,77 Data derived from common commercially available NP assays were pooled to calculate interval LRs. Data from the six unique studies14,19,35,50,51,55,58,60,67,73 that shared patient-level NT-proBNP results was 0.86 (95% CI = 0.83 to 0.89) for BNP and 0.76 (95% CI = 0.74 to 0.78) for NT-proBNP (Data Supplement S2).

According to the QUADAS-2 analysis performed by Hill et al.,29 the only category rated as high risk of bias in the majority of included studies was patient selection. In four,5,25,46,79 of the five studies published after Hill et al.'s systematic review that evaluated more than one index test, patients with renal disease were excluded. By eliminating patients with this common comorbidity from the sample population, there are likely to be fewer false-positive NP results; spectrum bias in these studies results in inflated specificity.12 Two studies40,83 not included in Hill et al.'s systematic review evaluated a single index

<table>
<thead>
<tr>
<th>Assay</th>
<th>Cutoff (pg/mL)</th>
<th>Sensitivity % (95% CI)</th>
<th>Specificity % (95% CI)</th>
<th>PPV % (95% CI)</th>
<th>NPV % (95% CI)</th>
<th>LR- (95% CI)</th>
<th>LR+ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triage Biosite</td>
<td>100</td>
<td>45.0 (42.4–47.5)</td>
<td>84.8 (82.8–86.7)</td>
<td>9.1 (4.1–20.2)</td>
<td>1.8 (1.2–2.9)</td>
<td>0.11 (0.07–0.16)</td>
<td>0.08 (0.04–0.12)</td>
</tr>
<tr>
<td>AxSym, Abbott</td>
<td>100</td>
<td>45.0 (42.4–47.5)</td>
<td>84.8 (82.8–86.7)</td>
<td>9.1 (4.1–20.2)</td>
<td>1.8 (1.2–2.9)</td>
<td>0.11 (0.07–0.16)</td>
<td>0.08 (0.04–0.12)</td>
</tr>
<tr>
<td>iSTAT, Abbott</td>
<td>100</td>
<td>45.0 (42.4–47.5)</td>
<td>84.8 (82.8–86.7)</td>
<td>9.1 (4.1–20.2)</td>
<td>1.8 (1.2–2.9)</td>
<td>0.11 (0.07–0.16)</td>
<td>0.08 (0.04–0.12)</td>
</tr>
<tr>
<td>Dimension, Dade Behring</td>
<td>100</td>
<td>45.0 (42.4–47.5)</td>
<td>84.8 (82.8–86.7)</td>
<td>9.1 (4.1–20.2)</td>
<td>1.8 (1.2–2.9)</td>
<td>0.11 (0.07–0.16)</td>
<td>0.08 (0.04–0.12)</td>
</tr>
</tbody>
</table>
test. The QUADAS-2 evaluation of these two studies is presented in Data Supplement S3.

### Lung US

A total of 268 citations were identified by the PubMed and EMBASE searches, 26 of which were selected for full-text review. One ED-based study,53 was excluded because cardiologists rather than EPs performed and interpreted the lung US (Data Supplement S2). Studies that took place in European countries,23,24 in which emergency medicine was emerging as a new specialty at the time of patient enrollment were included if the physicians performing and interpreting lung US were considered EPs in their countries. A total of eight studies,23,24,48,62,63,80,90 (N = 1,918 patients) were selected for inclusion in this review. All studies included adults presenting to the ED with dyspnea. Study characteristics are described in Data Supplement S3. Russell et al.63 selected only those patients in whom dyspnea was truly undifferentiated; patients in whom a diagnosis of AHF seemed clinically obvious were excluded. Training was limited to didactic and workshop sessions for most studies; in two of the studies48,63 sonographers were fellowship-trained in emergency US. Study characteristics are presented in Data Supplement S3.

A positive lung US was defined in every study by the presence of at least three B lines in two bilateral lung zones. In six of the included studies,23,24,48,62,63,81 the US protocol was based on scanning eight thoracic lung zones (four anterior and four lateral, as described by Volpicelli13,82). Two studies50,91 modified this protocol to interrogate six anterior-lateral thoracic lung zones.

Diffuse pulmonary edema identified on lung US proved to be a diagnostic variable with discriminatory value (positive LR 7.4, 95% CI = 4.2 to 12.8; negative LR 0.16, 95% CI = 0.05 to 0.51). Statistical heterogeneity was high for these pooled estimates (I² = 78% and P = 99%, respectively). Positive LRs among the eight included studies ranged from 2.619 to 10.21 (forest plot, Data Supplemental 2). Pleural effusions visualized on lung US were less helpful in diagnosing or excluding AHF (positive LR 2.0, 95% CI = 1.4 to 2.8).

The risk for bias in the criterion standard diagnosis was rated as high in the four studies23,24,63,90 that could be evaluated by QUADAS-2 (Data Supplement S3) because standard clinical criteria were not applied to the criterion standard diagnosis in these studies. In only one of these studies24 were lung US results incorporated into the criterion standard diagnosis. Risk of decreased applicability in the domain of patient selection was rated as high in three studies that excluded patients with comorbid conditions24 or patients requiring mechanical ventilation.23,63 Agreement between the two investigators applying the QUADAS-2 criteria was high71 (κ = 0.93, 95% CI = 0.79 to 1.0).

### Beside Echocardiography

After reviewing the full-text version of 18 studies, four20,21,48,63 (N = 675 patients) were selected for inclusion in this review (Data Supplement S3). Three ED-based studies18,19,53 were excluded from this review because cardiologists rather than EPs performed or interpreted the sonograms (Data Supplement S2). EPs fellowship-trained in US performed the US examinations in the studies by Anderson et al.48 and Russell et al.63. The sonographers in the studies by Nazerian et al.20, Russell et al.53, and Wang et al.21 were blinded to clinical data. Inter-observer agreement was reported only by Wang et al.21 (0.889). In all four included studies20,21,48,63 the physicians who made the final diagnosis were blinded to the echocardiograms performed as index tests. Three studies20,21,63 analyzed patients with suboptimal imaging as false negatives. To maintain consistency, these patients were not included in pooled estimates in our review.

A summary of the echocardiographic test characteristics analyzed in the four included studies20,21,48,63 are reported in Table 5. Ejection fraction (EF) was determined by visual estimation in the studies by Anderson et al.48, Nazerian et al.,20 and Russell et al.63. Wang et al.21 measured EF based on left ventricular dimensions at end-diastole and end-systole but did not report the test characteristics of EF as an index test for AHF in their study. Elevated left ventricular end-diastolic dimension, defined as > 28.6 mm/mm² by Wang et al.,21 did not improve diagnostic accuracy when compared to visual estimation of EF. Only one study, by Nazerian et al.20 evaluated diastolic function, which found that a

---

**Table 4**

Interval LRs of BNP and NT-proBNP Values

<table>
<thead>
<tr>
<th>BNP Value (pg/mL)</th>
<th>Interval LR</th>
<th>N (%)</th>
<th>NT-proBNP (pg/mL)</th>
<th>Interval LR</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–100</td>
<td>0.14 (0.12–0.18)</td>
<td>617 (28)</td>
<td>0–100</td>
<td>0.09 (0.05–0.17)</td>
<td>150 (7.5)</td>
</tr>
<tr>
<td>100–200</td>
<td>0.29 (0.23–0.38)</td>
<td>308 (14)</td>
<td>100–300</td>
<td>0.23 (0.16–0.33)</td>
<td>205 (10.2)</td>
</tr>
<tr>
<td>200–300</td>
<td>0.89 (0.67–1.17)</td>
<td>188 (9)</td>
<td>300–600</td>
<td>0.28 (0.20–0.39)</td>
<td>212 (10.5)</td>
</tr>
<tr>
<td>300–400</td>
<td>1.34 (0.98–1.83)</td>
<td>148 (7)</td>
<td>600–900</td>
<td>0.63 (0.46–0.87)</td>
<td>151 (7.5)</td>
</tr>
<tr>
<td>400–500</td>
<td>2.05 (1.47–2.84)</td>
<td>148 (7)</td>
<td>900–1,500</td>
<td>0.84 (0.67–1.06)</td>
<td>249 (12.4)</td>
</tr>
<tr>
<td>500–600</td>
<td>3.50 (2.30–5.35)</td>
<td>115 (5)</td>
<td>1,500–3,000</td>
<td>1.49 (1.19–1.86)</td>
<td>273 (13.6)</td>
</tr>
<tr>
<td>800–1,000</td>
<td>4.13 (3.01–5.68)</td>
<td>218 (10)</td>
<td>3,000–5,000</td>
<td>2.36 (1.81–3.09)</td>
<td>225 (11.2)</td>
</tr>
<tr>
<td>1,000–1,500</td>
<td>5.00 (3.21–7.89)</td>
<td>130 (6)</td>
<td>5,000–10,000</td>
<td>2.48 (1.91–3.21)</td>
<td>239 (11.9)</td>
</tr>
<tr>
<td>1,500–2,500</td>
<td>7.12 (4.53–11.18)</td>
<td>160 (7)</td>
<td>10,000–15,000</td>
<td>2.84 (1.90–4.23)</td>
<td>112 (5.6)</td>
</tr>
<tr>
<td>2,500–5,001</td>
<td>8.33 (4.60–15.12)</td>
<td>105 (5)</td>
<td>15,000–20,000</td>
<td>2.93 (1.95–4.39)</td>
<td>111 (5.5)</td>
</tr>
<tr>
<td></td>
<td>8.91 (4.09–19.43)</td>
<td>65 (3)</td>
<td>30,000–20,000</td>
<td>3.30 (2.05–5.31)</td>
<td>86 (4.3)</td>
</tr>
</tbody>
</table>

| N (%) |

BNP = B-type natriuretic peptide; LR = likelihood ratio; NT-proBNP = N-terminal proB-type natriuretic peptide.
restrictive pattern on pulsed Doppler analysis of mitral inflow most accurately predicted AHF with positive and negative LRs of 8.3 (95% CI = 4.0 to 16.9) and 0.21 (95% CI = 0.12 to 0.36), respectively.

The QUADAS-2 criteria could only be applied to the study by Russell et al. because these criteria do not apply to studies that compare multiple index tests. Risk of bias in patient selection was rated as high due to their exclusion of patients with causes of dyspnea presumed to be obvious (a patient with known heart failure who was not compliant with taking medications) and those who had received prior treatment (Data Supplement S3).

### Biopendence

Four studies that evaluated the diagnostic accuracy of bioimpedance on ED patients (N = 1,039 patients) were selected from the 65 studies that were screened (Data Supplement S2). One study evaluated conventional bioelectrical impedance analysis, while three used bioelectrical impedance vector analysis (BIVA) in which resistance and reactance are plotted as a bivariate vector on a nomogram. Characteristics of the included studies are listed in Data Supplement S3. One of the four studies was designed as a case-control study in which study groups were divided based on BNP value, clinical evidence of AHF, and respiratory symptoms; however, for our pooled analyses the author provided us with the final criterion standard diagnoses of AHF/not AHF and the associated hydration index values for each patient.

Segmental resistance measures thoracic fluid status using sensing and output electrodes placed on the anterior thigh and suprasternal notch. Segmental resistance values lower than the cutoff determined by ROC curve intercept offered the highest positive LR (10.6, 95% CI = 5.8 to 19.2; Table 6). BIVA, which accounts for age, sex, and body mass index, offered lower positive LRs (Table 6).

QUADAS-2 criteria could be applied to the studies that evaluated a single index test (Data Supplement S3). Risk of bias in patient selection was rated as high in these studies because they excluded patients with comorbid conditions that caused ascites and peripheral edema.

### Test–Treatment Threshold Estimates

Test–treatment threshold models contextualize a diagnostic test in a clinical setting by linking the discriminatory value of a test with the benefits and risks associated with a given treatment. Evidence regarding the clinical benefits and risks of pharmacologic treatments for AHF, however, is very limited. No pharmacologic therapy for the treatment of acute AHF has been given a class I/level of evidence A recommendation from the American Heart Association or Heart Failure Society of America. Randomized controlled trials investigating the effect of loop diuretics on clinically relevant outcomes are lacking. In the DOSE (Diuretic Optimization Strategies Evaluation in Acute Heart Failure) trial, high-dose intravenous diuretic therapy was associated with a higher risk of increasing serum creatinine within 72 hours of treatment initiation compared with low-dose diuretic therapy, but this increase was not observed at 60 days.

A recent Cochrane review on nitrate vasodilator therapy showed no difference in outcomes compared with alternative interventions, but this review reflects a paucity of available evidence. In the Vasodilation in the Management of Acute CHF (VMAC) trial comparing intravenous nitroglycerin with nesiritide and with placebo, 10% of the 216 patients receiving intravenous nitroglycerin experienced symptomatic hypotension; mean doses 3 hours after treatment initiation were 42 µg/min in catheterized patients and 29 µg/min in noncatheterized patients. These studies, however, may not provide an accurate estimation of the benefits and risks of diuretics and nitrates when administered early in the course of managing a patient in the ED. Some of the patients included in these studies were enrolled well after their initial presentation and early management. Doses studied in the VMAC trial may be lower than those administered to hypertensive patients with AHF in the ED. A test–treatment threshold model based on these numbers might be less relevant to the EP.

Table 5

<table>
<thead>
<tr>
<th>Lung US</th>
<th>N</th>
<th>n</th>
<th>% AHF (95% CI)</th>
<th>Sensitivity % (95% CI)</th>
<th>Specificity % (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>LR− (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive B-line scan²³, ²⁴, ⁴⁶, ⁶³, ⁸¹, ⁹¹, ⁹⁵</td>
<td>8</td>
<td>1914</td>
<td>48.2 (46.0–50.5)</td>
<td>85.3 (82.8–87.5)</td>
<td>92.7 (90.9–94.3)</td>
<td>7.4 (4.2–12.8)</td>
<td>0.16 (0.05–0.51)</td>
</tr>
<tr>
<td>Pleural effusion(s)⁵³,⁹⁰</td>
<td>2</td>
<td>155</td>
<td>40.7 (33.2–48.5)</td>
<td>63.5 (50.4–75.3)</td>
<td>71.7 (61.4–80.6)</td>
<td>2.0 (1.4–2.8)</td>
<td>0.49 (0.22–1.10)</td>
</tr>
<tr>
<td>Bedside echocardiography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restrictive mitral pattern²⁰</td>
<td>1</td>
<td>125</td>
<td>43.2 (34.9–52.0)</td>
<td>81.5 (68.6–90.7)</td>
<td>90.1 (80.7–95.9)</td>
<td>8.3 (4.0–16.9)</td>
<td>0.21 (0.12–0.36)</td>
</tr>
<tr>
<td>Reduced EF²⁴,⁴⁸,⁶³</td>
<td>3</td>
<td>325</td>
<td>41.2 (36.0–46.7)</td>
<td>80.6 (72.9–86.9)</td>
<td>80.6 (74.3–86.0)</td>
<td>4.1 (2.4–7.2)</td>
<td>0.24 (0.17–0.35)</td>
</tr>
<tr>
<td>Increased LV end-diastolic dimension²¹</td>
<td>1</td>
<td>84</td>
<td>58.3 (47.7–68.3)</td>
<td>79.6 (65.7–89.7)</td>
<td>68.6 (50.7–83.1)</td>
<td>2.5 (1.5–4.2)</td>
<td>0.30 (0.16–0.54)</td>
</tr>
</tbody>
</table>

Lung ultrasound: *defined as ≥ 2 bilateral lung zones with ≥ 3 B-lines per intercostal space.

Bedside echocardiography: *defined as E/A ratio > 2 or E/A between 1 and 2 and deceleration time (DT) < 130 msec; DT < 130 msec alone if atrial fibrillation. Defined as LVEDD > 28.6 mm/m².

AHF = acute heart failure; EF = ejection fraction; LV = left ventricular; LR = likelihood ratio; N = number of studies; n = number of patients; US = ultrasound.
The estimated benefits and risks of common pharmacologic interventions such as intravenous diuretics or nitrate vasodilator therapy that we chose to apply to a test–treatment threshold model, therefore, are hypothetical and used for illustrative purposes. Mathematically, the test–treatment threshold is based on the relative costs of treating patients without disease C (i.e., hypotension from vasodilator and diuretic therapy) and failing to treat those with disease B. Higher estimated risk of heart failure treatment C relative to treatment benefit B results in a higher treatment threshold. If the risk of treating dyspneic patients without underlying AHF were estimated to be 0.45 and the risk of failing to treat AHF patients expeditiously were 0.19, then the treatment threshold, or posterior probability of disease (PTT), at which the costs of these two risks (B and C) would be balanced is 70%. This assumes that the cost of NP testing (including harm to the patient) is negligible.

Figure 2 displays the posterior probabilities yielded by different diagnostic tests when starting at a pretest probability of 46%. Those diagnostic tests with positive LRs large enough to yield a posterior probability greater than 70% would guide the Bayesian clinician to treat for AHF.

Applying the Pauker and Kassirer threshold model to a diagnostic test with a dichotomous outcome also allows one to start with a posterior probability (treatment threshold; PTT) and using the positive and negative LRs associated with the test, calculate the range of pretest probabilities for which the diagnostic test has the potential to change the decision to treat for AHF. This range is defined by the limits of the no treat–test threshold and the test–treatment threshold. Starting with a posterior treatment threshold probability of 70% and using the positive and negative LRs of lung US, the no treat–test threshold is 24% and the test–treatment threshold is 94% (Figure 3). When the pretest probability lies between these thresholds, lung US has the potential to affect management. Different values for estimated risks and benefits of administering AHF treatment can be

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Pooled Test Performance Characteristics for Bioimpedance Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Studies</td>
<td>No. of Patients</td>
</tr>
<tr>
<td>Segmental BIA (≤54 Ohms) 60</td>
<td>1</td>
</tr>
<tr>
<td>Whole body BIA (≤441 Ohms) 60</td>
<td>1</td>
</tr>
<tr>
<td>BIVA (Z(Xc) – 1SD) 24</td>
<td>1</td>
</tr>
<tr>
<td>BIVA, HI (73.4%) 25,78</td>
<td>2</td>
</tr>
</tbody>
</table>

BIA = bioelectrical impedance analysis; BIVA = bioelectrical impedance vector analysis; HI = hydration index; Z(Xc) – 1SD = 1 standard deviation below mean Z-score vector (reactance).

Figure 2. Posterior probabilities yielded by different diagnostic tests when starting at a pretest probability of 46%. BNP = B-type natriuretic peptide; CXR = chest radiograph; LR = likelihood ratio; NT-proBNP = N-terminal proB-type natriuretic peptide.
incorporated into the Pauker and Kassirer model using a Microsoft Excel calculator (Data Supplement S4).

**DISCUSSION**

The diagnosis of AHF in the adult ED population remains challenging. This systematic review demonstrates that clinical elements such as past medical history, presenting symptoms, and physical examination findings, on their own, cannot be relied on for excluding or establishing the diagnosis. It also demonstrates that symptoms commonly sought in a clinical history such as orthopnea, paroxysmal nocturnal dyspnea, and weight gain fail to distinguish patients with AHF. S3 gallop, the physical examination finding most suggestive of AHF, is only 13% sensitive. Rales and peripheral edema are even less suggestive of AHF. This review did not evaluate the diagnostic accuracy of historical elements, symptoms, or examination findings in combination. Clinical gestalt, based on an aggregate effect of the history and physical examination, likely outperforms these diagnostic elements in isolation and plays an important role in determining the pretest probability of AHF. However, none of the studies included in this review investigated QRS amplitude, a parameter that has been shown to attenuate with worsening heart failure. Chest radiography is considered a fundamental component of the ED workup for AHF but radiographic signs of pulmonary edema and vascular redistribution are often absent in AHF patients. The poor sensitivity of CXR findings for diagnosing pulmonary edema has been previously described. Even in the presence of severely elevated pulmonary capillary wedge pressures in patients with heart failure, radiographic pulmonary congestion is absent 39% of the time.

The data in this systematic review are consistent with that of prior studies demonstrating that at recommended respective cut-points of 100 and 300 pg/mL, BNP and NT-proBNP testing are most useful for excluding AHF. Calculation of interval LRs from patient-level NP results, however, is unique to this review and helps provide a more clinically intelligible interpretation of NP test performance. This review highlights the shortcomings of dichotomizing continuous variables into binary outcomes above and below a single cutoff point. The positive LR associated with a BNP of 150 pg/mL, when applying pooled binary data relating to the single cutoff value of 100 pg/mL (Triage, Biosite, Table 3), is 2.2 (95% CI = 1.8 to 2.7). Based on this value, a diagnosis of AHF would be favored because it is statistically grouped with substantially higher BNP values in AHF patients. Using interval LR data, however, a
BNP value of 150 pg/mL is associated with an LR of 0.29 (95% CI = 0.23 to 0.38) and thus favors an alternative diagnosis (Table 4). Only above a cut-point of 800 pg/mL for BNP does the interval LR substantially increase the posttest probability of AHF. The interval LRs for NT-proBNP only modestly favor the diagnosis of AHF even at extremely elevated levels.

This review focuses on the diagnostic accuracy of NPs. The utility of these biomarkers, however, is based on whether or not their use results in clinical benefit. Several randomized controlled trials that have evaluated the clinical performance of NPs have failed to demonstrate differences in patient-centered clinical outcomes. One possible explanation is that NPs add little value beyond the clinical judgment of an EP. When the diagnosis of AHF is relatively certain, the performance of NPs is suboptimal. NPs may prove to have little value beyond the clinical judgment of an EP. When the diagnosis of AHF is uncertain, the performance of NPs is suboptimal. NPs may prove to have greater clinical utility when used with a Bayesian approach and interpreted as a continuous rather than as a dichotomous variable.

Given the limitations of NP values above recommended “rule-out” cut-points, point-of-care lung US can play a potentially significant role in the ED diagnosis of AHF in the ED. The diagnostic performance of lung US in this review was superior to any other diagnostic test that was studied in more than one cohort of patients. With a pooled positive LR of 7.4 (95% CI = 4.2 to 12.8) and negative LR of 0.16 (95% CI = 0.05 to 0.51), lung US demonstrates potential to both rule in and exclude the diagnosis of AHF. Statistical heterogeneity among these studies was substantial, and this should be taken into account while interpreting these summary estimates. Spectrum bias resulting from the exclusion of patients with alternative causes of dyspnea and peripheral edema should also be considered.

Perhaps the most robust data in our pooled lung US sample comes from the multicenter study by Pivotta et al. (N = 1,005). Patient exclusions were limited to those with initially obvious causes of dyspnea (traumatic pneumothorax) and intubated patients. Lung US in this study was performed by 62 EPs across community and academic hospitals, and the criterion standard diagnosis of AHF at hospital discharge was determined with high inter-rater agreement (κ = 0.93). The positive and negative LRs determined in this study were 14.0 (95% CI = 10.2 to 19.3) and 0.10 (95% CI = 0.08 to 0.14), respectively. Incorporation of lung US into the classification of AHF in this study led to a net reclassification improvement of 19% (95% CI = 14.6 to 23.6%). The authors of this study report that the vast majority of lung US examinations were performed within 40 minutes of ED presentation. The feasibility of lung US to rapidly identify pulmonary edema in real time shortly after ED presentation and before therapeutic intervention may increase the sensitivity of this test.

In a recent meta-analysis, Al Deeb et al. evaluated the test performance of lung US in diagnosing AHF. The pooled positive and negative LRs from this meta-analysis were 12.4 (95% CI = 5.7 to 26.8) and 0.06 (95% CI = 0.02 to 0.22), respectively. However, only three of the seven studies included in their analysis took place in the ED. Two of the seven included studies took place in the intensive care unit (ICU) where spectrum bias might increase the sensitivity of pathologic B-lines. While the prevalence of other causes of diffuse sonographic B-lines such as interstitial lung disease, multifocal pneumonia, and severe acute respiratory distress syndrome might decrease the specificity of this pattern for AHF in the ICU, higher estimates of specificity in these studies may be related to study design. The test characteristics of B-lines in one ICU study were evaluated in predefined groups of patients with pulmonary edema, COPD, and no cardiopulmonary disorder. In the other ICU study, sonography was performed by experts in thoracic US who used sonographic artifacts to classify other causes of dyspnea including pneumonia, venous thromboembolism, and pneumothorax.

Echocardiography is considered key to the diagnosis of chronic heart failure, but limited availability and echocardiographic experience among EPs have limited its role in the diagnostic evaluation of dyspnea in the ED setting. Echocardiography is more technically challenging and complex than lung US, but perhaps the simplest echocardiographic assessment relating to heart failure is the visual estimation of EF. EF data was based on visual estimation in the three included studies that reported this variable. Visual estimation of EF by both cardiologists and EPs has been shown to correlate well with quantitative assessments of EF. However, use of reduced EF alone as an echocardiographic variable for predicting AHF would result in the failure to detect the 50% of heart failure patients with preserved EF. Further, many patients also have a prior EF available in the medical record, which may limit the amount of new information provided by an EF recorded in the ED.

Pulsed Doppler analysis of mitral inflow enables the evaluation of relative velocities during early and late diastolic filling and provides a surrogate measure of elevated left ventricular filling pressures. A restrictive pattern of diastolic filling in the single included study was associated with a positive LR of 8.3 (95% CI = 4.0 to 16.9) for the diagnosis of AHF. However, mitral inflow analysis cannot reliably be applied to patients who are tachycardic or have permanent pacing or mitral valve prostheses, and standard definitions cannot be applied to patients in atrial fibrillation. Accurate mitral inflow analysis is also highly dependent on preload, leading to an increasingly common use of combined approaches that incorporate tissue Doppler imaging techniques. Perhaps most importantly, acquiring mitral inflow and tissue Doppler data may be beyond the scope of many EPs who do not have formal fellowship training. Given this, further studies investigating the ability of a restrictive pattern to predict AHF and the ability of EPs with minimal echocardiographic training to accurately perform and interpret Doppler analysis are needed.

Bioimpedance analysis has emerged as a potential noninvasive assessment of pulmonary congestion and peripheral edema, based on the theory that intrathoracic fluid decreases resistance to an applied electrical current. Both thoracic and whole body impedance...
devices have been evaluated. While many implantable devices incorporate a data channel to measure thoracic impedance, EPs rarely have the output from the device while the patient is being evaluated in the ED. Therefore, this review focuses on measurements taken by external monitoring devices. Barriers to routine adoption thus far have included the time that it takes to acquire the data and requirement for the patient to remain relatively motionless for up to several minutes.\textsuperscript{118}

**Implications for Future Research**

Given the suboptimal accuracy with which AHF is discriminated from other causes of dyspnea in the ED, clinical decision aids based on multiple logistic regression modeling may have a role in assisting with the clinical diagnosis of AHF. Useful models will likely require several diagnostic components and should incorporate rather than take the place of clinical gestalt. One such model based on age, pretest probability, and NT-proBNP intervals has demonstrated diagnostic accuracy in both internally derived and externally validated cohorts.\textsuperscript{87} Lung US and bedside echocardiography are diagnostic tests to consider including in different decision models and NPs are likely best utilized to exclude the diagnosis of AHF. We could foresee other biomarkers and diagnostic tests quantifying pathophysiology and assisting with the diagnosis of AHF by excluding other causes such as infection, acute kidney injury, or obstructive pulmonary disease. The clinical utility of these new diagnostic adjuncts would need to be evaluated alongside current diagnostic modalities.

Future diagnostic studies should adhere to the STARD reporting guidelines to make more explicit the potential for bias.\textsuperscript{47} Ideally, diagnostic studies for AHF would avoid overt spectrum bias by including patients with comorbidities such as renal failure. Reporting test characteristics as they pertain to discrete intervals of biomarker values in future studies would be more informative than reporting data relating to a single cutoff value.\textsuperscript{47} As shown in this review, dichotomization of continuous variables can produce distorted and exaggerated LRs. Dichotomous results from cutoff values in future biomarker and bioimpedance studies may be reported for expediency. Results based on cutoff values that are determined a priori rather than retrospectively derived from intercepts of ROC curves are likely to be more generalizable. Diagnostic test results that are based on interpreter judgment should be reported with statistical data reflecting rater reliability. Until a better criterion standard for the diagnosis of AHF becomes available, a clinical diagnosis made by two reviewers blinded to the index test will have to serve as the criterion standard.

**LIMITATIONS**

The most important limitations of this systematic review apply to the criterion standard diagnosis of AHF. The criterion standard for AHF in the included studies was a clinical diagnosis made by two or more clinicians after medical record review. In the absence of a better criterion standard for the diagnosis of AHF, a clinical diagnosis that combines a subjective assessment of a patient with some combination of objective data points must be relied on as a criterion standard. Estimations of the diagnostic accuracy of an index test for AHF are only as valid as the criterion standard diagnoses are accurate. The credibility of the criterion standard diagnosis is likely to vary among the studies included in this review. Incorporation of echocardiographic data, for example, into the diagnostic workup and criterion standard diagnosis among the included studies was variable.

Although our search strategies for the electronic databases used in this systematic review were constructed by an experienced health sciences librarian using accepted principles of health information science, it is still possible that they may have missed some studies eligible for inclusion. However, the comprehensive nature of our literature search makes it unlikely that we have excluded any important studies whose findings would significantly alter the conclusions of this review. In addition, there is the possibility of language bias in the findings of this review because non-English publications were excluded. However, this type of exclusion has not been shown to significantly affect the results of meta-analyses.\textsuperscript{119}

Data relating to all diagnostic tests reported by a study were included in the meta-analysis, even if some of these diagnostic tests were not the primary index test evaluated by that study. We analyzed multiple tests derived from a single study as independent diagnostic variables to aggregate results according to diagnostic category. Coming from the same sample of patients, which has a distinct composition and spectrum of disease, it is unlikely that the individual variables investigated in this sample behave independently with respect to test performance. Likewise, to think of the summary LRs for each diagnostic test as stand-alone values would overlook their interdependence and derivation from the same studies. Also, data collection relating to other diagnostic tests included in a study may not have been as rigorous as that applied to the study’s primary index test. Data from those diagnostic tests other than the primary index test under a study’s investigation likely factored into the criterion standard diagnosis. Incorporation bias challenges the validity of these data, potentially increasing the sensitivity and specificity of the diagnostic test.\textsuperscript{42,43}

The quality of studies that were included in the meta-analysis was variable. Differences in inclusion and exclusion criteria among the included studies put them at varying degrees of risk for spectrum bias.\textsuperscript{42} Appraisal of pooled results should factor in the clinical as well as statistical heterogeneity among included studies.

A limitation specific to the NP analyses in this review was our lack of consideration of age as a variable affecting BNP and NT-proBNP values. Data relating to the specificity of these tests may have been more accurate if age-based cutoffs were evaluated.\textsuperscript{7,12,120,121} Other variables known to affect NP values, such as renal function,\textsuperscript{75,77,80,122,123} were also not factored in to our analysis of diagnostic accuracy.

Last, the test–treatment threshold model suggested in this review is a conceptual model. It is limited by the
paucity of available data on which to base estimates for the clinical benefits and risks of treatment. However, it may be useful for the EP to apply this model to patient disposition; the posterior treatment threshold can be thought of as a threshold above which the physician, confident that AHF is the underlying cause of dyspnea, can stop testing. Below this threshold, the physician would be compelled to seek further evidence in the ED favoring the diagnosis of AHF.

CONCLUSIONS

Elements of clinical history, symptoms, physical examination, chest radiography, and electrocardiography, on their own, lack discriminatory value in making or excluding the diagnosis of acute heart failure in ED patients. B-type natriuretic peptide and N-terminal proB-type natriuretic peptide are most valuable in ruling out acute heart failure when values are lower than the suggested cutoff points of 100 and 300 pg/mL, respectively. Values above these thresholds may be less helpful for establishing the diagnosis than previously described. Lung ultrasound appears to have the best combination of test characteristics with the presence or absence of diffuse B-lines providing reliable information to confirm or exclude the ED diagnosis of acute heart failure. Other diagnostic modalities such as bioimpedance may have future utility in the ED setting, although more study is needed.

We thank Christopher Stewart, senior assistant librarian at SUNY Downstate Medical Center, for his help in conducting the literature searches for this systematic review. We thank the following authors for providing us with patient-level natriuretic peptide data: Tommy Chung, Camille Chenevier-Gobeaux, Joel Coste, Christopher R. deFilippi, Salvatore Di Somma, Andrew S. Liteplo, Simcha R. Meisel, and Thomas Mueller.

References


44. Dickstein K, Cohen-Solal A, Filippatos G, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM). Eur Heart J 2008;29:2388–442.
86. Rogers RK, Stoddard GJ, Greene T, et al. Usefulness of adjusting for clinical covariates to improve the ability of B-type natriuretic peptide to distinguish cardiac from noncardiac dyspnea. Am J Cardiol 2009;104:689–94.


103. McMurray JJ, Adamopoulos S, Anker SD, et al. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. Eur. J Heart Failure 2012;14:803–69.


Supporting Information

The following supporting information is available in the online version of this paper:
Announcing Usus – A community website on usage

Usus (Latin for usage) is a new, independent, community-run website (http://www.usus.org.uk/) for all those interested in the usage of online content. It is designed to support a productive conversation among librarians, publishers, aggregators, and repository managers so that we can all get the best possible usage reports for our electronic resources.

The Usus website provides:
- A source of hints and tips on solving known problems
- A list of vendors with problems that are affecting the credibility and/or usefulness of the COUNTER reports
- A collection point for suggestions for new COUNTER usage reports and metrics

Supervisory Board
The Usus Supervisory Board will ensure that the website is editorially independent and will serve the needs of the community. Chaired by Anne Osterman, Deputy Director of VIVA (the Virtual Library of Virginia), the members of the Supervisory Board are:

Anne Osterman, VIVA, USA (Chair)
Simon Bevan, Cranfield University, UK
Melissa Blaney, ACS Publications, USA
Anna Creech, University of Richmond, USA
Lorraine Estelle, JISC, UK
Oliver Pesch, EBSCO, USA
Kari Schmidt, Montgomery College, USA
Mark Tullos, ProQuest, USA

Travel Award
Thank you to Project COUNTER for the financial support needed to get Usus off the ground. COUNTER has also offered to provide a travel award worth £1,000/$1,500 to a librarian who contributes the best opinion piece for the News & Opinions section of the Usus site. The award can be used to travel to the Charleston Conference, UKSG Conference, or Electronic Resources & Libraries Conference. Please send your submissions of 1,000 words or less to usus.stats@gmail.com by December 31, 2014.

For more information, please contact Usus at: usus.stats@gmail.com