Improving the Diagnosis of Acute Heart Failure Using a Validated Prediction Model

Brian Steinhart, MD,*§ Kevin E. Thorpe, MMATH,‡** Ahmed M. Bayoumi, MD,*‡¶ Gordon Moe, MD,*¶‖ James L. Januzzi, Jr, MD,‡‡ C. David Mazer, MD†‡#

Toronto, Ontario, Canada; and Boston, Massachusetts

Objectives
We sought to derive and validate a prediction model by using N-terminal pro–B-type natriuretic peptide (NT-proBNP) and clinical variables to improve the diagnosis of acute heart failure (AHF).

Background
The optimal way of using natriuretic peptides to enhance the diagnosis of AHF remains uncertain.

Methods
Physician estimates of probability of AHF in 500 patients treated in the emergency department from the multicenter IMPROVE CHF (Improved Management of Patients With Congestive Heart Failure) trial recruited between December 2004 and December 2005 were classified into low (0% to 20%), intermediate (21% to 79%), or high (80% to 100%) probability for AHF and then compared with the blinded adjudicated AHF diagnosis. Likelihood ratios were calculated and multiple logistic regression incorporated covariates into an AHF prediction model that was validated internally by the use of bootstrapping and externally by applying the model to another 573 patients from the separate PRIDE (N-Terminal Pro-BNP Investigation of Dyspnea in the Emergency Department) study of the use of NT-proBNP in patients with dyspnea.

Results
Likelihood ratios for AHF with NT-proBNP were 0.11 (95% confidence interval [CI]: 0.06 to 0.19) for cut-point values <300 pg/ml; increasing to 3.43 (95% CI: 2.34 to 5.03) for values 2,700 to 8,099 pg/ml, and 12.80 (95% CI: 5.21 to 31.45) for values ≥8,100 pg/ml. Variables used to predict AHF were age, pre-test probability, and log NT-proBNP. When applied to the external data by use of its adjudicated final diagnosis as the gold standard, the model appropriately reclassified 44% of patients by intermediate clinical probability to either low or high probability of AHF with negligible (<2%) inappropriate redirection.

Conclusions
A diagnostic prediction model for AHF that incorporates both clinical assessment and NT-proBNP has been derived and validated and has excellent diagnostic accuracy, especially in cases with indeterminate likelihood for AHF. (J Am Coll Cardiol 2009;54:1515–21) © 2009 by the American College of Cardiology Foundation

Heart failure has the second-highest disease burden for length of stay in hospitalized patients (1) and is rapidly becoming the most expensive disease to manage worldwide (2,3). Decompensation occurs frequently, and the emergency department (ED) setting is where most of these acute heart failure (AHF) patients present (4). Correct early diagnosis and treatment is essential to reduce the rate of morbidity and mortality, yet the accurate clinical diagnosis in this setting occurs <80% of the time (5).

See page 1522

Biomarkers have been used to assist in the diagnosis of AHF for the undifferentiated patient, with B-type natriuretic peptide (BNP) and N-terminal pro–B-type natriuretic peptide (NT-proBNP) being the 2 biomarkers most studied. When used as dichotomous variables with set positive/negative threshold values, the authors of prospective studies (6,7) have found good sensitivities but relatively lower positive predictive values for the disease and have demonstrated a benefit of combining their results with clinical judgment (8). When treated as categorical variables with negative, positive, and indeterminate ranges, results show modest improvement for diagnosing the disease (6–8).

Although national guidelines (9–11) suggest that these tests may be of value in diagnosing AHF when the clinician
Diagnosis of AHF Using a Model

Methods

Data for the present analysis were obtained from a previous observational, prospective, blinded cohort study of a convenience sample of 534 patients presenting to 1 of 7 Canadian urban EDs with undifferentiated shortness of breath. Enrollment occurred between December 2004 and December 2005. Exclusion criteria were either acute myocardial infarction (defined as any ischemic presentation with either electrocardiogram), renal failure (serum creatinine >250 μmol/l; >2.8 mg/dl), malignancy or a case of dyspnea that clearly was without suspicion for AHF (e.g., wheezing in a 26-year-old patient with known asthma).

Consent was obtained, blood tests including NT-proBNP were drawn, and the patient underwent a history and physical examination performed by the emergency physician. After the chest radiograph and electrocardiogram were reviewed, the emergency physician estimated the probability of AHF (from 1% to 100%) without knowledge of the drawn NT-proBNP value. Standard clinical management was performed. The NT-proBNP was immunoasayed by the use of Roche Diagnostics Elecsys proBNP (Indianapolis, Indiana) on Elecsys 1010, 2010, or E170 instruments. Cross-reactivity and coefficient of variation of the assays were acceptable.

After completion of the study, adjudication for AHF was determined independently by 2 cardiologists by use of the Framingham Heart Score and NHANES I (National Health and Nutrition Examination Survey) Heart Failure Criteria as guides (19,20). Each cardiologist had available all index ED medical charts, further hospital admission test results and records (e.g., echocardiogram, computed tomography thorax, pulmonary function studies; patient response to diuretic or broncholytic therapy), and results of a 60-day follow-up telephone conversation but were blinded to the NT-proBNP levels. Because AHF is a clinical diagnosis, their final diagnoses were considered the "gold standard" reference for outcome measurements. Full details of the parent study methodology were described elsewhere (8).

On the basis of the clinician’s assessment, the study population was divided into low pre-test probability (≤20%), high pre-test probability (≥80), and intermediate pre-test probability (21% to 79%) subgroups for AHF (15). We then compared the clinician’s impression to the adjudicated diagnosis and NT-proBNP value.

The performance characteristics of NT-proBNP as both a categorical and continuous variable were analyzed by computing likelihood ratios (LRs) for the diagnosis of AHF over a number of NT-proBNP ranges (14). First we considered the current standard ranges in practice of AHF of <300 pg/ml, ≥300 and <900 pg/ml, and ≥900 pg/ml (8). Then we considered the values as continuous, creating the logarithmic ranges of <300 pg/ml, ≥300 and <900 pg/ml, ≥900 and <2,700 pg/ml, ≥2,700 and <8,100 pg/ml, and ≥8,100 pg/ml. The LR was calculated as the proportion of patients with AHF who had a result in the range divided by the proportion of patients without AHF who had a result in the range.

We fit a multiple logistic regression model (21) to determine whether pre-test probability could be combined with NT-proBNP and other covariates to predict AHF. Because of the very long right tails in the NT-proBNP distributions, we used the common logarithm of NT-proBNP values in the regression model, which "pulled" the extreme values closer to the rest of the data. Incorporation into a post-test model for AHF was performed by expressing the regression model on the probability scale by the inverse logit transformation. The concordance index c statistic (equivalent to the area under an ROC curve) was calculated as a measure of discrimination and the bootstrap method was used to validate the model for overfitting (R Language 2.5.0, Design Package Version 2.0-12, R Foundation for Statistical Computing, Vienna, Austria).

The calibrated model was examined on an external data set from the prospective American PRIDE (N-Terminal Pro-BNP Investigation of Dyspnea in the Emergency Department) trial of 573 undifferentiated dyspneic ED patients (6). The inclusion/exclusion enrollment criteria and
methodology were similar to our IMPROVE-CHF (Improved Management of Patients With Congestive Heart Failure) study. These patients were compared with the IMPROVE-CHF cohort by the use of the $t$ test for age and the Wilcoxon rank sum test for pre-test probability and NT-proBNP variables. All $p$ values were 2-sided, with values $<0.05$ considered significant. The percentage of PRIDE patients appropriately redirected was determined as was analysis of model improvement.

Our analysis of model improvement used the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) as developed by Pencina et al. ($22$). For a model to improve diagnostic accuracy, patients with the condition (AHF) should be reclassified to a greater risk group whereas those without the outcome (no AHF) should be reclassified to a lower risk group. These improvements need to be adjusted for the respective reclassification errors (e.g., a patient with the condition is reclassified to a lower-risk group). This is what the NRI measures but is dependent on the choice of risk group categories. The IDI is not dependent on risk groups; it can be interpreted as "the difference between improvement in average sensitivity and any potential increase in average 'one minus specificity'" ($22$).

### Results

Five hundred thirty-four patients were enrolled, and 34 patients were excluded because of protocol violation or not fulfilling inclusion/exclusion criteria. Of the 500 patients remaining, 17 data forms for pre-test probability were incomplete, resulting in 483 patient data analyzed. The average age was 70 years, with no significant difference in sex; further baseline characteristics and demographics are fully described elsewhere ($8$).

Adjudication for AHF resulted in discordance for diagnosis in 10 cases of 483; all 10 cases were ultimately agreed upon. The prevalence of adjudicated AHF in the 3 pre-test probability groups is shown in Table 1. The largest group was the intermediate clinical probability group with 38% of total patients (184 of 483) and an adjudicated AHF prevalence of 43%. The overall study prevalence for adjudicated AHF was 46%.

Distribution of NT-proBNP values within each pre-test group (Fig. 1) showed a trend toward lower values in the adjudicated no AHF subgroup compared with the AHF one but with significant overlap between the 2; this relationship continued between all 3 pre-test groups. Overall median NT-proBNP values for "no AHF" and "AHF" patients were 320 and 3,820 pg/ml, respectively. The values were skewed with overlap between the 2 groups. Use of the logarithm of the value, as expected, brought considerably more symmetry to the distribution.

Table 2 and Figure 2 depict the LR$s$ and post-test probabilities by the use of standard threshold cut points and multiple cut points for ruling in/out AHF. The LR for the

<table>
<thead>
<tr>
<th>Pre-Test Probability of AHF (Group)</th>
<th>Adjudication Diagnosis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AHF</td>
<td>No AHF</td>
</tr>
<tr>
<td>≤20%</td>
<td>26 (16.0)</td>
<td>137 (84.0)</td>
</tr>
<tr>
<td>21%–79%</td>
<td>80 (43.5)</td>
<td>104 (56.5)</td>
</tr>
<tr>
<td>≥80%</td>
<td>115 (84.6)</td>
<td>21 (15.4)</td>
</tr>
<tr>
<td>Total</td>
<td>221 (45.8)</td>
<td>262 (54.2)</td>
</tr>
</tbody>
</table>

Values are n (%). 
AHF = acute heart failure.
The lowest NT-proBNP range of <300 pg/ml had a value of 0.11 with narrow confidence intervals (CIs). The LRs for the other two ranges of 300 to 899 pg/ml and ≥900 pg/ml were 0.34 (95% CI: 0.19 to 0.60) and 2.75 (95% CI: 2.29 to 3.30), respectively. Analyzing NT-proBNP values by expanding the number of intervals ≥900 pg/ml from 1 to 3 allowed much more information to become available—the LRs for AHF with NT-proBNP ranged from 0.11 (95% CI: 0.06 to 0.19) for values <300 pg/ml to 12.80 pg/ml (95% CI: 5.21 to 31.45) for values ≥8,100 pg/ml; for intermediary ranges of 300 to 899 pg/ml, 900 to 2,699 pg/ml, and 2,700 to 8,099 pg/ml LR values remained modest. Only the ≥8,100 range had an acceptable LR (12.80) for diagnosing AHF (23).

When patient age, pre-test probability of AHF, and log(NT-proBNP) were used, the logistic regression fit yielded a model for estimating the probability that a patient had AHF (Online Appendix). This model had a concordance index of 0.905, which indicated excellent discrimination. Bootstrap validation indicated negligible overfitting, and the model was well calibrated internally. External validation on 573 cases confirmed excellent discrimination (c = 0.97) but did show some unreliability. Specifically the model tended to underestimate the true probability of AHF in the external data despite applying linear logistic calibration for prediction. We have used the uncalibrated model for simplicity.

Of 600 patients previously reported in the PRIDE study from Boston, 573 had complete data analyzed for this study. When compared with the IMPROVE-CHF study, age, pre-test probability, and NT-proBNP values were all found to be significantly different (Table 3) (p < 0.001). Most of the reclassification improvement occurred in the intermediate-probability category, as shown in Table 4. When applied to the pre-test low-probability risk group of 343 patients where the clinician was quite sure the diagnosis was not AHF, the model confirmed what the clinician already suspected in 276 of 282 patients but inappropriately redirected several patients to the intermediate probability category (30 of 58 patients), with few appropriately to high probability (3 of 3 patients). When the model was applied to the pre-test high-probability risk group of 91 patients where the clinician was quite sure the diagnosis was AHF, the same trend occurred. With most pre-test high-probability patients, the model verified what the clinician already thought (72 of 73 patients) but inappropriately redirected most of the remaining patients to the intermediate-probability category (14 of 18 patients).

When applied to the pre-test intermediate probability risk group where the clinician was most in need of diagnostic help, the model redirected the clinician to either low or...
high probability with an extremely high rate of appropriateness; in fact 44% (95% CI: 40% to 49%) of the initial high probability risk group and only 1 patient inappropriately reclassified. The model was an improvement to clinical judgment alone as measured by NRI and IDI (22). The NRI was 0.23 (p < 0.001), and the IDI was 0.11 (p < 0.001), which are both considered large degrees of improvement.

### Discussion

We have derived and externally validated a new mathematical model for improving the diagnosis of AHF in patients with undifferentiated dyspnea. Although more complex than the use of NT-proBNP as a dichotomous variable, the use of the marker as a continuous variable provided further enhancement of the diagnosis for AHF than when used as a categorical test. The use of a mathematical model for diagnostic medicine is innovative, as is its analysis by NRI and IDI for appropriately redirecting the clinician.

The authors of previous biomarker studies (6,8,15,24) have emphasized sensitivity, specificity, and ROC curve analyses that are test performance measures but were not necessarily designed to be used for decision making on individual patients (14). Positive and negative predictive values may enable clinicians to interpret biomarker values in a more meaningful way but they are influenced by the prevalence rates, which vary between studies (14). To minimize prevalence effect and facilitate decision making at the bedside, a more appropriate analysis is to calculate likelihood ratios (23,25,26). A recent systematic review of the English language literature found a pooled positive LR of 3.3 for BNP or NT-proBNP for diagnosing AHF (27).

Our resultant findings support a strategy for treating natriuretic peptide values as continuous, which obviates the need for clinicians to remember cut points or stratify them by age, which further strengthens the notion that although cut points are necessary to define boundaries of disease states, interpretation of biomarkers such as BNP or NT-proBNP as continuous variables (in conjunction with combining clinical variables and expert judgment) is a superior approach (28).

The patients in whom the clinicians are indecisive for the diagnosis may very well be the best target for natriuretic peptide testing because these patients are those in whom worse clinical outcomes have been observed (29). Indeed it is with these “gray zone” cases that clinicians most often need ancillary tests, and our data lend further support of this observation in that the pre-test LR for AHF in our intermediate group was 0.9, which approximates the LR of 1, suggesting clinical impression alone in this cohort holds little significance for the final diagnosis (14). Application of our model to the intermediate pre-test study group redirected close to one-half of patients, with a 99% accuracy. Referencing the McCullough et al. (15) landmark study analysis demonstrating 74% appropriate diagnosis using cut points in this patient cohort, it would appear use of the model would correctly diagnose one-quarter fewer of these patients but would also avoid inadvertently misdiagnosing the same amount. Clinical discretion would dictate the preferred approach to be used.

Incorporating a diagnostic test value into clinical assessment to create a diagnostic prediction rule for the presence or absence of a disease has shown to be of clinical benefit in other disease states (30); however, few studies on AHF have used this type of analysis. Baggish et al. (16) analyzed a

### Table 3 Comparison of Study Patient Populations

<table>
<thead>
<tr>
<th>Study Population</th>
<th>IMPROVE-CHF</th>
<th>PRIDE</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yrs (SD)</td>
<td>70.2 (14.3)</td>
<td>62.5 (17.1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pre-test probability</td>
<td>0.10</td>
<td>0.10</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>First quartile</td>
<td>0.10</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>0.40</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Third quartile</td>
<td>0.80</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>0.999</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>NT-proBNP (pg/ml)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>First quartile</td>
<td>206</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>1,299</td>
<td>461.4</td>
<td></td>
</tr>
<tr>
<td>Third quartile</td>
<td>4,219.5</td>
<td>2,643.5</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>35,585</td>
<td>117,390</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4 Model Versus Clinical Impression for Diagnosing Acute HF Based on External Pre-Test Risk Group

<table>
<thead>
<tr>
<th>Pre-Test Risk Group</th>
<th>Post-Test Risk Group</th>
<th>Adjudicated No HF</th>
<th>Adjudicated HF</th>
<th>Percent Appropriately Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP, n = 343 (accuracy = 89%)</td>
<td>LP (n = 282)</td>
<td>276</td>
<td>6</td>
<td>(2.1)</td>
</tr>
<tr>
<td></td>
<td>IP (n = 58)</td>
<td>30</td>
<td>28</td>
<td>48.3</td>
</tr>
<tr>
<td></td>
<td>HP (n = 3)</td>
<td>0</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>IP, n = 139</td>
<td>LP (n = 38)</td>
<td>37</td>
<td>1</td>
<td>97.3</td>
</tr>
<tr>
<td></td>
<td>IP (n = 77)</td>
<td>25</td>
<td>52</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>HP (n = 24)</td>
<td>0</td>
<td>24</td>
<td>100</td>
</tr>
<tr>
<td>HP, n = 91 (accuracy = 95%)</td>
<td>LP (n = 0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>IP (n = 18)</td>
<td>4</td>
<td>14</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>HP (n = 73)</td>
<td>1</td>
<td>72</td>
<td>(1.4)</td>
</tr>
</tbody>
</table>

*Inappropriately not reclassified (i.e., falsely not redirected). HF = heart failure; HP = high probability; IP = intermediate probability; LP = low probability.
study of 599 dyspneic patients to create an effective scoring tool for diagnosing AHF that required the integration of 7 clinical factors with the dichotomous NT-proBNP value and have recently published the derivation of this and a dyspnea risk score into an electronic format for personal digital assistant-type devices (31). These data and ours both support the concept that NT-proBNP values should be combined with clinical variables into a prediction calculation to be most useful to clinicians.

We believe our study makes intuitive sense to those clinicians who have experience in the use of BNP or NT-proBNP. Although discrete cut points allow for ease of use, they are often derived from clinical trials that may not have a typical cross section of those patients who are tested in “real life” situations, which leads to LRs that are less statistically robust, and aggregate clinical experience with natriuretic peptides would indicate that in fact many diagnoses, including acute coronary syndrome or pulmonary embolism, may result in elevations of these markers into the range that is consistent with AHF. Accordingly clinicians often approach markers such as BNP and NT-proBNP as continuous variables, recognizing that greater values are more consistent with AHF (28); this approach has been recommended as superior by recent consensus guidelines (32) but until now had not been analyzed in a rigorous fashion. By analyzing NT-proBNP as a continuous variable, we found that the model supplants this unstudied approach.

The incorporation of clinical impression as part of a prediction model in this study reaffirms most clinicians’ belief that their impressions should directly contribute toward any prediction rule (33,34). The model also does not require elaborate clinical information and is therefore relatively simple to use. Thus, for many undifferentiated dyspneic cases the model appears to quickly and reliably redirect the undecided clinician for diagnosing AHF. It could ultimately lead to improved health outcomes and streamlined research in this very challenging patient population.

A recent study (29) has shown that clinicians may be inaccurate for ruling in or ruling out AHF despite high levels of certainty and that the addition of biomarker testing in this setting may improve judgment. The application of our model on the external database showed excellent appropriate reclassification in the intermediate-probability clinical groups. In the other risk groups the model showed modest appropriate reclassification; it should be emphasized that in situations of reclassification, clinical judgment is necessary to correctly interpret the results. Future studies in various settings will clarify the model utility across all levels of clinical certainty.

**Study limitations.** The model uses the clinician assessment of pre-test probability, which makes it potentially vulnerable to interobserver variability for any given case. There may be differences across institutions as well, such as between academic and nonacademic settings or those with differing AHF rates. A study (35) of another disease shows greater agreement rates than expected, and standardized explicit data forms (36) have demonstrated more inter-rater disagreement than anticipated. As well, with our approach there is less reason for “clinician override” of the model as can occur with other rules, because clinical impression is weighted heavily in the model calculation. Future studies are needed to prospectively analyze inter-rater error for AHF pre-test probability (37). It is possible that more objective variables could replace the clinical estimate. Although AHF can accompany acute coronary syndromes, caution should be exercised when extrapolating results to patients with abnormal troponin values because these patients were outside of the spectrum of patients studied. We omitted age stratification in our analysis of standard NT-proBNP cut points as the numbers of study patients <50 years and >75 years of age were too few to draw meaningful conclusions; however, age as a significant variable for AHF was incorporated into the model. Another issue relates to the external validation of the model, which while confirming excellent model discrimination (c = 0.97), did show some unreliability in that the model tended to underestimate the true probability of AHF. As such, the model’s rate of appropriate redirection calculated on the external data is likely conservative.

Although we did not analyze BNP in this study, BNP suffers from the same issues regarding cut points as NT-proBNP (38). Although currently not generalizable to all settings, the fact that the 2 study cohorts were from different countries and so different (Table 3) suggests the model may perform well in other patient populations. Finally, the complex mathematics of the model formula can be seen as a deterrent for its use; methods for optimal implementation need to be studied including websites, PDA devices and hospital laboratory reporting systems.

**Conclusions**

A diagnostic prediction model for AHF that uses clinical assessment and NT-proBNP value has been derived and externally validated to appropriately direct the physician in a significant number of indeterminate cases. Further studies of implementation, cost, and impact analyses will help define the model’s general utility across all levels of clinical certainty and may foster similar analyses of other categorical diagnostic tests.

**Acknowledgment**

The authors thank Ms. Carolyn Ziegler, MA, MISt, Information Specialist, Health Sciences Library, St. Michael’s Hospital, for her invaluable assistance in preparing the paper.

**Reprint requests and correspondence:** Dr. Brian Steinhart, 1-008c Shuter Wing, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada. E-mail: steinhartb@smh.toronto.on.ca.
REFERENCES

1. Tsuyuki RT, Shibata MC, Nilsson C, Hervas-Malo M. Contempo-

32. 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.

Key Words: acute heart failure diagnosis natriuretic peptide prediction model Bayesian theorem.

APPENDIX
For the prediction model, please see the online version of this article.