Population Risk Prediction Models for Incident Heart Failure

A Systematic Review

Justin B. Echouffo-Tcheugui, MD; Stephen J. Greene, MD; Lampros Papadimitriou, MD; Faiez Zannad, MD, PhD; Clyde W. Yancy, MD, MSc; Mihai Gheorghiade, MD; Javed Butler, MD, MPH

Background—The prevalence of heart failure is expected to significantly rise unless high-risk patients are effectively screened and appropriate, cost-effective prevention interventions are implemented.

Methods and Results—We performed a systematic review to evaluate the prediction characteristics of the published heart failure risk prediction models as of August 2014 using MEDLINE and EMBASE databases. Eligible studies reported the development, validation, or impact assessment of a model. Two investigators performed independent review to extract data on study design and characteristics, risk predictors, discrimination, calibration, and reclassification ability of models, as well as validation and impact analysis. We included 13 publications reporting on 28 heart failure risk prediction models. Models had acceptable-to-good discriminatory ability (c-statistics, >0.70) in the derivation sample. Calibration was less commonly assessed, but was acceptable when it was. Only 2 models were externally validated more than once, displaying modest-to-acceptable discrimination (c-statistics, 0.61–0.79). When assessed, novel blood and imaging markers modestly improved risk prediction. One model assessed the prediction properties in race-based subgroups, whereas 2 models evaluated sex-based subgroups. Impact analysis found none of the models recommended for use in any clinical practice guideline.

Conclusions—Incident heart failure risk prediction remains at an early stage. The discrimination ability of current models is acceptable in derivation data sets but most models have not been externally validated. It remains unclear which models are cost-effective and best suit population screening needs. The effects of models on clinical and preventative care requires further study.

Key Words: calibration ■ discrimination ■ heart failure ■ prevention

Heart failure (HF) affects >5 million Americans.¹ Unless effective preventive interventions are implemented, it is estimated that by 2030, HF prevalence will grow by 25% and annual costs of care from $21 to $53 billion.² The prognosis of patients with HF remains poor with a projected 50% 5-year mortality, underscoring the importance of prevention.³ Prevention interventions, however, are unlikely to be cost-effective unless targeted toward reliably identified high-risk populations. Because of the epidemiological paradox, more HF cases occur in patients with the absence of any given individual risk factor, which makes it important to develop a more comprehensive risk assessment scheme than to rely solely on individual risk factors. Unfortunately, the best approach for quantifying overall incident HF risk in individuals remains unclear. In this study, we sought to systematically assess the performance characteristics and the use of the available HF risk models for predicting incident HF in the general population.

Clinical Perspective on p 447

Methods

Data Sources

We searched PubMed and EMBASE databases from January 1990 to August 2014 for HF risk prediction models published in English language. We used search terms related to HF and disease prediction (Appendix A in the Data Supplement) and reviewed the reference lists of studies and identified their citations through the ISI Web of Science for additional data. Eligible studies had to report a risk assessment model in a cohort study either by an equation or by a score. Two evaluators (J.B.E. and J.B.) independently identified articles and screened them for the inclusion criteria (Figure) by reviewing text and the Data Supplement. Disagreements were solved by consensus. We excluded studies for models developed in specific...
population (eg, postoperative patients, patients already diagnosed with HF, studies that reported measures of risk without information on model characteristics, and simulation studies). We extracted information on the area under the receiver-operating characteristic curve (AUC) or c-statistic, reclassification percentage, net reclassification improvement (NRI), and integrated discrimination improvement index (IDI).4

Bias, Clinical Usefulness, and External Validity
Risk for bias in models was assessed per Hayden et al5 recommendations, including 7 categories, such as participation (sampling bias), attrition (attrition bias), prognostic factor selection, prognostic factor measurement, outcome measurement (ascertainment bias), statistical analysis, and reporting model performance. Extent and methods of dealing with missing data, (Table in the Data Supplement), appropriateness of statistical methods, and clinical usefulness as a combination of clinical use and usability were assessed. Clinical use was assessed as whether predictive models proposed clinical scenarios or decision nodes linked to a risk category or threshold where the model may be useful for guiding diagnostic or therapeutic interventions. For usability, we noted whether a calculator or nomogram that enables easier knowledge translation was added. To examine generalizability and external validity, we evaluated whether models had been validated in an independent population in the original or a subsequent publication.

Data Extraction and Quality Assessment
Two reviewers (J.B.E. and J.B.) performed quality assessment and extracted information about design, setting, population characteristics, number of patients in the derivation and validation cohorts, number of outcomes of interest, number of candidate predictor variables, and the type of regression model used. For predictors, we specifically assessed for age, sex, hypertension, left ventricular hypertrophy, myocardial infarction, diabetes mellitus, valvular disease, and overweight/obesity, based on previously reported association of these factors with HF. For discriminative ability, information on the c-statistic, and for calibration, the difference between the observed and predicted rates, as well as the P value of the corresponding test statistic, were extracted. Data relevant to reclassification included NRI and IDI values, as well as their respective 95% confidence intervals and P values, when available. Given the heterogeneity in the type and number of metrics, the number and nature of risk factors included, and the study designs used, our analysis consisted mainly of a narrative synthesis of the evidence.

Effect of HF Risk Models
Effect of studies was identified through scanning the references of all publications and applying search strategies for impact studies proposed by Reilly and Evans,6 which combines the model’s acronym or name of the cohort or first author with specific search term combination (Appendix B in the Data Supplement). We searched relevant practice guidelines to assess the implementation of HF prediction models in countries or regions, in which such models were developed. We also targeted English-language guidelines compiled by a selection of professional organizations known to regularly issue recommendations for the management of HF, including the American College of Cardiology,7 the American Heart Association,7 Heart Failure Society of America,8 and the European Society of Cardiology.9
### Table 1. Overview of HF Risk Prediction Model

<table>
<thead>
<tr>
<th>Reference and Cohort</th>
<th>Population</th>
<th>Risk Factors</th>
<th>n</th>
<th>Outcome Follow-Up, y</th>
<th>Definition</th>
<th>Discrimination</th>
<th>Calibration</th>
<th>Validation</th>
<th>Reclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kannel et al10</td>
<td>US: whites</td>
<td>Age, LVH, HR, SBP, vital capacity, CHD, valve disease, DM, cardiomegaly</td>
<td>15267</td>
<td>486</td>
<td>4</td>
<td>Framingham criteria</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Butler et al11</td>
<td>US: whites and blacks</td>
<td>Age, CHD, smoking, SBP, HR, glucose, creatinine, albumin, and LVH</td>
<td>2935</td>
<td>258</td>
<td>4–9</td>
<td>Physician diagnosis and 0.73 (0.72 after treatment for HF</td>
<td>HL $\chi^2 = 6.24$ (P=0.621)</td>
<td>Apparent</td>
<td>Bootstrapping</td>
</tr>
<tr>
<td>Goyal et al12</td>
<td>US: mixed but mainly whites</td>
<td>Age, CHD, HTN, DM, AF, and valvular disease</td>
<td>359947</td>
<td>4001</td>
<td>9.5</td>
<td>ICD-9 code defined HF</td>
<td>Women: 0.886 and men: 0.877</td>
<td>NR</td>
<td>Apparent</td>
</tr>
<tr>
<td>Velagaleti et al13</td>
<td>US: mainly whites</td>
<td>Age, CHD, HTN, DM, CAD, AF, and valvular disease</td>
<td>2754</td>
<td>95</td>
<td>9.4</td>
<td>Framingham criteria</td>
<td>0.84 (0.838 after jacknife)</td>
<td>NR</td>
<td>Apparent</td>
</tr>
<tr>
<td>Velagaleti et al13</td>
<td>US: mainly whites</td>
<td>Age, HTN, DM, CAD, AF, valvular disease, BNP</td>
<td>2754</td>
<td>95</td>
<td>9.4</td>
<td>Framingham criteria</td>
<td>0.85 (0.862 after jacknife)</td>
<td>NR</td>
<td>Apparent</td>
</tr>
<tr>
<td>Velagaleti et al13</td>
<td>US: mainly whites</td>
<td>Age, HTN, DM, CAD, AF, BNP, albumin/creatinine ratio, valve disease</td>
<td>2754</td>
<td>95</td>
<td>9.4</td>
<td>Framingham criteria</td>
<td>0.86 (0.862 after jacknife)</td>
<td>NR</td>
<td>Apparent</td>
</tr>
<tr>
<td>deFilippi, et al14</td>
<td>US: mainly whites</td>
<td>Age, sex, race, SBP, ECG, glucose, CHD, smoking, creatinine, albumin, HR</td>
<td>4221</td>
<td>1279</td>
<td></td>
<td>Diagnosis+symptoms/signs+HF therapy</td>
<td>0.752</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>deFilippi, et al14</td>
<td>US: mainly whites</td>
<td>Age, sex, race, SBP, ECG, glucose, CHD, smoking, creatinine, albumin, HR</td>
<td>4221</td>
<td>1279</td>
<td></td>
<td>Diagnosis+symptoms/signs+HF therapy</td>
<td>0.767</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al15</td>
<td>Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI antihypertensives</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.815</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al15</td>
<td>Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI, antihypertensives, MR-proANP</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.824</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al15</td>
<td>Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI, antihypertensives, NT-proBNP</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.837</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al15</td>
<td>Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI, antihypertensives, CRP</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.823</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al15</td>
<td>Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI, antihypertensives, MR-proANP, NT-proBNP</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.838</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al15</td>
<td>Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI, antihypertensives NT-proBNP, CRP</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.842</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Reference and Cohort</th>
<th>Population</th>
<th>Risk Factors</th>
<th>n</th>
<th>Follow-Up, y</th>
<th>Definition</th>
<th>Discrimination</th>
<th>Calibration</th>
<th>Validation</th>
<th>NRI, %</th>
<th>IDI, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Smith et al</strong>&lt;sup&gt;15&lt;/sup&gt; Sweden: mainly whites</td>
<td>Age, sex, SBP, DBP, LDL, MI, antihypertensives, MR-proANP, NT-proBNP, CRP</td>
<td>5187</td>
<td>112</td>
<td>14</td>
<td>ICD-8, 9, and 10 codes</td>
<td>0.842</td>
<td>NR</td>
<td>NR</td>
<td>22% (P&lt;0.001)</td>
<td>0.05 (P&lt;0.001)</td>
</tr>
<tr>
<td><strong>Kalogeropoulos et al</strong>&lt;sup&gt;14&lt;/sup&gt; Health ABC score 2 US: whites and blacks</td>
<td>Age, CHD, smoking, SBP, HR, glucose, LVH, creatinine, albumin, IL-6</td>
<td>2610</td>
<td>311</td>
<td>9.4</td>
<td>Diagnosis+clinical or echo findings+medical therapy</td>
<td>0.734 (0.706–0.763)</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>Kalogeropoulos et al</strong>&lt;sup&gt;14&lt;/sup&gt; Health ABC score 2 US: whites and blacks</td>
<td>Age, CHD, smoking, SBP, HR, glucose, LVH, creatinine, albumin, IL-6, TNF-α, CRP</td>
<td>2610</td>
<td>311</td>
<td>9.4</td>
<td>Diagnosis+clinical or echo findings+medical therapy</td>
<td>0.737 (0.709–0.765)</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>Pfister et al</strong>&lt;sup&gt;17&lt;/sup&gt; PR Oactive trial score 19 European countries: whites</td>
<td>Age, CKD, diuretic use, Hba1c, DM, LDL, HR, R/LBBB, pioglitazone, microalbuminuria</td>
<td>4951</td>
<td>233</td>
<td>2.9</td>
<td>Medical record diagnosis</td>
<td>0.75 (0.751 after bootstrapping correction)</td>
<td>NR</td>
<td>Apparent</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>Kalogeropoulos et al</strong>&lt;sup&gt;14&lt;/sup&gt; CHS US: mixed but 86.5% whites</td>
<td>Age, CHD, smoking, SBP, HR, glucose, creatinine, albumin LVH, echo score (EF, LA, and E/A) and NT-proBNP</td>
<td>3752</td>
<td>286</td>
<td>5</td>
<td>Diagnosis+clinical or echo findings+medical therapy</td>
<td>0.777</td>
<td>NR</td>
<td>NR</td>
<td>10.6%</td>
<td>0.044</td>
</tr>
<tr>
<td><strong>Kalogeropoulos et al</strong>&lt;sup&gt;14&lt;/sup&gt; CHS US: mixed but 86.5% whites</td>
<td>Age, CHD, smoking, SBP, HR, glucose, creatinine, albumin LVH, echo score (EF, LA, and E/A) and NT-proBNP</td>
<td>3752</td>
<td>286</td>
<td>5</td>
<td>Diagnosis+clinical or echo findings+medical therapy</td>
<td>0.773</td>
<td>NR</td>
<td>NR</td>
<td>11.3%</td>
<td>0.045</td>
</tr>
<tr>
<td><strong>Kalogeropoulos et al</strong>&lt;sup&gt;14&lt;/sup&gt; CHS US: mixed but 86.5% whites</td>
<td>Age, CHD, smoking, SBP, HR, glucose, creatinine, albumin LVH, echo score (EF, LA, and E/A) and NT-proBNP</td>
<td>3752</td>
<td>286</td>
<td>5</td>
<td>Diagnosis+clinical or echo findings+medical therapy</td>
<td>0.789</td>
<td>NR</td>
<td>NR</td>
<td>16.3%</td>
<td>0.064</td>
</tr>
<tr>
<td><strong>Wang et al</strong>&lt;sup&gt;19&lt;/sup&gt; FHS US: mainly whites</td>
<td>Age, sex, BMI, SBP, HTN, DM, smoking, total and HDL-cholesterol, AF, CVD, EKG LVH, murmur</td>
<td>3428</td>
<td>162</td>
<td>11.3</td>
<td>Framingham criteria</td>
<td>0.846</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>Wang et al</strong>&lt;sup&gt;19&lt;/sup&gt; FHS US: mainly whites</td>
<td>Age, sex, BMI, SBP, HTN, DM, smoking, total and HDL-cholesterol, AF, CVD, EKG LVH, murmur, and sST2, GDF-15, hsTnT, hsCRP, and BNP</td>
<td>3428</td>
<td>162</td>
<td>11.3</td>
<td>Framingham criteria</td>
<td>0.870</td>
<td>NR</td>
<td>NR</td>
<td>0.39 (0.21–0.57)</td>
<td>0.04 (0.02–0.06)</td>
</tr>
<tr>
<td><strong>Agarwal et al</strong>&lt;sup&gt;20&lt;/sup&gt; ARIC US: biethnic—whites and blacks</td>
<td>Age, race, sex, prevalent CHD, SBP, antihypertensives, DM, smoking, HR, BMI</td>
<td>13555</td>
<td>1487</td>
<td>10</td>
<td>HF hospitalization or ICD-9 code on the death certificate</td>
<td>NR</td>
<td>0.7937 after bootstrapping</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>Agarwal et al</strong>&lt;sup&gt;20&lt;/sup&gt; ARIC US: whites and blacks</td>
<td>Age, race, sex, NT-proBNP</td>
<td>13555</td>
<td>1487</td>
<td>10</td>
<td>HF admission or ICD-9 code on death certificate</td>
<td>NR</td>
<td>0.745</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>Schnabel et al</strong>&lt;sup&gt;21&lt;/sup&gt; FHS US: whites</td>
<td>Age, BMI, EKG LVH, DM, MI significant murmur</td>
<td>725</td>
<td>161</td>
<td>10</td>
<td>Framingham criteria</td>
<td>0.71 (0.67–0.75)</td>
<td>HL χ²=7.29, (P=0.61)</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

(Continued)
### Results

#### Prediction Models

We screened 413 abstracts, retrieved 48 citations for full-text review, and selected 13 studies that described 28 HF prediction risk models (Figure). The characteristics of the risk prediction models are presented in Table 1.

#### Populations, Risk Factors, and Outcomes

Eleven models were developed from US cohorts, 1 in Sweden, 1 in Britain, and 2 multicenter cohorts (19 European countries in 1 study\(^\text{17}\) and 38 countries from America, Europe, Asia, and Africa in another study\(^\text{22}\)). Models were mostly developed in whites; the models based on US populations included non-white participants, mainly blacks. The number of participants ranged from 725 to 359,947 and 18 to 99 years of age. The length of follow-up ranged from 2.9 to 14 years. Three models were developed in specific populations, including diabetes mellitus,\(^\text{17}\) impaired glucose tolerance,\(^\text{22}\) and atrial fibrillation.\(^\text{23}\) The method of HF identification varied, but was mainly based on clinical presentation, hospital discharge records, and cause of death. One study defined HF using outpatient records in addition to hospital and death discharge records, and cause of death. One study published the numbers of variables tested for inclusion.\(^\text{22}\) Thus for most studies, the number of outcomes of interest per candidate variable could not be assessed. The predictors most commonly included in the final models were age, sex, body mass index, diabetes mellitus, systolic blood pressure, creatinine, serum albumin or total protein, LV hypertrophy, and coronary artery disease. Biomarkers were evaluated in several studies, including B-type natriuretic peptide (BNP) or N-terminal pro-BNP (NT-proBNP), midregional proatrial natriuretic peptide, interleukin (IL)-6, tumor necrosis factor (TNF-α), C-reactive protein (CRP), soluble IL-2 receptor, growth differentiation factor-15, high-sensitivity troponin I (hsTnI), and urinary albumin/creatinine ratio.

All the models were developed using a Cox regression model with reporting of the model with \(\beta\) coefficients, with 2 studies providing additional point-based scoring systems.\(^\text{10,11}\)

#### Performance of the Models

Table 1 displays the model performance. The AUC ranged from 0.71 to 0.87, indicating a modest-to-good discrimination. Risk scores were internally validated through bootstrapping in \(\text{4,11,17,20,22}\) and jackknife in 1 study.\(^\text{13}\) Five models had calibration estimated with the Hosmer–Lemeshow test. Most studies did not test for sex-based\(^\text{10–13,16–18,21,22}\) or race-based\(^\text{10–13,15–19,21,22}\) specific performance, except the Health Aging, and Body Composition (ABC) study HF score that predicted HF risk similarly in both sex and white/black race in development cohort,\(^\text{11}\) but the risk was underestimated in white men, especially in the 5% to 10% risk category, in the validation cohort.\(^\text{24}\) The AUCs were 0.71, 0.75, 0.74, and 0.72 in white men, white women, black men, and black women, respectively.\(^\text{24}\)

#### Model Improvement

In the Framingham Study, BNP and the urinary albumin/creatinine ratio increased the AUC from 0.84 to 0.86\(^\text{10}\) and 13% of subjects were appropriately reclassified. In the Health ABC study, the addition of IL-6, TNF-α, and CRP increased AUC from 0.717 to 0.737, mainly because of IL-6 that improved AUC to 0.734 (\(P<0.001\)).\(^\text{10}\) In the Cardiovascular Health Study, NT-proBNP increased AUC from 0.667 to 0.719 (\(P<0.001\)), but discrimination remained moderate.\(^\text{25}\) In the Malmö Diet and Cancer study, midregional proatrial...
natriuretic peptide, NT-proBNP, and CRP together improved AUC from 0.815 to 0.842 with an IDI of 0.05 (P<0.001) and an NRI of 22% (P<0.001). In the Atherosclerosis Risk in Communities study, adding NT-proBNP increased the AUC by 0.032, NRI by 13%, and IDI by 0.057; NRI was higher in women (17.9%) than in men (14.4%).

Echocardiographic score (left ventricular ejection fraction and mass, E/A ratio, and left atrium), and NT-proBNP improved AUC (echocardiography, 0.031; NT-proBNP, 0.027; combined, 0.043; all P<0.01), and yielded robust IDI (echocardiography, 43.3%; NT-proBNP, 42.2%; combined, 61.7%; all P<0.001), and NRI (echocardiography, 11.3%; NT-proBNP, 10.6%; combined, 16.3%; all P<0.01), when added to the Health ABC model. Participants at intermediate risk (5%–20% 5-year risk; 35.7% of the cohort) derived the most reclassification benefit. Echocardiography yielded modest reclassification when used sequentially after NT-proBNP.

A multimarker score with soluble ST2, growth differentiation factor-15, hsTnI, hsCRP, and BNP increased AUC (0.846–0.870), NRI (0.39), and IDI (0.04) in the Framingham study. In the Atherosclerosis Risk in Communities study, cardiac troponin T and NT-proBNP improved AUC by 0.040 and 0.057; and NRI by 50.7% and 54.7% in women and in men, respectively. In the Multi-Ethnic Study of Atherosclerosis cohort, and NRIs by 50.7% and 54.7% in women and in men, respectively. In the Cardiovascular Health Study, a comparison of the Health ABC score, NT-proBNP, and echocardiographic data showed an NRI of 22% (P<0.001) and an AUC of 0.765 to 0.770. With growth differentiation factor-15 and hsTnI, the NRI was 0.212 (P<0.001). In the Framingham cohort, a composite of LV hypertrophy and LV dysfunction increased AUC from 0.765 to 0.770. With growth differentiation factor-15 and hsTnI, the NRI was 0.212 (P<0.0001). In the British Heart Regional Study, the addition of NT-proBNP to the Health ABC model improved AUC from 0.706 to 0.764 (P<0.0001).

In the Prevention of Vascular and Renal End stage Disease (PREVEND) study, NT-proBNP, midregional proatrial natriuretic peptide, hsTnT, hsCRP, cystatin-C and urinary albumin excretion individually modestly improved the AUC of 0.826 by 4.7, 2.2, 2.5, 1.9, 1.0, and 1.6%, respectively.

External Validity
Five models were externally validated (Table 2). Of these, the Framingham and Health ABC models were generally lower and with a poorer calibration, although the latter was not assessed in most of validation studies.

Single Biomarker or Imaging Data
In the Cardiovascular Health Study, a comparison of the Health ABC score, NT-proBNP, and echocardiographic data for detecting LV dysfunction indicated >5% risk of HF at 5 years in 44% of the cohort and identified 76% of those who developed HF by 5 years. An abnormal NT-proBNP was present in 29.5% of the cohort and identified 45% of those who developed HF by 10 years; and baseline echocardiography identified reduced ejection fraction in 7.2% of the population, but identified only 12% of those who developed HF by 10 years.

Depending on the method of indexing LV mass (body surface area, fat-free mass, or height), the AUC for the Health ABC HF score (0.73, 0.75, and 0.64, respectively) was greater than that for BNP (0.62, 0.70, and 0.57, respectively) and NT-proBNP (0.62, 0.69, and 0.56, respectively; P<0.01 for all). Similar results were seen in the Framingham score except when LV mass was indexed to fat-free mass (0.65, 0.69 and 0.66, and 0.64, respectively). Addition of BNP resulted in an improvement in discrimination of stage B HF (difference in AUC, 0.01–0.03; P<0.05 for all). In an elderly at high risk, NT-proBNP (AUC, 0.74) was superior to the Framingham HF score (AUC, 0.63) in predicting asymptomatic LV systolic dysfunction.

Model Effect
There were no recommendations for using any risk prediction model to estimate the risk of incident HF in either the clinical or the community settings in any of the existing guidelines. There were no studies assessing the effect of adopting HF risk models in clinical practice on the process of care and outcomes of patients.

Discussion
This systematic review demonstrates the feasibility of global HF risk prediction in the general population. Compared with coronary artery disease and stroke, prediction of incident HF risk is at an early stage and is not incorporated into contemporary clinical or public health practice. Defining risk can be reliably done using a combination of common variables with an acceptable-to-good discrimination, although correcting for overfitting (internal validation) or testing in a new population (external validation) may attenuate discrimination. Implementation of prediction models may identify a wider segment of at-risk population compared with single risk factors. In addition, biomarkers may be used to further categorize patients into different risk strata in concert with the clinical variables.

HF guidelines advocate prevention of progression from at-risk HF stages (A and B) to symptomatic stages (C and D), thus requiring identification of susceptible patients likely to benefit from interventions, including lifestyle changes or more intense risk factor modification. Risk communication to patients may motivate adherence; indeed healthy lifestyle is associated with a lower lifetime risk of HF. Furthermore, risk scores can be incorporated into electronic medical records, allowing continual passive screening of high-risk subjects with alerts for patients entering higher-risk strata. These risk models may also circumvent uncertainties surrounding the stage B HF definition, which is currently dictated by what can be measured with imaging rather than a complete understanding of risk. In stepwise screening with models as the first step followed by imaging or biomarkers, those with stage B HF may be more reliably referred for interventions known to halt and possibly reverse structural changes that precede clinical HF. With increasing HF risk, LV mass and concentric remodeling increases, supporting the traditional belief that HF risk correlates with structural heart disease. HF models may be useful in assessing the use of novel biomarkers for risk prediction or for selecting patients for inclusion in clinical trials.
Indeed, investigating novel markers is warranted, especially given the relatively modest additional discriminative ability of natriuretic peptides (BNP or NT-proBNP) beyond that of clinical variables. This finding is not surprising because factors included in models (eg, plasma creatinine, female sex, and age) may increase these biomarkers independently of left ventricular ejection fraction and filling pressures.33 Moreover, several screening studies have found use of natriuretic peptides (BNP or NT-proBNP) beyond that of these characteristics to be assessed in concert. Attempts to update models by extra variables have focused on discrimination, providing an idea about their potential performance when applied to a new population. Calibration was not always assessed. A poorly calibrated model may have the same discrimination as a well-calibrated model, necessitating both of these validations of existing models are needed to guarantee their performance.

<table>
<thead>
<tr>
<th>Author, Reference</th>
<th>Validated Risk Score</th>
<th>Validation Population / Country</th>
<th>Ethnicity</th>
<th>Design</th>
<th>Sample Size</th>
<th>Age, y</th>
<th>Time-Horizon, y</th>
<th>Discrimination</th>
<th>Calibration</th>
<th>NRI, % (95% CI)</th>
<th>IDI, % (95% CI)</th>
<th>Reclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butler et al11</td>
<td>ARIC/CHS cohort</td>
<td>US</td>
<td>Mixed</td>
<td>Prospective cohort</td>
<td>12038</td>
<td>51.2 (SD, 10.5)</td>
<td>0.610</td>
<td>NR</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Kalogeropoulos et al30</td>
<td>ARIC cohort/ USA</td>
<td>US</td>
<td>Mixed</td>
<td>Prospective cohort</td>
<td>4298 for NHANES and 21, 221 for ARIC/CHS</td>
<td>5 (ARIC/CHS cohort)</td>
<td>0.762</td>
<td>NR</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Gupta et al30</td>
<td>ARIC cohort/ USA</td>
<td>US</td>
<td>Mixed</td>
<td>Prospective cohort</td>
<td>2145 for ENRICHD and 3640 for VISP</td>
<td>NA</td>
<td>0.783</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Agarwal et al20</td>
<td>ARIC cohort/ USA</td>
<td>US</td>
<td>Mixed Mainly white</td>
<td>Prospective cohort</td>
<td>5168</td>
<td>4</td>
<td>0.785</td>
<td>NR</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Wannamethe et al22</td>
<td>ARIC cohort/ USA</td>
<td>US</td>
<td>Mixed</td>
<td>Prospective cohort</td>
<td>5168</td>
<td>Mean: 51.2 (SD, 10.5)</td>
<td>0.797</td>
<td>NR</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

ARIC indicates Atherosclerosis Risk in Communities Study; AUC, area under the curve; BSA, body surface area; CHS, Cardiovascular Health Study; CI, confidence interval; Health ABC, Health Aging, and Body Composition study; ENRICHD, Enhancing Recovery in Coronary Heart Disease Patients Trial; FFM, fat-free mass; HF, heart failure; HL, Hosmer–Lemeshow; IDI, Integrative Discriminative Index; NA, not applicable; NHANES, National Health and Nutrition Examination Survey; NR, not reported; NRI, Net Reclassification Index; TIHN, The Health Improvement Network; US, United States; and VISP, Vitamin Intervention for Stroke Prevention Trial.

Some of the existing models have been statistically underpowered, especially as consideration was not given to the ratio of outcomes to predictors. Models with an inappropriate ratio (<10 events per variable) are less robust, with a performance that shrinks when applied to other populations. Although the overall study size and the prognostic ability of individual predictors is important, a key driver of statistical power is the number of events. These performance metrics mostly originated from the derivation sample, which are optimistic. Some models were internally validated (bootstrap or jackknife validation), providing an idea about their potential performance when applied to a new population. Calibration was not always assessed. A poorly calibrated model may have the same discrimination as a well-calibrated model, necessitating both of these characteristics to be assessed in concert. Attempts to update models by extra variables have focused on discrimination, and none of the studies reported change in the calibration. Finally, validation is an important step before a model can be recommended for widespread use. This was limited to only 3 models in US populations. Hence, further validations of existing models are needed to guarantee their generalizability.

HF models have mainly been published as mathematical equations. Complex formulas may not be suitable for clinical application. Indeed, previous research suggests that many
clinicians find risk calculation too time consuming and are not convinced of the value of information derived from these models. Additional translation efforts are needed to convert predictions equations into user-friendly tools to improve uptake. Efforts are needed to derive the appropriate context for defining high-risk status to facilitate their use. It is also important to study whether implementing HF risk prediction affects the behavior of healthcare providers and improves clinical outcomes. Few studies examined the incremental predictive value of novel biomarkers (eg, collagen matrix and omic markers) that may be useful for risk stratification. Clinical risk models provide good basis to assess the incremental value of new discoveries.

Further limitations to existing models exist. Misclassification of HF events may have occurred as there are no universally agreed specific criteria for diagnosing HF, thus affecting the performance of the models. HF is a heterogeneous syndrome, representing a spectrum of pathogenesis, as well as normal to depressed ejection fraction. Existing studies use multiple criteria to define HF with little agreement. Inconsistencies were indeed observed in the presently reviewed studies, which may have affected model performance. In most studies, diagnosis did not rely on imaging. Consequently, specific forms of HF, such as HF with reduced ejection fraction and HF with preserved ejection fraction, were not considered. Although there is an overlap between predictors of HF with reduced ejection fraction and HF with preserved ejection fraction, there are differences that merit consideration. Moreover, HF with reduced ejection fraction and HF with preserved ejection fraction are regarded as distinct pathophysiological entities, as exemplified by the lack of evidence-based disease modifying therapies for the latter. Accordingly, it stands to reason that differing preventative strategies and interventions may exist. Furthermore, the majority of participants included were whites. However, HF risk is known to be heightened in specific groups (eg, blacks), thus supporting the need to incorporate more patients of different ethnic backgrounds.

Although existing models may lead to implementation of prevention interventions, such efforts may be premature, particularly in the absence of external validation studies and impact analysis. The lack of studies evaluating the effect of model use on outcomes should be addressed in future studies. Such an undertaking could potentially strengthen the case for their use in routine clinical or public health practice. The absence of existing studies may simply be an indication of the embryonic nature of the HF prediction field, and of HF prevention in general. This limitation is not specific to HF and does not in itself invalidate the future use of the models. Indeed, even with coronary heart disease, for which prediction models have been in existence because the early 1990s, only a limited number of studies have assessed the effect of their implementation, although their use has been consistently recommended in guidelines. We did not explicitly rank or categorize the quality of existing HF models, as there is no agreed-on scientific system for rating risk prediction model quality. Some argue minimizing risk for potential bias is of critical importance, whereas others contend that a risk score should be judged on ability to perform accurately across diverse settings. Finally, our ability to assess publication bias was limited.

In conclusion, risk models for predicting incident HF have acceptable-to-good discriminative ability. These tools need to be better calibrated and externally validated, and the effect of their use on outcomes assessed before incorporation into clinical practice guidelines. With advancement of our understanding of systemic conditions and cardiac and noncardiac structural and functional abnormalities that predispose to HF risk, better approaches for HF prevention are needed. HF prevention extends beyond simply addressing risk factors or precursors, such as diabetes mellitus, hypertension, and coronary artery disease, especially as the latter is not a sine qua non for the development of HF. HF scores may be viewed as a means for selecting a targeted population that would undergo imaging for detection of asymptomatic structural disease (eg, left ventricular systolic dysfunction), a potential therapeutic target in itself. Successful prevention approaches mandate the correct identification of the target population, underscoring the importance of further research in the performance and use of incident HF risk models.

Disclosures
Dr Gheorghieade has been a consultant for Abbott Laboratories, Astellas, AstraZeneca, Bayer HealthCare AG, CorThera, Cytokinetiks, DebioPharm S.A., ErrekaTera Therapeutici, GlaxoSmithKline, Ikaria, Johnson & Johnson, Medtronic, Merck, Novartis Pharma AG, Otsuka Pharmaceuticals, Palatin Technologies, Percor Therapeutics, Protein Design Laboratories, Sanofi-Aventis, Sigma Tau, Solvay Pharmaceuticals, Takeda Pharmaceutical and Trevena Therapeutics. The remaining authors report no conflicts.

References


Risk prediction models for incident heart failure (HF) in asymptomatic populations have the potential to improve risk stratification and augment the practitioner’s ability to prevent HF or to influence its natural history in several ways. These might include communicating risk to patients, offering appropriate patient education, proactive risk factor management, and timely use of imaging to identify asymptomatic structural heart disease for therapeutic intervention, with consequent referral to a cardiologist. Incident HF risk models may also be useful for selecting high-risk patients for inclusion in HF prevention trials or in trials assessing the benefits of therapies for patients with asymptomatic structural heart disease. However, incident HF risk prediction using combinations of clinical and biological variables remains at an early stage in development. Although the discriminative ability of current incident HF prediction models is acceptable, in general, these tools need to be better calibrated and externally validated before incorporation into clinical practice guidelines. It remains unclear which models are cost-effective and best suit population screening needs. The effects of models on clinical and preventative care require further study.
Population Risk Prediction Models for Incident Heart Failure: A Systematic Review
Justin B. Echouffo-Tcheugui, Stephen J. Greene, Lampros Papadimitriou, Faiez Zannad, Clyde W. Yancy, Mihai Gheorghiade and Javed Butler

Circ Heart Fail. 2015;8:438-447; originally published online March 3, 2015;
doi: 10.1161/CIRCHEARTFAILURE.114.001896
Circulation: Heart Failure is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-3289. Online ISSN: 1941-3297

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circheartfailure.ahajournals.org/content/8/3/438

Data Supplement (unedited) at:
http://circheartfailure.ahajournals.org/content/suppl/2015/03/03/CIRCHEARTFAILURE.114.001896.DC1
http://circheartfailure.ahajournals.org/content/suppl/2016/12/26/CIRCHEARTFAILURE.114.001896.DC2

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Heart Failure can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Heart Failure is online at:
http://circheartfailure.ahajournals.org//subscriptions/
# Supplemental Material
## Supplemental Table. Reporting and management of missing data

<table>
<thead>
<tr>
<th>Author, Reference</th>
<th>Name of the risk score</th>
<th>Outcome(s)</th>
<th>Variables included in risk models</th>
<th>Management of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kannel et al, 1999(^1)</td>
<td>Framingham Heart Study failure risk score</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Butler et al, 2008(^2)</td>
<td>The Health ABC (Aging, and Body Composition) Heart Failure Score</td>
<td>Reported</td>
<td>82 had missing data on HF status at baseline; these participants were excluded</td>
<td>Reported</td>
</tr>
<tr>
<td>Goyal et al, 2010(^3)</td>
<td>Kaiser Permanente Georgia risk score</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Velagaleti et al, 2010(^3)</td>
<td>Framingham Heart Study Risk score 2</td>
<td>Reported</td>
<td>Participants were excluded</td>
<td>Reported</td>
</tr>
<tr>
<td>deFilippi, et al, 2010(^5)</td>
<td>Cardiovascular Health Study score 1</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Smith et al, 2010(^5)</td>
<td>The Malmö Diet and Cancer Study (MDCS) risk score 6</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Kalogeropoulos et al, 2010(^7)</td>
<td>Health ABC (Aging, and Body Composition) Study score 3</td>
<td>Reported</td>
<td>Missing values of covariates were imputed using the multiple imputation by chained equations procedure</td>
<td>Reported</td>
</tr>
<tr>
<td>Pfister et al, 2013(^8)</td>
<td>PRoactive trial score</td>
<td>NR</td>
<td>Participants excluded</td>
<td>NR</td>
</tr>
<tr>
<td>Kalogeropoulos et al, 2012(^9)</td>
<td>Cardiovascular Health Study</td>
<td>NR</td>
<td>Patients were excluded from the study</td>
<td>NR</td>
</tr>
<tr>
<td>Study</td>
<td>Study Type</td>
<td>Patients Excluded</td>
<td>Study Type</td>
<td>Patients Excluded</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Wang et al, 2012&lt;sup&gt;10&lt;/sup&gt;</td>
<td>Framingham Heart Study</td>
<td>Reported</td>
<td>Framingham Heart Study</td>
<td>Reported</td>
</tr>
<tr>
<td>Agarwal et al, 2012&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Atherosclerosis Risk in Communities (ARIC) Study cohort</td>
<td>NR</td>
<td>Atherosclerosis Risk in Communities (ARIC) Study cohort</td>
<td>NR</td>
</tr>
<tr>
<td>Schnabel et al, 2012&lt;sup&gt;12&lt;/sup&gt;</td>
<td>Framingham Heart Study</td>
<td>NR</td>
<td>Framingham Heart Study</td>
<td>NR</td>
</tr>
<tr>
<td>Wong et al, 2013&lt;sup&gt;13&lt;/sup&gt;</td>
<td>Nateglinide and Valsartan in Impaired Glucose Tolerance Outcomes Research Study (NAVIGATOR)</td>
<td>NR</td>
<td>Nateglinide and Valsartan in Impaired Glucose Tolerance Outcomes Research Study (NAVIGATOR)</td>
<td>NR</td>
</tr>
<tr>
<td>Wannamethe et al, 2014&lt;sup&gt;14&lt;/sup&gt;</td>
<td>British Heart Regional Study</td>
<td>NR</td>
<td>British Heart Regional Study</td>
<td>NR</td>
</tr>
</tbody>
</table>

NR: Not reported
### Database: PubMed (January 1990 until August 2014)

#1 "heart failure " OR "congestive heart failure" OR "cardiac insufficiency" OR "cardiac failure"
#2 "decompensation" OR "progression"
#3 #1 NOT #2
#4 (validat* OR predict*[tiab] OR rule*) OR (predict* AND (outcome* OR risk* OR model*)) OR ((history OR variable* OR criteria OR scor* OR characteristic* OR finding* OR factor*) AND (predict* OR model* OR decision* OR identif* OR prognos*)) OR (decision* AND (model* OR clinical* OR logistic models)) OR (prognostic AND (history OR variable* OR criteria OR scor* OR characteristic* OR finding* OR factor* OR model*)) OR "stratification" OR "ROC Curve"[MeSH] OR "discrimination" OR "discriminate" OR "c statistic" OR "area under the curve" OR "Calibration" OR "Indices" OR "algorithm" OR "Multivariable")
#5 "cohort" OR "observational" OR "prospective" OR "trial" OR "epidemiology"
#6 #3 AND #4 AND #5
#7 (Animals[MeSH] NOT Humans[MeSH])
#8 #6 NOT #7

Limits: age>19 years, English language

### Database: EMBASE (January 1990 until August 2014)
#1 "heart failure" OR "congestive heart failure" OR "cardiac insufficiency" OR "cardiac failure"
#2 "decompensation" OR "progression"
#3 #1 NOT #2
#4 (validat* OR predict* OR rule*) OR (predict* AND (outcome* OR risk* OR model*)) OR ((history OR variable* OR criteria OR scor* OR characteristic* OR finding* OR factor*) AND (predict* OR model* OR decision* OR identif* OR prognos*)) OR (decision* AND (model* OR clinical* OR logistic models)) OR (prognostic AND (history OR variable* OR criteria OR scor* OR characteristic* OR finding* OR factor* OR model*)) OR ("stratification" OR "ROC Curve" OR "discrimination" OR "discriminate" OR "c statistic" OR "area under the curve" OR "Calibration" OR "Indices" OR "algorithm" OR "Multivariable")
#5 "cohort analysis" OR "observational study" OR "prospective study" OR "clinical trial" OR "randomized controlled trial" OR "epidemiology"
#6 #3 AND #4 AND #5 AND #6 AND ([adult]/lim OR [aged]/lim) AND [humans]/lim AND [english]/lim AND [1-1-1990]/sd NOT [23-8-2014]/sd
Appendix B – Search terms for impact studies

Search terms used to identify all impact studies, which are combined with each specific risk scores acronym, or if not applicable the name of the cohort in which the score was developed or first author:


EMBASE: (Effectiveness:ti,ab OR Comparing:ti,ab OR Compared:ti,ab OR Evaluate:ti,ab) AND (Algorithm:ti,ab OR Strategy:ti,ab OR Managed:ti,ab OR Management:ti,ab OR Decision:ti,ab) AND (heart failure " OR "congestive heart failure" OR "cardiac insufficiency" OR "cardiac failure") AND [humans]/lim AND [1-1-1990]/sd NOT [8-23-2014]/sd
Supplemental References


일반인에서의 심부전 발생 위험 예측 모델
A Systematic Review

유 병수 교수 · 원주세브란스 기독병원 심장내과

초록

배경
고위험군 환자들을 효율적으로 선별하여, 적절하면서도 비용 효율이 높은(cost–effective) 예방적 조치를 시행한다면, 심부전의 유병률을 낮출 수 있을 것이다.

방법 및 결과
연구진은 발표된 심부전 위험 예측 모델들의 특성을 평가하기 위해, 2014년 8월 현재의 MEDLINE과 EMBASE 데이터베이스를 이용하여 체계적인 문헌 검색과 검토를 수행하였다. 연구에 적합한 논문들은 예측 모델의 개발, 적합성, 영향 평가 등을 보고하였다. 연구진은 적합성과 영향 분석뿐만 아니라, 데이터를 추출하기 위해 연구의 설계 및 특성, 위험 예측인자, 식별, 계측, 재분류 능력을 독립적으로 검토하였다. 최종적으로 28가지의 심부전 예측 모델에 대해 언급한 13개의 발표 연구가 포함되었는데, 모델들은 용인될 만큼 좋은(acceptable-to-good) 식별력을 갖추고 있었다(c-statistics, 0.70). 그리고 계측은 흔하지 않게 평가되었고, 용인할 만큼은 되었다. 2개의 모델만이 외부에서 한 번 이상 검증되었으며, 적당하게 용인할 만한(modest-to-acceptable) 식별력을 보여주었다(c-statistics, 0.61–0.79). 평가가 진행되는 동안, 새로운 혈액 및 영상 표지자들은 경미하게 위험 예측을 개선시켰다. 한 개의 모델이 인종을 기초로 하위 그룹에서 예측 특성을 평가한 반면, 두 개의 모델은 성별을 기초로 하위 그룹에서 예측 특성을 평가하였다. 영향 분석에서는 어떠한 모델도 임상 진료지침에서의 사용이 추천되지 않을음을 확인하였다.

결론
심부전 위험의 예측은 초기 단계에 머물러 있다. 현존 모델의 예측 능력은 주어진 자료에서는 효과적이었으나, 타집단을 대상으로는 검증되지 않았다. 또한, 어떤 모델이 비용 효율적이 고 일반 검진에 필요한지는 명확하지 않다. 모델들이 임상적, 예방적 조치에 대한 효과는 앞으로 추가적인 연구가 필요하다.