Identifying Heart Failure Hospitalizations
Can an Automated Algorithm Become a Gold Standard?

Peter Carson, MD

The article is of particular relevance given the increasing importance of accurately assessing events, such as heart failure hospitalizations which, as the authors point out, are a crucial part of clinical research and epidemiology in heart failure. In clinical research, heart failure hospitalizations have become a component of the primary outcome of recent clinical trials to assess the efficacy of new therapies. Of course, it was not always so because early clinical heart failure trials focused on all-cause mortality. Therapies, such as converting enzyme inhibitors, β blockers, and aldosterone antagonists, became standard of care because of their demonstration of life-saving potential. Because the field has matured in knowledge and treatment, it has become more difficult to prove efficacy in all-cause mortality. Therapies, such as converting enzyme inhibitors, β blockers, and aldosterone antagonists, became standard of care because of their demonstration of life-saving potential.

The current article compares an automated algorithm with a physician panel in the ARIC population. The automated algorithm was based on published (although untested) criteria from a clinical trials workshop and required that elements be present from 3 sets of criteria for an acute decompensated heart failure (ADHF) hospitalization: (1) signs and symptoms of heart failure; (2) intravenous treatment of heart failure medications; (3) elevated natriuretic hormones, or left ventricular dysfunction, or moderately elevated natriuretic hormones and cardiac dysfunction, particularly abnormal diastolic function. As described by Loehr et al, the algorithm had moderate success: sensitivity 0.68, specificity 0.75, positive predictive value 0.85, and negative predictive value 0.53. These results were slightly better than those of previously published classification systems, but perhaps also disappointing because the current algorithm contained additional parameters of biomarkers and imaging to add to assessment that previous systems lacked.

Why was the automated algorithm less successful than expected in categorizing heart failure hospitalizations? In accepting 1 premise of the article that the physician reviewer panel be accepted as the gold standard, it would certainly be possible to surmise that the clinical judgment of the physician reviewers allowed them to more accurately identify heart failure hospitalization results but neutral mortality. The accurate identification of such events takes on even greater prominence because one considers that they are increasingly seen as sentinel events independent of other factors that influence patient prognosis. Furthermore, heart failure hospitalizations are now serving as a metric in evaluating success of programs and, in the short term, even influence reimbursement schemas.

In clinical trials, clinical events committees or, in the parlance of this article, a physician review panel have most commonly provided the adjudication of heart failure hospitalizations. These panels have associated expense and time consumption, which prompt sponsors to take pause, but then support because adjudication committees are seen as part of the structure of the modern clinical trials. However, such factors might make them impractical in evaluating nonfatal events in large registries, government databases, or from projects such as in the Loehr article, The Atherosclerosis Risk In Communities (ARIC) study. Then, there is the case to be made for the use of automated systems as the need for accurately categorizing heart failure hospitalizations grows. Among the barriers to the use of an automated algorithm is that of the definition of heart failure, which is a constellation of signs, symptoms, and laboratory findings that can be quite conclusive and can be present in conditions in which heart failure is not present or is only part of the clinical presentation.

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Division of Cardiology, the Veterans Affairs Medical Center, Washington, DC.

Correspondence to Peter E. Carson, MD, Division of Cardiology, Veterans Affairs Medical Center, 50 Irving St, Washington, DC 20422. E-mail peter.carson@va.gov


Circ Heart Fail is available at http://circheartfailure.ahajournals.org
DOI: 10.1161/CIRCHEARTFAILURE.113.000449
failure hospitalizations. The data presented that contributed to reduced specificity would lend some support to this explanation. The baseline characteristics of the group considered as false positives, automated algorithm positive but reviewer panel negative, indicate a group with a modest overrepresentation of COPD (chronic obstructive pulmonary disease) and end-stage renal disease patients compared with the false-negative group or the group with agreement for the presence of heart failure. The physician reviewer panel would have been able to interpret the presentation of such patients (signs and symptoms), which might look like heart failure but yet could be ascribed to these comorbid conditions. The use of intravenous agents usually used for heart failure might also be considered as part of an unfocused therapeutic approach again attributable to a comorbid condition mimicking ADHF. The physician group might also have been able to interpret the elevated natriuretic hormone levels as not attributable to decompensated heart failure, although the degree of elevation of these biomarkers is striking in the false-positive group. It should be noted that the natriuretic peptide data are presented as mean values, but median levels would be more informative. In these circumstances, an algorithm would be disadvantaged because of its fixed criteria. The structure of the algorithm itself may be the predominant reason for the even lower sensitivity in that each of 3 elements were required to be present: the first 2 elements (signs and symptoms of heart failure, IV medications for heart failure) were invariably present (although with low specificity), but components of the third element (elevated natriuretic peptides or evidence of cardiac dysfunction) weaken the overall performance of the algorithm. One component, natriuretic peptides, was similar to the overall algorithm in its sensitivity and specificity but other components, evidence of systolic dysfunction (left ventricular ejection fraction <40%) or a mix of mildly elevated natriuretic peptides with diastolic dysfunction, performed poorly. The low sensitivity of the systolic dysfunction component (0.46) could be expected because unselected heart failure databases uniformly report at least half of patients with heart failure have an left ventricular ejection fraction >40%. The performance of the third element with moderately elevated natriuretic hormones, including diastolic dysfunction, was particularly problematic. Diastolic dysfunction is neither invariably present in heart failure with preserved ejection fraction (HF-P EF) patients, 31% prevalence in a recent report of echocardiographic findings in a HF-P EF cohort, nor always reported, likely accounting for the very low sensitivity (0.20) in the current report.

The automated algorithm is compared with the ARIC physician reviewer panel; therefore, it is reasonable to evaluate the definitions and performance of this referent standard. The ARIC classification protocol of this panel has been previously published in comparison with other diagnostic systems, including Framingham, National Health and Nutrition Examination Survey (NHANES), Modified Boston, and Gothenburg. Operationally, 2 members reviewed the materials with a third member involved if there was lack of agreement, a commonly used system in end point adjudication. The article provides some insight on the workings of the ARIC group, in that it presents data for baseline characteristics presented as groupings by agreement or lack of agreement between the automated algorithm and the panel: the groups with agreement both for the heart failure hospitalization or not contain characteristics that suggest accuracy for these 2 groups. However, as described above, the false-positive group (automated algorithm positive, panel negative) has remarkably high level of natriuretic peptides for a group that is not decompensated heart failure. Furthermore, the false-negative group (automated algorithm negative, panel positive) is also puzzling in that the characteristics do not suggest what would lead the panel to conclude ADHF was present, whereas the algorithm was not. Perhaps the explanation then is that, for the panel, they knew it when they saw it. The further question of how physician review panel adjudication compares with investigator assessment in clinical trials, or to other approaches such as diagnostic-related group categorization in large databases, is not a subject of Loehr et al1 and is, therefore, outside the scope of this editorial.

Can an algorithm perform as well as a panel such that it could be used to evaluate events such as heart failure hospitalizations in large databases? The current comparison indicated only moderate agreement, with the majority of disagreements involving false negatives. It is worth noting that in the current comparison, definition of ADHF in the ARIC classification system included both definite ADHF (clear evidence of HF with certainty as to the cause of the presentation) and possible ADHF (without as much certainty that HF is the cause of the presentation). An algorithm could not interpret this nuance and, if the comparison is confined to definite ADHF events, then the sensitivity improves modestly (0.76). Could the automated algorithm still perform better? The Framingham, modified Boston, NHANES, and Gothenburg classification systems had fewer false negatives in comparison with the automated algorithm in these classification systems and are based on clinical criteria alone without biomarkers or imaging. They do however have poor specificity. There are components of the automated algorithm that need to be improved on, particularly in element 3. It is very likely that if the algorithm were confined to addressing heart failure with reduced ejection fraction, it would perform well as the left ventricular ejection fraction <40% component would have enabled better sensitivity to be achieved, and the natriuretic peptide levels would allow there to be a reasonable specificity. It is the HF-P EF cases that present the most difficulty for the automated algorithm because they do for physician in practice as well as reviewers. The automated algorithm uses natriuretic peptide cutoff levels that have been published and include adjustment for age (for NT-pro BNP), an important issue in that heart failure is a condition that is increasingly prevalent with advancing age. However, the natriuretic peptide levels have largely been developed from decompensated heart failure with reduced ejection fraction (HF-REF) patients and further work needs to be done to define appropriate levels for those with HF-P EF. The moderately reduced levels in element 3 might be useful to replace the current higher cutoff levels. Is there any parameter of diastolic dysfunction that could be included in an automated algorithm? This is an area of controversy, but parameters such as left atrial volume and left ventricular mass are the best indicators of morbidity for diastolic heart failure. However, these parameters would be difficult to extract from a
chart review in large databases. Therefore, at the current time, no structural or functional criteria to aid in identifying decompened cases can be recommended.

The phrase “I know it when I see it” became even more popular shortly after the Supreme Court case when, in the movie Goldfinger, James Bond was asked “what do you know about gold” and answered “I know it when I see it.” From the data in the current article, a physician event review panel remains the gold standard for heart failure hospitalizations but the investigators have made an excellent effort to evaluate an automated algorithm. In large unselected databases and registries, such a tool would likely be beneficial for assessing events in HF-REF subjects and may, with some modification (and hopefully new data), be beneficial also in HF-PEF, thereby providing a way to assess ADHF events with usability across the spectrum of heart failure.

Disclosures

The author has served as a chairman or member of previous and current clinical events committees.

References


Key Words: Editorials ◼ algorithm ◼ heart failure hospitalization
Identifying Heart Failure Hospitalizations: Can an Automated Algorithm Become a Gold Standard?
Peter Carson

Circ Heart Fail. 2013;6:621-623
doi: 10.1161/CIRCHEARTFAILURE.113.000449
Circulation: Heart Failure is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2013 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/6/4/621

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/