Supplementary Online Content

Blecker S, Katz SD, Horwitz LI, et al. Comparison of approaches for heart failure case identification from electronic health record data. *JAMA Cardiol.* Published online October 5, 2016. doi:10.1001/jamacardio.2016.3236

eMethods. Machine Learning Model Development

eTable 1. Classifiers of Heart Failure, Using Logistic Regression of Structured Data (Algorithm 3)

eTable 2. Top 25 Features for Heart Failure Classification, Using a Machine-Learning Algorithm on Unstructured Data (Algorithm 4)

eTable 3. Top 25 Features for Heart Failure Classification, Using a Machine-Learning Algorithm on Both Structured and Unstructured Data (Algorithm 5)

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Machine Learning Model Development

To develop the algorithm using machine learning on free text (algorithm 4), we first combined the text from admission notes, physician progress notes, echocardiogram reports, chest imaging reports, and consult notes. The text was divided into individual words, with each word being a potential independent feature in the model for heart failure identification. The feature was 1 if the word appeared in any note and 0 otherwise. We used all words that appeared in the notes of 10 or more patients within the development set.

Using these features, we developed an L1-regularized logistic regression model with the dependent variable of a discharge diagnosis of heart failure. L1-regularization, which introduces a penalty term to model parameters, was used to limit the final number of features and avoid overfitting of the model. In model development, we used a tenfold cross-validation on the development set to choose the L1-regularization hyperparameter and found that C=0.1 optimized the area under the receiver operating curve (AUC).

We developed algorithm 5 using the same approach, but included the structured variables from algorithm 3 along with the free text as potential features in the L1-regularization logistic regression model.

These analyses were performed in python (python.org) using the ipython notebook (https://ipython.org/notebook.html) and the scikit-learn (http://scikit-learn.org) machine learning algorithm library. Specifically, we used the method sklearn.linear_model.LogisticRegression with penalty='l1'.

eTable 1. Classifiers of Heart Failure, Using Logistic Regression of Structured Data (Algorithm 3)

Characteristic		Beta coefficient
Age		0.04 (0.03,0.04)
Female		-0.19 (-0.30,-0.08)
Black/African American Race		0.16 (-0.02,0.34)
Hispanic/Latino Ethnicity		-0.20 (-0.43,0.02)
Medicaid		0.21 (0.07,0.35)
Heart failure in problem list		3.90 (3.69,4.12)
Prior diagnosis of any heart failure		1.19 (0.99,1.38)
Prior diagnosis of primary heart failure		0.09 (-0.56,0.75)
Prior echocardiography		0.70 (0.58,0.81)
Inpatient loop diuretics		1.50 (1.37,1.63)
Outpatient loop diuretics		0.66 (0.52,0.79)
Inpatient ACE inhibitors or ARB		-0.08 (-0.23,0.07)
Outpatient ACE inhibitors or ARB		0.34 (0.20,0.48)
Inpatient beta-blockers		0.26 (0.12,0.41)
Outpatient beta-blockers		-0.01 (-0.17,0.14)
Inpatient heart failure beta-blockers		-0.09 (-0.26,0.09)
Outpatient heart failure beta-blockers		0.86 (0.70,1.02)
Systolic blood pressure		-0.01 (-0.01,0.00)
Diastolic blood pressure		0.00 (0.00,0.01)
Creatinine		0.06 (0.03,0.10)
Sodium		-0.01 (-0.02,0.00)
BNP (reference group: no BNP)		
	<500	0.20 (-0.03,0.43)
	500-999	0.64 (0.38,0.89)
	1000-4999	1.34 (1.18,1.51)
	5000-9999	1.66 (1.38,1.93)
	10000-19999	1.80 (1.41,2.19)
	≥20000	1.68 (1.32,2.05)
Acute MI in problem list		0.46 (0.19,0.74)
Atherosclerosis in problem list		0.30 (0.17,0.43)

eTable 2. Top 25 Features for Heart Failure Classification, Using a Machine-Learning Algorithm on Unstructured Data (Algorithm 4)

Free text feature	Beta coefficient
chf	1.764
hf	1.459
nyha	1.233
failure	0.824
congestive	0.817
lasix	0.611
hypokinetic	0.578
diuresis	0.509
pacemaker	0.412
mvr	0.396
interstitial	0.386
furosemide	0.368
afib	0.349
icd	0.340
repair	0.330
ef	0.329
no	-0.298
hydrocodone	0.281
past	-0.278
avoid	0.275
overload	0.274
allergies	-0.274
comment	-0.268
weights	0.257
conjunctiva	0.253

eTable 3. Top 25 Features for Heart Failure Classification, Using a Machine-Learning Algorithm on Both Structured and Unstructured Data (Algorithm 5)

Characteristic or free text feature	Beta coefficient
Heart failure in problem list*	2.491
chf	1.827
hf	0.914
Prior diagnosis of any heart failure*	0.910
congestive	0.866
hfpef	0.808
nyha	0.749
lasix	0.626
primacor	0.524
failure	0.493
428	-0.478
Outpatient heart failure beta-blockers*	0.468
hypokinetic	0.447
diuresis	0.423
pvc	0.416
coumadin	0.394
diastolic	0.376
icd	0.360
unlikely	-0.353
Outpatient loop diuretics*	0.337
ef	0.327
systolic	0.312
bumex	0.312
Inpatient loop diuretics*	0.306
valve	0.299

*Structured data element.