Received: 14 June 2018 | Revised: 20 October 2018 | Accepted: 23 October 2018

DOI: 10.1111/echo.14220

ORIGINAL INVESTIGATION

WILEY Echocardiography

Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography

Joon-myoung Kwon MD[1] | Kyung-Hee Kim MD, PhD[2] | Ki-Hyun Jeon MD, MS[2] | Jinsik Park MD, PhD[2]

[1]Department of Emergency Medicine, Mediplex Sejong Hospital, Incheon, Korea

[2]Department of Cardiology, Cardiovascular Center, Mediplex Sejong Hospital, Incheon, Korea

Correspondence
Joon-myoung Kwon, Department of Emergency Medicine, Mediplex Sejong Hospital, Incheon, Korea.
Email: kwonjm@sejongh.co.kr

Background: Heart disease (HD) is the leading cause of global death; there are several mortality prediction models of HD for identifying critically-ill patients and for guiding decision making. The existing models, however, cannot be used during initial treatment or screening. This study aimed to derive and validate an echocardiography-based mortality prediction model for HD using deep learning (DL).

Methods: In this multicenter retrospective cohort study, the subjects were admitted adult (age ≥ 18 years) HD patients who underwent echocardiography. The outcome was in-hospital mortality. We extracted predictor variables from echocardiography reports using text mining. We developed deep learning-based prediction model using derivation data of a hospital A. And we conducted external validation using echocardiography report of hospital B. We conducted subgroup analysis of coronary heart disease (CHD) and heart failure (HF) patients of hospital B and compared DL with the currently used predictive models (eg, Global Registry of Acute Coronary Events (GRACE) score, Thrombolysis in Myocardial Infarction score (TIMI), Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) score, and Get With The Guidelines-Heart Failure (GWTG-HF) score).

Results: The study subjects comprised 25 776 patients with 1026 mortalities. The areas under the receiver operating characteristic curve (AUROC) of the DL model were 0.912, 0.898, 0.958, and 0.913 for internal validation, external validation, CHD, and HF, respectively, and these results significantly outperformed other comparison models.

Conclusions: This echocardiography-based deep learning model predicted in-hospital mortality among HD patients more accurately than existing prediction models and other machine learning models.

KEYWORDS

artificial intelligence, coronary artery disease, deep learning, echocardiography, heart disease, heart failure

Kwon and Kim are take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

1 | INTRODUCTION

Cardiovascular disease (CVD) is the number 1 cause of death globally, an estimated 17.7 million people died from CVD in 2015, representing 31% of all global deaths. Of these deaths, an estimated 7.4 million were due to coronary heart disease (CHD). And approximately 26 million adults worldwide are living with heart failure (HF).[1] An estimated 1.8 million people die of CHD in Europe every year, representing the most common cause mortalities of CVD.[2] HF is the leading cause of hospitalization in Europe and United states, resulting in over 1 million admissions as a primary diagnosis and representing 1%–2% of all hospitalizations.[3]

Risk prediction of mortality among heart disease (HD) patients is crucial for identifying those in need of critical care and for guiding clinical decision making. There are several mortality prediction models for CHD (eg, the Global Registry of Acute Coronary Events [GRACE] score, and the Thrombolysis in Myocardial Infarction [TIMI] score)[4] and HF (eg, the Meta-Analysis Global Group in Chronic Heart Failure [MAGGIC] score, and the Get With the Guidelines-Heart Failure [GWTG-HF] score).[5] However, these models cannot be used for initial treatment or screening, because they use the results of various modalities, including laboratory tests. And, due to their disease-specific nature, they can be used after diagnosis.

To overcome these limitations, we developed and validated a mortality prediction model for HD using only the results of echocardiography. Because echocardiography is a non-invasive bedside examination whose results can be confirmed immediately, prediction models based on only echocardiography results will be helpful in the clinical setting.[6–8] We used deep learning (DL) to derive a high-performance prediction model.[9] Recently, DL has achieved state-of-the-art performance in several domains, including medical imaging and prognosis prediction.[10,11] To the best of our knowledge, this study is the first to predict mortality risk based on echocardiography results using DL.

2 | METHODS

This multicenter retrospective cohort study was conducted in two hospitals. The study subjects were adult (age ≥ 18 years) patients who were admitted with HD (International statistical Classification of Disease and related health problem (ICD)-10 codes: I00-I09, I11, I13, and I20-I51) during the study period and underwent echocardiography during admission. We excluded patients with missing values. The Sejong General Hospital Institutional Review Board (2018-0385) and Mediplex Sejong Hospital Institutional Review Board (2018-025) approved the study protocol and waived the need for informed consent due to the general impracticability and minimal harm.

The characteristics of both hospitals were different (hospital A: a cardiovascular teaching hospital, hospital B: a community general hospital). Data from hospital A were split by date into model derivation data (July 2010–February 2017) and internal validation data (March 2017–February 2018). Data from hospital B (March 2017–February 2018) were only used for external validation (Figure 1).

The primary outcome was in-hospital mortality. We used only echocardiography result as predictor variables (Figure 2). Echocardiography report includes patient basic information (age, sex, weight, height, and heart rate) and echocardiography results. We selected 11 continuous predictor variables and 54 categorical predictor variables from echocardiography report (Table 1).

And we used the "grep" function of base R (R Development Core Team, Vienna, Austria) to get the value of the categorical variable from the text of echocardiography report.[12,13] The "grep" function searched the text for the keywords. For example, to determine the value of mitral valve stenosis, 5 level categorical predictor variable, we checked whether the text of the mitral valve description column contained "stenosis" or "MS", and confirmed that the text contained "normal", "no_", "absent", "trace", "mild", "moderate", "severe", "grade I", "II", "III" and "IV".

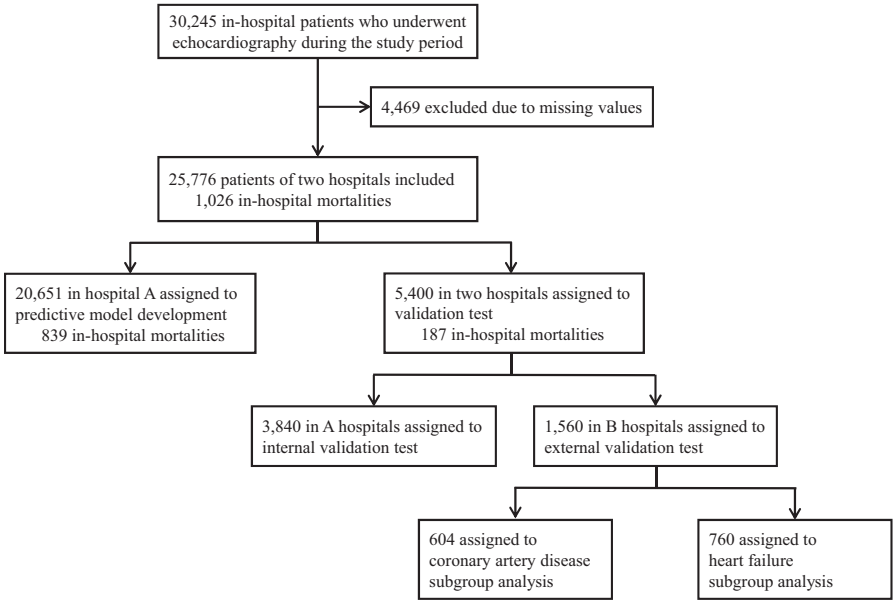

FIGURE 1 Study flow chart

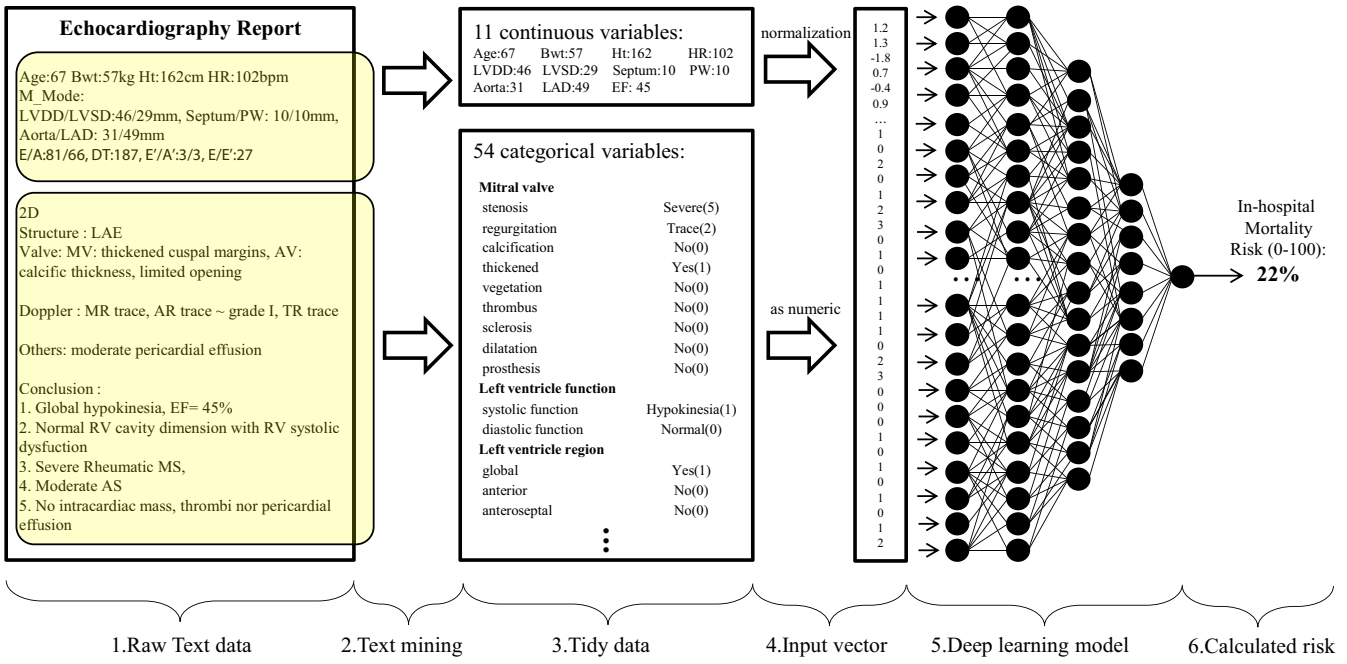

FIGURE 2 The process of deep learning based mortality prediction model using echocardiography result

The objective of this study was to derivate prediction model for in-hospital mortality. We derived 3 models, DL, logistic regression (LR) and random forest (RF), using only derivation data of hospital A. In the previous studies, logistic regression and random forest are the most commonly used machine learning methods and show better performance than traditional methods in several medical domains.[14,15] The DL model consisted of 3 hidden neural network layers with 362 nodes, batch normalization and dropout layers using TensorFlow (The Google Brain Team).[16] And we used the Adam optimizer with the default parameters and binary-cross entropy as a loss function.[17] RF and LR were derived using the "randomForest" and "glmulti" package in R (R Development Core Team).[18]

TABLE 1 Predictor variables from echocardiography report

Continuous predictor variables (11)	
Baseline information (4)	Age (year), weight (Kg), height (cm), heart rate (bpm)
Echocardiography result (7)	LVDD, LVSD, septum thickness, PWT, aorta dimension, LAD, Ejection fraction
Categorical predictor variables (54)	
Baseline information (1)	Rhythm[3: sinus/atrial fibrillation/other]
Mitral valve description (10)	Stenosis[5: absent/trace/mild/moderate/severe], regurgitation[5: absent/trace/mild/moderate/severe], calcification[2: Y/N], thickened[2: Y/N], vegetation[2: Y/N], thrombus[2: Y/N], sclerosis[2: Y/N], dilatation[2: Y/N], prosthesis[2: Y/N], tethering[2: Y/N]
Aortic valve description (9)	Stenosis[5: absent/trace/mild/moderate/severe], regurgitation[5: absent/trace/mild/moderate/severe], calcification[2: Y/N], thickened[2: Y/N], vegetation[2: Y/N], thrombus[2: Y/N], sclerosis[2: Y/N], dilatation[2: Y/N], prosthesis[2: Y/N]
Mitral valve description (9)	Stenosis[5: absent/trace/mild/moderate/severe], regurgitation[5: absent/trace/mild/moderate/severe], calcification[2: Y/N], thickened[2: Y/N], vegetation[2: Y/N], thrombus[2: Y/N], sclerosis[2: Y/N], dilatation[2: Y/N], prosthesis[2: Y/N]
Left ventricle regional description (12)	Global[2: Y/N], anterior[2: Y/N], anteroseptal[2: Y/N], septal[2: Y/N], inferoseptal[2: Y/N], inferior[2: Y/N], lateral[2: Y/N], inferolateral[2: Y/N], anterolateral[2: Y/N], apical[2: Y/N], mid[2: Y/N], basal[2: Y/N]
Left ventricle functional description (2)	Systolic function[4: normal/hypokinesia/akinesia/dyskinesia], diastolic dysfunction[5: normal/grade I/II/III/IV]
Left ventricle other description (4)	Dyssynchronized[2: Y/N], thinning[2: Y/N], dilatation[2: Y/N], aneurysm[2: Y/N]
Pericardium description (3)	Effusion[2: Y/N], fat[2: Y/N], thickened[2: Y/N]
Inferior vena cava (2)	Dilatation[2: Y/N], plethora[2: Y/N]
Right ventricle description (2)	Dilatation[2: Y/N], dysfunction[2: Y/N]

First, we compared the performances of the three models using internal validation data which were not used for the model derivation. Second, we applied these prediction models to echocardiography report of hospital B (external validation data) and validated that the model is applicable to other hospital. As a comparative measure, we used the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) to measure the performance of the model. The AUROC is one of the most commonly used metrics for prediction model and shows sensitivity against 1-specificity. Compared to the AUROC, the AUPRC is suitable for verifying false alarm rate with varying sensitivity and to assess precision (ie, 1-false positive rate) against recall (ie, sensitivity).[19,20]

We performed subgroup analyses of CHD (ICD-10 codes: I20–I25) and HF (ICD-10 codes: I50 and I255) with hospital B patients. Patients who belonged to both groups, such as those with ischemic cardiomyopathy, were study subjects in both subgroup analyses. We compared the performance of DL with TIMI and GRACE scores in the CHD group and with MAGGIC and GWTG-HF scores in the HF group. We evaluated 95% confidence interval using bootstrapping (10 000 times re-sampling with replacement).[21]

3 | RESULTS

We included 30 245 patients in this study and excluded 4469 patients with missing values. The study subjects comprised 25 776 patients with 1026 mortalities. DL was developed using 20 651 derivation data. The performance test was conducted using 3840 internal validation of hospital A and 1560 external validation data (hospital B). Subgroup analyses were performed among 604 CHD and 760 HF patients using external validation data, with 15 and 42 in-hospital mortalities, respectively (Figure 1 and Table 2).

Deep learning (AUROC: 0.912, AUPRC: 0.143) outperformed RF (AUROC: 0.893, AUPRC: 0.134) and LR (AUROC: 0.875, AUPRC: 0.160) for predicting in-hospital mortality for HD during internal validation of hospital A. As a result of applying predictive model to hospital B (external validation), DL (AUROC: 0.898, AUPRC: 0.280) outperformed RF (AUROC: 0.848, AUPRC: 0.186) and LR (AUROC: 0.841, AUPRC: 0.205) for predicting in-hospital mortality.

Deep learning (AUROC: 0.958, AUPRC: 0.458) significantly outperformed GRACE (AUROC: 0.881, AUPRC: 0.137) and TIMI (AUROC: 0.806, AUPRC: 0.139) for predicting in-hospital mortality for CHD during external validation (Figure 3). DL (AUROC: 0.913,

TABLE 2 Baseline characteristics of study subjects

	Survival discharge patients (n = 25 025)	In-hospital mortality patients (n = 1026)	*P*-value
Baseline characteristics			
Age, year	63.9 ± 13.9	71.7 ± 11.7	<0.001
Female, n	11 436 (45.7%)	502 (48.9%)	0.240
Body surface area, kg/m^2	1.71 ± 0.22	1.82 ± 0.27	<0.001
Atrial fibrillation/flutter, n	398 (1.6%)	54 (5.2%)	<0.001
Heart failure, n	11 097 (44.3%)	763 (74.4%)	<0.001
Coronary artery disease, n	10 844 (43.3%)	298 (29.0%)	<0.001
Heart rate, bpm	72.0 ± 17.6	94.6 ± 25.3	<0.001
Echocardiographic findings			
Ejection fraction, %	55.8 ± 13.9	44.2 ± 17.9	<0.001
Left ventricular diastolic dimension, mm	49.6 ± 6.8	49.3 ± 9.3	0.165
Left ventricular systolic dimension, mm	32.5 ± 8.5	35.4 ± 11.2	<0.001
Interventricular septum thickness, mm	10.1 ± 2.1	10.5 ± 3.9	0.001
Posterior wall thickness, mm	10.0 ± 7.5	10.3 ± 3.6	0.010
Aortic dimension, mm	31.8 ± 7.4	31.6 ± 4.8	0.098
Left atrium dimension, mm	42.4 ± 8.5	44.6 ± 12.9	<0.001
Early diastolic velocity of mitral inflow(E), cm/s	68.0 ± 22.7	70.3 ± 23.8	0.001
Late diastolic velocities of mitral inflow, cm/s	76.4 ± 23.6	77.8 ± 29.1	0.267
Deceleration time of mitral E velocity ms	214.1 ± 53.1	201.8 ± 48.9	<0.001
Early diastolic mitral annulus velocity (Ea), cm/s	5.6 ± 1.9	5.0 ± 1.5	<0.001
Late diastolic mitral annulus velocity, cm/s	8.2 ± 2.3	7.0 ± 3.5	<0.001
E/Ea	13.4 ± 7.6	17.1 ± 9.2	<0.001
Peak TRPG, mm Hg	23.6 ± 9.0	29.4 ± 12.3	<0.001
Estimated PA pressure, mm Hg	30.5 ± 10.8	38.8 ± 14.3	<0.001

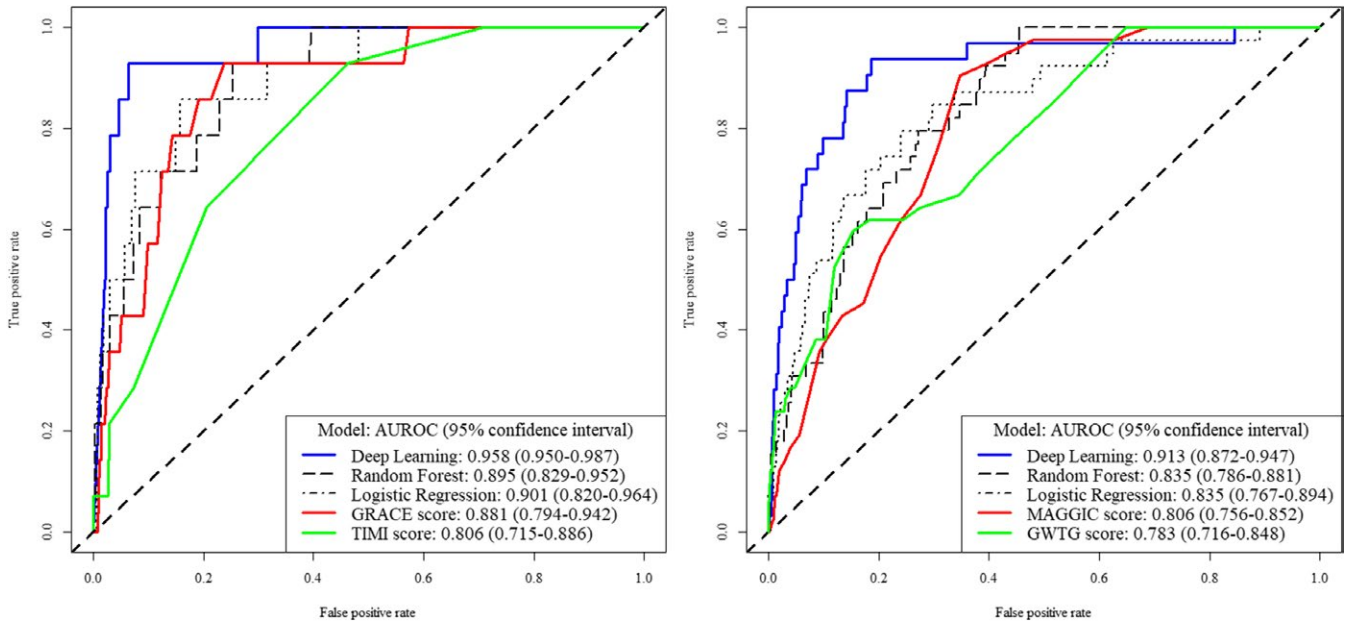

FIGURE 3 †ROC curve and areas under the receiver operating characteristic curve (AUROC) for Coronary heart disease and Heart failure during External validation. †ROC denotes receiver operating characteristic curve, AUROC = area under the recevier operating characteristic curve; GRACE = Global Registry of Acute Coronary Events; GWTG-HF = Get With The Guidelines-Heart Failure; MAGGIC = Meta-Analysis Global Group in Chronic Heart Failure; TIMI = Thrombolysis in Myocardial Infarction

AUPRC: 0.351) significantly outperformed MAGGIC (AUROC: 0.806, AUPRC: 0.154) and GWTG-HF (AUROC: 0.783, AUPRC: 0.285) for HF during external validation (Figure 3).

4 | DISCUSSIONS

In this study, we found that DL predictive model based on echocardiography predicted in-hospital mortality of HD patient more accurately than the existing predictive models (GRACE, TIMI, MAGGIC, and GWTG-HF scores) and other machine learning methods (LR and RF). The reasons for the high performance of DL are that DL evaluates the relationship between variables and automatically extracts features which classify events and non-events for prediction through many layers compared to the LR and RF models.[9] This is also why DL shows better results than traditional machine learning in several domains such as vision and prediction.[10,11]

As DL and machine learning are not derived from medical knowledge-based rules but the relationship between the given data and results, the models memorize the characteristics of the derivation data. Due to this, the performances of prediction models are not guaranteed in other situations without external validation. Wolpert explains the "no free lunch theorem"; if optimized in one situation, a model cannot produce good results in other situations.[22] This study validated that DL showed the best performance in other situations through external validation and subgroup analysis.

With imbalanced data, in which the number of negatives outweighs the number of positives, the AUROC has a limitation for evaluating the performance because the false positivity rate (number of false positives/total number of real negatives) does not decrease dramatically when the total number of negatives is large.[19] The AUPRC, on the other hand, is suitable for imbalanced data, as it considers the fraction of true positives among positive predictions. Most patients do not experience rare events such as in-hospital mortality (ie, imbalanced data). Therefore, the AUPRC is a more appropriate measure than AUROC in this study.

Imbalanced data is a significant problem in model derivation, too. When the data are very imbalanced, the trained model tends to perform poorly on minority classes (ie, low sensitivity). To overcome this problem, we multiplied the count of events (ie, mortality cases) and adjusted the ratio of non-events to events in the training data.[23] This increased the accuracy of the DL model. Unfortunately, this often occurs in medical data and the clinical setting, since most patients are non-events (ie, normal or non-mortality data). Knowledge of AUPRC and data processing methods in such studies are helpful to medical researchers planning to conduct studies in the area of machine learning or deep learning.

There are several limitations to our study. First, DL is known as a "black box." We cannot interpret the DL model, in terms of variable importance or approach to decision risk.[9] Furthermore, we cannot confirm that this developed prediction model was the simplest and most accurate method. However, it is important that our study shows the research method and possibility using text mining and deep learning to other researchers. In recent times, interpretable deep learning has been studied and is our next area of focus for research.[24] Second, as we conducted this study in two hospitals, these

study results cannot be generalized to all hospitals worldwide. And there is the potential weakness relates to the issue of variability in risk factor and disease impact among ethnicities. We plan to conduct a multicenter study to include more hospitals and more ethnicities. Third, an echocardiography report is the human verbal and numerical interpretations of the primary echocardiography image. Applying deep-learning directly to the echocardiography image might be a way to increase accuracy and reveal new knowledge and this will be our next area of study.

Despite several limitations, DL has achieved high performance in prognosis prediction in several medical domains. Medical researchers should be interested in the applicability and future development of DL in the domains of medicine.

Conclusion: A deep learning model based on echocardiography results predicted in-hospital mortality among HD patients more accurately than the existing prediction models and other machine learning models.

ORCID

Joon-myoung Kwon https://orcid.org/0000-0001-6754-1010

REFERENCES

1. Cardiovascular diseases (CVDs) Fact Sheet. World Health Organization. http://www.who.int/mediacentre/factsheets/fs317/en/. Published 2015.
2. Townsend N, Wilson L, Bhatnagar P, Wickramasinghe K, Rayner M, Nichols M. Cardiovascular disease in Europe: epidemiological update 2016. *Eur Heart J*. 2016;37(42):3232–3245.
3. Ambrosy AP, Fonarow GC, Butler J, et al. The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries. *J Am Coll Cardiol*. 2014;63(12):1123–1133.
4. Poldervaart JM, Langedijk M, Backus BE, et al. Comparison of the GRACE, HEART and TIMI score to predict major adverse cardiac events in chest pain patients at the emergency department. *Int J Cardiol*. 2017;227:656–661.
5. Ferrero P, Iacovoni A, D'Elia E, Vaduganathan M, Gavazzi A, Senni M. Prognostic scores in heart failure - Critical appraisal and practical use. *Int J Cardiol*. 2015;188(1):1–9.
6. Phillips CT, Manning WJ. Advantages and pitfalls of pocket ultrasound vs daily chest radiography in the coronary care unit: a single-user experience. *Echocardiography*. 2017;34(5):656–661.
7. Carlino MV, Paladino F, Sforza A, et al. Assessment of left atrial size in addition to focused cardiopulmonary ultrasound improves diagnostic accuracy of acute heart failure in the Emergency Department. *Echocardiography*. 2018;35(6):785–791.
8. Ghany R, Palacio A, Chen G, et al. A screening echocardiogram to identify diastolic dysfunction leads to better outcomes. *Echocardiography*. 2017;34(8):1152–1158.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;304(6):649–656.
11. Kwon JM, Lee Y, Lee Y, et al. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc*. 2018;7(13):e008678.
12. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. *J Stat Softw*. 2008;25(5):1–54.
13. Kim YS, Yoon D, Byun JH, et al. Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents. *PLoS ONE*. 2017;12(8):1–14.
14. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes*. 2016;9(6):629–640.
15. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016;44(2):368–374.
16. Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning. 12th USENIX Symp Oper Syst Des Implement (OSDI '16). 2016:265–284.
17. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2017 IEEE Int Conf Consum Electron ICCE 2017. December 2014:434–435.
18. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak*. 2011;11(1):51.
19. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. 2015;68(8):855–859.
20. Weng CG, Poon J. A new evaluation measure for imbalanced datasets. *Conf Res Pract Inf Technol Ser*. 2008;87:27–32.
21. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000;19(9):1141–1164.
22. Wolpert DH. The supervised learning no-free-lunch theorems. *Proc 6th Online World Conf Soft Comput Ind Appl*. 2001;1:10–24.
23. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–1284.
24. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. 2016.

How to cite this article: Kwon J-M, Kim K-H, Jeon K-H, Park J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography*. 2018;00:1–6. https://doi.org/10.1111/echo.14220