REVIEW ARTICLE—EDUCATIONAL TRACK

Nuclear Cardiology Data Analyzed Using Machine Learning

Kenichi Nakajima, MD, PhD[1] and Koji Maruyama, PhD[2, 3]

Received: June 6, 2022/Revised manuscript received: June 21, 2022/Accepted: June 24, 2022

Abstract

Machine learning has become popular in clinical practice, and the amount of research that uses artificial intelligence is rapidly increasing. In contrast to conventional statistical and rule-based methods, machine learning creates algorithms based only on combinations of input and output databases. Basic understanding of the internal workings of artificial intelligence, its structures and need for appropriate databases, as well as its strengths and weaknesses is important for efficient machine learning application. The cardiological applications of machine learning include diagnosing coronary artery diseases and heart failure, and examples are addressed herein. A preliminary application of machine learning to a ^{123}I-metaiodobenzylguanidine-based risk model appears promising, and further studies using similar approaches are anticipated. Nuclear medicine physicians and cardiologists should play key roles in developing machine learning-based methods to ensure practical and reliable decisions.

Keywords: Artificial intelligence, Classification, Database, Overfitting, Prediction
Ann Nucl Cardiol 2022; 8 (1): 80–85

Machine learning (ML), is a type of artificial intelligence that has become popular for analyzing phenomena in societies and life sciences. Medical classifications and predictions derived from multiple components of data and medical images are usually laborious and cannot be realized by simple decision-making algorithms. Medical journals have recently published reviews and featured topics on ML (1–3) such as what is ML, what is the difference between ML and standard statistical approaches, what are cautions when using ML, and how to judge its applicability to specific data analyses. This article does not describe precise algorithms or methods, but readers who are interested in algorithms will easily find many resources. While targets of ML include classification, prediction, image segmentation, and the creation and analysis of images, we focus on classification (for example, the diagnosis of a specific disease) derived from multiple databases, which will become a common theme of medical practice.

Classical diagnostic approach and machine learning

Various diagnostic criteria have been established in clinical guidelines, such as sex, categorized symptoms and signs, laboratory test results, medical imaging findings and some parameters measured using various imaging modalities. For example, typical diagnostic criteria in the guidelines for cardiac sarcoidosis published by the Japanese Circulation Society include a definitive diagnosis based on ≥ 2 of 5 major criteria or 1 major criterion plus ≥ 2 of 3 minor criteria (4). Some flowcharts of the diagnosis are also convenient for clinical practice. Understanding of the logical pathways used by experts in specific fields is applied to the final diagnosis, which is probably the best or standardized approach to diagnosis and therapy. Diagnostic decisions have been based on statistical methods, pattern recognition, or the intuition of experienced physicians. However, an underlying database with a complicated structure and unknown interactions among multiple factors renders the task of optimally extracting a diagnostic rule or algorithm overwhelming, if not impossible.

Machine learning could help to resolve such complex circumstances because the methods do not require predefined calculation rules; only combinations or associations among data with labels are needed. In supervised learning, a large dataset of such combinations or associations is provided and the task is to determine the rules between inputs and outputs.

doi: 10.17996/anc.22-00164

1) Department of Functional Imaging and Artificial Intelligence, Kanazawa University Graduate School of Advanced Preventive Medical Sciences, Kanazawa, Japan
2) Wolfram Wolfram Research Inc., Champaign, IL, USA
3) Department of Chemistry, Osaka Metropolitan University, Osaka, Japan

The input data could consist of numeric values, classes (such as sex, medical history, etc.) images, and/or even sounds. The output can also be numeric, such as the probability of a disease, or a classification. Clinical databases comprising thousands of pieces of data have been used to create a model of heart failure (5). Although large repositories of data are preferable, well-curated datasets with reliable labels from multiple institutions might not be available. The algorithm at the training stage searches for the best possible rule that infers an output for a given input with minimal error. Unsupervised training can also be used to find hidden patterns; for example, some groups with abnormalities can be identified among numerous normal groups. Groups or clusters might be anticipated by preceding expert knowledge, but they could also be novel and so far undiscovered.

Machine learning can also recognize specific image patterns. Consider the example of diagnosing whether or not resting myocardial perfusion is abnormal. Our clinical experience has shown that several factors can be determined from nuclear images in Digital Imaging and Communications in Medicine (DICOM) format such as a single large perfusion defect, multiple perfusion defects, left ventricular dilation, and heterogeneous distribution. Conventional diagnostic methods systematically integrate these factors quantitatively and qualitatively to determine whether the diagnosis is myocardial infarction or cardiomyopathy. In contrast, ML typically attempts to identify relationships between combinations of some features (input) and final diagnoses (output).

Algorithms used in machine learning

The algorithms frequently used by ML in practice include linear and logistic regression, tree-based methods, nearest neighbors, support vector machines, and artificial neural networks. Logistic regression is a method of classification that finds a separating (hyper) plane in the data space with minimal error probability and calculates probabilities using the sigmoid function; hence the name. Despite the simplicity, it works well under various clinical conditions because the nature of many clinical variables is that of a monotonous increase or decrease. A decision tree is the simplest of the tree-based models. It recursively constructs a tree structure by repeatedly splitting data in a process known as recursive partitioning. Trees can be combined using ensemble learning to create more potent classifiers such as random forests and boosted trees. The support vector machine is a robust model in which training data are differentiated by a hyperplane that maximizes the width of gaps between two classes. These methods can also be generalized to classify more than two classes.

Whereas classification can be based on linear separation, a nonlinear decision boundary can also be identified and used in practice. In the method called k-nearest neighbors, each new given datum is classified by finding its k-nearest training data points using a distance function. The label (diagnosis or outcome) is then assigned by majority vote among the k data. This does not work well in imbalanced datasets with small fractions of events.

An artificial neural network consists of a set of nodes and edges; nodes form input, hidden, and output layers, and those in neighboring layers are connected by edges, each of which has a factor called “weight”. Input data vectors are sequentially processed at layers through linear and non-linear transformations, and the results are compared with corresponding output data. The weights and other parameters are adjusted during training to bring the output of the network as close as possible to the given data.

Application of machine learning to numeric data

Figure 1 shows a simple example of models for predicting cardiac events from left ventricular ejection fraction (LVEF) and a myocardial perfusion defect. We created 200 datapoints with random variables, assuming normal (Gaussian) distribution, in which 50 and 150 patients were with and without cardiac events, respectively. Figure 1B shows the distribution of LVEF vs. perfusion defect (%). Training in the Wolfram language for example, proceeds in a single line:

Classify [training dataset → output dataset, Method], where the training dataset is a combination of two-dimensional input vectors (LVEF and defect%) and either “events” or “no events” comprise output, and Method specifies the method/algorithm described in the preceding section. The probability of events is calculated internally, and Figure 1C shows examples of probability density plots, depending on the method. The features of each method can be seen in these plots. For instance, two classes in the logistic regression plot are separated by a straight line with sigmoid-like probability distribution around it.

Overfitting

Overfitting is the most common problem during training. It is metaphorically similar to the situation in which students memorize all answers to given problems, instead of learning a generally applicable rule to solve them. A student who has learned from 10 myocardial perfusion scans can faultlessly deliver the correct diagnosis for specific patients, but might not necessarily interpret new images appropriately. Overfitting typically occurs when a complex algorithm is applied to a small dataset. Hence, accuracy might be very high for training, but low for validation. Figure 2 shows an example. We created 200 datapoints in which 20 patients had cardiac events and 180 did not, then applied an artificial neural network with 10, 100, and 5,000 training rounds. The accuracy increased from 84% to 94% as the number of rounds increased. However, the

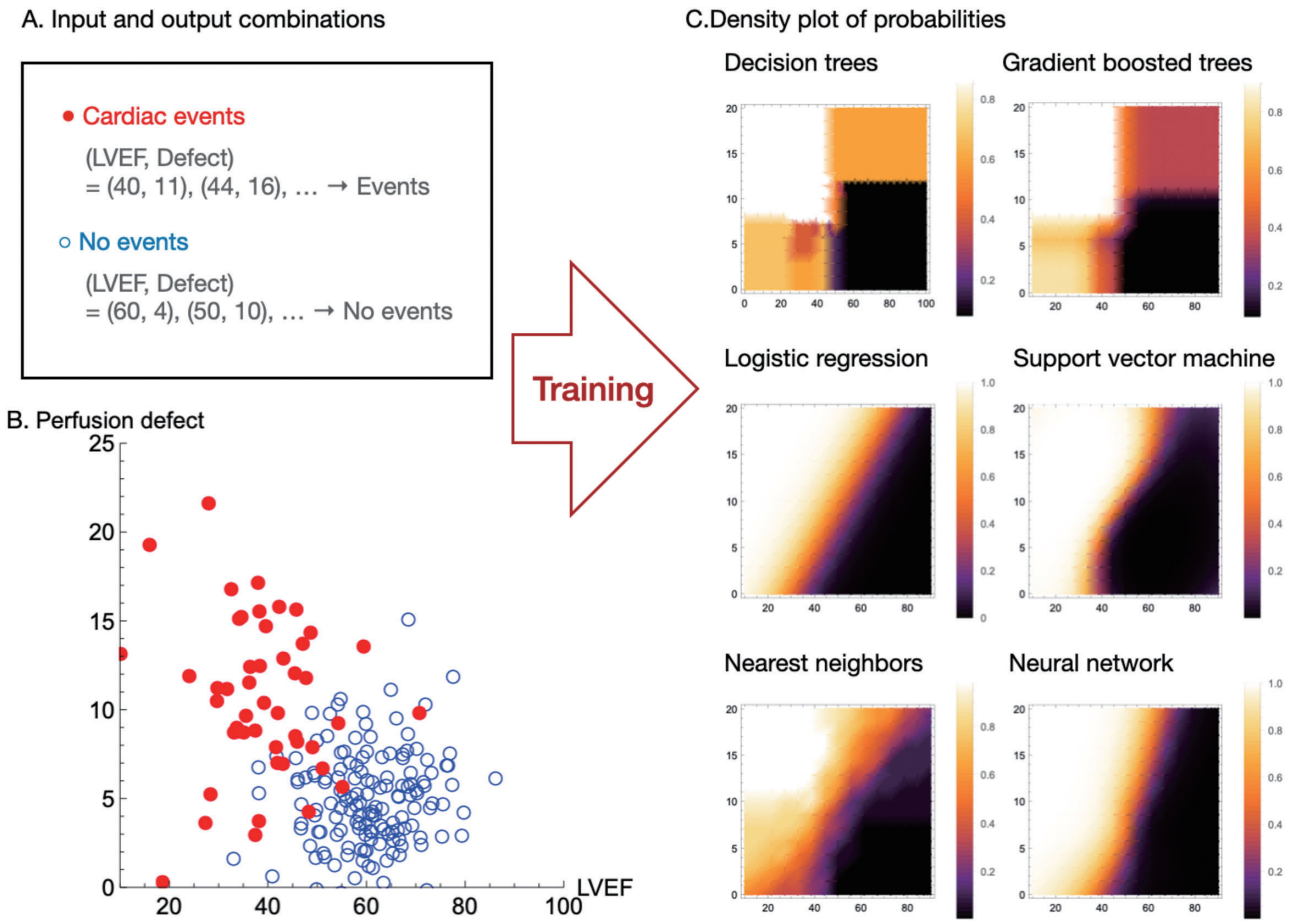

Figure 1 Steps for creating database, training and density plots of probabilities.
These samples created by random numbers and combinations of left ventricular ejection fraction (LVEF) and defect (%) were used to determine events (red) or no events (blue). Probabilities for events are plotted on the right.

classes labeled “events” and “no events”, were distinguished with unnatural sharpness (probability of 0% or 100%; Figure 2D). The problem of overfitting can be overcome by reducing model complexity, increasing the numbers of training data, limiting the numbers of training rounds, and balancing over- and under-fitting by adjusting appropriate parameters.

Application of machine learning to prognostic analysis

Clinical prognostic evaluations require a multi-factorial understanding of various types of data. Although some variables might be statistically significant, interactions among parameters can be complex and not simply explained by pathophysiological considerations. We therefore assumed that ML could play a role in extracting unknown causes and interactions for cardiac events. ^{123}I-metaiodobenzylguanidine (MIBG) is a potent predictor of cardiac events such as progressive heart failure, heart failure death and sudden cardiac or arrhythmic death. When the heart to mediastinum ratio (H/M) is an indicator of sympathetic nerve activity, patients with low MIBG H/M are more likely to have a poor prognosis. This raises the following issue. Although an H/M of 1.6 was a cutoff for good and poor prognoses in the ADMIRE-HF study (6), the actual difference in prognosis between an H/M of 1.5 and 1.7 remained unknown. We therefore created a statistical multivariable risk model to estimate 2- and 5-year mortality (%unit) based on MIBG H/M, age, sex, LVEF, and New York Heart Association (NYHA) functional class (7, 8), then the model was validated in a subsequent cohort of patients with chronic and acute decompensated heart failure (9, 10).

Clinical backgrounds, which were not included in the above analyses, might also be important, since patients who are more likely to end up with heart-failure (or pump-failure) death and arrhythmic (or sudden cardiac) death significantly differ. Therapeutic options also depend on the likelihood and mode of cardiac death. Thus, we considered that ML might be useful for differentiating the mode of cardiac death in chronic heart failure. We selected 13 factors including age, sex, NYHA functional class, estimated glomerular filtration rate, left ventricular ejection fraction, and MIBG parameters (11). Four-fold cross-validation of 512 patients using a random forest and logistic regression showed that the area under the receiver-operating characteristic curve (AUC) was 0.92 and 0.72 for heart-failure death and fatal arrhythmic events, respectively.

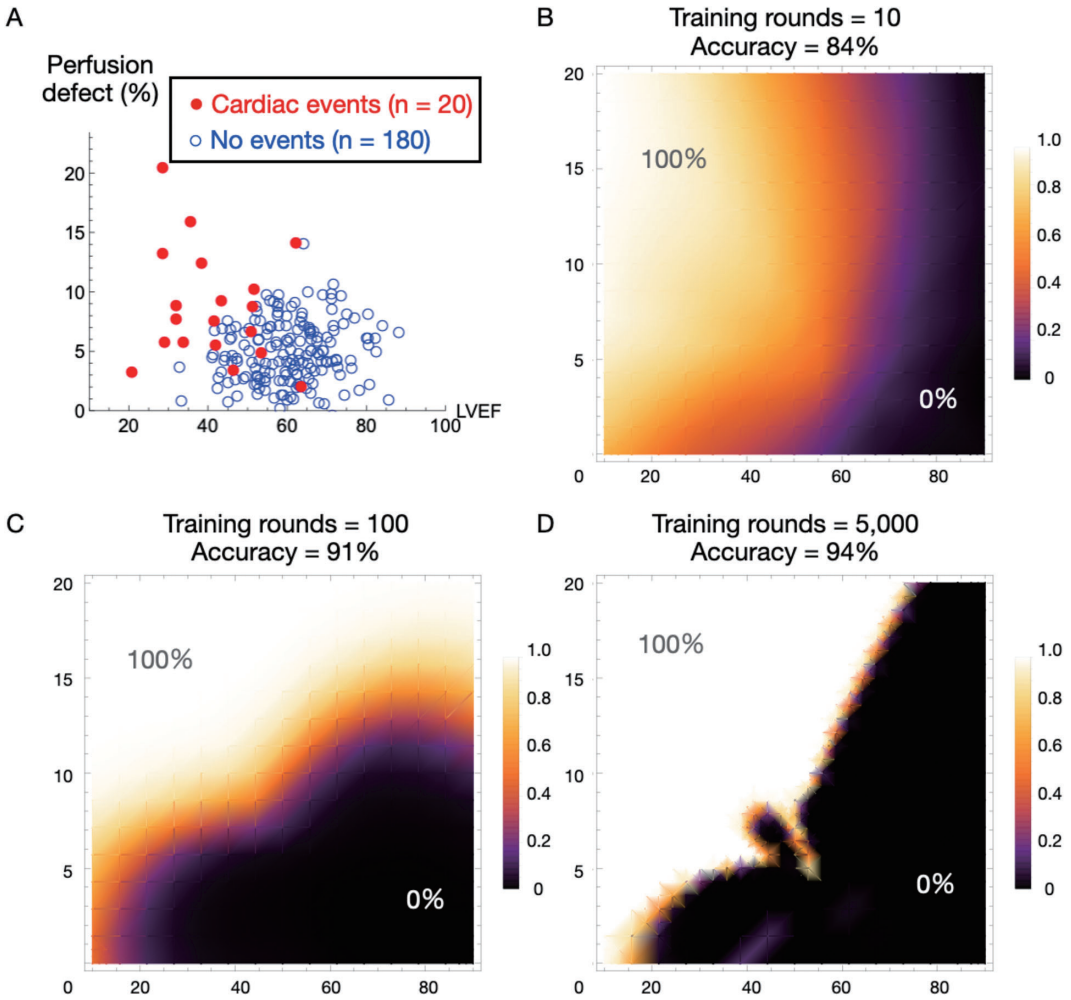

Figure 2 Application of neural network and 10, 100, and 5,000 training rounds.
(A) Cardiac events (red) developed in 20 of 200 patients (B–D). Density plots of probabilities for events are shown.

The probability of each mode of death can be simulated under specific conditions. Figure 3 shows the application of the model to two scenarios. One is a 70-year-old male with NYHA class 3, LVEF 30%, H/M 1.3, diabetes, and a high likelihood of heart-failure death. The other is a 50-year-old male with NYHA class 2, LVEF 60%, H/M 1.3, no diabetes, and a relatively high likelihood of arrhythmic death. The adequacy of such simulated conditions requires further validation.

Caution is required when popular ML models are used as long-term survival models in various cohorts, because such models do not consider patients censored during followup. Since patients who were censored and alive during a specific term might have been excluded from analysis, careful handling is required to create risk models even when ML can do this automatically.

The role of physicians

From a technological viewpoint, ML could be used more frequently in various research studies and subsequently tested in clinical practice. Automated ML-based interpretation will become a convenient clinical tool that can provide guidance to non-specialists and a second opinion to trained, experienced specialists. However, the curation of appropriate numbers of databases suitable for application to clinical problems, the execution of good training and validation, the adjustment of algorithms to avoiding under- and over-fitting, and subsequent clinical trials to test methodological validity are indispensable. Although ML technology essentially requires no clinical knowledge, a clinical rationale should be incorporated into numerous steps. The ability to explain a decision might also be important from clinical perspective, since clinicians might not trust suggestions from a black box. Even if we had an excellent ML-based software that works with 99% accuracy, but the reasons for judgement are hidden, we may still rather prefer conventional software of 80% accuracy based on interpretable algorithms for clinical decision-making (2). In medical practice, the decision-making always involves explainability and responsibility for the patient's prognosis. We need to explain the reasons and expected outcomes as part

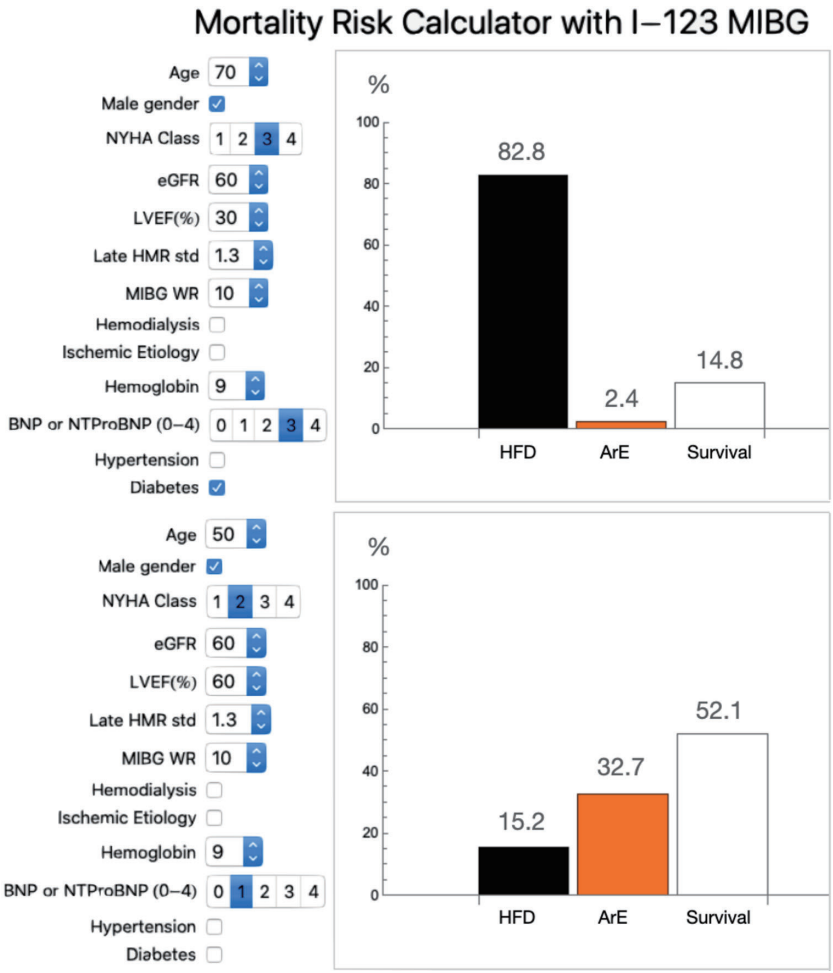

Figure 3　Cardiac mortality risk calculator using ^{123}I-MIBG.

Risk calculator is based on ML model (11). Upper and lower panels respectively indicate patients at high risk for HFD and ArE. Categorical BNP and NTProBNP values are expressed from low to high (0–4).

ArE: fatal arrhythmic events, BNP: brain natriuretic peptide, HFD: heart-failure death, NTProBNP: N-terminal pro B-type natriuretic peptide.

of informed consent to patients. For instance, a function to visualize the artificial intelligence decision process, such as the information gain ranking or the Shapley values to indicate the priority of the use of input parameters, may be helpful for clinical applications. Cooperation among nuclear medicine physicians, radiologists, and cardiologists in addition to software programmers and developers is key to the sound development of artificial intelligence technology.

Acknowledgments

All ML processing presented herein proceeded in Wolfram language. The authors appreciate Norma Foster for editorial assistance.

Sources of funding

JSPS Grants-in-Aid for Scientific Research C (No. 20K07990) to principal investigator, K. Nakajima.

Conflicts of interest

K. Nakajima has collaborative research funds from PDRadiopharma, Tokyo, Japan, Nihon MediPhysics, Tokyo, Japan, Siemens Healthcare, Tokyo, Japan, and Spectrum Dynamics, Israel. K. Maruyama is an employee of Wolfram Research Inc., Champaign, IL, USA.

Reprint requests and correspondence:
Kenichi Nakajima, MD, PhD
Department of Functional Imaging and Artificial Intelligence, Kanazawa University Graduate School of Advanced Preventive Medical Sciences,
13-1 Takara-machi, Kanazawa, 920-8640, Japan
E-mail: nakajima@med.kanazawa-u.ac.jp

References

1. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK,

Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J 2019; 40: 1975–86.
2. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC State-of-the-Art Review. J Am Coll Cardiol 2021; 77: 300–13.
3. Otaki Y, Miller RJH, Slomka PJ. The application of artificial intelligence in nuclear cardiology. Ann Nucl Med 2022; 36: 111–22.
4. Terasaki F, Yoshinaga K. New guidelines for diagnosis of cardiac sarcoidosis in Japan. Ann Nucl Cardiol 2017; 3: 42–5.
5. Mpanya D, Celik T, Klug E, Ntsinjana H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review. Int J Cardiol Heart Vasc 2021; 34: 100773.
6. Jacobson AF, Senior R, Cerqueira MD, Wong ND, Thomas GS, Lopez VA, et al. Myocardial iodine-123 meta-iodobenzylguanidine imaging and cardiac events in heart failure. Results of the prospective ADMIRE-HF (AdreView Myocardial Imaging for Risk Evaluation in Heart Failure) study. J Am Coll Cardiol 2010; 55: 2212–21.
7. Nakata T, Nakajima K, Yamashina S, Yamada T, Momose M, Kasama S, et al. A pooled analysis of multicenter cohort studies of ^{123}I-mIBG imaging of sympathetic innervation for assessment of long-term prognosis in heart failure. JACC Cardiovasc Imaging 2013; 6: 772–84.
8. Nakajima K, Nakata T, Yamada T, Yamashita S, Momose M, Kasama S, et al. A prediction model for 5-year cardiac mortality in patients with chronic heart failure using ^{123}I-metaiodobenzylguanidine imaging. Eur J Nucl Med Mol Imaging 2014; 41: 1673–82.
9. Nakajima K, Nakata T, Doi T, Kadokami T, Matsuo S, Konno T, et al. Validation of 2-year ^{123}I-meta-iodobenzylguanidine-based cardiac mortality risk model in chronic heart failure. Eur Heart J Cardiovasc Imaging 2018; 19: 749–56.
10. Tamaki S, Yamada T, Watanabe T, Morita T, Kawasaki M, Kikuchi A, et al. Usefulness of the 2-year iodine-123 metaiodobenzylguanidine-based risk model for post-discharge risk stratification of patients with acute decompensated heart failure. Eur J Nucl Med Mol Imaging 2022; 49: 1906–17.
11. Nakajima K, Nakata T, Doi T, Tada H, Maruyama K. Machine learning-based risk model using ^{123}I-metaiodobenzylguanidine to differentially predict modes of cardiac death in heart failure. J Nucl Cardiol 2022; 29: 190–201.