Tree-structured Subgroup Analysis of Receiver Operating Characteristic Curves for Diagnostic Tests

Caixia Li, PhD, Claus-C. Glüer, PhD, Richard Eastell, MD, Dieter Felsenberg, MD, David M. Reid, MD, Christian Roux, MD, Ying Lu, PhD

Rationale and Objectives: Multiple diagnostic tests are often available for a disease. Their diagnostic accuracy may depend on the characteristics of testing subjects. The investigators propose a new tree-structured data-mining method that identifies subgroups and their corresponding diagnostic tests to achieve the maximum area under the receiver-operating characteristic curve.

Materials and Methods: The Osteoporosis and Ultrasound Study is a prospectively designed, population-based European multicenter observational study to evaluate state-of-the-art diagnostic methods for assessing osteoporosis. A total 2837 women underwent dual x-ray absorptiometry (DXA) and quantitative ultrasound (QUS). Prevalent vertebral fractures were determined by a centralized radiology laboratory on the basis of radiographs. The data-mining algorithm includes three steps: defining the criteria for node splitting and selection of the best diagnostic test on the basis of the area under the curve, using a random forest to estimate the probability of DXA being the preferred diagnostic method for each participant, and building a single regression tree to describe subgroups for which either DXA or QUS is the more accurate test or for which the two tests are equivalent.

Results: For participants with weights ≤54.5 kg, QUS had a higher area under the curve in identifying prevalent vertebral fracture. For participants whose weights were >58.5 kg and whose heights were ≤167.5 cm, DXA was better, and for the remaining participants, DXA and QUS had comparable accuracy and could be used interchangeably.

Conclusions: The proposed tree-structured subgroup analysis successfully defines subgroups and their best diagnostic tests. The method can be used to develop optimal diagnostic strategies in personalized medicine.

Key Words: Receiver-operating characteristic curve; random forest; classification and regression tree; subgroup analysis; personalized medicine.

Multiple diagnostic tests are commonly available for the same disease. Their diagnostic accuracy may depend on the characteristics of the testing subjects. For example, bone mineral density (BMD) measured by dual x-ray absorptiometry (DXA) and the speed of sound (SOS) by quantitative ultrasound (QUS) devices are continuous diagnostic markers for osteoporosis. Compared to DXA, QUS has the advantages of low cost, portability, and absence of radiation exposure, but it may be less accurate. A recent prospective multicenter epidemiologic study (1) pointed out that age may influence the choice of quantitative bone assessment techniques in elderly women. In the era of personalized medicine, proper methods are needed to find subgroups with their corresponding optimal diagnostic strategies.

The area under the receiver-operating characteristic (ROC) curve (AUC) is a measure of diagnostic accuracy (2,3). A higher AUC reflects higher diagnostic accuracy. Differences between AUCs depend not only on the tests themselves but also on the population tested. A recent regression approach to ROC analysis (4) detects the interactions between diagnostic performance and covariates and assesses diagnostic utility after adjusting for covariate effects. Because of possible complex interactions, particularly when the number of covariates is large, modeling on the basis of regression approaches may make it difficult to answer the question of who should undergo which test. Ciampi et al

Acad Radiol 2012; 19:1529–1536

From the School of Mathematics and Computational Science, Sun Yet-Sen University, Guangzhou, Guangdong, China (C.L.); the Section of Biomedical Imaging, Department of Diagnostic Radiology, Universitätsklinikum Schleswig-Holstein, Campus Kiel, Kiel, Germany (C.-C.G.); the University of Sheffield Clinical Sciences Centre, Sheffield, United Kingdom (R.E.); Diagnostic Radiology, Free University Berlin, Berlin, Germany (D.F.); the School of Medicine & Dentistry, University of Aberdeen, Aberdeen, United Kingdom (D.M.R.); Centre d'Evaluation des Maladies Osseuses, Service de Rhumatologie, Assistance-Publique, Hopitaux de Paris, René Descartes University, Paris, France (C.R.); the Department of Health Research and Policy, Stanford University, Stanford, CA 94305-5405 (Y.L.); and the Veterans Affairs Cooperative Studies Program Palo Alto Coordinating Center, VA Palo Alto Health Care System, Mountain View, California (Y.L.). Received August 31, 2012; accepted September 18, 2012. This work is supported by grant R01EB 004079 from the National Institutes of Health (Bethesda, MD) to the University of California, San Francisco, (Y. Lu, principal investigator), while the first and last authors were employed at the Department of Radiology and Biomedical Imaging, University of California, San Francisco. The Osteoporosis and Ultrasound Study was sponsored by Eli Lilly (Indianapolis, IN), Sanofi-Aventis (Paris, France), Procter & Gamble Pharmaceuticals (Cincinnati, OH), Hoffmann-La Roche (Basel, Switzerland), Pfizer (New York, NY), and Novartis (Basel, Switzerland). **Address correspondence to:** Y.L. e-mail: ylu1@stanford.edu

http://dx.doi.org/10.1016/j.acra.2012.09.007

(5) proposed a tree-structured subgroup analysis for survival data on the basis of a Cox model with interaction terms to find subgroups of patients for whom one treatment is preferable to the other. Negassa et al (6) investigated a model selection in tree-structured subgroup analysis on the basis of Ciampi et al's work. These tree-based methods demonstrated efficiency to handle large numbers of covariates and identify operational subgroups of patients.

In this paper, we extend tree-based methods to the development of evidence-based decision rules to choose the most accurate diagnostic test according to easily collected covariates of a subject and apply this new approach to BMD and QUS in the diagnosis of prevalent osteoporotic fractures.

MATERIALS AND METHODS

Description of the Study Data

The Osteoporosis and Ultrasound Study (OPUS) (7) is a prospectively designed, population-based European multicenter observational study to evaluate state-of-the-art diagnostic methods for assessing osteoporosis. Population-based random samples were selected from five participating study centers in the United Kingdom, France, and Germany. All investigations were conducted in accordance with the Declaration of Helsinki and were approved by the appropriate institutional human research committee at each participating center. Of 2837 women participating, 463 (16%) were 20 to 39 years old, and 2374 (84%) were 55 to 79 years old. Techniques evaluated in this study included spine and hip BMD by DXA using bone densitometers manufactured by GE Lunar (Madison, WI) and Hologic (Bedford, MA), broadband ultrasound attenuation, and SOS measured using the DTU-one (OSI/Osteometer Meditech, Hawthorne, CA) and UBIS 5000 (Diagnostic Medical Systems, Montpellier, France). Baseline x-ray films were obtained for all study participants and used to evaluate prevalent vertebral fractures at a centralized radiology laboratory. Women with $\geq$20% height reduction from the young population mean height were considered fractured.

In our particular application, the gold standard of a clinical event was a prevalent vertebral fracture defined by the baseline radiographs. The diagnostic tests to be compared were the baseline DXA measurement of hip BMD (test 1 [T_1]) and QUS measurement of SOS measured using the DTU-one (test 2 [T_2]). Of 2322 elderly participants with complete hip BMD and DTU-one SOS information, 371 (16%) had prevalent vertebral fractures, whereas 1951 (84%) had no fractures. The characteristic variables for subgroup construction included age, height, weight, and body mass index. Table 1 summarizes the data.

TABLE 1. Summary Statistics of Diagnostic Measurements and Covariates

	Nonfractured Subjects (n_0 = 1951)	Fractured Subjects (n_1 = 371)
Diagnostic test		
Hip BMD (DXA)	878.85 ± 140.03	802.11 ± 149.59
SOS (QUS)	1546.33 ± 10.53	1541.98 ± 10.29
Continuous covariates		
Age (y)	66.42 ± 6.86	69.18 ± 7.10
Height (cm)	160.64 ± 6.31	159.61 ± 6.30
Weight (kg)	68.71 ± 12.38	67.72 ± 12.41
BMI (kg/m^2)	26.61 ± 4.51	26.56 ± 4.42

BMD, bone mineral density; BMI, body mass index; DXA, dual x-ray absorptiometry; QUS, quantitative ultrasound; SOS, speed of sound. Data are expressed as mean ± standard deviation.

Description of Recursive Partitioning Tree Algorithm and Random Forest

A recursive partitioning tree algorithm, also known as a classification and regression tree (CART) (8), consists of a sequence of splits of a group of subjects into two subgroups according to values of covariates. These splits form branches to generate a tree. Subjects before a split form a parent node. The resulted subgroups are its daughter nodes. Because there are many ways to split a node into two daughter nodes, a utility function needs to be defined for selecting the best split among all the possible binary splits. Typically, subjects in a study are divided randomly into the training and validation data. A tree grows on the basis of the training data such that the utility is homogenous within nodes but maximally different between nodes. The splitting step grows a large tree. Using a cost-complex function of CART (8), a nested sequence of subtrees

$$\mathrm{Tr}_0 \succ \mathrm{Tr}_1 \succ \mathrm{Tr}_2 \succ \cdots \succ \mathrm{Tr}_l = \mathrm{Root} \quad (1)$$

is identified that represents the optimal choices of trees at different size. Here, Tr_0 is the largest tree, and Tr_l is the smallest tree that has everyone in it. The validation data are used to determine which one of these subtrees has the best utility value in an independently collected data set. The use of validation data is to ensure that the final tree does not overly fit the training data because the splitting step is data dependent.

An alternative method to consider sampling variation is the random-forest approach (also known as bootstrap aggregating or bagging) proposed by Breiman (9). In this approach, m bootstrap samples are generated from original data as new training sets and m trees are fully grown on the basis of the bootstrap samples. Thus, there are m trees generated to form a random forest. Such a forest accounts for the effect of sampling variations in tree constructions. As a result, different trees may have different decision rules that choose different diagnostic tests as the best one for the same subject. Note that a random forest is a "committee of experts." Because trees in the forest have different decision rules, there is not a simple way to explain the underlying rationale driving the combined predictions. Rather, a forest predicts the best decision for each individual.