DOI:10.3969/j.issn.1005-202X.2018.11.017

医学生物物理

基于机器学习的骨质疏松性骨折预测研究

余锦娟,林勇
上海理工大学医疗器械与食品学院,上海 200093

【摘 要】复杂疾病的预测是遗传学研究的一个重要课题。本文引入机器学习的方法,将临床变量与遗传变量作为特征,对骨质疏松性骨折进行预测研究。对临床表型和遗传变异数据进行特征选择后分别使用Logistic回归分析法、XGBoost算法对临床因子特征变量、临床因子+遗传因子特征变量进行预测;最后,使用十折交叉验证法,对预测结果进行验证。实验结果表明,相较单独使用临床因子进行预测,加入遗传因子变量,XGBoost、Logistic方法的预测准确率均得到提高;另外,XGBoost方法较Logistic回归模型预测效果更好。

【关键词】骨质疏松性骨折;机器学习;XGBoost算法;分类预测;十折交叉验证;LASSO降维
【中图分类号】R318;R683 【文献标志码】A 【文章编号】1005-202X(2018)11-1329-05

Prediction of osteoporotic fractures based on machine learning

YU Jinjuan, LIN Yong
School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Prediction of complex diseases is an important topic in genetics research. Herein a machine learning method is introduced, taking clinical variables and genetic variables as features to predict osteoporotic fractures. After that the features of clinical phenotypes and heritable variations were selected, Logistic regression analysis and XGBoost algorithm were used to predict the characteristic variables of clinical factors, clinical factors and genetic factors. Finally, ten-fold cross validation method was used to verify the prediction results. The experimental results show that both the prediction accuracies of XGBoost and Logistic methods are improved after adding genetic factor variation as compared with using clinical factor alone. In addition, the XGBoost method is superior to Logistic regression model in the prediction of osteoporotic fractures.

Keywords: osteoporotic fracture; machine learning; XGBoost algorithm; classification prediction; ten-fold crossvalidation; LASSO dimension reduction

前 言

骨质疏松症是由于多种原因导致的骨密度和骨质量下降,骨微结构破坏,造成骨脆性增加,从而容易发生骨折的全身性骨病。由骨质疏松症引起的骨折叫骨质疏松性骨折,作为骨质疏松最严重的并发症,骨质疏松性骨折会导致长期残疾、患病个体的高死亡率和医疗保健系统的巨大经济负担[1]。所以,建立有效的骨质疏松性骨折预测模型是十分必要的。目前广泛应用于骨折风险预测的办法是:选择两组人群,一组是骨折组、一组是非骨折组,两组人群有不同的临床表型;因变量为是否骨折的两分类变量,自变量包括年龄、性别、体质量、骨密度等;再通过Logistic回归分析,构建预测模型。如Van Hemert等[2]通过骨折风险评分预测一般人群骨质疏松性骨折,指出部分临床因素对于骨质疏松性骨折的影响程度。陈超等[3]通过Logistic回归分析探讨了绝经后骨质疏松症发生骨折的影响因素,研究结果表明仅仅针对临床表现因素使用传统的Logistic回归分析对骨折的预测效果并不理想。将遗传因素纳入骨质疏松性骨折的研究目前还不多。本文研究遗传因子对骨质疏松性骨折预测的改善作用,并引入XGBoost模型来提高预测精度。XGBoost模型目前被机器学习、数据挖掘、统计学等专家广泛应用于人工智能、数据分析和统计学习等领域[4]。XGBoost是极端梯度上升(Extreme Gradient Boosting)的简称,是一种

【收稿日期】2018-07-13
【基金项目】国家自然科学基金(31301092)
【作者简介】余锦娟,硕士在读,研究方向:医学信息工程,E-mail: jinjuan_yu@163.com
【通信作者】林勇,博士,副教授,研究方向:医学信息工程,E-mail: yong_lynn@163.com

基于梯度Boosting的集成学习算法，其原理是通过弱分类器的迭代计算实现准确的分类效果[5]。它是兼具线性模型和Boosted Tree模型的一种优化模型。本文结合临床表型数据及基因型数据探讨及评价应用XGBoost模型预测骨质疏松性骨折发生的风险，较Losgistic方法具有更好的预测结果。

1 基于机器学习骨质疏松性骨折分类预测方法

本文对骨质疏松性骨折分类预测的流程包括关联基因选择、基因位点补缺、特征提取、建立模型、数据验证5个步骤，如图1所示。下面对流程中每一步骤的实现进行详细描述。

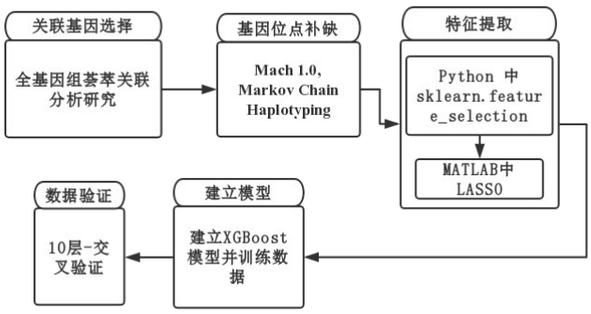


图1 分类预测方法流程图

Fig.1 Flowchart of classification prediction

1.1 关联基因选择

人类基因组中遗传位点信息巨大且大多数与骨质疏松性骨折无关。本文选用的基因型数据从Richards等[6-8]近几年来的全基因组荟萃关联分析研究中获取到17个相关基因，分别是C17orf53、RPS6KA5、CTNNB1、SLC25A13、DCDC5、SOST、DKK1、SPTBN1、FAM210A、STARD3NL、FUBP3、WNT4、LRP5、WNT16、MBL2、ZBTB40、MEPE。

1.2 基因SNP位点补缺

本文采用的700个中国西安人样本做预测研究，在基因分型过程中存在少量缺失。在此，采用单核苷酸多态(SNP)的基因型数据，通过基因位点补缺软件对基因的遗传位点进行补缺。其原理是通过建立SNP数据补缺问题与隐马尔可夫模型(Hidden Markov Model, HMM)的映射关系，然后利用SNP位点间的连锁不平衡度反映出SNP位点间的关系特点，再将数据补缺问题转化为HMM解码问题，最后解码HMM[9-10]。

1.3 特征向量提取

在建立模型的过程中，使用的基因位点数据越多，计算量越大。为降低计算的复杂度，同时又保证预测的准确性，提取其中的特征位点显得尤其重要。本研究提取基因位点特征向量分为两步：

(1)使用Python sklearn.feature_selection包中特征提取函数SelectKBest，选择卡方统计量排名前10的特征。此方法的原理是单变量特征提取，分别计算每个特征的卡方统计量，根据卡方统计量相关联的P值来选取特征。对17个基因分别提取卡方统计量排名前10的特征，并将获取的基因数据进行横向拼接，得到一个含有700个样本、170维变量的自变量数据矩阵$\boldsymbol{A}$。

(2)采用MATLAB中LASSO(Least Absolute Shrinkage And Selection Operator)包实现特征变量选择。LASSO是一种压缩估计。它通过构造一个罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计，能够有效地实现特征选择[11]。LASSO中的LogisticR使用L1范数正则化Logistic回归。L1范数是指向量中各个元素绝对值之和，也叫稀疏规则算子。简而言之，即使参数值接近于零。在原始的代价函数后面加上一个L1正则化项，即所有权重ω的绝对值的和，乘以λ/n。如下：

$$C = C_0 + \frac{\lambda}{n}\sum_{\omega}|\omega| \tag{1}$$

同样计算导数得：

$$\frac{\partial C}{\partial \omega} = \frac{\partial C_0}{\partial \omega} + \frac{\lambda}{n}\mathrm{sgn}(\omega) \tag{2}$$

上式中sgn(ω)表示ω的符号。那么权重ω的更新规则为：

$$\omega \rightarrow \omega' = \omega - \frac{\eta\lambda}{n}\mathrm{sgn}(\omega) - \eta\frac{\partial C_0}{\partial \omega} \tag{3}$$

比原始的更新规则多出了η * λ * sgn(w)/n这一项。当ω为正时，更新后的ω变小；当ω为负时，更新后的ω变大。因此它的效果就是让ω往0靠，使网络中的权重尽可能为0，也就达到了降维的效果。通过调整λ值可以改变降维的程度，在函数LogisticR中取λ=0.01，迭代2 000次，将170维矩阵$\boldsymbol{A}$降到40维，再根据权重大小分别选取权重系数绝对值排名前10、20的特征，将临床因子特征向量和遗传因子特征向量进行拼接，最终得到临床因子特征向量加上10个遗传因子特征向量(CLINIC+GENETIC10)、临床因子特征向量加上20个遗传因子特征向量(CLINIC+GENETIC20)的两类数据。

1.4 XGBoost建模过程

XGBoost是一个监督模型，由一堆CART树组合而成，CART树的叶子节点对应的值是一个实际分

数，而非一个确定类别，这将有利于实现高效的优化算法[12]。XGBoost模型可表示为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{4}$$

其中K表示树的棵数，F表示所有可能的CART树，f表示一棵具体的CART树。模型的目标函数可表示为：

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

通过优化目标函数，找到最佳参数组最终确定模型。目标函数的优化过程是通过加法训练完成，即分步骤优化目标函数，首先优化第1棵树，在此基础上添加一棵最优的CART树优化第2棵树，在第t步时，在现有的t-1棵树的基础上添加一棵最优的CART树$f(t)$，直至优化完K棵树。将目标函数做泰勒二阶展开：

$$obj(\theta) = \sum_{i}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_k) + \text{constant} \tag{6}$$

其中：

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), \quad h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{7}$$

y_i是第i个样本的真实标签，$\hat{y}_i^{(t-1)}$是t-1棵树组成的模型对第i个训练样本的预测值，g_i、h_i为两个衡量损失函数的参数。本研究有700个样本，通过并行运算求取700次g_i、h_i，每一次都朝着梯度下降最快的方向，最终得到最优的目标函数。

1.5 十折交叉验证

交叉验证是一种评估统计分析、机器学习算法对独立于训练数据的数据集的泛化能力。交叉验证一般要尽量满足：训练集的比例要足够多，一般大于一半；训练集和测试集要均匀抽样[13]。

十折交叉验证用来测试算法准确性。将数据集分成10份，轮流将其中9份作为训练数据，1份作为测试数据。每次试验均会得出相应的正确率（或差错率）。10次结果的正确率（或差错率）的平均值作为对算法精度的估计，一般还需要进行多次十折交叉验证，再求其均值，作为对算法准确性的估计。之所以选择将数据集分为10份，是因为通过利用大量数据集、使用不同学习技术进行的大量试验，表明十折是获得最好误差估计的恰当选择，而且也有一些理论根据可以证明这一点。

为保证验证的可靠行，本文使用的是十折分层交叉验证，即每次抽取的测试数据骨折与非骨折的比例与所有数据集中两者的比例相等。

2 实验设计与结果分析

2.1 实验数据

本文研究对象由700名不相关的个体组成，其中350名为骨质疏松性（低外伤性）髋部骨折患者，350名为健康的居住在西安市及其周边地区的对照者。这700名志愿者生活的地理位置相似，年龄相仿，并提供了性别、身高、年龄、体质量、髋部骨密度数据。所有参与者均由各自的机构伦理审查委员会批准，均提供了书面知情同意书。在中国汉族病例研究中，针对骨质疏松性骨折具有潜在、显著性影响的基因位点（SNPs）进行全关联组分析（GWAS）时，该研究对象数据被初次使用[14-15]。研究挑选出17个与骨折相关的基因位点，本研究将这17个基因上的位点数据通过特征提取作为遗传因子特征向量部分，将对骨折影响程度较高的5个变量：性别、身高、年龄、体质量、髋部骨密度及它们的平方，共10项数据作为临床因子特征向量部分。经过特征提取及数据拼接形成CLINIC+GENETIC10、CLINIC+GENETIC20两类数据。

2.2 验证方法

为验证本研究的有效性，本文对采用十折交叉验证得出的准确率和标准差、ROC曲线面积进行分类模型评估；采用精确率、召回率和F-score进行分类模型对不同类别识别表现的评估。准确率是分类器正确分类的样本数与总样本数之比：

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8}$$

其中TP=true positive，TN= true negatives，FP= false positive，FN= false negatives，ROC曲线越靠近左上角边界，即曲线下面积（AUC）越大，表示分类器性能越好。精确率是分类为真实正例样本数与分类为正例样本数之比；召回率是分类为真实正例样本数与所有真实正例样本数之比。计算公式分别为：

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

F-score是准确率的延伸，结合了精确率和召回率：

$$F - \text{score} = \frac{2\text{Recall Precision}}{\text{Recall} + \text{Precision}} \tag{10}$$

2.3 实验结果与分析

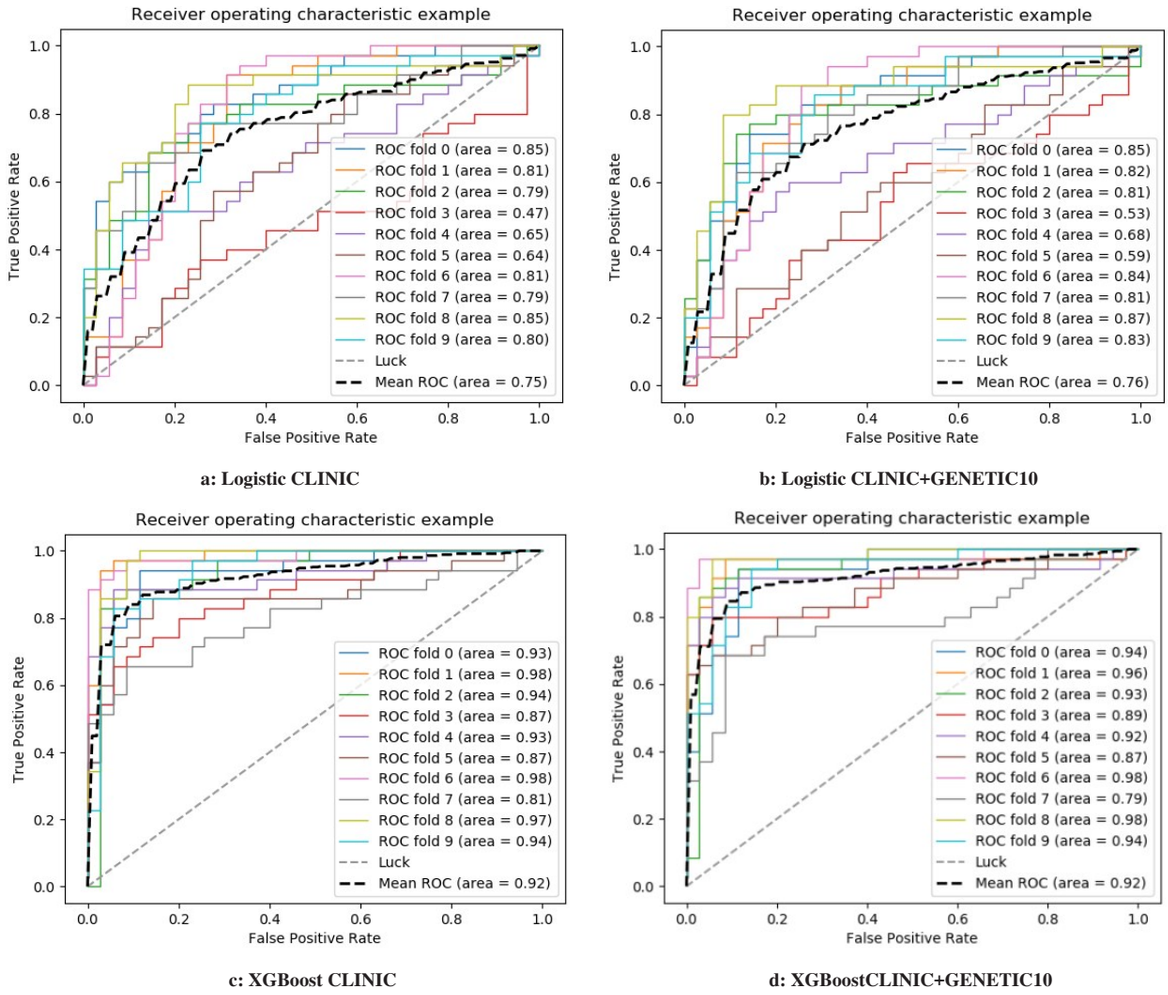
2.3.1 CLINIC+GENETIC10结果　为验证本文提出方法的有效性，将Logistic回归分析和XGBoost两种方法的分类表现相比较。设骨折为正类，未骨折为负类，单独使用临床因子数据（CLINIC）、使用临床因子+遗传因子数据（CLINIC+GENETIC10）在两种方法下的分类表现如表1所示。两种方法下测试集生

成的ROC曲线如图2所示。通过表1和ROC曲线图可以直观地发现，XGBoost方法在平均准确率、标准差、精确率、召回率、F-score分数上的表现均比Logistic方法要好。同时，在两种方法中加入遗传因子的预测效果比单独使用临床因子做预测，准确率分别提高6%、20%，且XGBoost方法提高的效果更明显。另外，在ROC曲线中XGBoost方法的面积也明显大于Logistic方法，两种方法中加入遗传因子的模型面积也明显更大。以上数据说明，引入遗传因子、使用XGBoost方法，均能够提高预测准确率。

表1 两种分类方法下加入10个遗传因子预测结果比较

Tab.1 Comparison of prediction results for test sets adding 10 SNPs

Method		Accuracy±*SD*	Precision	Recall	F-score
Logistic	CLINIC	0.703± 0.078	0.683	0.800	0.737
	CLINIC+GENETIC10	0.709 ±0.078	0.737	0.800	0.767
XGBoost	CLINIC	0.863± 0.067	0.800	0.914	0.853
	CLINIC+GENETIC10	0.883±0.055	0.829	0.971	0.895



a: Logistic CLINIC **b: Logistic CLINIC+GENETIC10**

c: XGBoost CLINIC **d: XGBoostCLINIC+GENETIC10**

图2 两种方法在测试集中生成的ROC曲线对比图

Fig.2 Comparison of receiver operating characteristics curves generated by test sets using two methods

2.3.2 CLINIC+GENETIC20 结果引入20个基因特征变量，Logistic方法预测准确率达到0.707，XGBoost方法预测准确率达到0.880，相较单独使用临床因子进行预测，准确率分别提高4%、17%。显然加入遗传因子能够改善骨质疏松性骨折的预测；同时，XGBoost方法相较传统的Logistic方法预测效果更好，预测结果见表2。

表2 两种分类方法下加入20个遗传因子预测结果比较

Tab.2 Comparisons of prediction results for test sets adding 20 SNPs

Method		Accuracy±*SD*	Precision	Recall	F-score
Logistic	CLINIC	0.703 ± 0.078	0.683	0.800	0.737
	CLINIC+GENETIC20	0.707±0.086	0.730	0.771	0.750
XGBoost	CLINIC	0.863 ± 0.067	0.800	0.914	0.853
	CLINIC+GENETIC20	0.880±0.056	0.810	0.971	0.883

3 总结与展望

本文创新地采用基于机器学习理论的XGBoost方法并加入遗传因子特征变量对骨质疏松性骨折进行分类预测。首先，提出了基于机器学习理论的XGBoost方法模型构建过程；其次，使用Logistic回归分析方法、XGBoost方法分别对临床因子特征变量、临床因子+遗传因子特征变量进行训练；最后，通过十折交叉验证得出的准确率、标准差及精确率、召回率、F-score、ROC曲线比较各模型在测试集上预测的效果，验证了本文提出方法的有效性。骨质疏松性骨折的预测研究有助于人们对自身骨骼健康及骨折风险进行合理地评估，并通过改变生活习惯或者药物干预等方式预防骨折发生。

研究过程中，作者尝试使用神经网络模型对研究数据进行训练并预测，由于本研究的样本量较少，在神经网络模型中预测结果并不理想。本研究中选用10个遗传因子的特征变量和20个遗传因子的特征变量，目的是为了降低模型的复杂程度，同时平衡样本数据少导致的欠拟合。随着样本量的增加，我们可以增加特征向量的数量，并使用更加优化的预测模型，可以得到更好的预测效果。骨质疏松症的研究通常从骨折和骨密度两方面展开。本文预测研究使用的骨折信息是离散因变量，通过训练分类器实现。在未来的工作中我们将针对连续变量骨密度展开预测研究。

【参考文献】

［1］ JOHNELL O, KANIS J. Epidemiology of osteoporotic fractures［J］. Osteoporosis Int, 2005, 54(Suppl.1): 58-63.

［2］ VAN HEMERT A M, VAN DENBROUCKE J P, BIRKENHÄGER J C, et al. Prediction of osteoporotic fractures in the general population by a fracture risk score: a 9-year follow-up among middle-aged women［J］. Am J Epidemiol, 1990, 132(1): 123-135.

［3］ 陈超, 李前龙, 邱乐, 等. 骨质疏松性骨折风险性预测进展［J］. 中国老年学杂志, 2008, 28(12): 1144-1146.
CHEN C, LI Q L, QIU L, et al. Development in risk prediction of osteoporotic fracture［J］. Chinese Journal of Gerontology, 2008, 28 (12): 1144-1146.

［4］ TORLAY L, PERRONE-BERTOLOTTI M, THOMAS E, et al. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy［J］. Brain Inform, 2017, 4(3): 159-169.

［5］ CHEN W, FU K, ZUO J, et al. Radar emitter classification for large data set based on weighted-XGboost［J］. Iet Radar Sonar Navigat, 2017, 11(8): 1203-1207.

［6］ RICHARDS J B, KAVVOURA F K, RIVADENEIRA F, et al. Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture［J］. Ann Intern Med, 2009, 151 (8): 528-547.

［7］ RICHARDS J B, RIVADENEIRA F, INOUYE M, et al. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study［J］. Lancet, 2008, 371(9623): 1505-1512.

［8］ RICHARDS J B, ZHENG H F, SPECTOR T D. Genetics of osteoporosis from genome-wide association studies: advances and challenges［J］. Nat Rev Genet, 2012, 13(8): 576-588.

［9］ RABINER L R. A tutorial on Hidden Markov model and selected applications in speech recognition［J］. Read Speech Recogn, 1990, 77 (2): 267-296.

［10］李昂, 温琪, 顾星博, 等. 单核苷酸多态性数据缺失值填补方法研究［J］. 中国公共卫生, 2014, 30(12): 1576-1582.
LI A, WEN Q, GU X B, et al. Researching on the missing value filling method of single nucleotide polymorphism data［J］. Chinese Journal of Public Health, 2014, 30(12): 1576-1582.

［11］刘晓宁. 基于Lasso特征选择的方法比较［J］. 安徽电子信息职业技术学院学报, 2014, 13(1): 26-30.
LIU X N. Method comparison based on Lasso feature selection［J］. Journal of Anhui Vocational College of Electronics & Information Technology, 2014, 13(1): 26-30.

［12］BABAJIDE M I, SAEED F. Bioactive molecule prediction using extreme gradient boosting［J］. Molecules, 2016, 21(8): 983-994.

［13］李艳芳, 王钰, 李济洪. 几种交叉验证检验的可重复性［J］. 太原师范学院学报(自然科学版), 2013, 12(4): 46-49.
LI Y F, WANG Y, LI J H. Repeatability of several cross validation tests［J］. Journal of Taiyuan Normal University (Natural Science Edition), 2013, 12(4): 46-49.

［14］CHO Y S, GO M J, KIM Y J, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits［J］. Nature Genet, 2009, 41(5): 527-534.

［15］KIEL D P. Deep resequencing of promising osteoporosis loci from GWAS［J］. Bone, 2011, 48(7): S53-S54.

（编辑：黄开颜）