In Silico Strategy for Diagnosis of Chronic Kidney Disease

Nikita Basant, ETRC, Lucknow, India
Shikha Gupta, CSIR-IITR, Lucknow, India

ABSTRACT

Chronic kidney disease (CKD) is the third deadliest reason for mortality worldwide. An early detection of CKD would help to decelerate the loss of kidney function. Computational approaches provide opportunities to screen large populations for diagnosis of CKD. In this study, qualitative and quantitative models were developed to discriminate CKD and non-CKD subjects and to predict serum creatinine (SC) levels in populations using three simple clinical attributes as the predictors. The models were rigorously validated using stringent statistical coefficients, and applicability domains were also determined. The qualitative models yielded a binary classification accuracy >94% in test data, whereas, the quantitative models rendered a correlation (R^2) of >0.94 in the test data. Values of all the statistical checks were within their respective thresholds, thus putting a high confidence in the proposed models. The proposed models can be used as the tools for screening large populations for their renal status.

KEYWORDS
Applicability Domain, Chronic Kidney Disease, Clinical Attributes, Ensemble Machine Learning, Qualitative Model, Quantitative Model, Serum Creatinine

1. INTRODUCTION

The kidneys are among the most vital organ of the body and their major role is to purify the blood by extracting waste products of metabolism, and also help to control the osmolality, volume, acid–base status and ionic composition of the extracellular environment. The kidneys also play important roles in controlling the production of red blood corpuscles and regulating blood-pressure. Chronic kidney disease (CKD) is a growing health problem and has been recognized as the third deadliest reason for mortality worldwide (Dedhia, 2007). Dash and Agarwal (2006) have reported approximately 7.85 million patients suffering from CKD in India. Patients with CKD lose kidney function and over time are at risk of developing end-stage kidney disease (ESKD). The ESKD patients are more prone to develop various complications, including malnutrition (Sathishbabu and Suresh, 2012). Identification of the individuals at risk for CKD is an important first step in modifying the progressive course of CKD. Early identification of CKD would provide the best opportunity to implement strategies known to decelerate the loss of kidney function. Clinical investigations for the identification of renal malfunctioning are tedious, time and cost effective. Moreover, several of the patients reports for exhaustive renal function screening at a later stage, generally beyond the point of reversal treatment. Therefore, there is a need for some alternate methods for screening the vast populations for their renal function tests, which could help in early detection of the problem (Echouffo-Tcheugui and Kengne, 2012). In recent years, the computational approaches have emerged as the appropriate methods to facilitate a rapid and inexpensive screening of large populations for several diseases. These methods not only offer possibilities for the qualitative and quantitative prediction of disease, but also help in
identifying the input attributes which are quantitatively related with the endpoint. The qualitative method can provide a tool for initial screening of the populations for detection of any CKD, whereas, the later one can predict the level of disease in a quantitative manner. The serum creatinine (SC) has been considered as an important indicator of the kidney function (Perazella and Reilly, 2003). Globally acceptable normal ranges of SC in adult male (0.7 - 1.3 mg/dl) and female (0.4-1.0 mg/dl) have been established (Verma et al., 2006). A value higher than the normal range has been considered as an indicator of the malfunctioning of the renal system. Recently, some reports on computational modeling have been published in the literature to predict the CKD prevalence among the masses (Kshirsagar et al., 2008, Echouffo-Tcheugui and Kengne, 2012, Vijayarani and Dhayanand, 2015). However, due to their low prediction accuracies, these may find low acceptability in future predictions. A low predictive performance of both the qualitative and quantitative models could be due to the selection of less relevant input variables, inappropriate modeling methods, and incorrect usage or lack of external validation of the models. This emphasizes for a need of the computational methods ensuring higher predictive accuracies required for their wider acceptability.

In recent years, ensemble machine learning (EML) methods (Snelder et al., 2009) have emerged as unbiased tools for modeling the complex relationships between a set of independent and dependent variables and have been applied successfully in various research areas (Yang et al., 2010). In general, these methods overcome the problems with weak predictors (Hancock et al., 2005) and reduce the over-fitting the training data (Dietterich, 2000). Decision tree forest (DTF) and decision treeboost (DTB) implementing the bagging and boosting techniques, respectively, are relatively new methods for improving the accuracy of a predictive function (Yang et al., 2010).

The main goal of the present study is to identify the measurable clinical attributes that could be considered as the discriminating factors among the populations with and without onset of CKD, and could relate quantitatively with the SC, the CKD marker to develop an in silico strategy for predicting the SC level in the CKD populations. Accordingly, EML based qualitative and quantitative models (DTB) were developed using a large CKD population database. The reliability of the constructed models was tested using a separate test dataset and through deriving several statistical coefficients recommended in the literature, determining the applicability domain of the proposed models for screening new populations. The results yielded by the qualitative DTB model was also compared with those obtained using the linear discriminant analysis (LDA) method.

2. MATERIALS AND METHODS

2.1. Dataset and Data Processing

In this study, a database reporting CKD status among a population of 400 individuals was collected from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease). According to the information available in the repository, the dataset was collected from the hospital in Tamil Nadu (India) over a period of two months and reports 25 health/clinical attributes (Table 1) measured in 400 odd subjects of age between two and ninety years. In this dataset, a total 250 subjects were diagnosed with CKD and the remaining 150 were without any kidney problem. Prior to modeling, the data were screened and all the categorical and nominal health attributes were removed. All the cases with missing values were also removed. A total of 203 cases were left with 12 independent attributes. The qualitative (CKD and non-CKD) and quantitative (SC) models were then constructed using these attributes and relatively insignificant attributes were removed in successive steps following the criteria of misclassification rate (MR) and root mean squared error (RMSE). Finally, three significant attributes, namely, hemoglobin (Hb), blood glucose random (BGR),
In Silico Strategy for Diagnosis of Chronic Kidney Disease
[www.igi-global.com/article/in-silico-strategy-for-diagnosis-of-chronic-kidney-disease/171147?camid=4v1a](www.igi-global.com/article/in-silico-strategy-for-diagnosis-of-chronic-kidney-disease/171147?camid=4v1a)

Molecular Spaces Quantum Quantitative Structure-Properties Relations (QQSPR): A Quantum Mechanical Comprehensive Theoretical Framework