A Novel Approach for Prediction of Vitamin D Status Using Support Vector Regression

Shuyu Guo1*, Robyn M. Lucas1, Anne-Louise Ponsonby2, the Ausimmune Investigator Group*

1 National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia, 2 Murdoch Childrens Research Institute, Melbourne, Australia

Abstract

Background: Epidemiological evidence suggests that vitamin D deficiency is linked to various chronic diseases. However, direct measurement of serum 25-hydroxyvitamin D (25(OH)D) concentration, the accepted biomarker of vitamin D status, may not be feasible in large epidemiological studies. An alternative approach is to estimate vitamin D status using a predictive model based on parameters derived from questionnaire data. In previous studies, models developed using Multiple Linear Regression (MLR) have explained a limited proportion of the variance and predicted values have correlated only modestly with measured values. Here, a new modelling approach, nonlinear radial basis function support vector regression (RBF SVR), was used in prediction of serum 25(OH)D concentration. Predicted scores were compared with those from a MLR model.

Methods: Determinants of serum 25(OH)D in Caucasian adults (n = 494) that had been previously identified were modelled using MLR and RBF SVR to develop a 25(OH)D prediction score and then validated in an independent dataset. The correlation between actual and predicted serum 25(OH)D concentrations was analysed with a Pearson correlation coefficient.

Results: Better correlation was observed between predicted scores and measured 25(OH)D concentrations using the RBF SVR model in comparison with MLR (Pearson correlation coefficient: 0.74 for RBF SVR; 0.51 for MLR). The RBF SVR model was more accurately able to identify individuals with lower 25(OH)D levels (<75 nmol/L).

Conclusion: Using identical determinants, the RBF SVR model provided improved prediction of serum 25(OH)D concentrations and vitamin D deficiency compared with a MLR model, in this dataset.

Introduction

There have been increasing concerns about vitamin D deficiency around the world. Epidemiological evidence suggests that hypovitaminosis D is linked to various chronic diseases such as colorectal, prostate and breast cancers[1,2,3], as well as cardiovascular diseases and diabetes[4,5,6]. Vitamin D status is assessed by the serum concentration of 25-hydroxyvitamin D (25(OH)D), an accepted biomarker[7]. However measuring 25(OH)D requires blood sampling and laboratory resources for quantitative assays. This approach may not be feasible for testing hypotheses of vitamin D status as a risk factor for chronic disease in large epidemiological studies.

An alternative approach for estimating vitamin D status is to derive a predictive model based on measurements of 25(OH)D concentration and questionnaire data on known determinants, from a subset of the study cohort. Values for the remainder of the cohort are then predicted, based on their questionnaire data[8,9,10]. Past studies have used multiple linear regression (MLR) modelling to develop these predictive models. However, the final models typically explain only a small proportion of the total variability in 25(OH)D concentration, that is, the coefficient of determination (R²) values from such predictive models have ranged from 0.13 to 0.42[8,9,10,11,12,13]. In some publications, predicted and actual 25(OH)D levels have been compared in a validation sample, with Spearman(9,10) or Pearson(12) correlation coefficients ranging from 0.23 to 0.51.

Recent studies on vitamin D status prediction are shown in Table 1. These models, based on MLR, have a number of potential limitations. For example, outliers can be highly influential in MLR models, with large differences in parameters dependent on inclusion or exclusion of these values. Moreover, MLR reflects a relationship between the means of the dependent variable and the independent variables[14], although in chronic
SVM methods include support vector classification (SVC) for classification and support vector regression (SVR) for prediction.

The SVR method differs from that of MLR in the underlying theoretical settings. The basic idea of regression methods is to construct an optimal regression hyperplane with n-1 dimensions that best fits the data in an n-dimensional space. If we take the simplest example, a two-dimensional data space can be generated by two variables in a dataset; the regression hyperplane is a straight line (with one dimension). As for other conventional methods, the MLR algorithm fits a model using the least mean squares approach to define the linear hyperplane[26,27]. However, the real world is much more complicated than a linear correlation. Furthermore, the regression hyperplane based on a least mean squares approach is greatly affected by outliers. In the SVR method, these problems are solved by 1) using integrating kernel functions (i.e polynomial, sigmoid and radial basis functions) to add more dimensions to lower dimensional space or add nonlinearity to the model; and 2) introducing user-specified parameters to control the trade-off of prediction errors and flatness of the regression plane (see Methods section). Figure 1 illustrates the difference between MLR and SVR prediction models.

In this paper, we examine the utility of an SVR algorithm, in comparison with a MLR algorithm, in predicting serum 25(OH)D concentration based on the determinants of vitamin D status already identified in a population of Australian Caucasian adults.

Materials and Methods

Study population

Data included here are from 494 participants from the control group of the Ausimmune Study[28]. The Ausimmune Study is a multi-centre, case-control study examining risk factors for multiple sclerosis. The control group was randomly selected from the Australian Electoral Roll in four different study regions. Participants completed a questionnaire including self-reported recent sun exposure and sun protection behaviours, physical activity, smoking history, diet and the use of supplements. Skin types were defined by spectrophotometric measurements of skin reflectance to calculate melanin density for exposed skin sites (dorsum of hand, shoulder) and non-exposed skin sites (upper inner arm, buttock) using a spectrophotometer (Minolta 2500d)[29]. Height, weight, waist and hip circumference were also measured. Serum 25(OH)D levels were determined by liquid chromatography dual mass spectrometry at a central laboratory. Because the number of non-Caucasian participants was small (n = 26), only data from the Caucasian participants in the control group were included for the purpose of developing the vitamin D prediction model.

Statistical analysis

The MLR model. The important determinants of vitamin D status were defined using MLR and forward purposeful selection of covariates, as previously described[30]. Briefly, 12 variables were retained in the MLR environmental and phenotypic determinants model: latitude, ambient ultraviolet radiation levels, ambient temperature, hours in the sun 6 weeks before the blood draw (log transformed to improve the linear fit), frequency of wearing shorts in the last summer, physical activity (three levels: mild, moderate, vigorous), sex, hip circumference, height, left back shoulder melanin density, buttock melanin density and inner upper arm melanin density. A square root transformation of the dependent variable (serum 25(OH)D concentration) in the MLR model was performed because of heteroscedasticity of the residuals[30].

The SVR model. Given a dataset with n independent variables and m observations, the MLR model can be written as \( y = f(x) = W^T X + b \) where \( W \) represents the vector of the coefficients, \( X \) represents the vector of the independent variables, and \( b \) is the intercept. To estimate the best fit, we minimize the sum of the squared errors:

\[
\min \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{m} (y_i - (W \cdot X_i + \hat{b}))^2,
\]

(where \( i \) represents the \( i \)th observation).

When the correlation between \( x \) and \( y \) is linear, the form of the SVR algorithm is similar to that of MLR: \( y = f(x) = W^T X + b \). However, the SVR method has two additional parameters: \( C \) and \( \varepsilon \). The parameter \( C \) is introduced to adjust the error sensitivity of the training data in order to avoid over-fitting; setting \( C \) to a high value results in fewer prediction errors in the training data:

\[
\min \sum_{i=1}^{m} (y_i - (W \cdot X_i + \hat{b}))^2 + C \sum_{j=1}^{n} |W_j|^2,
\]

(where \( j \) represents the \( j \)th variable). The second parameter \( \varepsilon \) is the regularization constant, which controls the flatness of the final model [31]. The goal of SVR is to determine an optimal function that has less than \( \varepsilon \) deviation from the target values for the training data, so that we do not count errors that are less than \( \varepsilon \), and at the same time the regression hyperplane needs to be as flat as possible.

By using different kernel functions, which transform data into a high dimensional space or add non-linearity, the SVR algorithm allows application of nonlinear regression[32]. The Radial Basis
<table>
<thead>
<tr>
<th>Reference</th>
<th>Cohort</th>
<th>Sample</th>
<th>Model covariates</th>
<th>R² for the model</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giovannucci et al, 2006 [8]</td>
<td>Health Professionals Male</td>
<td>Geographical region</td>
<td>Dietary vitamin D intake deciles of predicted 25(OH)D score (p&lt;0.001)</td>
<td>28%</td>
<td>Measured plasma 25(OH)D level rose across increasing</td>
</tr>
<tr>
<td>Chan et al., 2010 (AHS-2)</td>
<td>Adventist Health Study-2 Male &amp; Female</td>
<td>Race White: 22%</td>
<td>US, Canada</td>
<td>Skin type Total: 42%</td>
<td>25.75%</td>
</tr>
<tr>
<td>Millen et al., 2010</td>
<td>Women's Health Initiative Female</td>
<td>Langley's</td>
<td>21%</td>
<td>Pearson correlation coefficient for measured plasma</td>
<td></td>
</tr>
<tr>
<td>Peiris et al., 2011</td>
<td>Veterans Administration Male</td>
<td>Triglyceride</td>
<td>12.9%</td>
<td>The model correctly classified vitamin D deficiency status</td>
<td></td>
</tr>
</tbody>
</table>

DOI: 10.1371/journal.pone.0079970.t001
Function (RBF) SVR method adopts the RBF kernel function, also known as the Gaussian kernel, which is the same as a Gaussian distribution function. Compared to linear SVR, the RBF SVR method has one more parameter, $c$, which determines the degree of nonlinearity\[33\].

For the RBF SVR modelling, the data were randomly separated into two independent samples: the ‘training sample’ (n = 294) was used to develop the parameters of the vitamin D prediction model and the ‘validation sample’ (n = 174) was used for all statistical analyses noted below. The same 12 variables were included in the model as for the MLR modelling, described above. Parameters were determined by grid search, i.e. exhaustive searching through a set of parameters, followed by cross validation. The parameters with the best model performance were selected.

**Model comparison.** Predicted values from the MLR model were derived by summing coefficients multiplied by the individual values of the covariates\[8\]. Predicted values from the SVR model were derived by running the model with the individual values of the covariates. We compared the predictions from the RBF SVR and MLR models to measured 25(OH)D values in the “validation sample” Results were reported as means, standard deviations (SDs), minima and maxima. Mean absolute differences, i.e. the mean of the absolute differences between the individual predicted and measured 25(OH)D values, were calculated as an indication of the magnitude of error. Differences between results from the RBF SVR and MLR models were analysed with the Wilcoxon signed rank test. The correlation between predicted and measured serum 25(OH)D concentrations was analysed using a Pearson correlation coefficient ($r$). Bland-Altman plots were used to provide the mean bias (the average of the difference between measured 25(OH)D and prediction scores from the two compared modelling methods) across the range of 25(OH)D levels, and 95% limits of agreement between the methods.

We tested the accuracy of classification into categories of vitamin D status using predicted 25(OH)D scores. Data in the validation sample were analysed by generating the receiver operating characteristic (ROC) curve. Sensitivities and specificities were generated for a range of cut off for the ROC curve. In chronic disease epidemiology studies, “exposures” are often categorised into quintiles. Thus, here individuals in the validation set were also classified according to quintile of predicted 25(OH)D scores and measured 25(OH)D concentration, for the purpose of testing the performance of the two models.

Data analysis for the RBF SVR model was performed using Matlab R2001b. Analyses for the MLR model, Pearson correlation, Wilcoxon signed rank test, Bland-Altman plots and ROC curves were performed using Stata 12.0 (Statacorp, Texas).

**Results**

Means, SDs, minima and maxima of predicted 25(OH)D scores for the two models are presented in Table 2. A summary, as the mean absolute difference between measured and predicted 25(OH)D for the two models, is also given. The mean absolute difference between measured and predicted 25(OH)D concentrations generated by the RBF SVR model was significantly smaller than that for the MLR model ($p = 0.012$). Figure 2 demonstrates the correlation between the measured and predicted 25(OH)D concentration for the MLR (Figure 2A) and RBF SVR (Figure 2B) models. Consistent with this, the Pearson correlation coefficients indicated better correlation between predicted scores and measured 25(OH)D concentrations for the RBF SVR model ($r = 0.74$).
than for the MLR model ($r = 0.51$). Bland Altman plots showed that there was tighter agreement between measured 25(OH)D concentration and predicted scores for the RBF SVR model than for the MLR model. 95% limits of agreement were $-24.92$, $48.37$ (Figure 3A) and $-38.26$, $31.03$ (Figure 3B) for the MLR and RBF SVR models, respectively. There was a slight negative bias across the range of measured 25(OH)D concentrations that was greater for the RBF SVR than the MLR predicted scores ($-3.62$ nmol/L, $-0.37$ nmol/L, respectively). Predicted scores from both models showed a greater tendency to negative bias at higher 25(OH)D concentrations.

We compared the sensitivity of the two modelling techniques for correctly classifying individuals as being vitamin D deficient vs. sufficient, using different cut-points. When vitamin D deficiency...
was defined as 25(OH)D level of <75 nmol/L (vs. ≥75 nmol/L), both models had reasonable sensitivity, but the RBF SVR model performed significantly better (P<0.01, Figure 4). The sensitivity for the RBF SVR model was 81.6% compared to the MLR model of 67.1%. The area under the curve (AUC) for the MLR ROC curve was 0.79 (95% confidence interval (CI) 0.73–0.86) compared with an area under the curve of 0.87 (95%CI, 0.82–0.92) for RBF SVR. Using a 25(OH)D level of 50 nmol/L as the cut off point, the AUC for the MLR ROC curve was 0.79 (95%CI, 0.68–0.89) compared with an AUC of 0.86 (95%CI, 0.79–0.94) for RBF SVR, P = 0.064. Notably, however, only 13% of the test sample were vitamin D deficient according to this cut off point (25(OH)D < 50 nmol/L) with 25(OH)D levels measured using an LC-MS/MS assay. The superior performance of the RBF SVR model was less apparent with the limited number of ‘positive’ cases. As previously reported, 25(OH)D levels from a Diasorin Liaison assay were also available for these samples[34] with the results negatively biased compared to results from the LC-MS/MS assay, i.e. a greater proportion of the sample <50 nmol/L. We thus also tested the performance of the two modelling methods using the Liaison 25(OH)D results. Here the AUC for the curve generated from the MLR results was 0.69 (95%CI, 0.62–0.76), compared to that for the RBF SVR of 0.83 (95%CI, 0.77–0.89). That is, the RBF SVR model performed significantly better than the MLR model, P<0.0001.

In epidemiological studies, exposures are often categorised into quintiles for analysis, so we classified predicted 25(OH)D scores and measured 25(OH)D concentration by quintile to determine how well the two prediction models performed in each quintile group. For the MLR model 50.2% of the predicted 25(OH)D scores, compared to 66.1% of predicted scores for the RBF SVR model, fell into the same quintile as the measured 25(OH)D values. Figure 5 shows the percentage of correct classification in each quintile. As is illustrated in Figure 5, both MLR and RBF SVR models performed well in predicting 25(OH)D concentration

**Table 2.** Predicted 25(OH)D concentration and mean absolute difference between predicted and measured 25(OH)D level (nmol/L).

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured 25(OH)D level</td>
<td>81.71</td>
<td>28.33</td>
<td>14.2</td>
<td>163.3</td>
</tr>
<tr>
<td>Predicted level MLR</td>
<td>81.3</td>
<td>20.41</td>
<td>34.54</td>
<td>121.71</td>
</tr>
<tr>
<td>Predicted level RBF SVR</td>
<td>78.10</td>
<td>18.87</td>
<td>28.01</td>
<td>129.91</td>
</tr>
<tr>
<td>Mean absolute difference MLR</td>
<td>19.04</td>
<td>15.23</td>
<td>0.18</td>
<td>76.39</td>
</tr>
<tr>
<td>Mean absolute difference RBF-SVR</td>
<td>15.65</td>
<td>8.91</td>
<td>0.05</td>
<td>49.33</td>
</tr>
</tbody>
</table>

RBF SVR, radial basis function support vector regression (nonlinear support vector regression).
MLR, multiple linear regression.
Mean absolute difference is the average of the absolute differences between the predicted and measured values.

DOI: 10.1371/journal.pone.0079970.t002

![Figure 3. Bland – Altman plots of measured 25(OH)D concentration compared to predicted scores from (A) a MLR model; (B) a RBF SVR model. The solid lines indicate the mean bias (middle line) and 95% limits of agreement (top and bottom lines). All measurements are in nmol/L.](doi:10.1371/journal.pone.0079970.g003)

![Figure 4. ROC curves of MLR and RBF SVR. ROC curves showing true-positive rates (sensitivity) plotted against the false-positive rate for different cut off points of the quantified components of MLR (gray diamonds) and RBF SVR (black circles). The points highlighted are 25(OH)D scores of 75 nmol/l for MLR and RBF SVR. The area under the curve is 0.79 and 0.87 for MLR and RBF SVR respectively.](doi:10.1371/journal.pone.0079970.g004)
in the second and third quintile (Q2 and Q3). Although both prediction models were limited in their detection of extreme values the RBF SVR model had superior performance compared to the MLR model for correct prediction in quintiles 1, 4 and 5. The MLR model had very poor performance in predicting the highest serum 25(OH)D score; the prediction accuracy for Q5 was 0%.

Figure 5. Accuracy of predicted 25(OH)D score in each quintile of 25(OH)D concentration.
doi:10.1371/journal.pone.0079970.g005

Figure 6 illustrates the percentage of individuals classified into each quintile according to actual and predicted 25(OH)D concentration. The quintile distribution of predicted 25(OH)D concentration derived from RBF SVR model is much more accurate than the MLR model, according to the quintile distribution of measured 25(OH)D concentration.

Figure 6. Percentage of individuals classified by quintiles of measured 25(OH)D concentration and predicted 25(OH)D score.
doi:10.1371/journal.pone.0079970.g006
Discussion

We compared the performance of MLR and RBF SVR models for the prediction of vitamin D status, using a set of predetermined explanatory variables. Using the RBF SVR for prediction of serum 25(OH)D concentration resulted in lower mean absolute error in comparison with the MLR model. In the validation sample we observed better correlation between predicted scores and measured 25(OH)D concentration for the RBF SVR model compared to the MLR model. Furthermore, the RBF SVR method demonstrated higher sensitivity in classifying vitamin D status as deficient/insufficient and the AUC for the RBF SVR ROC curve was significantly larger than that for the MLR ROC curve.

This is the first study in which serum 25(OH)D concentration has been modelled using RBF SVR, with previous studies focussing on MLR models. For example, Bertrand et al.[10] reported a MLR model using data from three US cohorts, with Spearman correlation coefficients between predicted and measured 25(OH)D of 0.23, 0.40, and 0.24, respectively. In the Women’s Health Initiative, Millen et al.[12] reported a comparable correlation (0.45), using a MLR model. In the Framingham Offspring Study, Liu et al.[9] observed a correlation of 0.51 between predicted and measured levels. Using the results from these prediction models imposes several limitations on the accurate estimation of “exposure” in chronic disease epidemiology. Such models have substantial unexplained variability (R^2 = 0.15–0.42) and the predicted scores are only moderately correlated with actual 25(OH)D levels. In previous studies, the predicted scores were based on data that were incomplete for known determinants of vitamin D status, such as sun sensitivity characteristics (e.g. skin colour, ability to tan), actual sun exposure and sun exposure behaviours (e.g. time spent outdoors and protective clothing). Proxies such as physical activity and ethnicity were used instead of actual sun exposure and skin colour, allowing considerable measurement error and misclassification on key determinants.

In our study, time spent outdoors and direct measurements of untanned skin colour were included as predictors in the MLR model. But even so, the MLR model using these environmental and phenotypic factors explained only a modest proportion of the total variability in serum 25(OH)D levels (R^2 = 0.36) and the Pearson correlation coefficient (for predicted vs. measured values) was 0.51. The performance of our MLR model was consistent with the prediction models reported in the previous studies, suggesting intrinsic limitations of the MLR models.

Here we did not use the R^2 value to evaluate the performance of the RBF SVR model, because this method is not based on a least mean squares approach. However, using the RBF SVR model, we observed a correlation of 0.74 between predicted scores and measured 25(OH)D concentration. Moreover, the RBF SVR model had higher sensitivity and performed better than MLR in correctly identifying individuals with vitamin D deficiency. Interestingly, the difference in sensitivity and AUC between the two models was less when the prevalence of vitamin D deficiency was low, i.e. with a cut-off of 50 nmol/L using the 25(OH)D results from the LC-MS/MS assay.

Millen et al.[12] concluded that predicted 25(OH)D scores do not adequately reflect serum 25(OH)D concentrations, and Peiris et al.[13] argued that vitamin D status cannot be reliably predicted and that common laboratory tests are required, especially for high-risk groups. Our study indicates that 25(OH)D scores developed using an RBF SVR model much better reflect actual serum 25(OH)D concentration. Although the RBF SVR model had some limitations in predicting extreme values, generally, the estimated vitamin D status was consistent with the measured 25(OH)D concentration. One limitation of our analyses was that only one validation dataset was available. Future studies testing the RBF SVR model in a range of other populations would further advance the understanding of its utility as a tool in epidemiological studies. After validation in population-based datasets, tools developed from SVM models could also be of value to primary care physicians and others to assess the risk of vitamin D deficiency to provide a more rational basis for vitamin D testing.

Conclusion

Our results demonstrated a statistically significant superiority of an RBF SVR model in comparison with a MLR model for the prediction of serum 25(OH)D concentrations in the Ausimmune Study dataset. The accuracy of 25(OH)D scores from the RBF SVR model was greater. Thus the RBF SVR method has considerable promise for the prediction of vitamin D status for use in chronic disease epidemiology and potentially other situations.

Acknowledgments

The Ausimmune Investigator Group includes: Caron Chapman (Barwon Health, Geelong, Australia), Alan Coulthard (The University of Queensland and Royal Brisbane and Women’s Hospital, Brisbane, Australia), Keith Dear (National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia), Terry Dwyer (Murdoch Childrens Research Institute, Melbourne, Australia), Trevor Kilpatrick (Centre for Neuroscience, The University of Melbourne, Melbourne, Australia), Robyn Lucas (National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia), Tony McMichael (National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia), Michael P Pender (The University of Queensland and Royal Brisbane and Women’s Hospital, Brisbane, Australia), Anne-Louise Ponsonby (Murdoch Childrens Research Institute, Melbourne, Australia), Bruce Taylor (Menzies Research Institute, Hobart, Australia), Patricia Valery (Menzies School of Health Research, Darwin, Australia), Ingrid van der Mei (Menzies Research Institute, Hobart, Australia), David Williams (John Hunter Hospital, Newcastle, Australia).

Author Contributions

Conceived and designed the experiments: SG RL AP AIG. Performed the experiments: SG. Analyzed the data: SG. Contributed reagents/materials/analysis tools: RL. Wrote the paper: SG RL. Obtained permission for use of Ausimmune dataset: RL AP. Contributed comments to drafts of the paper: AIG.

References


Author/s: 
Guo, S; Lucas, RM; Ponsonby, A-L

Title: 
A Novel Approach for Prediction of Vitamin D Status Using Support Vector Regression

Date: 
2013-11-26

Citation: 

Persistent Link: 
http://hdl.handle.net/11343/265624

File Description: 
Published version

License: 
CC BY