Annals of Biomedical Engineering (© 2008)
DOI: 10.1007/s10439-008-9492-1

Partial Least Squares Regression: A Valuable Method for Modeling Molecular Behavior in Hemodialysis

E.A. FERNÁNDEZ,[1,2] R. VALTUILLE,[3] P. WILLSHAW,[4] and M. BALZARINI[2,5]

[1]Faculty of Engineering, Catholic University of Córdoba, Camino Alta Gracia Km 10, Cordoba 5000, Argentina; [2]National Council of Scientific and Technological Research (CONICET), Buenos Aires, Argentina; [3]Fresenius Medical Care, Buenos Aires, Argentina; [4]School of Health Science, Swansea University, Swansea, UK; and [5]Biometric Department, National University of Córdoba, Cordoba, Argentina

(*Received 13 July 2007; accepted 27 March 2008*)

Abstract—The aim of this work was to use the Partial Least Squares Regression (PLS) technique to fit simple models for the interpretation of an underlying complex process. In this study, the technique was used to build a statistical model for molecular kinetic data obtained from hemodialyzed patients. By using PLS we derived statistical linear models for the prediction of the equilibrated urea concentration which would be reached 30–60 min after the end of the dialysis session. Models with an average relative prediction error (RPE) of less than 0.05% were achieved. The model predictive accuracy was evaluated in a cross-center study yielding an RPE < 3%. The chosen model was robust to variations such as sampling extraction time demonstrating a high capacity for modeling kinetics. It also was found to be useful for bedside monitoring. Finally, the PLS technique allowed identification of the most important co-variables in the model and of those patients with outlier patterns in their molecular dynamics.

Keywords—Kinetic modeling, Equilibrated urea, Hemodialysis adequacy, Statistical modeling.

INTRODUCTION

Over the last few years, clinical trials related to diverse molecules as biomarkers of the dialysis process have been carried out worldwide.[9,19,20] There has been an outstanding technical growth in hemodialysis technique as well as in the development of mathematical approaches, which have contributed to our understanding of the kinetics of hemodialysis.[7,16] However, statistical modeling of molecular biomarker behavior, urea being one of the most important, has so far not been fully addressed.

One of the main problems in dialysis kinetic modeling is the increase of the blood molecular concentrations following the dialysis session ("molecular rebound"). In the case of urea, the most popular target biomarker, its equilibrium concentration is reached between 30 and 60 min after the end of the session. Most hemodialysis adequacy indices used in practice are based on the molecular concentration at the end of the session,[5,14,15,23,27–29] although this could lead to inadequate hemodialysis (HD) dose estimation. By contrast, the use of an equilibrated concentration of the biomarker produces a better evaluation of treatment since it is independent of marker kinetic behavior.[14–16,27] In clinical practice, waiting for the achievement of the equilibrated urea concentration is usually impractical. Therefore, the availability of a model able to predict subject-specific equilibrated concentration will be very helpful.

Although the kinetics of urea is non-linear, its extraction from blood follows some exponential family model as a function of time, we predict its equilibrated concentration after the end of the treatment session by means of a linear statistical model. In this study, we have employed a statistical approach based on PLS regression to predict equilibrated urea as a function of three timed samples of urea concentration (measured at 0, 120, and 240 min into the dialysis session) and anthropometric data from patients. It is important to appreciate that the underlying non-linear kinetics of urea drive the statistical model but that the statistical model itself may be linear, i.e., the non-linear urea kinetics only provide data for input into the model; the model does not attempt to describe those kinetics, only to predict an outcome.

A linear statistical model is based on linear combinations of unknown parameters which must be estimated from patient data. Statistical models involve two stages: model identification and estimation. The first one relies on prior knowledge or basic assumptions about the problem at hand resulting in a hypothesized mathematical structure. The model can

Address correspondence to E.A. Fernández, Faculty of Engineering, Catholic University of Córdoba, Camino Alta Gracia Km 10, Cordoba 5000, Argentina. Electronic mail: elmerfer@gmail.com

be expressed as $E(\mathbf{Y}) = F(\mathbf{X}, \mathbf{B})$, where $E(\mathbf{Y})$ is the expected value of the output vector, F is a function, $\mathbf{X}$ is a matrix of input variables and $\mathbf{B}$ is a vector of parameters that need to be estimated. In this way a set of potential mappings has been defined. The second stage, estimation, includes the selection of a specific mapping (a 'proper' $\mathbf{B}$) from the set of possible ones, choosing the parameter vector $\mathbf{B}$ that performs with least error ($\mathbf{Y} - F(\mathbf{X}, \mathbf{B})$) on the available dataset.[18]

The basic assumption in our approach for modeling kinetics was that F was a linear function, i.e., $E(\mathbf{Y}) = b_1x_1 + b_2x_2 + \cdots + b_px_p$. There are several techniques to find a proper $\mathbf{B}$ when using a linear model. However, many of them, such as ordinary least squares, are very sensitive to correlated input variables.[8] In the case of equilibrated urea concentration prediction, the input variables were intradialysis urea concentrations and anthropometric data, which strongly correlate within a patient. The multiple co-linearity problem can be overcome by the technique known as Partial Least Squares regression (PLS).[30] PLS regression not only generalizes but also combines features from principal component analysis, to deal with correlated co-variables, and multiple regression to fit linear models.[1,26] It is particularly useful when one or several dependent variables (outputs) must be predicted from a large and potentially highly correlated set of independent variables (inputs). In this study, PLS regression was applied to estimate the $\mathbf{B}$ coefficients of several linear models in order to predict the equilibrated urea concentration at the end of the dialysis session from variables correlated by patient-specific effects. The input variables were the intradialysis urea concentrations ($\mathbf{U}_0$, $\mathbf{U}_{120}$, $\mathbf{U}_{240}$), the body weight and ultrafiltration patient data. The linear approach was good in terms of prediction errors. In addition, it was easy to apply, particularly when prior knowledge is limited.

METHODS

The Knowledge Discovery in Data Base (KDD) strategy was used as the analysis framework.[10] Thus, several steps involving different KDD stages such as problem/data understanding, collection, cleaning, preprocessing, analysis-modeling, and results interpretation were implemented.

Data Understanding and Collection

The Patients

One hundred and nine stable patients were selected from two dialysis units as follows: 61 from Unit 1 (mean age 56 ± 3.5 years and mean time on dialysis (MTD) 32 ± 12.3 months) and 48 from Unit 2 (mean age 58 ± 18.0 years and MTD of 42 ± 23.5 months). All patients were from Buenos Aires, Argentina and were subjected to chronic HD treatment for at least 3 months. The selection criteria to include patients in the study were: (1) patients without infection or hospitalization in the last 30 days; (2) patients with an A-V fistula (70% autologous fistula and 30% prosthetic fistula) with a blood flow rate (QB) of ≥300 mL/min; and (3) patients having consented to participate in the study. The study protocol complied with the Helsinki Declaration and was approved by the Ethical Committee of the Catholic University of Córdoba.

All patients received HD three times a week with Baxter® machines (model 1550) using variable bicarbonate and sodium. Hollow-fiber polysulfone (Fresenius F6 and F8) and cellulose diacetate (FB170 and 210, Nissho Corp.) dialyzers were used. For the purpose of this study, all patients were dialyzed over 240 min and the flows of blood (QB) and dialysate (QD) were fixed at 300 and 500 mL/min, respectively.

It is known that hemodialysis dose is influenced by several factors including dialysis time, hemodialysis schedule, and blood and dialysate flow.[6] To decrease the complexity, such variables were handled externally, fixing their values to control their effects on the equilibrated urea prediction model.

The Input and Output Variables

Blood samples were obtained at the mid-week HD session. They were taken from the arterial line at different times to obtain urea determinations: (1) predialysis urea ($\mathbf{U_0}$), at the beginning of the procedure; (2) intradialysis urea ($\mathbf{U_{120}}$), in the middle of the HD session (at 120 min from the beginning); (3) postdialysis urea ($\mathbf{U_{240}}$), at the end of the HD session.

For the intradialysis urea ($\mathbf{U_{120}}$) and postdialysis urea ($\mathbf{U_{240}}$), QB was slowed to 50 mL/min and blood was sampled 15 s later. At this point, access recirculation ceased and the dialyzer inlet blood reflected the arterial urea concentration.

Regarding the protocols for intradialysis samples, it is worth noting that originally Smye *et al.*[28] proposed taking them within 60 min from the beginning of the session and at 20 min before its finalization. We, however, decided to take the intradialysis sample 120 min after the beginning of the HD session ($\mathbf{U_{120}}$), which allowed us to compare our results with those reported by Ghu *et al.*[15]

Urea (U) determinations were performed in triplicate on each blood sample, using two autoanalyzers (Hitachi 704 in Unit A and Technicon RA 1000 Bayer in Unit B). The urea averages were calculated and recorded with an accuracy of 1% for both machines. For information about the pre- and posttreatment

TABLE 1. Summary statistics of the patient data distribution.

	U_0	U_{120}	U_{240}	Bw	UF	Ueq
Min.	59	31	21	45.3	0	23
1st Qu.	127	64	40	59.4	1.9	50
Median	149	77	49	71.3	2.7	59
Mean	149.5	79.65	53.18	71.81	2.666	62.69
3rd Qu.	169	96	62	83.8	3.3	76
Max.	221	144	98	119	5.5	112

The concentration units are mg/dL and anthropometric units are kilograms. Min: minimum; 1st Qu: first quantile; 3rd Qu: third quantile; Max: Maximum.

status of the patient, we used the pre- and postdialysis body weights ($\mathbf{BW_0}$, $\mathbf{BW_{240}}$). Both variables are commonly used in clinical practice to decide the treatment schedule as well as to calculate the treatment dose.[16] These variables were recorded in the same dialysis session when the blood samples were taken.

The output variable was the equilibrated urea. For the purpose of this study, the patients were retained one hour in the dialysis center and the equilibrated urea levels (**Ueq**) were extracted 60 min after the end of HD. The summary statistics for the input and output variables are shown in Table 1.

Data Preprocessing

The collected input variables ($\mathbf{U_0}$, $\mathbf{U_{120}}$, $\mathbf{U_{240}}$, $\mathbf{BW_0}$, and **Uf**, where $\mathbf{Uf} = \mathbf{BW_0} - \mathbf{BW_{240}}$) have different units and ranges so, for the sake of comparison, they were all standardized.

Data Analysis-Modeling

Model Fitting

We can express the equilibrated urea concentration for the ith patient in a linear model as:

$$\mathrm{U}^{\mathrm{i}}_{\mathrm{eq}} = \mu + b_1 \cdot \mathrm{U}^{i}_{0} + b_2 \cdot \mathrm{U}^{i}_{120} + b_3 \cdot \mathrm{U}^{i}_{240} + b_4 \cdot \mathrm{Bw}^{i}_{0} + b_5 \cdot \mathrm{Uf}^{i} + e^{i} \quad (1)$$

where μ is the model intercept (overall mean), b_j ($j = 1...5$) represents the influence of each input variable on the concentration of urea at equilibrium which needs to be estimated, and $\mathbf{e^i}$ is the error term associated to the equilibrated urea recorded in the ith patient.

In matrix form the linear model in Eq. (1) is expressed as

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{X} \cdot \mathbf{B} + \mathbf{E} \quad (2)$$

where $\mathbf{Y}^{N\times1}$ is a $N \times 1$ vector of observed equilibrated urea from the "N" patients, $\boldsymbol{\mu}^{N\times1}$ is the vector of a constant value representing the model intercept, $\mathbf{X}^{N\times p}$ is the standardized input variable matrix, in this specific case $p = 5$ and $\mathbf{X}^{N\times5} = \left[\mathbf{U}_0^{N\times1}, \mathbf{U}_{120}^{N\times1}, \mathbf{U}_{240}^{N\times1}, \mathbf{BW}^{N\times1}, \mathbf{Uf}^{N\times1}\right]$. The parameter vector $\mathbf{B}^{p\times1}$ contains the regression coefficients (b_j) and $\mathbf{E}^{N\times1}$ is the vector of error terms.

Generally, **B** is estimated by means of Ordinary Least Squares (OLS). However, when the input variables are correlated, as in this case, the PLS technique is more appropriate since correlation among input variables (multicollinearity) could produce instability in the estimated regression coefficients. Here the coefficients were learned from the data by means of the PLS regression which described the common structure between the standardized output (**Y**) and input variables (**X**) in such a way that **B** maximizes the covariance between them.[1] In the PLS algorithm,[1,30] **X** and **Y** are expressed as:

$$\mathbf{X}^{N\times p} = \mathbf{T}^{N\times A} \cdot \left(\mathbf{P}^{p\times A}\right)' + \mathbf{H}^{N\times p} \quad (3)$$

$$\mathbf{Y}^{N\times1} = \mathbf{U}^{N\times A} \cdot \left(\mathbf{C}^{1\times A}\right)' + \mathbf{R}^{N\times1} \quad (4)$$

where $A \leq p$ and **H** and **R** are error matrices. The columns of **T** and **U** (**X** and **Y** are "score" matrices) provide a new representation of the **X** and **Y** variables in an orthogonal space that allows us to estimate properly the coefficients of **B** taking into account the multicollinearity problem.

The columns (factors) of the score matrices are useful for representing the variability among patients both in the input and output variables.

The matrices **P** and **C** are the projections ("loadings") of the **X** and **Y** columns into the new set of variables in **T** and **U**. The **T** matrix is calculated as $\mathbf{T} = \mathbf{X} \cdot \mathbf{W}$ where $\mathbf{W} = \mathbf{U}(\mathbf{P'U})^{-1}$. In the PLS algorithm, **U** and **P** are built iteratively[26] by means of matrix products between consecutive deflations of the original matrices **X** and **Y**. Thus, the **T** matrix is also a good estimator of **Y**, so

$$\mathbf{Y}^{N\times1} = \mathbf{T}^{N\times A} \cdot \left(\mathbf{C}^{1\times A}\right)' + \mathbf{E}^{N\times1} \quad (5)$$

where $\mathbf{C}^{1\times A}$ is the "loadings" matrix of **Y** that projects it over the new space represented by **T**. The error term in **G** represents the deviations between the observed and predicted responses.

Replacing **T** in the above equation yields:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{W} \cdot \mathbf{C}' + \mathbf{E} = \mathbf{X} \cdot \mathbf{B} + \mathbf{E} = \hat{\mathbf{Y}} + \mathbf{E} \quad (6)$$

where $\hat{\mathbf{Y}}$ is the predicted output.

The number of factors chosen has an impact on the estimation of the regression coefficients (see below). In a model with "A" factors, the **B** coefficients are calculated as follows:

$$\mathbf{B}^{p\times1} = \mathbf{W}^{p\times A}\left[\mathbf{C}^{1\times A}\right]' \quad (7)$$

Model Evaluation

The database was divided in two sets: the training set with 65% of the patients (chosen at random), and the validation set, which contained the remainder. The training set was used to fit the model via a cross-validation technique[21] with 20 partitions of size n of this training dataset for choosing the optimal number of factors. At each run and for models with none, then one to five PLS factors the Root Mean Square Error of Prediction was calculated as $RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(U_{eq}^{i} - \hat{U}_{eq}^{i}\right)^2}$ where U_{eq}^{i} is the observed equilibrated urea and $\hat{U}_{eq}^{i}$ the predicted one. The RMSEP reported for each model is the average value across the 20 runs. To evaluate the relative importance of the regression coefficients in **B**, the Variable Importance on Prediction (VIP) scores[4] were analyzed by means of the leave-one-out method.

Once the model was fitted, the validation dataset was used to evaluate the adjusted model in a "real environment." The following statistics were used as measures of prediction accuracy: the prediction error $PE^{i} = U_{eq}^{i} - \hat{U}_{eq}^{i}$ and the relative prediction error $RPE^{i} = 100\left(PE^{i} \middle/ U_{eq}^{i}\right)$ where U_{eq}^{i} is the observed equilibrated urea from the validation dataset.

In clinical practice, the intradialysis extraction time and the duration of the session may fluctuate. As a consequence, some variation in the input variables $\mathbf{U_{120}}$ and $\mathbf{U_{240}}$ could be expected. To simulate this variation and to evaluate its impact on the prediction, the adjusted model was tested with perturbed input variables ($\mathbf{U_{120}}$ and/or $\mathbf{U_{240}}$). The perturbation was obtained by adding to the original input variables a normal random noise with zero mean and a standard deviation equal to 1% or 5% of overall $\mathbf{U_{120}}$ and $\mathbf{U_{240}}$ means. A RPE obtained from the differences between the prediction with and without perturbation was calculated.

RESULTS

Model Evaluation

By carrying out an analysis of the first PLS factors, it was possible to observe the patient variability in a new orthogonal space, corrected for the multicollinearity among input variables. In Fig. 1, the two **X-scores** or factors with the highest variance are shown. They represented **71%** of the total **X** variation related to **Y**. In the figure, each point represents a patient and the label represents the value of the patient's equilibrated urea. The first factor (the one with the highest variance) positively correlated with the equilibrated urea since the patients with high equilibrated urea values were those with high positive values of the factor. In Fig. 1 the dotted lines split the representation space into meaningful quadrants since they hold patients with different relationships between input variable profiles and the Ueq.

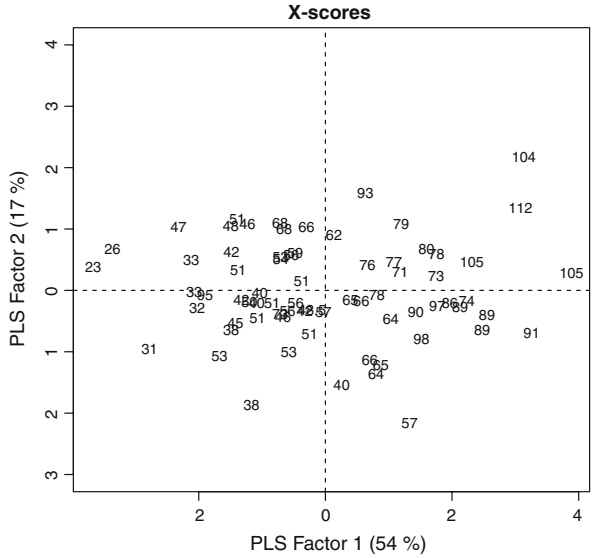

FIGURE 1. The plane spanned by the first two PLS factors of the X-score matrix. Each point represents a patient and the point-label for a given Ueq level.

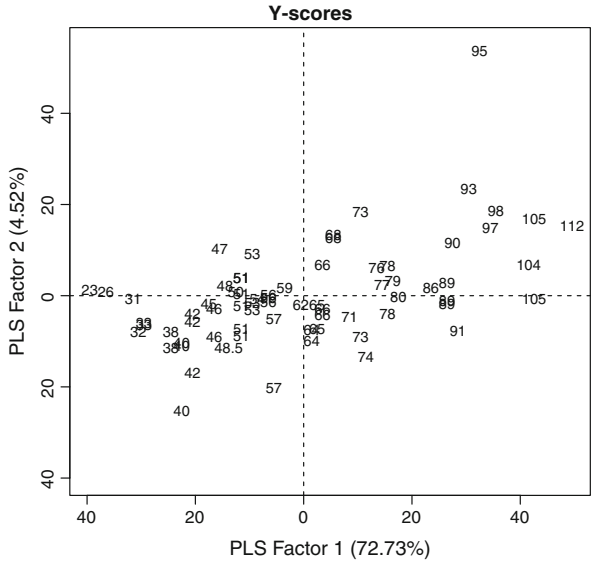

FIGURE 2. The plane spanned by the first two PLS factors of the Y-score matrix. Each point represents a patient and the point-label for a given Ueq level.

Similar to the **X-scores** (Eq. 3), the **Y-scores** (Eq. 4) also provide a view of the observed equilibrated urea in a new, output, space. The first two factors (Fig. 2) hold **77.6%** of the total variance relating **Y** and **X**. The first factor from the **Y-scores** also correlates with Ueq in a

positive way. However, in this space it is easy to observe that one patient's score behaves very differently from the rest (patient with Ueq = 95).

Plotting the first **X-score** against the first **Y-score** (Fig. 3), the linear relationship between both spaces became clear. This observation is in agreement with the linear relationship assumed between **Y** and **X** except for the patient with Ueq = 95. This patient also behaved differently in the **Y-score** space, suggesting that he could be treated as an outlier in **Y** for the linear model.

A new model was fitted leaving out this particular patient. The variation in the input and output variables explained by models with one to five factors is shown for both the dataset containing all patients and the dataset without the patient regarded as an outlier (Table 2). The elimination of the patient with Ueq = 95 only affected the percentage of explained variance of **Y**.

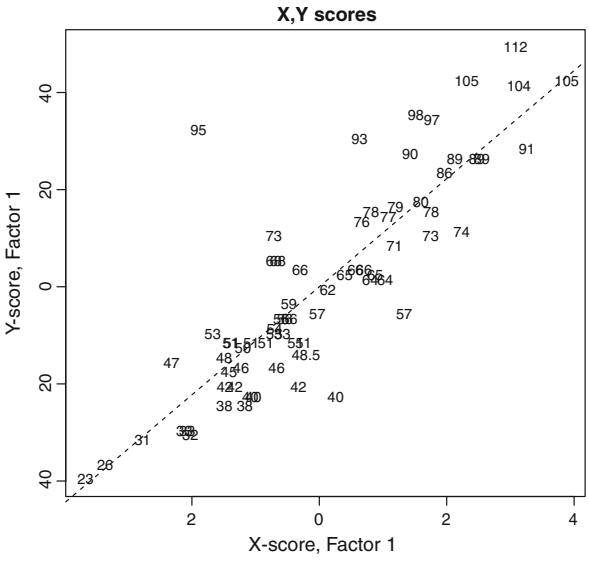

FIGURE 3. The plane spanned by the first PLS factor of the responses (Y) against inputs variables (X).

Model Selection

As stated above, diminishing the number of factors in the model ($A < p$) could benefit the prediction process by avoiding the inclusion of noise in the **B** coefficients. Table 3 shows the RMSEP across all cross-validation runs. The row named "Diff" contained the expected difference in RMSEP between two models: one with "A" factors and the other with "$A + 1$" where $A = 0, 1, 2, 3, 4$. The number of factors for the final model was chosen based on the minimum value of Diff. For both datasets (including or not the patient with Ueq = 95), the model with **3** components was chosen.

Prediction Accuracy

Results in Table 4 show the performance for all the models (the chosen one and the remainders) using data from the validation set. It is important to note that the model with the first three PLS factors generalized very well, achieving a PE of 0.75 ± 7.68 mg/dL and RPE of 0.05 ± 12.95%.

The three PLS factor model yields the following expression to predict equilibrated urea for a specific patient:

TABLE 2. Percent of variability in input and output variables explained by PLS models with one to five factors.

	PLS model trained with all data					PLS model trained without the Ueq = 95 in the training set				
PLS factors	1	2	3	4	5	1	2	3	4	5
Cumulative % of variation for input variables (**X**)	54.47	71.95	83.45	97.59	100	54.35	71.59	83.43	97.64	100
Cumulative % of variation for output variable (**Ueq**)	72.73	77.56	79.32	76.64	79.66	81.44	86.41	88.56	88.91	88.93

Models fitted from the whole dataset and excluding an outlier case (patient with Ueq = 95).

TABLE 3. The RMSEP for Ueq prediction models built with none to five PLS factors.

	All data						PLS model trained without the Ueq = 95 in the training set					
PLS factor	None	1	1 & 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5	None	1	1 & 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5
RMSEP	2.15	1.32	1.10	1.04	1.04	1.05	2.13	1.17	0.85	0.77	0.77	0.78
Diff	0.839	0.22	0.06	0.001	0.006		0.96	0.32	0.08	0.002	0.003	

Models fitted from the whole dataset and excluding an outlier case (patient with Ueq = 95).
Diff: difference between models built with A and $A + 1$ PLS factors.

TABLE 4. RMSEP (mg/dL), PE, and RPE for the validation set with models built with one to five PLS factors.

Factors	1	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5
RMSEP	7.74	7.59	6.98	7.10	7.12
PE ± Sd	0.60 ± 8.53	0.02 ± 8.39	0.75 ± 7.68	0.96 ± 7.79	0.96 ± 7.79
RPE ± Sd (%)	−1.12 ± 14.35	−1.55 ± 14.31	0.05 ± 12.95	0.35 ± 13.02	0.32 ± 13.09

Sd: Standard Deviation.

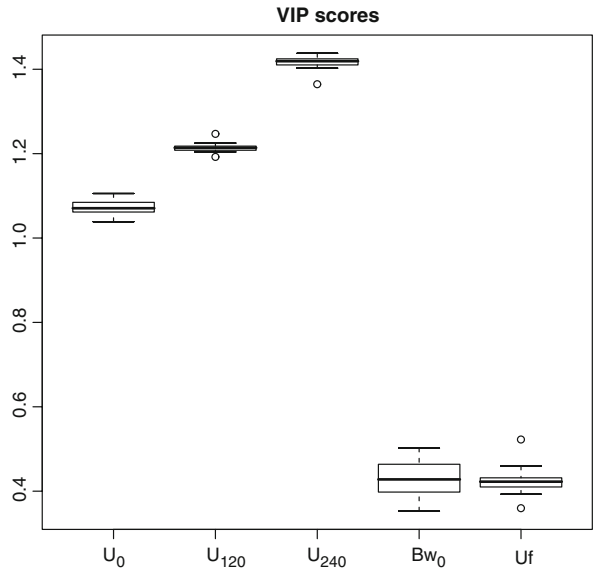

FIGURE 4. VIP scores distributions for predictor variables (U_0, U_{120}, U_{240}, B_w, and Uf) under a cross-validation test.

$$\hat{U}^i_{eq} = 1.1547 + 0.0012 \cdot U^i_0 + 0.2071 \cdot U^i_{120} + 0.8047 \cdot U^i_{240} + 0.0418 \cdot Bw^i_0 - 0.4465 \cdot Uf^i \quad (8)$$

In Fig. 4, the boxplots of the VIP-scores for each input variable are shown. The boxplots represent the empirical distributions of the scores measuring the importance of each variable. They were obtained by a cross-validation procedure.[4] The higher the VIP-score for a variable, the more important the variable is for the prediction.[4] The box-plots also suggested that $\mathbf{U_{240}}$ was the most important variable in explaining the equilibrated urea, followed by $\mathbf{U_{120}}$ and $\mathbf{U_0}$. By means of the cross-validation strategy it was demonstrated that the VIP-scores were stable (small variability) across different training datasets.

The extraction of an intradialysis sample could be troublesome since it needs a trained technician and the implementation of a pump stopping procedure.[22,24] For this reason, a model without the $\mathbf{U_{120}}$ concentration was also built. The summary statistics of the prediction errors for the reduced model (without U_{120}, Eq. 9), with different numbers of PLS factors also suggested that a three PLS factor-based model should be chosen (Table 5).

The three PLS factors-based models yield the following equation:

$$\hat{U}^i_{eq} = 2.2058 + 0.0691 \cdot U^i_0 + 0.9673 \cdot U^i_{240} - 0.042 \cdot Bw^i_0 + 0.543 \cdot Uf^i \quad (9)$$

The model showed above achieved a PE = 0.84 ± 8.05 mg/dL and RPE = 0.04 ± 13.90%, being highly competitive with the full model.

Many authors include in the adequacy equations the normalized **Uf** (**Uf** divided by the $\mathbf{Bw_0}$ weight).[3,5,28] In this study a model with the normalized **Uf** was also fitted by means of a three factor PLS model with the normalized **Uf**:

$$\hat{U}^i_{eq} = 2.9058 + 0.0057 \cdot U^i_0 + 0.1722 \cdot U^i_{120} + 0.858 \cdot U^i_{240} - 0.0257 \cdot Bw^i_0 + 28.3644 \cdot \frac{Uf^i}{Bw^i_0}$$

The model yielded similar results (PE = 1.04 ± 7.78 mg/dL and RPE = 0.45 ± 13.09%.) to previous models. Thus, the normalization of **Uf** seems to be irrelevant in this context.

TABLE 5. Fitting evaluation criteria for models built with one to four PLS factors, removing the U_{120} input variable.

	PLS factors			
Fitting criteria	1	1, 2	1, 2, 3	1, 2, 3, 4
RMSEP	7.58	7.79	7.32	7.36
PE ± Sd	0.73 ± 8.34	0.13 ± 8.61	0.84 ± 8.05	0.98 ± 8.08
RPE ± Sd (%)	−0.94 ± 14.79	−1.55 ± 15.34	0.04 ± 13.90	0.23 ± 13.91

Sd: Standard Deviation.

TABLE 6. Relative Prediction Error under different noise levels for U_{120}, U_{240} input variables independently and simultaneously.

Noise level (%)	U_{120} Mean ± Sd	U_{240} Mean ± Sd	Both Mean ± Sd
1.00	0.00 ± 0.33%	0.07 ± 0.77%	0.20 ± 0.85%
5.00	0.08 ± 1.59%	0.54 ± 3.99%	0.17 ± 4.44%

Sd: Standard Deviation.

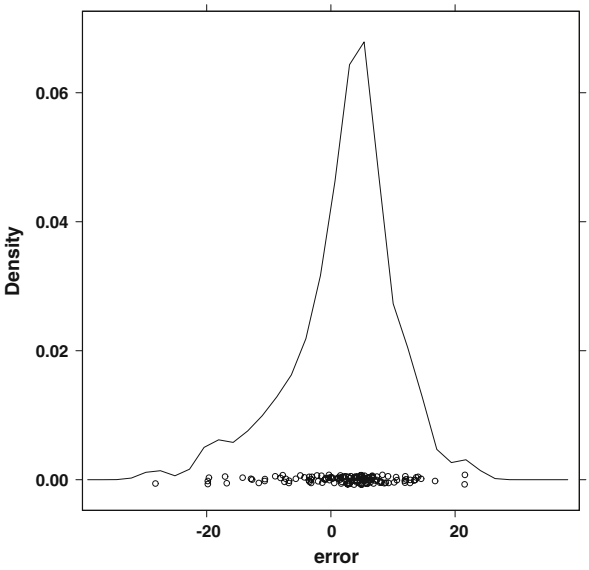

FIGURE 5. Residuals density distribution for PLS model fitted with three PLS factors.

TABLE 7. Evaluation of the PLS model with three factors (Eq. 8) for the mixed and cross-center data analysis.

	Mixed data	Unit 1	Unit 2
PE ± Sd	0.75 ± 7.68	2.56 ± 10.73	0.0 ± 9.91
RPE ± Sd (%)	0.05 ± 12.95	1.95 ± 14.78	–2.68 ± 14.78

Sd: Standard Deviation.

From the data presented in Table 6 it is possible to note that the fitted model was robust to variations in the intradialysis time extraction and session length (average RPE lesser than **0.6**).

The residual density distribution of the fitted model (Eq. 8) is also shown in Fig. 5. The ''normal-like'' distribution suggested a random behavior of them.

In Table 7, the PLS model with three PLS factors (Eq. 8) was compared against a PLS model built with data from Unit 1 and tested with data from Unit 2 and vice-versa. It is possible to observe that the PLS model with three PLS factors was able to generalize well achieving, in the cross-center test, a maximum RPE of 2.68 ± 14.78% (expressed in absolute units).

DISCUSSION

Molecular dynamics occurring during the dialysis process are very complex since they can be affected by different phenomena such as recirculation and volume sequestration.[2,16,17,25] Multiple variables are involved in the underlying process and most of them are patient-dependent. These variables display high variability and co-variability, making their use difficult in the development of linear models by the ordinary least squares (OLS) estimation process. In this paper, we demonstrate an application of Partial least squares regression[30] to obtain a linear statistical model to predict the equilibrated urea concentration. The model was based on correlated hemodialysis variables and anthropometric patient data.

The linear modeling strategy to predict equilibrated molecular concentrations is not very restrictive on biological assumptions. It is simpler and easier to apply than kinetics models which strongly depend on prior knowledge. Although the underlying molecular kinetics could well be non-linear, the linear statistical approach for predicting the equilibrated urea concentration from kinetic and anthropometric patient data may be good enough in terms of practical prediction errors.

PLS regression is a powerful tool for building linear models with correlated input variables such as kinetic studies where patients are observed several times. PLS handles multicollinearity among predictor variables by means of their transformation into a new space (built by the factors). The linear combination coefficients for these transformations are obtained in such a way that the first factors have most of the information (percent of variation) relating the input variables with the output variables. Moreover, the PLS factors are orthogonal among themselves, so the interaction terms between factors do not need to be accounted for in the modeling. In addition, the visual representation of patient variability is free of multicollinearity problems. The plotting of the first factors allows an analysis of patient variability based on the simultaneous study of all predictors and the response. From the analysis of the VIP scores it is also possible to identify those variables with greater contributions to the output explanation. Several statistical approaches can be used to evaluate linear fitting with regard to its postdictive and predictive capabilities. The application of the PLS method is not restricted just to equilibrated urea as a dependent variable since it is general enough to be applied to any other biological molecules with an unknown behavior.

In this study, we have derived a linear statistical model to estimate the molecular equilibrated concentration at the end of the dialysis session without a need

to retain the patients for a longer time to take the equilibrated concentration measure. The VIP analysis suggested that $\mathbf{U_{240}}$ followed by $\mathbf{U_{120}}$ were the most important input variables for **Ueq** prediction. We had previously reached a similar conclusion using an artificial neural network (ANN) approach.[26] Therefore, the VIP scores of sample urea concentrations suggest a correlation over time between urea concentrations. The last measured ureas have the most important role in predicting the equilibrated urea, and this is probably due to the time-dependent kinetic process underlying the dialysis.[13] However, the PLS regression already accounts for this type of correlation.

The selected model used three PLS factors, which suggested that in the new space the problem can be adequately represented with three dimensions. These factors explained **83%** and **79%** of the **X** and **Ueq** variance, respectively. On the other hand, it is highly probable that the residual variance, considering the random distribution of the residuals, could be attributed to inherent noise in the data (Fig. 5).

The selected model with all input variables (predictors) presented in Eq. (8) was very accurate (PE = 0.75 ± 7.68 mg/dL). The model was robust to variations of the input variables, especially those that could be affected by different times of measurement such as $\mathbf{U_{120}}$ and $\mathbf{U_{240}}$. The relative prediction error was small (RPE = 0.05 ± 12.95%) and had a standard deviation lower than that obtained when other prediction approaches were applied to the same dataset.[14,15,26] Thus, on applying non-linear deterministic[27] and ANN[14] approaches to the same dataset to obtain the predicted Ueq for each patient,[10,14] the achieved RPEs were −12.68 ± 34.3% and −1.88 ± 13.46%, respectively.

Some authors have used kinetics or adequacy models with ultrafiltration normalized by body weight.[3,5,6,29] However, such normalization in the fitted linear model did not produce any improvement of the actual prediction accuracy obtained by using the raw ultrafiltration data.

Important medical knowledge has been gained from empirical models using sample data that represent well the target population. These approaches can also be statistically validated in several ways. Here, the generalization capability of the proposed linear model was also tested by means of a cross-unit test. A PLS-based model was built, trained with the patients from Unit1 and tested with the patients from Unit 2, and vice-versa. The relative prediction errors of the PLS-based model built with patients from Unit 1 and Unit 2 were RPE = 1.95 ± 14.78% and RPE = −2.68 ± 14.78%, respectively. As expected, these RPE were slightly higher than those achieved using data from both centers mixed together. However, they were smaller than those obtained from the neural network approach in a cross-unit test.[14]

Using the new space provided by the PLS regression test, data from patients who behaved differently from the rest were easily recognized. This information could be very useful for clinical/treatment analysis of these patients.

The use of an intradialysis sample ($\mathbf{U_{120}}$) provided valuable information to predict the equilibrated urea. Smye *et al.*[27] were among the first to use an intradialysis sample to model Ueq. In clinical practice the extraction of an additional blood urea sample could be very problematic. To overcome this limitation, a PLS-based linear model excluding $\mathbf{U_{120}}$ as predictor can also be accurately used.

It has been reported that the adequacy of the session based on the Kt/V equation, recommended by the National Kidney Foundation Dialysis Outcome Quality Initiative (DOQI) guidelines[22] and the European Renal Association (ERA-guidelines)[24] conferred better results when urea equilibration has occurred or when good estimation is provided.[10–12,14] In this context, we demonstrate that the PLS-based model can provide a reliable estimate of the equilibrated urea, which in turn could be used in dose adequacy evaluation.

ACKNOWLEDGMENTS

This work was partially financed by grants from Catholic University of Córdoba and National Council of Science and Technology from Argentina (PIP CONICET 5338).

REFERENCES

[1]Abdi, H. Partial Least Squares (PLS) regression. In: Sage Encyclopedia of Social Science Research Methods, edited by M. Lewis-Beck, A. Bryman, and T. Futting. Thousand Oaks, CA: Sage Publications Inc., 2003.

[2]Canaud, B., J. Y. Bosc, L. Cabrol, *et al.* Urea as a marker of adequacy in hemodialysis: lesson from in vivo urea dynamics monitoring. *Kidney Int.* 76(S):S28–40, 2000.

[3]Cheng, Y. L., K. S. Chol, K. F. Chau, S. A. Li, A. U. Yung, A. W. Yu, and K. K. Wong. Urea reduction ratio that considers effects of ultrafiltration and intradialytic urea generation. *Am. J. Kidney Dis.* 37:544–549, 2001.

[4]Chong, I. G., and C. H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemomet. Int. Lab. Syst.* 78:103–112, 2005.

[5]Daugirdas, J. Simplified equations for monitoring kt/v, pcrn, ekt/v and epcrn. *Adv. Ren. Replace. Ther.* 2(4):295–304, 1995.

[6]Daugirdas, J. T., T. A. Depner, F. A. Gotch, T. Green, *et al.* HEMO study group 1997, Comparison of methods to

predict equilibrated Kt/V in the HEMO study. *Kidney Int*. 52:1395–1405, 1997.
[7]Depner, T. A. History of dialysis quantification. *Semin. Dial*. 12(S1):14–19, 1999.
[8]Draper, N. R., and H. Smith. Applied Regression Analysis, 3rd edn. USA: John Wiley, 1998.
[9]Eknoyan, G., G. J. Beck, A. K. Cheung, *et al*. Effect of dialysis dose and membrane flux in maintenance hemodialysis. *N. Engl. J. Med*. 347:2010–2019, 2002.
[10]Fernández, E. Data mining and neural networks for dialysis kinetic analysis. PhD Thesis, Santiago de Compostela University, Spain (in Spanish), 2003.
[11]Fernández, E. A., R. Valtuille, J. Presedo, and P. Willshaw. Comparison of different methods for hemodialysis evaluations by means of ROC curves: from artificial intelligence to current methods. *Clin. Nephrol*. 64(3):205–213, 2005.
[12]Fernández, E. A., R. Valtuille, J. Presedo, and P. Willshaw. Comparison of standard and artificial neural network estimators of hemodialysis adequacy. *Artif. Organs* 29(2):159–165, 2005.
[13]Fernández, E. A., R. Valtuille, P. Willshaw, and M. Balzarini. Molecular kinetics modeling in hemodialysis: on-line molecular monitoring and spectral analysis. *ASAIO J*. 53(5):582–586, 2007.
[14]Fernández, E. A., R. Valtuille, P. Willshaw, and C. A. Perazzo. Using artificial intelligence to predict the equilibrated blood urea concentration. *Blood Purif*. 19(3):271–285, 2001.
[15]Ghu, J., J. Yang, I. U. Chen, and Y. Lai. Prediction of equilibrated BUN by an artificial neural network in high efficient hemodialysis. *Am. J. Kidney Dis*. 3:638–646, 1998.
[16]Gotch, F. A. Kinetic modeling in hemodialysis. In: Clinical Dialysis, 2nd edn., edited by A. R. Nissenson, R. N. Fine, and D. Gentile. Norwalk, CT: Appleton and Lange, 1990.
[17]Kaufman, A. M., D. Schneditz, S. Smye, H. D. Polaschegg, and N. W. Levin. Solute disequilibrium and multicompartmental modeling. *Adv. Ren. Replace. Ther*. 2(37):319–329, 1995.
[18]Kennedy, R. L., L. Yuchun, B. Van Roy, C. D. Reed, and R. P. Lippman. Solving Data Mining Problems Through Pattern Recognition. New Jersey: Prentice Hall, 1998.
[19]Locatelli, F. Dose of dialysis, convection and hemodialysis patients outcome – what the HEMO study doesn't tell us: the European viewpoint. *Nephrol. Dial. Transplant*. 18:1061–1065, 2003.
[20]Locatelli, F., T. Hannedouche, S. Jacobson, *et al*. The effect of membrane permeability on ESRD: design of a prospective randomized multicentre trial. *J. Nephrol*. 12:85–88, 1999.
[21]Mevik, B. H., and H. R. Cederkvist. Mean Squared Error of prediction (MSEP) estimates for principal component regression (PCR and partial least squares regression (PLSR). *J. Chemomet*. 18(9):422–429, 2004.
[22]NKF-K/DOQI Clinical Practice Guidelines for Hemodialysis Adequacy: update. *Am. J. Kidney Dis*. 48(S1):S2–90, 2006.
[23]Roa, L. M., and M. Prado. The role of urea kinetic modeling in assessing the adequacy of dialysis. *Crit. Rev. Biomed. Eng*. 32(5–6):461–539, 2004.
[24]Section II. Haemodialysis adequacy. *Nephrol. Dial. Transplant*. 17(S7):16–31, 2002.
[25]Sharma, A., P. Espinosa, L. Bell, A. Tom, and C. Rodd. Multicompartmental urea kinetics in well-dialyzed children. *Kidney Int*. 58:2138–2146, 2000.
[26]Shawe-Taylor, J., and C. N. Kernel. Methods for Pattern Analysis. Cambridge: Cambridge University Press, 2005.
[27]Smye, S., J. Tattersal, and E. Will. Modeling the postdialysis rebound: the reconciliation of current formulas. *ASAIO J*. 45(6):562–569, 1999.
[28]Smye, S. W., E. J. Will, and E. J. Lindley. Postdialysis and Equilibrium urea concentrations. *Blood Purif*. 20:189, 2002.
[29]Tattersal, J., D. Detakats, P. Chamney, R. Greenwood, and K. Farrington. The post dialysis rebound: predicting and quantifying its effect on Kt/V. *Kidney Int*. 50(6):2094–2102, 1996.
[30]Wold, S., M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemomet. Intell. Lab. Syst*. 58:109–130, 2001.