Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Determination of total organic carbon and soluble solids contents in Tanreqing injection intermediates with NIR spectroscopy and chemometrics

Wenlong Li, Haibin Qu*

Pharmaceutical Informatics Institute, Zhejiang University, Hangzhou 310058, China

Abstract

Near infrared spectroscopy combined with chemometrics was investigated for the fast determination of total organic carbon (TOC) and soluble solids contents (SSC) of Tanreqing injection intermediates. The NIR spectra were collected in transfective mode, and the TOC and SSC reference values were determined with Multi N/C UV HS analyzer and loss on drying method. The samples were divided into calibration sets and validation sets using the Kennard–Stone (KS) algorithm. The Dixon test, leverage and studentized residual test were studied for the sample outlier analysis. The selection of wavebands, spectra pretreated method and the number of latent variables were optimized to obtain better results. The quantitative calibration models were established with 3 different PLS regression algorithms, named linear PLS, non-linear PLS and concentration weighted PLS, and the net result was defined as the average of the predicted values of the different calibration models. The overall results indicated that the presented method is more powerful than single multivariable regression method, characterized by higher mean recovery rate (MRR) of the validation set, and can be used for the rapid determination of TOC and SSC values of Tanreqing injection intermediates, which are two important quality indicators for the process monitoring.

1. Introduction

Tanreqing injection is a widely used Chinese patent drug, which has good efficiency in treating upper respiratory infection, and has played a great part in the prevention and treatment of Severe Acute Respiratory Syndrome (SARS) and influenza A (H1N1) [1–3]. It is made from five kinds of Traditional Chinese Medicine (TCM) materials, namely: Radix Scutellariae, Forsythia suspensa, Flos Lonicerae, Bear gall powder, and Cornu gorali. During the manufacture of Tanreqing injection, lots of technological processes and intermediates need online monitoring or rapid analysis, such as extraction, separation, purification, and liquid preparation process, for the process control and real time release (RTR).

Near infrared (NIR) spectroscopy is an ideal process analytical tool characterized by high speed, multi-index simultaneous determination, and chemical reagent free. The NIR spectroscopy is viewed as a green analytical technology and is more and more frequently applied in the quality control and process analysis of TCM productions [4–6]. The test targets include physics, chemistry and biology indices, which were selected according to the specific requirements for the industrial production.

During the NIR spectral analysis, the most important step is to build the calibration models, which correlate the independent variable x(spectral data) and dependent variable y(test targets). Various regression algorithms have been studied, including linear regression methods, such as multiple linear regression (MLR), principal component regression (PCR), as well as nonlinear regression methods, such as artificial neural network (ANN), support vector machine (SVM) and Gaussian process. Among these regression algorithms, partial least-square regression (PLSR) is the mostly widely used one, due to its clear and definite physical meaning, and also the compatibility of linear and nonlinear correlation between the independent variables and dependent variables. In view of the different types of relationship between X and Y, a series of PLS based regression algorithms were developed, such as linear PLS, non-linear PLS, concentration weighted PLS [7–9], and so on. As for the NIR spectral analysis, the linear and non-linear correlations between the independent variables and dependent variables always coexist, in these cases, different kinds of regression models can be established respectively, and the net result can be obtained through the consensus of these models. It can also be viewed as consensus strategy, which differs from the consensus modeling in the traditional sense [10–12].

In this paper, the NIR spectroscopy was adopted for the rapid determination of Total organic carbon (TOC) and soluble solids contents (SSC) in 4 kinds of Tanreqing injection intermediates, and during the
calibration process, different chemometrics methods were applied, and the final result was calculated through the consensus strategy of linear PLSR, non-linear PLSR, and concentration weighted PLSR models, which were established independently.

2. Materials and methods

2.1. Experimental materials

The samples were all collected from a TCM pharmaceutical factory (Shanghai Kaibao Pharmaceutical CO., LTD., Shanghai, China) in the liquid preparation process of Tanreqing injection. In every batch, after the purified TCM extracts were put into the solvent feed tank, 10 ml of the mixture was taken as sample 1, then the pH of the mixture in the tank was adjusted with NaOH solution, after decolored with active carbon, 10 ml of the filtrate was taken as sample 2; the rest filtrate was again filtrated with triple filter, and 10 ml of the filtrate was taken as test sample 3; the rest filtrate was then hyperfiltrated and made into semi-finished product, and 10 ml of semi-finished product was collected as test sample 4. For privacy, the detailed technical parameters are omitted here. The 4 kinds of intermediate test samples above mentioned were collected from the product line of 30 batches intermediates. Samples were scanned with a 2 mm path length and equilibrated at 25 °C for 10 min before scanning to ensure that the samples were analyzed at the same temperature, and each spectrum was obtained by averaging 64 scans. The NIR instrument was controlled by a compatible PC, and a RESULT Operation workstation was used for data acquisition. Under the selected instrument conditions, the spectra of all the samples were collected as shown in Fig. 1.

2.2. Instruments

The NIR spectra of the samples were collected at 4 cm⁻¹ interval over the spectral region 4000–10,000 cm⁻¹ with an Antaris MX FT-NIR System (Thermo Scientific, Madison, USA) equipped with an optical fiber transfectance adapter. Samples were scanned with a 2 mm path length and equilibrated at 25 °C for 10 min before scanning to ensure that the samples were analyzed at the same temperature, and each spectrum was obtained by averaging 64 scans. The NIR instrument was controlled by a compatible PC, and a RESULT Operation workstation was used for data acquisition. Under the selected instrument conditions, the spectra of all the samples were collected as shown in Fig. 1.

2.3. Measurements of the references values

The TOC values of the samples were determined with a Multi N/C UV HS analyzer (Analytik Jena AG, Germany) in NPOC mode. Before the determination, N = 66 samples were diluted 1000 times and digested using persulfate oxidation method with the assistance of a double-wavelength high-intensity ultraviolet lamp. The standard curve was established using series of phthalic acid solutions with different concentrations.

All the computations, including selection of the spectra wavebands, mathematical pretreatment, principal component analysis and partial least squares regression, were performed using TQ analyst software package (Version 8.0, Thermo Scientific, Madison, USA), The Unscrambler (Version 10.1, Camo Inc., Trondheim, Norway), and MatLab (Version 7.13, MathWorks, USA).

3. Results and discussion

3.1. Outlier sample eliminating

Outlier samples are the samples with abnormal spectra or reference values. Several factors may lead to unusable spectra, including the reliability of spectra collecting method, the fluctuation of the testing environment, the small mistake of the operating personnel, and the significant difference of the sample matrix [13]. Dixon’s test and leverage diagnostic method have been proven effective for the outlier identification [14].

In the Dixon’s test, the spectra data were firstly processed with PCA method, and then the Mahalanobis distance between the average spectrum and every spectrum was calculated with the first few PCs. Then the samples were sorted according to the Mahalanobis distance, and the samples having significant differences with the average spectra are considered as outlier samples. In this research, the results of Dixon’s test were shown in Fig. 2. The leverage diagnostic shows the relationship between the leverage and residual values for each component and each sample in a PLSR calibration model. The leverage values reflect how much the standard is influencing the model, and the studentized residual value is the residual value of a spectrum after it has been scaled by the leverage value. The residual values plotted by the Leverance diagnostic have been divided by their standard error to produce the studentized residual values, which places all samples on a similar scale, regardless of their leverage values, and makes it easier to identify standards that may be outliers. Fig. 3 shows how the leverage values compare with the studentized residual values. In the present study, we provisionally viewed the samples with significant differences as outlier samples, and then reinstated them one by one. If the model performance deteriorated after a certain sample was reinstated, then the sample was deemed to be outlier sample, and should be removed from the calibration set. Conversely, the sample was still viewed as a normal one.

After the Dixon’s test and leverage diagnostic test, 4 samples for the TOC, and 6 samples for the SSC had significant differences to the average of the calibration samples, therefore these samples were identified as outliers and were removed from the calibration set. The remaining 66 samples for TOC and 104 samples for SSC were used as normal samples.

3.2. Division of calibration and validation sets

All the remaining samples were split into calibration and validation sets by Kennard–Stone (KS) algorithm [15], the former was used to build regression models and the latter to evaluate model performance. The KS algorithm firstly calculates the Euclidean distances of every two samples, and the two samples with the greatest distances are firstly taken into the calibration set, and then follows a stepwise procedure in which new selections with the largest minimum distance with any sample already existed in the calibration set are taken in, until the number of samples specified by the analyst is achieved. With the algorithm,
about two thirds of the total samples were chosen to form the calibration set while the remaining one third of the total samples went into the validation set. The content ranges of TOC and SSC are listed in Table 1, from which it can be concluded that a calibration set with enough representative was selected from a pool of total samples.

### 3.3. Comparison of different PLSR regression method

In view of both linear and nonlinear relationships between the spectral data and the reference values of TOC and SSC that coexisted, three regression methods were adopted to establish the calibration models, including linear PLS models, non-linear PLS models and the concentration weighted PLS models. Table 2 lists the RMSEC, RMSECV, RMSEP, and related Rs for all the calibration models of TOC and SSC, from which it would be easy to see that the three PLS methods provide similar results. Detail computing formula of above mentioned parameters can be found in Ref. [16].

It is well known that, during the PLS calibration, the most important issues are the selection of wavebands, spectra pretreated method and the determination of the most suitable latent variable (LV) numbers. Therefore the optimization methods are discussed in the next sections.

### 3.4. Selection of wavelengths

Different methods, including Monte Carlo uninformative variable elimination method (MC-UVE) [17], correlation coefficients method [18], were compared. However, the performance parameters of the calibration models were not remarkably improved. So the wavebands with clear physical meaning and high noise–signal ratio were adopted ultimately [19–23], which are 8932.66–4818.64 cm\(^{-1}\) for the TOC calibration model, and 9002.08–4839.28 cm\(^{-1}\) for the SSC calibration model.

### 3.5. Spectra pre-processing approach

As for the spectral data preprocessing, different combinations were compared to optimize the calibration models, including Savitzky–Golay algorithm and wavelets for the noise removing, first-order derivative and second-order derivative for the baseline eliminating, MSC and SNV for the light-scatter effects correcting, and also OSC method [24–27]. The comparison of different spectra pretreated methods were listed in Table 3.

### Table 1

| Quality parameters | Total sets | Calibration sets | Validation sets |
|--------------------|------------|------------------|-----------------|
|                    | Mean | Range | Mean | Range | Mean | Range |
| SSC (%)            | 2.55 | 2.27–2.79 | 2.53 | 2.27–2.79 | 2.58 | 2.27–2.76 |
| TOC (mg/L)         | 55.4 e3 | 52.9–59.7 | 54.9 e3 | 52.9–59.7 | 55.6 e3 | 53.2–59.1 |

Fig. 2. The Dixon’s test result of the Tanreqing injection intermediate samples.

Fig. 3. The leverage and the Studentized residual values of the Tanreqing injection intermediate samples.
3.6. Determination of the optimum latent variable numbers

Due to the multicollinearity among the NIR spectroscopic data, and also the autocorrelation with the dependent variables, the dimensionality reduction is very necessary. In the PLSR calibration, the optimum latent variable (LV) numbers are always determined using Leave-one-out (LOO) cross validation method, in which the spectrum of one sample in the calibration set was removed, and a PLS regression model was built with the remaining spectra in the calibration set, the left-out sample was predicted with this model and the procedure was repeated with leaving out each of the samples in the calibration set. Predicted residual error sum square (PRESS), which represents the sum of square of deviation between predicted and reference values of all the samples in the LOO cross-validation decreases with the increasing of the LVs, and when the PRESS value tends to be constant or increases gradually, the optimum LV numbers is achieved [28]. For both SSC and TOC, the optimum LVs were determined according to the LOO cross-validation method, and the PRESS-LVs correlation diagrams were shown in Fig. 4.

3.7. Calibration, optimization and validation of the models

Using the optimum parameters selected according to the procedures above described, the PLSR quantitative calibration models of TOC and SSC were established. It can be concluded from Table 2 that the coefficients of determination for calibration, cross validation, and external validation are all above 0.9, which show a satisfied correlation between the true values and the model predicted values. For all the 6 models, the RMSEP/RMSECV values are close to 1, which indicates the calibration models have satisfactory robustness. The correlation diagrams and the residuals plots of the PLSR calibration models are shown in Fig. 5.

3.8. Application of the calibration models

In the earlier study beforehand, we have developed a series of NIR spectroscopy based method for the rapid analysis of Tanreqing injection and its intermediates [29–31]. However, the indices of the analysis are mostly organic compounds. Although those active pharmaceutical ingredients (API) influence directly the effects of Tanreqing injection, it would have been otherwise impractical to use them as monitoring control indices, because it is difficult to find the definite link between the operating parameters and them. Instead, SSC and TOC are also two Critical Quality Attributes (CQAs) of the Tanreqing injection intermediates, which indicates the physical property and the chemical composition, respectively. The two CQAs need to be determined promptly and accurately during the liquid preparation process of Tanreqing injection. However, traditional analytical technology cannot meet this demand. The present NIR spectroscopy based methods have been successfully applied on the manufacturing lines, preferably solving the problem.

In the course of industrial applications, we found that the analysis results can be more accurate with the consensus strategy of different PLSR models, in which the net result is defined as the average of the different PLSR models. To prove this, the mean recovery rate (MRR),

### Table 2

| QIs | Regression methods | LVs | Calibration | Cross-validation | Prediction |
|-----|--------------------|-----|-------------|------------------|-----------|
|     |                    |     | \( R_C \)   | \( R_{CV} \)     | \( R_P \)   | \( RMSEC \) | \( RMSECV \) | \( RMSEP \) |
| TOC | PLS                | 9   | 0.9975      | 0.9491 e3        | 0.9680 e3  | 0.9100 e3 |
|     | WPLS               | 9   | 0.9870      | 0.9487 e3        | 0.9507 e3  | 0.385 e3  |
|     | NPLS               | 9   | 0.9801      | 0.622 e3         | 0.379 e3   |           |
| SSC | PLS                | 10  | 0.9918      | 0.628 e3         | 0.9335     | 0.0413    |
|     | WPLS               | 10  | 0.9905      | 0.5578 e3        | 0.9509     | 0.0431    |
|     | NPLS               | 10  | 0.9908      | 0.9464 e3        | 0.9519     | 0.0428    |

### Table 3

| Pretreated methods | TOC     | SSC     |
|--------------------|---------|---------|
| \( R_C \) | \( RMSEC \) | \( R_P \) | \( RMSEP \) | \( R_C \) | \( RMSEC \) | \( R_P \) | \( RMSEP \) |
| Raw spectra        | 0.9999  | 21.1    | 0.9248   | 0.784e3   | 0.9612   | 0.0377   | 0.9044   | 0.0498   |
| OSC                | 0.9477  | 0.188e3 | 0.9630   | 0.408e3   | 0.9001   | 0.0212   | 0.9289   | 0.0407   |
| Wavelet + 1d       | 0.9322  | 0.129e3 | 0.9083   | 0.499e3   | 0.9218   | 0.0205   | 0.9387   | 0.0593   |
| Wavelet + 2d       | 0.9778  | 0.192e3 | 0.9365   | 0.582e3   | 0.9701   | 0.0222   | 0.9054   | 0.0613   |
| SNV + SG(7.3) + 1d | 0.9822  | 0.360e3 | 0.4678   | 0.167e4   | 0.6802   | 0.1000   | 0.1906   | 0.1200   |
| MSC + ND(15,5) + 2d| 0.9975  | 0.134e3 | 0.5689   | 0.410e3   | 0.9918   | 0.0172   | 0.9335   | 0.0413   |

* Bias = −0.0202.
which can be calculated with the following formula, was proposed to indicate the prediction accuracy of the individual models and the consensus strategy of different PLSR models.

\[
MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{predict value}}{\text{actual value}} \quad (1)
\]

Take the validation samples as example, the MRRs are listed in Table 4, from which, it can be concluded that the MRRs of the consensus strategy are always closer to 1, except the liner PLSR model of TOC. Therefore, in practical applications, after collecting the spectra of the samples, the TOC and SSC values are calculated separately with the three PLSR models, and then analysis of variance (ANOVA) is done to test whether there are significant differences among them. If no, the average values of the three models can be viewed as the final result, while if there are significant differences among them, standard analysis procedure should be done to obtain the standard values.

4. Conclusion

In the present study, NIR spectroscopy based method was developed for the rapid analysis of Tanreqing injection intermediates. The overall results indicated that NIR spectroscopy had the capability to determine the SSC and TOC, and with consensus modeling strategy, the determination results can be more accurate and reliable. TOC and SSC are two important quality indices in most liquid product lines of TCM, therefore, the presented method is fairly effective and deserves popularization.

Conflict of interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

Financial supports are from the Zhejiang Traditional Medical Science and Technology projects (2015ZB023), and China Postdoctoral Science Foundation Special Funded Project (2013T60604). The authors would also like to acknowledge the cooperators from Kaibao Pharmacy Co., Ltd., for supporting the works and for permission to publish the above results.

References

[1] T. Wu, X. Yang, X. Zeng, P. Poole, Traditional Chinese medicine in the treatment of acute respiratory tract infections, Respir. Med. 103 (2009) 102–104.
[2] http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohyzs/s3585/200905/40478.htm
[3] H. Guo, G. Zheng, J. Zhao, M. Jiang, X. He, A. Lu, Traditional Chinese medicine therapeutic characteristics in the treatment of Influenza A virus subtype H1N1 based on data mining, World Science and Technology-Modernization of Traditional Chinese Medicine, 772 2011, pp. 13–16.
[4] H. Huang, H. Qu, In-line monitoring of alcohol precipitation by near-infrared spectroscopy in conjunction with multivariate batch modeling, Anal. Chim. Acta 707 (2011) 47–56.
[5] H. Xiong, X. Gong, H. Qu, Monitoring batch-to-batch reproducibility of liquid-liquid extraction process using in-line near-infrared spectroscopy combined with multivariate analysis, J. Pharm. Biomed. Anal. 70 (2012) 178–187.
[6] B. Xu, Z.S. Wu, Z.Z. Lin, C.L. Sui, X.Y. Shi, Y.J. Qiao, NIR analysis for batch process of ethanol precipitation coupled with a new calibration model updating strategy, Anal. Chim. Acta 720 (2012) 22–28.
[7] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction, Chemom. Intell. Lab. Syst. 88 (2007) 183–188.

[8] Y. Wang, H. Cao, Y. Zhou, Y. Zhang, Nonlinear partial least squares regressions for spectral quantitative analysis, Chemom. Intell. Lab. Syst. 148 (2015) 32–50.

[9] L. Xu, J. Jiang, H. Wu, G. Shen, R. Yu, Variable-weighted PLS, Chemom. Intell. Lab. Syst. 85 (2007) 140–143.

[10] P. Shahbazikhah, J.H. Kalivas, A consensus modeling approach to update a spectroscopic calibration, Chemom. Intell. Lab. Syst. 120 (2013) 142–153.

[11] K. Liu, X. Chen, L. Li, H. Chen, X. Ruan, W. Liu, A consensus successive projections algorithm—multiple linear regression method for analyzing near infrared spectra, Anal. Chim. Acta 858 (2015) 16–23.

[12] J. Bocard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Anal. Chim. Acta 769 (2013) 30–39.

[13] Z.C. Liu, W.S. Cai, X.G. Shao, Outlier detection in near-infrared spectroscopic analysis by using Monte Carlo cross-validation, Sci. China, Ser. B: Chem. 51 (2008) 751–759.

[14] H.Y. Yu, Methods for Quality and Wine Age Determination of Chinese Rice Wine Based on Near Infrared Spectroscopy, Zhejiang University, Hangzhou, 2007.

[15] L. Ramirez-Lopez, K. Schmidt, T. Behrens, B. van Wesemael, J.A.M. Dematté, T. Scholten, Sampling optimal calibration sets in soil infrared spectroscopy, Geoderma 226–227 (2014) 140–150.

[16] C.G. Chan, C.C. Chu, D.K.W. Mok, F.T. Chau, Analysis of berberine and total alkaloid content in Cortex Phellodendri by near infrared spectroscopy (NIRS) compared with high-performance liquid chromatography coupled with ultra-visible spectroscopic detection, Anal. Chim. Acta 592 (2007) 121–131.

[17] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2008) 188–194.

[18] W. Li, Z. Cheng, Y. Wang, H. Qu, Quality control of Lonicerae Japonicae Flos using near infrared spectroscopy and chemometrics, J. Pharm. Biomed. Anal. 72 (2013) 33–39.

[19] H.Y. Cen, Y. He, Theory and application of near infrared reflectance spectroscopy in determination of food quality, Trends Food Sci. Technol. 18 (2007) 72–83.

[20] W.L. Li, H.B. Qu, Rapid quantification of phenolic acids in Radix Salviae Miltiorrhiza extract solutions by FT-NIR spectroscopy in transreflective mode, J. Pharm. Biomed. Anal. 52 (2010) 425–431.

[21] F. Liu, Y. He, L. Wang, Comparison of calibrations for the determination of soluble solids content and pH of rice vinegars using visible and short-wave near infrared spectroscopy, Anal. Chim. Acta 610 (2008) 196–204.

[22] M.C. Alamar, E. Bobelyn, J. Lammertyn, B.M. Nicolaï, E. Moltó, Calibration transfer between NIR diode array and FT-NIR spectrophotometers for measuring the soluble solids contents of apple, Postharvest Biol. Technol. 45 (2007) 38–45.

[23] G.M. Vasques, S. Grunwald, J.O. Sickman, Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra, Geoderma 146 (2008) 14–25.

[24] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing? Trends Anal. Chem. 50 (2013) 96–106.

[25] E. Szymańska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L.M.C. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, Trends Anal. Chem. 69 (2015) 34–51.

[26] A. Rinnan, F.V.D. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends Anal. Chem. 28 (2009) 1201–1222.

[27] S. Wold, H. Antti, F. Lindgren, J. Ohman, Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.

[28] X. Li, Y. He, C. Wu, D. Sun, D. Nondestructive measurement and fingerprint analysis of soluble solid content of tea soft drink based on Vis/NIR spectroscopy, J. Food Eng. 82 (2007) 316–323.

[29] W. Li, L. Xing, L. Fang, J. Wang, H. Qu, Application of near infrared spectroscopy for rapid analysis of intermediates of Tanreqing injection, J. Pharm. Biomed. Anal. 53 (2010) 350–358.

[30] W. Li, Z. Cheng, Y. Wang, H. Qu, Feasibility research on noninvasive analysis of Tanreqing injection with near infrared spectroscopy, J. Near Infrared Spectrosc. 20 (2012) 667–674.

[31] W. Li, Z. Cheng, Y. Wang, H. Qu, A study on the use of near-infrared spectroscopy for the rapid quantification of major compounds in Tanreqing injection, Spectrochim. Acta A Mol. Biomol. Spectrosc. 101 (2013) 1–7.