第 37 卷第 12 期
2012 年 6 月

中国中药杂志
China Journal of Chinese Materia Medica

Vol. 37, Issue 12
June, 2012

样本数据重复性对 NIR 校正模型的影响

隋丞琳，吴志生，林兆洲，徐冰，杜敏，史新元*，乔延江*
（北京中医药大学 中药信息工程研究中心，北京 100102）

［**摘要**］ **目的**：探讨样本数据重复采集对所构建近红外（NIR）定量校正模型稳健性的影响，初步阐释该影响产生的原因。**方法**：以银黄液为研究载体，采集样本的近红外光谱，并以高效液相测定值为参考值，采用偏最小二乘算法建立黄芩苷定量校正模型，对潜变量因子累积贡献曲线进行深入探讨，在潜变量空间阐述重复采样对所建立的定量校正模型的影响。**结果**：在对重复采集光谱平均后，以最优光谱预处理方法建立的定量预测模型达到理想预测结果（RMSECV = 1.824）。该模型潜变量因子累计贡献率曲线下的面积，明显大于其他光谱建模方式，即所得的模型更加稳健。**结论**：多次测量取平均能够显著提高模型的预测性能，使所得的模型更加稳健。

［**关键词**］ 近红外光谱；偏最小二乘法；重复性样本

近红外光谱（near infrared spectroscopy，NIRS）分析法是利用近红外谱区（波长为 780 ~ 2 526 nm）携带的物质信息，对物质定性、定量的一种分析方法。作为一种间接的分析技术，近红外分析法需采用化学计量学等手段提取有用信息，建立光谱特征与待测量之间的校正模型，才能实现对于未知样本的定性或定量分析[1]。构建稳健的校正模型是获得良好预测结果的根本。

建模样本集能够代表总体样本在应用过程中所涉及的所有信息，应涵盖足够的有效信息和背景信息[2]。但建模样本集包含的样本数并不是越多越好[3]，因为模型中每增加一个样本，在增加有效信息的同时也增加了干扰信息[4]，使得引入误差的机会提高。

本研究以清开灵注射液中间体银黄液为载体，探讨建模样本重复性对模型的干扰问题，即在相同近红外光谱条件下，同一操作者在极短时间间隔内，平行测量同一样本，对该样本采用不同的平行测量数据进行建模对于模型稳定性的影响。

1 材料

Antaris 傅里叶变换 NIR 光谱仪（美国 Thermo Nicolet 公司）及 TQ Analyst 8.0 数据处理软件；Agilent 1100 高效液相色谱仪，二极管阵列检测器，ChemStation 色谱工作站，SunFire C_{18} 色谱柱（4.6 mm × 150 mm，5 μm，Waters）。

银黄液（批号分别为 011402A，011403A，011601A，011608A，011706A，011708A，北京中医药大学药厂）；对照品黄芩苷（中国食品药品检定研究院，批号 110715-201016）；甲醇（色谱纯，美国 Fisher 公司）；水为自制纯净水；磷酸（分析纯，北京化工厂）。

2 方法

2.1 样品的制备　取银黄液 6 批次，蒸馏水分别稀释成以黄芩苷计 1，2，3，4，5，6 g · L^{-1} 6 个质量浓度水平，共 36 个银黄液样本。

根据银黄液样本所含黄芩苷的 HPLC 分析结果，选取样本中 3 个质量浓度（分别为 1，2，6 g · L^{-1}）作为校正集（也称训练集），用于建立校正模型。剩余 3 个质量浓度（分别为 3，4，5 g · L^{-1}）作为验证集，对模型进行外部验证。

2.2 HPLC 测定黄芩苷含量　按 2010 年版《中国药典》清开灵注射液项下的高效液相色谱法测定黄芩苷含量的方法配制对照品溶液[5]。

精密移取样品适量，置 100 mL 量瓶中，加 70% 乙醇稀释至刻度，摇匀，滤过，取续滤液，既得（每 1 mL 中约含黄芩苷 50 μg），即精确测定 **2.1** 项下 36 个样本浓度，作为 NIR 的参考值。

采用 SunFire C_{18} 色谱柱（4.6 mm × 150 mm，5

［稿件编号］ 20111205022
［基金项目］ 国家“重大新药创制”科技重大专项（2010ZX09502-002）；北京市支持中央在京高校共建项目
［通信作者］ * 乔延江，教授，Tel：（010）84738621，E-mail：yjqiao@263.net；* 史新元，E-mail：shixinyuan01@163.com

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

μm)，流动相甲醇-水-磷酸(47∶53∶0.2)，流速 1 mL·min^{-1}，柱温 30 ℃，检测波长 276 nm，进样量 10 μL。理论塔板数以黄芩苷峰计算，应不低于 3 000。

2.3 NIR 光谱采集方案　取银黄液样本，混合均匀，以内置背景作为参比。为提高方法的重复性，降低偶然因素对全局测量结果的影响，充分确保实验结果对于终端操作者的适用性[6]。采样过程分 2 d 完成，每天分别由 2 个操作者对所有样品进行光谱采集。

2.4 NIR 光谱采集条件　采用透射方式采集光谱，以仪器内置背景作为参比，光谱采集范围 10 000 ~ 4 000 cm^{-1}，分辨率 8 cm^{-1}，扫描次数 32 次，每个样本平行测定 3 次(大部分文献中光谱平均测量次数多为 3 次)，温度 25 ℃，相对湿度 35%。所得近红外原始光谱见图 1。

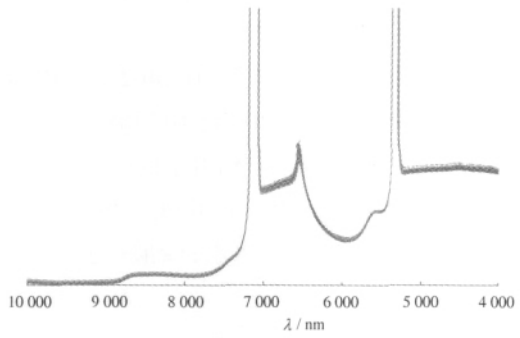

图 1　近红外原始光谱图

Fig. 1　Original NIR spectra

3　结果与讨论

3.1 校正模型的建立　运用 TQ Analyst 8.0 光谱分析软件，以偏最小二乘法(PLS)建立黄芩苷含量的校正模型，并用验证集样品进行外部验证。

建模样本选取时，采用如下 4 种方式：①3 条平行采集光谱全部用于建模；②3 条平行采集光谱的平均光谱建模，下文简称平均光谱建模；③3 条平行采集光谱中的任意 2 条光谱建模，下文简称任意 2 条光谱建模；④3 条平行光谱中的任意 1 条光谱建模。

3.2 光谱预处理方法考察　在建立模型前，对样品的原始吸收光谱进行预处理，以消除噪音、基线漂移、降低重叠峰的影响等，提高模型的预测精度，使所得模型更加稳健。本文比较了一阶导数、标准正则变换(standard normal variate, SNV)、savitzky-golay(9∶2)平滑法(下文简称 SG 平滑)等光谱预处理方法对模型性能的影响。以校正集样品的交叉验证均方差(root mean square error of cross validation, RMSECV)为指标优化建模参数，以留一交叉验证法选择合适的潜变量因子数；以预测误差均方根(root mean square error of prediction, RMSEP)考察模型对未知样本的预测效果。RMSECV 及潜变量因子数相对较小者所得预处理方法最优，见表 1。该数据经 TQ Analyst 8.0 数据处理软件计算获得。

表 1　4 种建模方式下不同预处理方法所得 PLS 模型参数

Table 1　Parameters of PLS models based on different pretreatment methods under 4 building model styles

建模方式	光谱预处理	Raw	SG 平滑	1st	1st + SG	1st + SG + SNV
平均光谱	RMSECV	5.17	4.36	7.49	1.15	0.992
	RMSEP	6.81	6.46	7.36	0.952	1.09
	LVs	9	10	8	9	9
全部光谱	RMSECV	4.39	4.77	6.55	0.566	0.599
	RMSEP	5.5	5.94	7.09	0.646	0.824
	LVs	12	7	12	11	10
任意光谱	RMSECV	5.29	4.41	7.61	0.719	0.77
	RMSEP	6.7	6.7	8.29	0.69	0.696
	LVs	10	11	4	11	7
任意 2 条光谱	RMSECV	4.55	4.21	6.33	0.604	0.659
	RMSEP	7.16	7.51	8.07	4.35	4.49
	LVs	11	11	13	11	10

注：1st. 一阶导数；Raw. 原始光谱；SG. SG 平滑；SNV. 标准正则变换。

由表 1 可知，对于银黄液样本，尽管样本选取方式有所不同，但分别采用一阶导数-SG 平滑和一阶导数-SG 平滑-SNV，这 2 种光谱预处理方法所得结果均较其他方法理想。

3.3 建模波段的选择　为剔除冗余信息，改善模型的预测能力，缩减运算时间，在建立近红外光谱校正模型前需对光谱波段进行筛选。经一阶导数-SG 平滑法和一阶导数-SG 平滑-SNV 法处理后，银黄液的近红外光谱在 6 005.25 ~ 5 395.85 cm^{-1}，6 471.93 ~ 6 009.10 cm^{-1} 及 7 050.47 ~ 7 015.76 cm^{-1} 附近光谱吸收差异相对较大，见图 2，包含了较多的过程信息，所以，确定上述 3 个波段为银黄液的建模波段。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

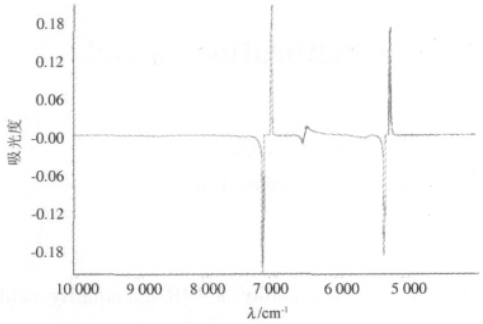

图 2　一阶导数-SG 平滑近红外光谱图

Fig. 2　NIR spectra of first derivative-SG smooth

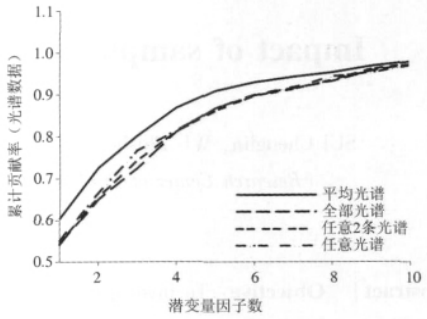

图 3　累计贡献率随不同潜变量因子数变化图

Fig. 3　Curve of cumulative percent with variable factors

3.4　不同建模方式结果评价　累计贡献率是前 m 个相互正交综合指标之和，一般当其大于 80% 时，则可保留前 m 个综合指标，即潜变量因子数。该参数可用于选择潜变量因子数，及模型预测能力的评价[8]。当潜变量因子数为 6 时，累计贡献率（本文中指光谱数据的方差贡献）均可达 90% 以上，见图 3，即已能够代表原始光谱数据中的绝大部分信息[7]，且平均光谱建模所得累计贡献率曲线潜变量因子为 6 ~9 时，图中位于下方的 3 种建模方式所得累计贡献率曲线缓慢上升，且幅度稍大于平均光谱建模，并呈现出向平均光谱靠拢的趋势，说明潜变量因子数进一步增加将涵盖一些光谱中的局部特征信息，当潜变量因子数达到 10 时，4 种建模方式的累计贡献率基本一致且较为稳定，达 97% 以上。故平均光谱建模的曲线下面积大于其他 3 种方式建模的曲线下面积。当潜变量因子数小于 6 时，包含所建定量模型的主要变异，从曲线的走势可以看出，平均光谱建模所得累计贡献率曲线明显较高，而其他 3 种建模方式的累计贡献率均低于平均光谱建模。可知，采用平均光谱所建模型具有较好的稳定性，且优势较其他方式显著。当潜变量因子数继续增加，曲线渐缓，即已包括构建定量模型之外的信息，而这些信息可能是操作天数，操作人员等随机变异，这些潜变量因子的引入可能会使模型出现过拟合现象，降低模型的预测能力。

由此结果可进一步获知，由于电子分析仪器一般由光路系统、电路系统等组成，内部结构复杂。仪器在正常运转条件下，也会受到各种条件（仪器环境、人员操作、系统间作用等）的影响，以致所得信号存在一定的随机误差。因此，求得重复测定数据平均值能够降低该类误差对结果的影响，保证实验结果的可重复性。

4　结论

本研究确证平均光谱建模的必要性及合理性，探索近红外光谱数据建模方式对所建预测模型的影响。

对同一样本进行平行近红外光谱采集，所得光谱与光谱之间不完全相同，存在一定差异，该差异会使建模结果受到轻微扰动。若期望克服重复样本所引入的模型扰动，可对平行采集光谱进行平均，该处理能够降低建模光谱数据内部差异，进而增强模型稳定性，改善并提高模型的预测能力。同时，能够适当降低由于仪器内部原因及操作程序等所造成的随机误差，提高实验结果可重复性。

[参考文献]

[1]　Yves Roggo, Pascal Chalus, Lene Maurer, et al. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies [J]. J Pharm Biomed Anal, 2007, 44: 698.

[2]　Shao Xueguang, Bian Xihui, Liu Jingjing, et al. Multivariate calibration methods in near infrared spectroscopic analysis [J]. Anal Methods, 2010, 2 (11): 1664.

[3]　Martens H A, Dardenne P. Validation and verification of regression in small data sets [J]. Chemom Intell Lab Syst, 1998, 44: 110.

[4]　刘绪平，胡昌勤，田克仁，等. 近红外光谱定量分析样本选择理论研究进展 [J]. 药物分析杂志，2010，30 (7) : 1340.

[5]　中国药典. 一部 [S]. 2010: 1111.

[6]　Max Feiberg. Validation of analytical methods based on accuracy profile [J]. J Chromatogr A, 2007, 1158: 178.

[7]　Özgür Yeniay, Atilla Gökas. A comparison of partial least squares regression with other prediction methods [J]. Hacet J Math Stat, 2002, 31: 106.

[8]　Saikat Maitra, Jun Yan. Principle component analysis and partial least squares: two dimension reduction techniques for regression [C]. Arlington: Casualty Actuarial Society, Discussion Paper Program, 2008: 79.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Impact of sample data repeatability on NIR calibration model

SUI Chenglin, WU Zhisheng, LIN Zhaozhou, XU Bing, DU Min, SHI Xinyuan*, QIAO Yanjiang*
(*Research Center of Traditional Chinese Medicine Information Engineering, Beijing University of Chinese Medicine, Beijing* 100102, *China*)

[**Abstract**] **Objective**: To investigate the impact of repeated data acquisition on the stability of NIR quantitative calibration model, and make a preliminary analysis on reasons for the impact. **Method**: Yinhuang decoction was used as the subject, and NIR spectrum samples were collected. By reference to HPLC´s determination value, the baicalin quantitative calibration model was established by using recursive least square algorithm to detect cumulative-LVs curve of latent variables. The impact of calibration model caused by repetitive samples was explained in latent variance space. **Result**: After averaging the repetitive spectrum samples, quantitative prediction model, which was built by optimal method of spectrum pretreatment, showed the ideal prediction result (RMSECV = 1.824). The area under the cumulative-LVs curve of latent variables was obviously larger than other modeling methods, i. e., this model is more stable. **Conclusion**: Averaging of multiple measurements can dramatically improve the predictive ability of the model and make the model more stable.

[**Key words**] near-infrared spectroscopy (NIRs); partial least square (PLS); repetitive samples

doi: 10.4268/cjcmm20121213

[责任编辑 马超一]

封面图片简介

穿心莲 *Andrographis paniculata*(Burm. f.) Nees,别名榄核莲、一见喜。爵床科一年生草本,高 50 ~ 100 cm,全株味极苦。茎直立,多分枝,具四棱,节稍膨大。叶对生,卵状矩圆形至矩圆形披针形,长 2 ~ 11 cm,宽 0.5 ~ 2.5 cm,先端渐尖,基部楔形,全缘或浅波状,上面深绿色,下面灰绿色,侧脉 3 ~ 4 对;叶柄短或近无柄。圆锥花序顶生或腋生;花萼 5 深裂,外被腺毛;花冠淡紫白色,唇形,上唇外弯,2 齿裂,下唇直立,3 浅裂,雄蕊 2,药室一大一小,大的被髯毛,花丝一侧有柔毛;子房 2 室。蒴果长椭圆形,长约 1.5 cm,两侧呈压扁状,中央具一纵沟。花期 8 ~ 9 月,果期 10 月。

药用干燥的地上部分。味苦,性寒。归心、肺、大肠、膀胱经。清热解毒,凉血,消肿。用于感冒发热,咽喉肿痛,口舌生疮,顿咳劳嗽,泄泻痢疾,热淋涩痛,痈肿疮疡,毒蛇咬伤。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net