Comparison of clinical studies: near infrared predictions of multiple analytes in human sera

Marilyn R. Gatin, James R. Long, Paul W. Schmitt, Paul J. Galley and John F. Price

Boehringer Mannheim Corporation, 9115 Hague Road, Indianapolis, IN 46250, USA.

Introduction

In Studies A and B, near infrared (NIR) spectra and analyte reference data were collected on serum samples from a mixed diabetic/non-diabetic population. The objective of Studies A and B was to compare calibrations developed from spectral data sets collected by different operators under the same experimental conditions but separated in time. The calibration comparison would thus provide valuable information on spectrometer stability and calibration robustness over a period of several months. Applying the model developed for one study to predict results from another study would also indicate the degree of difficulty expected in the future development of robust calibrations.

The primary target for these Studies was glucose in serum, a low concentration analyte. In addition, other larger concentration analytes were studied: total protein, albumin, globulin, cholesterol and triglycerides. Instrumental problems were encountered, identified and addressed in this comparison by use of Residuals Analysis and Outlier Detection analysis.[1] Once the data sets were screened to eliminate data collected under abnormal conditions, a fair comparison between the two data sets could be made.

Experimental

Serum samples were collected from 20 non-diabetics and 100 diabetics in each study. The two Studies were geared toward a multivariate prediction of glucose and, as such, each study was designed to have an even distribution of glucose concentrations across the range. Pre-selection of diabetic study participants was done from Accu-Chek III capillary blood glucose measurements. Type I and Type II diabetics were equally distributed, each comprising *c.* 40% of the population. Over 70% of the study populations were over the age of 60 years and *c.* 25% were under 50 years of age. This population is consistent with an aging diabetic population.

Near infrared spectra were collected on each fresh serum sample using an NIRSystem 6500 spectrometer (Perstorp Analytical, Inc.) using a temperature controlled (37ºC) 1 mm flow cell in stopped-flow mode. Transmittance spectra (100 co-added scans) were collected at 2 nm resolution over 1100–2500 nm range and referenced to a 0.9% saline solution. The detector gain was optimized for the upper wavelength region, therefore, the usable spectral region was limited to *c.* 1300–2500 nm. Two operators were used for the collection of spectral data, one for each study. Data collection in each study was spread over five weeks with six months separation between the two collection periods.

A panel of reference values was collected daily on a BM/Hitachi 717 Analyzer (Boehringer Mannheim Corporation). The clinical panel included serum glucose, total protein, albumin, globulin, cholesterol (HDL) and triglycerides. Several other clinically significant analytes were included in the panel but were not used in the development of calibration models. The clinical results were consistent with clinical data on an aging diabetic population.

Leave-one-out cross-validation PLS was performed on the data in both the mid and upper NIR regions for each analyte. The leave-one-out Standard error of validation (*SEV*) values from the individual Studies were used as the basis for comparison with the standard error of prediction (*SEP*) values for the cross-prediction between the two Studies. The *SEV* values describe the errors associated with building a calibration model (intra-study) while the *SEP* values describe the errors associated with predicting from that model from an independent data set (inter-study). As discussed more fully by J.R. Long *et al.*,[1] Study B data set was truncated based on the results of Residual Analysis and Outlier Detection.

Results

After Outlier Detection, 115 of Study A samples were retained. Truncation of Study B was quite severe in that only 57 of the original 120 samples remained. The range of values for each analyte can be seen in Table 1. In each study the analytes covered a wide range of clinical values and the ranges between the two studies were similar in most cases. Intra- and inter-study prediction performance statistics are listed in Table 2 for each analyte.

Glucose

The inter-study performance for glucose, in the case of Study A calibration model applied to Study B data, is seen in Figure 1 comparing predicted versus reference glucose values. The *SEP* (27 mg dL^{-1}) for Study B was lower than the *SEV* (30 mg dL^{-1}) from Study A. Study A had two samples outside of the Study B range that may have contributed to the higher *SEV* in Study A and the lower *SEP* for Study B.

Study B, in spite of having only 57 samples, performed as well as Study A, which had twice as many samples. The *SEV* of 25 mg dL^{-1} is significantly better than the *SEV* for Study A and when applied to Study A had an equal *SEP* of 27 mg dL^{-1} (see Figure 2). Performance of the Study B calibration model is noteworthy. For glucose, where it was possible to pre-select study

Table 1. Analyte ranges for human sera in Studies A and B.

Study	A	B
# Samples	115	57
Glucose range (mg dL^{-1})	40–530	37–419
Total Protein range (g dL^{-1})	5.7–8.2	6.2–8.3
Albumin range (g dL^{-1})	2.5–4.5	3.1–4.4
Globulin range (g dL^{-1})	2.3–4.6	2.7–4.6
Cholesterol range (mg dL^{-1})	109–348	119–360
Triglycerides range (mg dL^{-1})	42–1379	53–1528

Table 2. Results of intra- and inter-study comparison for several analytes in human sera.

Analyte	Calibration Set	Test Test	*SEV* (mg dL^{-1})	*SEP* (mg dL^{-1})
Glucose	A	A	30	—
	A	B	—	27
	B	B	25	—
	B	A	—	27
Total Protein	A	A	0.28	—
	A	B	—	0.29
	B	B	0.31	—
	B	A	—	0.28
Albumin	A	A	0.18	—
	A	B	—	0.16
	B	B	0.15	—
	B	A	—	0.18
Globulin	A	A	0.21	—
	A	B	—	0.27
	B	B	0.26	—
	B	A	—	0.23
Cholesterol	A	A	11	—
	A	B	—	13
	B	B	12	—
	B	A	—	16
Triglycerides	A	A	18	—
	A	B	—	27
	B	B	21	—
	B	A	—	32

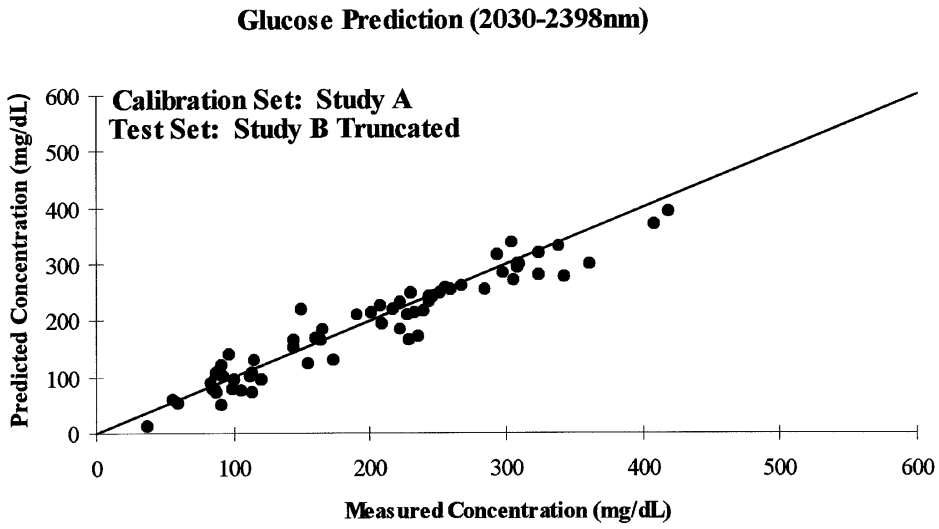

Figure 1. Prediction of glucose in Study B data set from Study A calibration model.

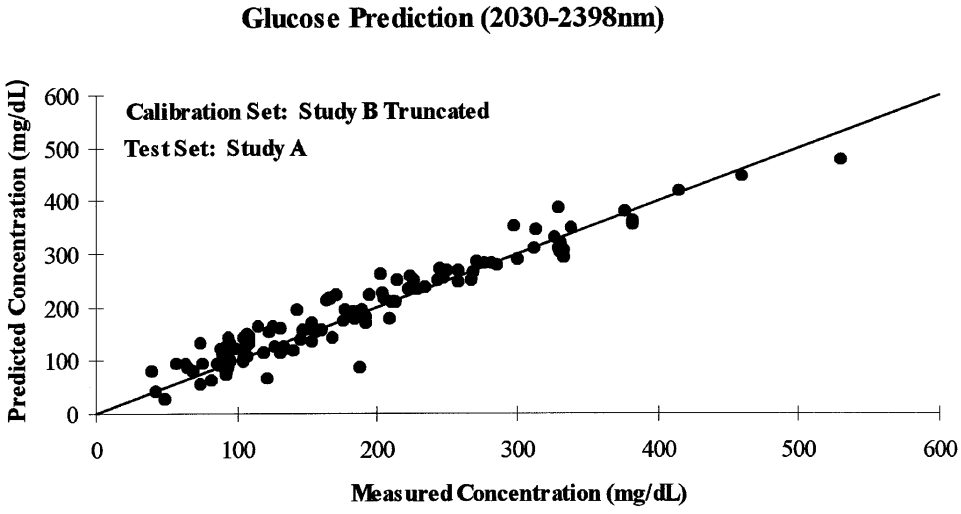

Figure 2. Prediction of glucose in Study A data set from Study B calibration model.

participants to create an even distribution across the range, the number of samples required for a robust calibration was lower than expected.

Total Protein

There was little difference between Studies A and B and how they performed in intra- or inter-study comparisons. The *SEV* for Study A was 0.28 g dL^{-1} and for Study B it was 0.31 g dL^{-1}.

When Study A calibration model was applied to Study B, the *SEP* was 0.29 g dL^{-1} and when Study B model was applied to Study A the *SEP* was 0.28 g dL^{-1} (lower than the *SEV*). Protein absorption peaks and the concentration of total protein are both large. Because of these conditions and the similar ranges of total protein in each data set, the models developed for total protein were robust.

Albumin

Albumin prediction performance, like that of total protein, is similar in intra- and inter-study comparisons. The *SEV* for Study A was 0.15 g dL^{-1} and for Study B was 0.18 g dL^{-1}. Study A calibration applied to Study B gave a slightly lower *SEP* (0.16 g dL^{-1}) than Study B calibration applied to Study A (0.18 g dL^{-1}). Similar arguments of absorbance strength and, to a lesser degree, concentration can be made for albumin as for total protein.

Albumin prediction performance mimics total protein because it is one of the two major contributors to the total protein content. The *SEV*s for intra-study predictions are larger for total protein than for albumin. The total protein prediction model uses the spectral features from both albumin and globulin and, thus, incorporates their individual errors into the *SEV*.

Globulin

Intra-study comparisons for globulin resulted in *SEV*s of 0.21 g dL^{-1} for Study A and 0.26 g dL^{-1} for Study B. The inter-study comparisons were similar. The *SEP* for Study A model applied to Study B was 0.27 g dL^{-1} and for Study B model applied to Study A the *SEP* was 0.23 g dL^{-1}. Prediction statistics for globulin are between total protein and albumin results. Because globulin is a calculated reference value (total protein minus albumin) the reference values incorporate measurement error from both total protein and albumin assays.

Globulin was better predicted from Study A data than from Study B data in both inter- and intra-study comparisons. The range of values in Study A was wider than in Study B and this fact may be responsible for the differences between the prediction results.

Cholesterol

The intra-study prediction results gave very low *SEV*s of 11 and 12 mg dL^{-1} respectively for Study A and Study B. Very little degradation was found when the inter-study comparisons were made. Study A calibration tested on Study B data resulted in an *SEP* of 13 mg dL^{-1}, a slight degradation of performance. Study B model applied to Study A had a slightly larger increase to 16 mg dL^{-1}.

The range for cholesterol in Study B included values higher than the range in Study A and it is likely that these higher cholesterol samples were not predicted as well from a model which did not include those values. In addition, cholesterol is present in relatively low concentrations compared with the proteins and also has a smaller spectral response. Robust calibrations are not as favored for such analytes but the above comparisons show that very good prediction models for cholesterol can be achieved.

Triglycerides

Prediction models for triglycerides result in intra-study *SEV*s of 18 mg dL^{-1} for Study A and 21 mg dL^{-1} for Study B. When Study A model is applied to Study B the *SEP* is 27 mg dL^{-1}, almost double the *SEV*. The same trend is observed when Study B model is applied to Study A—the *SEP* of 32 mg dL^{-1} is considerably larger than the *SEV* of 21 mg dL^{-1}. This degradation of model performance indicates that the models are sensitive to differences between the data sets. The higher *SEV* for Study B and the higher *SEP* when predicting Study B from the Study A model suggests that the Study B data set is responsible for the lack of robustness. Triglycerides calibrations are

subject to a similar limitation experienced by many serum components (including glucose): relatively low concentrations and low NIR spectral response.

Comparing the range of the two Studies shows that in Study B the highest triglyceride value is almost 200 mg dL^{-1} higher than in Study A. To compound the problem, the reference method is not as accurate above 1000 mg dL^{-1}. Both Study A and B have samples above that value and those high values undoubtedly have a large influence on the calibration models. In the calibration process uniform distribution across the range is the ideal. The tryglyceride range has a wide gap above 500 mg dL^{-1} that results in less than ideal calibration conditions. It is, therefore, expected that the calibration models for triglycerides on these data sets would not fare as well as for glucose (where study participants were selected to assure a uniform distribution).

Conclusions

When conditions were favorable for a robust calibration to be developed (high analyte concentration, strongly absorbing analytes or evenly distributed values over the concentration range) inter-study prediction performance was excellent. Serum glucose, total protein, albumin, globulin and cholesterol predictions each benefited from one or more of these favorable conditions. Triglycerides do not meet any of the above favorable conditions and the prediction performance suffered.

Aside from identifiable instrument problems, Studies A and B, while separated in time and conducted by different operators, were very similar and it was possible to accurately predict most analytes from new data using pre-determined calibration models. Screening out unusually high triglyceride samples would improve both the intra- and inter-study prediction performance. Therefore, it is expected that a calibration model could be developed which is robust over time for most of the analytes of interest and could be applied to future similar data collected under similar conditions.

Reference

1. J.R. Long, M.R. Gatin, P.W. Schmitt, P.J. Galley and J.F. Price,“Investigation of NIR Instrument Performance During Clinical Study of Human Sera”, in *Near Infrared Spectroscopy: The Future Waves*, Ed by A.M.C. Davies and P.C. Williams. NIR Publications, Chichester, pp. 153–157 (1996).