

Phantom Glucose Calibration Models from Simulated Noninvasive Human Near-Infrared Spectra

Mark A. Arnold,^{*,†} Jason J. Burmeister,[†] and Gary W. Small[‡]

Department of Chemistry and Optical Science and Technology Center, 230 Iowa Advanced Technology Laboratories, University of Iowa, Iowa City, Iowa 52242, and Center for Intelligent Chemical Instrumentation, Department of Chemistry, Ohio University, Athens, Ohio 45701

The validity of published reports claiming to have successfully measured in vivo blood glucose from noninvasive near-infrared spectra collected in a time-dependent manner is challenged on the basis of results obtained from a phantom glucose spectral data set. An in vitro model is used to simulate noninvasive human near-IR spectra. The phantom glucose data set is created by purposely omitting glucose in these modeled samples. Glucose values are then assigned to successive phantom glucose spectra, and multivariate calibration models are generated for glucose based on partial-least squares regression. As expected, calibration models are incapable of predicting glucose values when the glucose assignments are made randomly. Apparently functional models are obtained, however, when glucose assignments are made in a nonrandom, time-dependent manner. Prediction errors from these nonrandom models are essentially identical to those published by others as evidence of successful noninvasive blood glucose measurements. Chance temporal correlations between assigned glucose concentrations and some uncontrolled experimental parameter are responsible for this apparent model functionality.

The interaction of electromagnetic radiation in the near-infrared (near-IR) region of the spectrum with molecular species can provide valuable clinical and biomedical information in a noninvasive and nondestructive manner. Pulse oximetry, for example, is a widely used method for continuously measuring in vivo blood oxygen saturation.¹ In addition, near-IR photon diffusion methods are currently being developed for noninvasive imaging of tumors within soft tissue.^{2,3}

Near-IR spectroscopy has been proposed as a means for measuring in vivo blood glucose values noninvasively.^{4–8} The concept is to allow near-IR radiation to penetrate a vascular-equilibrated region of the body, followed by the acquisition of a

spectrum of the tissue through either a transmission or a reflectance measurement. Quantitative glucose information would then be extracted from the measured spectrum through the use of suitable data processing methods.

The presence of many overlapping spectral bands in the near-IR region requires the use of multivariate calibration models to relate near-IR spectral information to glucose concentrations. Methods such as partial least-squares (PLS) regression or artificial neural networks (ANN) use data at multiple wavelengths to model the components of the spectral background and subtract the background contribution from the overlapping glucose information. The calculation of these calibration models requires a set of tissue spectra and a corresponding set of actual blood glucose levels obtained through conventional invasive clinical measurements. The ease with which data at many wavelengths can be collected and the corresponding practical difficulties of asking a human volunteer to undergo many invasive glucose measurements combine to produce highly underdetermined data sets for use in computing the calibration models. Measurements at hundreds of wavelengths are typically combined with only tens of samples. This leads to a situation in which chance correlations between spectral information and glucose concentrations can produce fortuitous results. Assessing the validity of the calibration models is thus an extremely important part of the experimental procedure.

Although several research groups claim success in measuring blood glucose levels from near-IR spectra collected from human subjects,^{9–14} the source of spectral information used within their PLS or ANN algorithms cannot be rigorously assessed. The

- (6) Arnold, M. A. In *Handbook of Clinical Laboratory Automation, Robotics, and Optimization*; Kost, G. J., Ed.; John Wiley & Sons: New York, 1996; Chapter 26, pp 631–647.
- (7) Heise, H. M. *Horm. Metab. Res.* **1996**, *28*, 527–534.
- (8) Heise, H. M. In *Biosensors in the Body*; Fraser, D. M., Ed.; John Wiley & Sons: New York, 1997; Chapter 3, pp 79–116.
- (9) Heise, H. M.; Marbach, R.; Koschinsky, Th.; Gries, F. A. *Artif. Organs* **1994**, *18*, 439–447.
- (10) Jagemann, K.; Fischbacher, C.; Danzer, K.; Müller, U. A.; Mertes, B. Z. *Phys. Chem. Bd.* **1995**, *191*, 179–190.
- (11) Marbach, R.; Koschinsky, Th.; Gries, F. A.; Heise, H. M. *Appl. Spectrosc.* **1993**, *47*, 875–881.
- (12) Robinson, M. R.; Eaton, R. P.; Haaland, D. M.; Koepp, G. W.; Thomas, E. V.; Stallard, B. R.; Robinson, P. L. *Clin. Chem.* **1992**, *38/9*, 1618–1622.
- (13) Müller, U. A.; Mertes, B.; Fischbacher, C.; Jageman, K. U.; Danzer, K. *Int. J. Artif. Organs* **1997**, *20*, 285–290.
- (14) Fischbacher, Ch.; Jagemann, K.-U.; Danzer, K.; Müller, U. A.; Papenkordt, L.; Schüler, J. *Fresenius J. Anal. Chem.* **1997**, *358*, 78–82.

* Corresponding author: (e-mail) mark-arnold@uiowa.edu.

[†] University of Iowa.

[‡] Ohio University.

- (1) Flewelling, R. In *The Biomedical Engineering Handbook*; Bronzino, J. D., Ed.; CRC Press: Boca Raton, FL, 1995; Chapter 88, pp 1346–1356.
- (2) Yodh, A.; Chance, B. *Phys. Today* **1995**, *48*, 34–40.
- (3) Sevich-Muraca, E. M. *J. Biomed. Opt.* **1996**, *3*, 342–355.
- (4) Amato, I. *Science* **1992**, *258*, 892–893.
- (5) Arnold, M. A. *Curr. Opin. Biotechnol.* **1996**, *7*, 46–49.

glucose absorbance values are so small with respect to the background absorbance of the tissue matrix that no one to date has been able to display spectra of tissue samples with visually distinct glucose bands. The inability to visualize glucose-specific information makes it extremely difficult to verify the origin of the spectral information used in multivariate calibration models for predicting in vivo glucose concentrations. In fact, it is not known whether such glucose predictions are even based on the spectroscopic properties of glucose rather than on some correlative response between the glucose concentration and the spectral features of other matrix constituents.

In addition, it is critical to avoid temporal correlations within the data set^{8,9,11,15} in order to eliminate the possibility of building multivariate calibration models on the basis of time-dependent spectral features, such as instrumental intensity drifts or room-temperature variations. Indeed, Heise and co-workers, along with many others, have clearly acknowledged the potential pitfalls of using a nonrandomized run order for collecting spectroscopic data for subsequent multivariate analysis.^{8,9,11} Nevertheless, several attempts to demonstrate the feasibility of near-IR spectroscopy for noninvasive blood glucose measurements use an experimental protocol where near-IR spectra are collected from human subjects in a manner similar to a glucose tolerance test. In such a protocol, near-IR spectra are collected at some specified frequency while blood glucose levels are systematically altered by either ingestion of glucose or injection of insulin. Blood samples are collected simultaneously to provide the glucose information required for the development and evaluation of subsequent calibration models.

This report addresses the critical relationship between information within noninvasive near-IR spectra and the prediction accuracy obtained from multivariate calibration models. Our approach is to assess the prediction accuracy for glucose from PLS calibration models built with spectra completely void of glucose information. This assessment is accomplished by generating a phantom glucose data set where noninvasive human spectra are simulated and glucose is purposely omitted. PLS regression models for glucose are then developed by assigning glucose concentrations to each spectrum and submitting the resulting data set to a standard PLS calibration protocol. Results illustrate how nonrandom, time-dependent glucose variations during the data collection protocol can produce PLS models that erroneously appear to predict glucose when, in fact, absolutely no glucose information is present.

EXPERIMENTAL SECTION

In Vitro Model. Our analysis of near-IR spectra collected from human tissue reveals that water and fat within the tissue matrix are principally responsible for the absorption of near-IR light. This finding led to the development of an in vitro model to simulate the human body within the spectrometer.¹⁶ Sequential layers of animal fat and aqueous buffer solution result in near-IR spectra that accurately match those collected from human volunteers. This model permits examination of important experimental parameters that are difficult, or impossible, to investigate in a systematic manner with living human subjects.

A phantom glucose data set was prepared by use of our in vitro model. Noninvasive human spectra were simulated with a 1.6-mm-thick layer of blended beef fat and a 5.7-mm-thick layer of an aqueous protein solution. Eighty protein solutions were prepared by dissolving bovine serum albumin (BSA) in a 0.1 M, pH 7.4 phosphate buffer. BSA concentrations spanned the physiological range of 45.0–85.8 g/L.¹⁷ Near-IR spectra were collected for each protein solution combined with the fat layer. Three spectra were collected sequentially for each sample, and samples were measured randomly with respect to protein concentration. Spectra were collected over a six-day period. Specifically, the spectra for 11, 15, 20, 14, 10, and 10 samples were collected the first, second, third, fourth, fifth, and sixth days, respectively. A single background spectrum was collected at the midpoint of each day. The sample for each background spectrum included the fat layer and an aqueous layer with no BSA.

Spectra were collected with a Nicolet 740 Fourier transform spectrometer equipped with a 400-W tungsten–halogen source, CaF₂ beam splitter, room-temperature InGaAs detector, 1.49- μ m-long pass optical filter, 1.85- μ m short-pass optical filter, and gold-coated mirrors. Each spectrum was collected as 256 coadded 16K interferograms which were subsequently Fourier transformed to provide single-beam spectra from 7000 to 5000 cm⁻¹ with a point spacing of 1.9 cm⁻¹. Solution temperatures were maintained at 37.0 \pm 0.1 °C during the data collection by use of a VWR 1140 water bath in conjunction with a Wilmad model 118 jacketed cell holder and 1-in.-diameter sapphire windows. The tungsten–halogen lamp in use at the beginning of the experiment failed in the middle of the fourth day. A new lamp was installed and the data collection promptly resumed.

PLS Models. Unless noted otherwise, all PLS calibration models were based on single-beam spectra, and model optimization and verification were accomplished by using independent calibration, monitoring, and prediction subsets of the data.¹⁸ These subsets were established by assigning glucose concentrations to the first single-beam spectrum collected for a given protein sample solution and then splitting these spectra into calibration, monitoring, and prediction data sets. The second and third spectra for each protein solution were discarded to better simulate the conditions of a typical noninvasive experiment where only one spectrum can be obtained for each point in time and, therefore, for each glucose concentration.

The prediction data set was established first by extracting ~20 randomly selected spectra. The remaining spectra corresponded to the calibration data set that was used to develop the PLS models. Model performance during optimization was judged by comparing the standard error for a monitoring data set that corresponded to 10 randomly selected spectra from the calibration data set. The calibration set was used to build the model, the monitoring set was used to judge the predictive ability of potential models during optimization, and the prediction set was withheld from all optimizations and used simply to assess the performance of the final optimized model. Model performance was judged by comparing standard error values for each type of data set. The

(15) Martens, H.; Naes, T. *Multivariate Calibration*; Wiley: New York, 1989.

(16) Burmeister, J. J.; Chung, H.; Arnold, M. A. *Photochem. Photobiol.* **1998**, *67*, 50–55.

(17) Pesce J.; Kaplan, L. A. *Methods in Clinical Chemistry*; Mosby: St. Louis, MO, 1987.

(18) Small, G. W.; Arnold, M. A.; Marquardt, L. A. *Anal. Chem.* **1993**, *65*, 3279–3289.

standard error of calibration (SEC), standard error of monitoring (SEM), and standard error of prediction (SEP) were computed by the following equations:

$$\text{SEC} = [1/(n_c - f - 1) (\sum_i^{n_c} (c_a - c_p)^2)]^{1/2}$$

$$\text{SEM} = [1/n_m (\sum_i^{n_m} (c_a - c_p)^2)]^{1/2}$$

$$\text{SEP} = [1/n_p (\sum_i^{n_p} (c_a - c_p)^2)]^{1/2}$$

where n_c , n_m , and n_p correspond to the number of spectra in the calibration, monitoring, and prediction sets, respectively, f is the number of PLS factors used to build the model, and c_a and c_p represent the assigned glucose concentration and glucose concentration predicted by the model, respectively.

All PLS calibration models were generated without mean centering or spectral normalization. Unless noted otherwise, the following modified grid search was used to establish the optimum spectral range and number of PLS factors. For a given model size (number of PLS factors), SEM values were computed for a series of PLS models constructed over 100-cm⁻¹-wide regions at 100-cm⁻¹ intervals beginning with 5300–5200 cm⁻¹. The region yielding the lowest SEM was further examined by systematically increasing and decreasing the upper and lower limits while tracking the SEM. This process of modifying the maximum and minimum frequencies was repeated four times with a smaller step size with each iteration. This procedure was repeated for 1–15 PLS factors. The optimal model was taken as the combination of spectral range and model size that produced the lowest SEM. This entire optimization process was repeated three times with a unique, randomly selected set of spectra in the monitoring data set which permitted greater robustness in the final results.¹⁹ Finally, the spectral range that provided the lowest SEM with the fewest number of factors was selected as optimal. Working calibration models were then established by using all the calibration spectra in combination with this spectral range and number of factors.

Clarke Error Grid. The Clarke error grid²⁰ was used to judge analytical performance of the glucose models generated from each phantom glucose data set. This grid relates measurement error to the correctness of the corresponding clinical action. This grid system assigns predicted vs actual glucose values to five categories (A–E) based on the correctness of the clinical decision that would follow the glucose measurement. Ideally, all measurements should fall in the A region where the correct clinical action is taken on the basis of the measured value. Measurements in the E region represent the worst situation where the measurement is so inaccurate that the indicated clinical action is exactly opposite the required treatment.

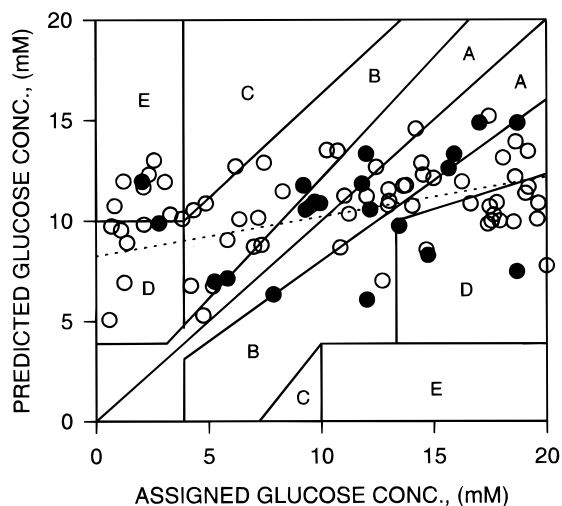


Figure 1. Concentration correlation plot for the best PLS model based on random glucose assignments. Open circles correspond to calibration points, closed circles correspond to prediction points, and the line corresponds to the regression line.

Spectral Noise. Root-mean-square (rms) noise levels on 100% lines were obtained by taking the ratio of replicate single-beam spectra, converting the resulting transmittance values to absorbance units (AU), fitting the resulting data to a second-order polynomial, and computing the standard deviation about this fit. Results are reported in microabsorbance units (μ AU) for the specified spectral region.

RESULTS AND DISCUSSION

Random Glucose Assignments. PLS models generated from randomly assigned glucose values are incapable of predicting glucose concentrations in the prediction data set. In this experiment, assigned glucose concentrations ranged from 0.59 to 20.00 mM with a mean of 10.82 mM and standard deviation of 6.03 mM. Values were assigned randomly relative to sample order. The mean assigned concentrations for the calibration and prediction sets were 10.70 and 11.21 mM, respectively, and the corresponding standard deviations were 6.37 and 4.87 mM, respectively. The best model corresponded to three PLS factors over the 5420–5200-cm⁻¹ spectral range.

The concentration correlation plot presented in Figure 1 reveals that this model predicts essentially the mean assigned concentration for every spectrum. In addition, the SEC and SEP values (6.17 and 4.61 mM, respectively) match the standard deviation of the assigned values in the calibration and prediction sets, respectively. An F -test was used to compare the SEP to the standard deviation of the glucose concentrations within the prediction data set (SDP). The computed F -value was significant only at the 60% probability level, leading to the clear conclusion that the calibration model is not explaining glucose information. This finding is consistent with the fact that no glucose-specific information is present in the spectra.

Time Profile Glucose Assignments. Because blood glucose concentrations are not always random with respect to time in published experiments designed to demonstrate the feasibility of noninvasive sensing with near-IR spectroscopy, we have also evaluated PLS calibration models where the assigned glucose

(19) Shaffer, R. E.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **1996**, *68*, 2663–2675.

(20) Clarke, W. L.; Cox, D.; Gonder-Frederick, L. A.; Carter, W.; Pohl, S. L. *Diabetes Care* **1987**, *10*, 622–628.

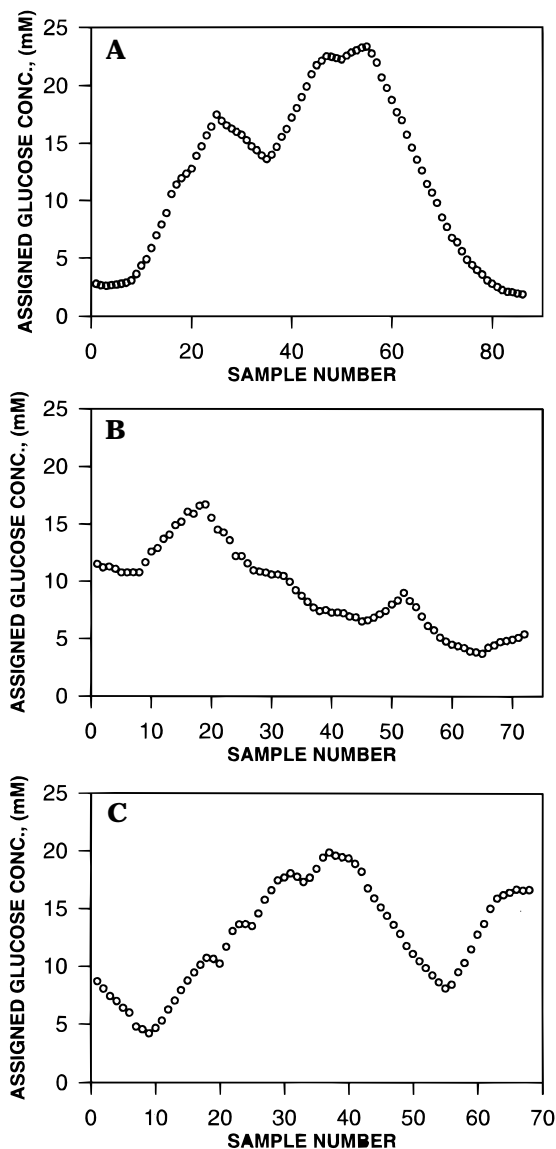


Figure 2. Glucose–time profiles used for phantom glucose concentration assignments where the data were taken from (A) Heise,⁹ (B) Danzer,¹⁰ and (C) Arnold.

concentrations are nonrandom. In this case, glucose assignments were made according to three unique glucose–time profiles that represent data from a typical single-day glucose tolerance experiment.

Figure 2 presents the three glucose–time profiles used in this work. The first (Figure 2A) was reported by Heise and co-workers as they evaluated noninvasive blood glucose measurements from near-IR spectra collected in a diffuse reflectance mode along the inner section of the lower lip of human volunteers.⁹ This glucose–time profile comes from one of three experimental designs explored by this research group. The first corresponds to a glucose tolerance protocol where 133 spectra were collected from the same volunteer over a two day period. The data in Figure 2A was taken from the first of these 2 days. Their second experimental design involved collecting 219 spectra from the same individual over a 14 day period, and the third design involved collecting a limited number of samples from many different individuals (399 spectra from 133 volunteers). Figure 2B was

reported by Danzer and co-workers as they compared PLS and ANN algorithms for processing noninvasive diffuse reflectance near-IR spectra collected from the middle finger of the right-hand of human volunteers.¹⁰ The Danzer group typically collects data by a glucose tolerance protocol.^{10,13,14} Plotted glucose values in Figure 2A and B were graphically interpolated directly from figures within these published reports. The third glucose–time profile (Figure 2C) was collected in our laboratory and corresponds to a single-day experiment where noninvasive spectra were collected by transmitting light through the webbing tissue between the thumb and forefinger.

A data set was established for each of the above-mentioned time profiles by assigning sequential glucose values to successive spectra in the phantom glucose data set. Glucose concentrations for Figure 2A provided 86 values ranging from 1.89 to 23.31 mM. After the samples were split into calibration and prediction data sets, the mean and standard deviation of the concentrations in the prediction set ($n = 20$) were 11.84 and 7.49 mM, respectively. Values in Figure 2B provided 72 values with glucose concentrations ranging from 3.67 to 16.67 mM. The mean and standard deviation of values in the prediction data set ($n = 21$) were 10.27 and 2.98 mM, respectively. Glucose concentrations for the 68 measurements in Figure 2C ranged from 4.20 to 19.84 mM with a mean and standard deviation of 11.52 and 4.23 mM, respectively, for values in the prediction data set ($n = 19$).

Optimized PLS Models with Glucose–Time Correlation.

Figure 3 presents concentration correlation plots for the resulting three optimized PLS calibration models. Inspection of these plots suggests that models from each time profile are capable of measuring glucose even though the spectra contain absolutely no glucose information. Although each plot displays considerable scatter, the degree of scatter is similar for both the calibration and prediction data points. More importantly, the majority of the prediction data points fall within the A region of the superimposed Clarke error grid. In fact, 70%, 85.7%, and 52.6% of the prediction points fall within the A region when the Heise, Danzer, and Arnold time profiles, respectively, are used. An additional 20%, 9.5%, and 36.8% fall within the B region for the Heise, Danzer, and Arnold time profiles. For all the time profiles combined, only one prediction value falls close to the E region (see Figure 3A), and this value technically falls within the C region. These observations imply the models are functioning properly, particularly when compared to the analogous plot in Figure 1 where glucose values were assigned randomly and all predictions for assigned glucose values below 4 mM are in the D and E regions of the Clarke error grid.

Model parameters and performance characteristics are listed in Table 1 for both the random and nonrandom models described above. Models based on the nonrandom time-dependent glucose assignments incorporate a wider spectral range and more PLS factors to provide significantly lower prediction errors compared to the standard deviation of the concentration values in the prediction data set. F -values computed from the comparison of SEP and SDP are significant at the 100%, 99%, and 89% levels for the Heise, Danzer, and Arnold time profiles, respectively. The values of SEP and SDP are clearly different. Results are also provided in Table 1 for linear regression analysis of the prediction points within each concentration correlation plot. In all cases,

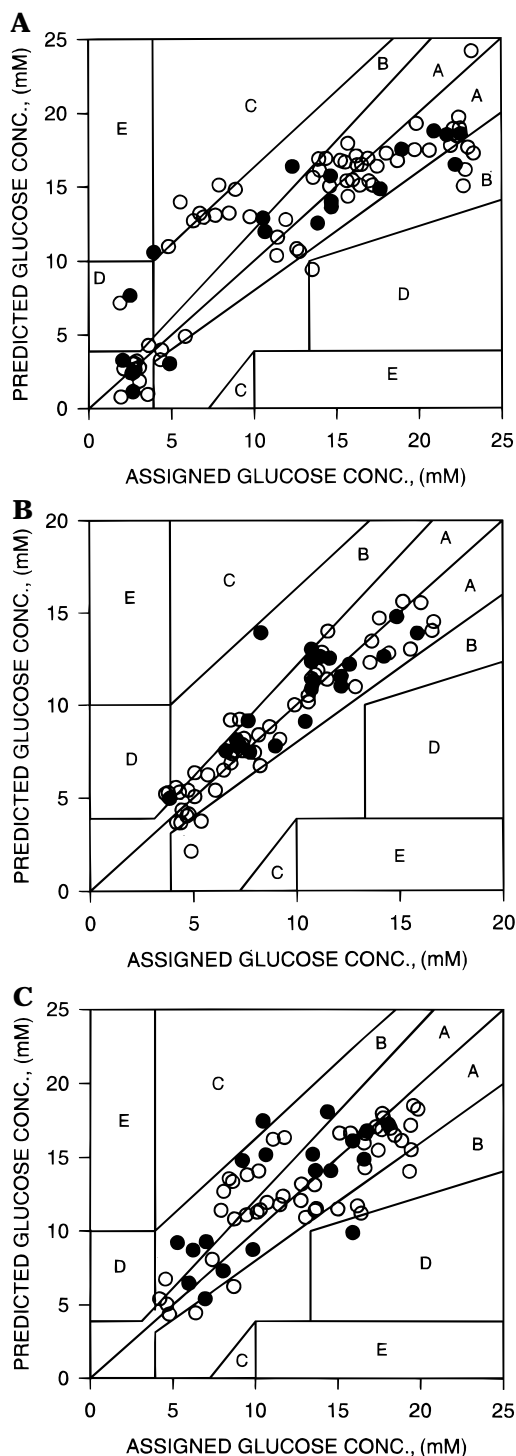


Figure 3. Concentration correlation plots, superimposed on the Clarke error grid,²⁰ for optimized phantom glucose calibration models based on assignments from the (A) Heise,⁹ (B) Danzer,¹⁰ and (C) Arnold glucose–time profiles. Open circles correspond to calibration points, and closed circles correspond to prediction points.

the y -intercepts are lower, slopes are closer to unity, and R^2 values are higher for models based on the time profile glucose assignments. Taken together, these results provide a clear indication that the glucose correlations embodied within the time profile calibration models are nonrandom.

Spectral Correlations. Correlation between the real concentration of BSA in the samples and the assigned glucose concentra-

tions would account for the apparent ability to predict glucose in the absence of glucose-specific information. Correlation plots of assigned glucose values vs BSA concentrations are random, however, for each data set. In fact, no correlation is expected between BSA concentration and assigned glucose level because the spectra were collected in random order relative to protein concentrations. Indeed, R^2 values are 0.011, 0.047, and 0.026 for assignments based on the Heise, Danzer, and Arnold time profiles, respectively. The random nature of the relationship between protein concentration and assigned glucose value precludes the possibility that subsequent glucose predictions are based on the spectral properties of BSA.

The apparent functionality of these phantom glucose models is based on chance correlations between assigned glucose concentrations and some uncontrolled experimental parameter that varies as a function of time. Instrument alignment, for example, can vary slightly during the course of a data collection session as the result of systematic changes in the ambient room temperature. Alternatively, incident source intensity can vary slightly as a function of time due to variations in the source power supply. Such temporal connections between assigned glucose concentration and instrument performance can provide the link necessary for model functionality, even though no glucose information is present in the spectra.

The source of temporal variation within our phantom glucose data set is not readily apparent. In fact, care was taken to minimize such variations. Ambient temperature was controlled to within ± 0.5 °C and the source power supply operated with 1% precision throughout the experiment. Although the spectrometer used in this case was not capable of dynamic alignment, the interferometer was realigned several times each day. The resulting spectral data set is characterized by a relative standard deviation (RSD) of 2.9% and a mean signal-to-noise ratio of 8019 (± 2263) at the maximum peak intensity of the single-beam spectra (6013.013 cm^{-1}). In addition, rms noise values of 100% lines are 8.08 (± 1.25) and 10.6 (± 3.1) μAU for the 6100 – 6000 - and 6000 – 5900 - cm^{-1} spectral regions, respectively. As expected, rms noise levels increase dramatically for wider spectral ranges as the optical throughput drops dramatically due to strong water absorbance. Figure 4 presents a representative 100% line superimposed on a typical sample spectrum. For the 100% line shown in Figure 4, computed rms noises are 100.9 and 2370 μAU over the 6500 – 5700 - and 6700 – 5400 - cm^{-1} spectral ranges, respectively.

Despite our attempts to maintain constant experimental conditions, a slight positive correlation exists between spectral intensity and the order in which the spectra were collected. Linear regression analysis for such a plot (intensity at 6013.013 cm^{-1} vs spectrum number) reveals a slope of 0.032, a y -intercept of 41.8, and an R^2 value of 0.396.

Effects of Spectral Range, Resolution, and Transformation. Numerous experimental and data processing parameters might affect the extent to which a time-dependent correlation within the data set can be linked to the analyte concentration. For example, the existence of more independent variables, or a more highly underdetermined system, will increase the probability of the PLS algorithm to find and exploit weak correlations of this nature. To explore such effects, we examined the impact of three common parameters: spectral range, spectral resolution, and

Table 1. Parameters and Performance Characteristics for Phantom Glucose Models

time profile	spectral range, (cm ⁻¹)	model size	SEC (mM)	SEP (mM)	<i>F</i> -prob (%)	regression analysis of prediction data		
						β_0 (mM)	β_1	<i>r</i> ²
random	5420–5200	3	6.17	4.61	60	8.238	0.1976	0.1251
Heise ⁹	6610–5800	9	3.70	3.00	100	2.748	0.7469	0.8425
Danzer ¹⁰	6000–5690	7	1.34	1.69	99	3.097	0.7386	0.6878
Arnold	6000–5600	6	2.87	3.16	89	4.258	0.7222	0.5393

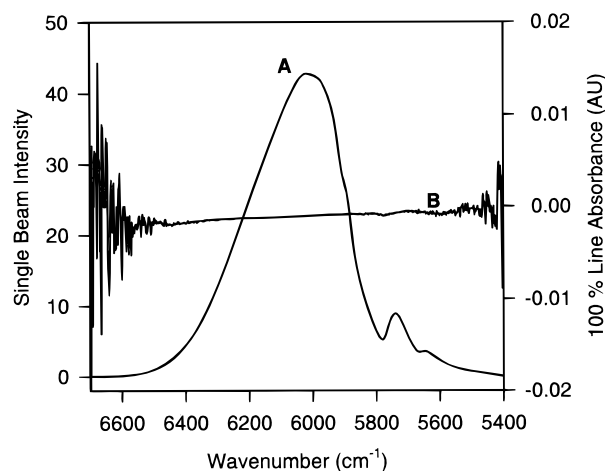


Figure 4. (A) Single-beam spectrum for the model and (B) a 100% line for two back-to-back spectra of a single sample.

logarithmic transformation of the spectral data.

Spectral range is a critical parameter for any PLS calibration model. Two general strategies can be employed in selecting this range. First, various nonbiased computational procedures have been described for selecting the spectral range that provides the best analytical performance. Modified grid searches, simulated annealing, and genetic algorithms have been evaluated as a means for defining the optimal spectral range. When applied to our phantom glucose data set, any search method will work to find the spectral region most sensitive to temporal correlations within the data set.

An alternate strategy involves the selection of the spectral range on the basis of an a priori knowledge of the glucose absorption spectrum.²¹ The spectral range could be selected as the widest range that incorporates all known analyte absorption bands while rejecting particularly noisy regions. Alternatively, narrower regions that isolate single absorption bands, or different combinations of multiple bands, can be tested to identify the spectral region that contains the best analytical information.

In practice, a combination of these approaches is most likely needed. The search method is used to define the best spectral range, and the user's knowledge of the glucose absorption spectrum is then used to evaluate the credibility of this range. In reality, most PLS calibration models based on near-IR spectra require a wide spectral range that encompasses known analyte absorption bands as well as information pertaining to interfering matrix components.

Glucose is known to possess three distinct absorption bands in the first overtone region under study here. These bands are

centered at 5775, 5920, and 6200 cm⁻¹.²¹ Therefore, the 6500–5700-cm⁻¹ spectral range would be appropriate for this analysis because it incorporates all the glucose spectral information while rejecting the noisy regions caused by strong water absorption (see Figure 4). In this analysis, the best model size was established by finding the number of factors that provided the lowest mean SEM for three unique monitoring data sets.

While investigating the effect of spectral range, we noticed that predictions from one particular spectrum were unusually poor. This spectrum corresponds to the first spectrum collected after the lamp was replaced on day 4 of the data collection period. By chance, this spectrum was included in the prediction data sets based on the Danzer and Arnold time profiles. SEP values are unusually high when this spectrum is included. With the Danzer time profile, for example, the PLS model predicts a value of 28.5 mM glucose for this spectrum when the assigned value was only 8.3 mM. The corresponding absolute error (20.2 mM) is significantly greater than any other error in this data set. Similar results are obtained with the Arnold time profile. With this spectrum, SEP values are 4.97 and 2.79 mM for the Danzer and Arnold time profiles, respectively. After removing this spectrum from both of these prediction data sets, the SEP values drop to 2.37 and 2.64 mM, respectively. Hence, this spectrum was removed from these prediction data sets, unless noted otherwise. Removal of this spectrum results in SDP values of 3.02 and 4.34 mM for the Danzer and Arnold time profiles, respectively. This spectrum is located in the calibration data set for the Heise time profile and, therefore, has little impact on the computed SEP. As will be discussed below, this spectrum is identified as an outlier on the basis of standard statistical testing methods.

Entries 1, 4, and 7 in Table 2 summarize model performance with the spectral range fixed to 6500–5700 cm⁻¹ and the other parameters remaining the same as those described above. Standard errors are greater when this fixed spectral range is used with the Heise and Danzer time profiles. For the Arnold time profile, the standard errors are smaller with the fixed spectral range, but considerably more factors are used. For all three time profiles, high levels of significance in the computed *F*-values indicate clear differences between SEP and SDP values. Analysis of the corresponding concentration correlation plots reveals nonrandom correlations in each case. In fact, the overall structure of these plots is similar to those presented in Figure 3. The following percentages of the prediction points fall into the indicated regions of the Clarke error grid. Heise: 55% A; 35% B; 5% C; 5% D; 0% E. Danzer: 55% A; 35% B; 0% C; 10% D; 0% E. Arnold: 66.7% A; 27.8% B; 5.6% C; 0% D; 0% E. When taken together, these values again indicate apparently functional glucose calibration

(21) Hazen, K. H. Ph.D. Dissertation, University of Iowa, August 1995.

Table 2. Characteristics of PLS Models Computed over 6500–5700 cm⁻¹

entry	time profile	spectrum type	spectral resolution	PLS factors	SEC (mM)	SEP (mM)	F-test ^a
1	Heise	single beam	2	6	4.55	3.69	100
2	Heise	single beam	32	6	4.45	3.47	100
3	Heise	log(1/I)	2	5	4.64	3.43	100
4	Danzer	single beam	2	7	2.35	2.37	86
5	Danzer	single beam	32	10	1.94	1.80	99
6	Danzer	log(1/I)	2	6	2.44	2.68	70
7	Arnold	single beam	2	11	1.17	2.64	98
8	Arnold	single beam	32	5	3.09	3.61	78
9	Arnold	log(1/I)	2	6	2.82	3.75	73

^a Level at which the *F*-value is significant when comparing the SEP to the standard deviation of the prediction (SDP) data set. SDP values are as follows: SDP_{Heise} = 7.49 mM, SDP_{Danzer} = 3.02 mM, and SDP_{Arnold} = 4.34 mM. SDP and SEP values for the Danzer and Arnold time profiles reflect removal of the first spectrum following replacement of the source lamp on day 4 of data collection (see text for details).

models. There is no indication that a fixed spectral range is more or less sensitive to producing phantom calibration models relative to the grid search method described above.

Spectral resolution controls the number of points available within a given spectral region. A higher resolution spectrum would result in more data points, thereby providing greater degrees of freedom for finding chance correlations within a data set. To examine the effect of resolution, the interferograms corresponding to our phantom glucose data set were reprocessed following degradation of the spectral resolution from 2 to 32 cm⁻¹. After reprocessing, spectral data sets with nominal point spacings of 2, 4, 8, 16, and 32 cm⁻¹ were generated and analyzed over the 6500–5700-cm⁻¹ spectral range by the procedures described above.

No trends in prediction error were observed on the basis of resolution. Table 2 lists the results obtained for the 2- and 32-cm⁻¹ data sets. Prediction errors varied from 3.44 to 3.89 mM with the Heise time profile, from 1.41 to 2.85 mM with the Danzer time profile, and from 2.62 to 3.72 mM with the Arnold time profile. Again, inspection of correlation plots reveals nonrandom models with no apparent trend as a function of spectral resolution. The following percentages of the prediction points fall into the indicated regions of the Clarke error grid for the 32 cm⁻¹ resolution data. Heise: 55% A; 35% B; 5% C; 5% D; 0% E. Danzer: 65% A; 30% B; 0% C; 5% D; 0% E. Arnold: 43.8% A; 50% B; 11.1% C; 0% D; 0% E.

Transformation of the spectral data before submission to the PLS regression may influence the likelihood of finding chance correlations. A logarithmic transformation (log(1/*I*)) is commonly used to approximate absorbance spectra for diffuse reflectance spectra or when a representative background spectrum is unavailable. The effect of this logarithmic transformation was investigated by transforming the 2-cm⁻¹ single-beam spectra. Again, the spectral range was fixed to 6500–5700 cm⁻¹ and the optimum number of PLS factors was established as before.

Models computed from logarithmic transformed spectra are similar to models based on single-beam spectra. The corresponding SEP values are listed as entries 3, 6, and 9 in Table 2 and can be compared to entries 1, 4, and 7 for the Heise, Danzer, and Arnold time profiles, respectively. There is no clear trend between the different time profiles as the computed SEP decreases for the Heise time profile and increases for the Danzer and Arnold time

profiles. Again, the concentration correlation plots are similar to those presented in Figure 3. The following percentages of the prediction points fall into the indicated regions of the Clarke error grid. Heise: 45% A; 40% B; 5% C; 10% D; 0% E. Danzer: 45% A; 45% B; 0% C; 10% D; 0% E. Arnold: 44.4% A; 44.4% B; 11.1% C; 0% D; 0% E. In all three cases, nonrandom correlations are demonstrated.

Cross Validation and Outlier Detection. The leave-one-out cross-validation procedure is a well-recognized method that is frequently used to judge the validity of multivariate calibration models. This method is particularly beneficial when a limited number of observations is available for assessing model relevance. As such, we have applied this processing scheme to our phantom glucose data set. In addition, numerous outlier detection schemes have been proposed for identifying spectra that do not statistically fall within the overall population of a given data set. Two representative methods of outlier detection were applied to our phantom data set to examine the extent to which the removal of identified outliers reduces prediction errors, thereby exaggerating the effects of chance correlations.

For our cross-validation work, all spectra for a given time profile were combined into a single data set. In addition, each data set included the first spectrum collected following replacement of the source lamp. The resulting standard deviations across all samples (SDS) based on the Heise, Danzer, and Arnold time profiles are 7.09, 3.66, and 4.23 mM, respectively. The leave-one-out cross-validation algorithm was applied over the 6500–5700-cm⁻¹ spectral range and the cross-validation standard error of prediction (CV-SEP) was computed. The optimum model size corresponded to the number of factors that produced the lowest CV-SEP.

Our initial cross-validation analysis was performed with single-beam spectra collected with 2-cm⁻¹ resolution. Analysis with the Heise time profile resulted in an 11-factor model with a CV-SEP of 3.79 mM. The analogous model developed with an independent prediction set (see Table 1) required only six factors and was characterized by a similar SEP of 3.69 mM. For the Danzer time profile, the cross-validation model required 11 factors to achieve an CV-SEP of 1.95 mM. In comparison, a seven-factor model produced an SEP of 2.37 mM with our independent prediction set. Similarly, the CV-SEP for the Arnold time profile was lower than the analogous SEP value (2.49 vs 2.64 mM), but in this case, the cross-validation model only required 10 factors compared to 11 factors for the model based on an independent prediction data set. Comparison of prediction errors reveals no clear trend marking a difference between these two processing schemes in terms of sensitivity to chance correlations.

A second cross-validation analysis was performed on logarithmic transformed (log(1/*I*)) spectra with 32-cm⁻¹ resolution. These conditions are analogous to those used by Heise and co-workers as they assessed their spectral data collected from human subjects.^{8,9,11} The CV-SEP values for the Heise, Danzer, and Arnold time profiles were 4.16, 3.05, and 3.13 mM, respectively. Optimized model size were 11, 12, and 13 factors, respectively. Each of these CV-SEP values is higher than the corresponding SEP and CV-SEP reported above for the 2-cm⁻¹ resolution spectra. In fact, these prediction errors are the largest found in our investigation. The resulting concentration correlation plots are presented in Figure 5. Although these plots illustrate the same

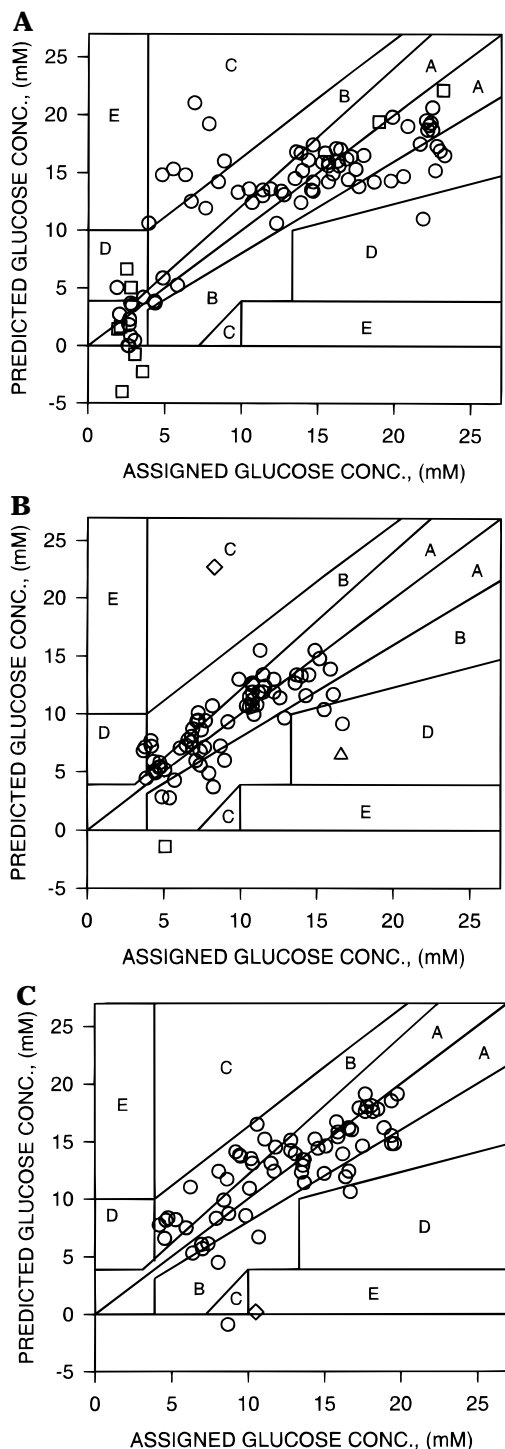


Figure 5. Concentration correlation plots superimposed on the Clarke error grid for calibration models with a fixed spectral range (6500–5700 cm^{-1}), $\log 1/I$ preprocessing, and cross validation based on assignments from the (A) Heise,⁹ (B) Danzer,¹⁰ and (C) Arnold glucose–time profiles. Circles, squares, triangles, and diamonds represent spectra that were not found to be outliers, outliers based upon high leverage values, outliers based upon Cooke’s distance, and outliers based on leverage and Cook’s distance, respectively.

type of nonrandom structure observed in our other correlation plots, the data points appear more scattered than before with points entering the D region of the Clarke error grid. In fact, negative concentration predictions are observed in these plots. Five negative predictions were obtained with the Heise time

profile, one with the Danzer time profile and one with the Arnold time profile. The high CV-SEP values for these models reflect these negative values.

A negative concentration in a leave-one-out cross-validation prediction is a strong indication that the corresponding spectrum differs significantly compared to all other spectra in the data set. For this reason, we examined each data set for outliers based on leverage values and Cook’s distances.^{15,22} A high leverage value for a given spectrum indicates that this observation is distant from the center of the data space defined by the PLS factor scores. Cook’s distance is an aggregate measure of the influence of a particular observation on the fitted values of all the other observations in the data set.

High leverage values were indicated for nine spectra when assignments were made from the Heise time profile. The corresponding predictions for these nine spectra are denoted as squares in Figure 5A. As expected, spectra for most of the negative predictions were identified as outliers by this method. In addition, one of the outlying spectra corresponds to the first spectrum collected after the lamp was replaced during the fourth day of data collection. As note above, predictions from this spectrum are consistently poor. After all indicated outlying samples are removed, the SDS drops to 6.71 mM. The corresponding CV-SEP is 4.39 mM for an 11-factor PLS model. With the Danzer time profile, high leverage values were noted for only two spectra. Again, these outlying spectra correspond to a negative prediction and the first spectrum collected after lamp replacement. After outlier removal, the SDS changes to 3.67 mM, and a 12-factor PLS model is found to be optimal with a CV-SEP of 2.39 mM. Only one spectrum is identified as an outlier with a high leverage value when the Arnold time profile is used. Again, this value corresponds to the first point following lamp replacement which results in the negative prediction shown in Figure 5C. Upon removal, the SDS increases to 4.58 mM and the resulting PLS model requires 11 factors to provide a CV-SEP of 2.86.

Fewer outlying spectra are detected by an inspection of Cook’s distances. In fact, no outliers are indicated with the Heise time profile, two spectra are marked as outliers with the Danzer time profile, and one spectrum is identified as an outlier with the Arnold time profile. The outlier identified with the Arnold time profile corresponds to the same spectrum noted above as an outlier based on leverages. For the Danzer time profile, one of the two outliers based on Cook’s distance corresponds to the first spectrum following lamp replacement. Removal of both spectra marked as outliers by Cook’s distance with the Danzer time profile results in an SDS of 3.60 mM and an optimized PLS model with 11 factors and a CV-SEP of 1.86 mM.

Relation to Previously Published Work. Of course, the results of our investigation say nothing directly about the validity of the results reported by Heise⁹ or Danzer.¹⁰ Indeed, we are only using the time profiles reported by these groups, not their actual spectroscopic data. The quality of their spectra may be better than that of the spectra used in our phantom glucose data set. It may also be possible that their data do not contain chance correlations between their assigned glucose values and some

(22) Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Regression Models*; Irwin: Chicago, 1996; pp 375–384.

uncontrolled, time-dependent experimental parameter. The available information does not permit a rigorous evaluation of these points.

Nevertheless, it must be recognized that the process of collecting spectra from human subjects adds considerable complexity and variability to the spectral measurements. In fact, more temporal correlations are likely with human subjects because of systematic variations in the physical and chemical integrity of the measurement site as spectra are collected. The mechanical means used to interface the human subject to the spectrometer will likely induce changes in tissue integrity, thereby creating temporal variations within the spectral data set. The results presented in this study clearly show that the occurrence of such an interaction between the spectral and glucose variations can lead to erroneous calibration models for blood glucose.

Heise and co-workers clearly understood the potential pitfalls of chance correlations within a multivariate data set. As such, they proposed two data collection protocols designed to minimize the temporal correlations inherent in a glucose tolerance protocol. In one protocol, near-IR spectra were collected from a single person over several days. In this particular example, measurements were made three times a day over a two-week period. Eight unique spectra and two blood samples were collected during each measurement. In a different protocol, the Heise group collected three spectra and one blood sample from 133 randomly selected subjects. Although each of these protocols will clearly be less prone to temporal correlations when compared to a glucose tolerance protocol, it must be recognized that chance correlations (temporal or otherwise) can still be a factor within these data sets.

Given the extreme sensitivity of multivariate calibration methods to chance correlations, the existence of any potential chance correlations must be rigorously investigated and reported. Clearly, a simple correlation between spectral intensity and assigned glucose value must be performed. Correlations between the time of day a spectrum is collected and the assigned glucose concentration must also be assessed. When large populations of human subjects are involved, correlations between blood glucose values and an array of spectral altering parameters (e.g., room temperature, body temperature, water content, protein concentration, hematocrit level, etc.) must be examined and shown to be negligible before one can claim to have successfully measured blood glucose from near-IR spectra collected noninvasively. Indeed, many of these parameters have been shown to significantly alter the scattering properties of light through a tissue-simulating phantom.^{23–25}

CONCLUSIONS

Clearly, the relationship between time and glucose concentration prohibits the use of a glucose tolerance protocol for collecting noninvasive near-IR spectra for possible blood glucose measure-

ments. When typical glucose tolerance time profiles are used for assigning glucose concentrations to spectra within our phantom glucose data set, PLS calibration models perform essentially as well as those published by others who claim to be measuring glucose noninvasively based on glucose tolerance protocols. Although prediction errors for such experiments differ considerably in magnitude and method of computation, errors proposed as evidence of accurate glucose sensing typically range from 2.0 to 3.5 mM. SEP and CV-SEP values from our phantom glucose experiments are within this range, thereby casting serious doubt on the validity of these previous claims.

No particular method of spectral processing appears to be more or less sensitive to temporal correlations within the data set. No significant differences were detected in model performance when a fixed spectral range was used compared to an optimized spectral range. No significant differences were apparent when models based on spectra with different resolutions were compared or when models based on single-beam spectra vs logarithmic transformed spectra were compared. In addition, models developed with the leave-one-out cross-validation method are similar to those based on the use of an independent prediction data set. Finally, the detection and removal of spectral outliers on the basis of leverages and Cook's distance tend to exaggerate false calibration models by lowering computed prediction errors. Overall, we found no evidence that one method of data processing is more prone to extracting chance correlations than another.

The burden of proof is clearly on future investigators to prove that their models are based on glucose-specific information and not chance temporal correlations. Such proof must certainly be a demand of the United States Food and Drug Administration as they evaluate putative noninvasive blood glucose sensing technology designed for hospital and/or home settings. Indeed, a recent public announcement suggests one company's device actually performs better in the hands of unsupervised diabetic volunteers in uncontrolled home environments compared to operation by trained experts under environmentally controlled laboratory conditions.^{26–28} Such findings are inconsistent with robust models based on glucose-specific spectral information and imply that these models are actually based on chance temporal correlations induced by variations in the measurement environment.

ACKNOWLEDGMENT

Financial support for this work was provided by the National Institutes of Health under Grant DK45126.

Received for review September 30, 1997. Accepted February 19, 1998.

AC9710801

(23) Kohl, M.; Essenpreis, M.; Cope, M. *Phys. Med. Biol.* **1995**, *40*, 1267–1287.

(24) Qu, J.; Wilson, B. C. *J. Biomed. Opt.* **1997**, *2*, 319–325.

(25) MacBride, D. M.; Malone, C. G.; Hebb, J. P.; Cravalho, E. G. *Appl. Spectrosc.* **1997**, *51*, 43–49.

(26) PRNewswire (Pittsburgh) Biocontrol compiling data from in-home study on Diasensor(R)1000. July 15, 1997.

(27) Reuters News Service (Pittsburgh) Biocontrol says submits test data to FDA. July 18, 1997.

(28) PRNewswire (Pittsburgh) Biocontrol submits preliminary data to the FDA. July 18, 1997.