

Commentary

Performance Standards for Continuous Glucose Monitors

PAUL G. HAYTER, M.S.,¹ MANOJ SHARMA, Ph.D.,¹ LOU DUNKA, Ph.D., F.A.C.B.,¹
PHIL STOUT, M.S.,² DAVID A. PRICE, M.D.,¹ DAVID L. HORWITZ, M.D., Ph.D.,¹
JOSEPH MARHOUL, Ph.D.,¹ and SHABNAM VAEZ-ZADEH, M.D., M.P.A.¹

THE FOLLOWING POINTS TO CONSIDER draw from cross-functional discussions at LifeScan between Medical & Clinical Affairs and R&D regarding performance standards for continuous glucose monitors. Some comments are meant to resolve questions, and some comments are intended to broaden the topic to stimulate a rich discussion.

How should fingerstick and interstitial fluid (ISF) continuous monitoring results be compared? What about other sample types (e.g., venous, arterial, etc.)?

This topic is a method comparison question with a caveat. In this case two different sample types are being measured, not a single sample type with two different methods. Different sample types may introduce artifacts not normally found in method comparisons, i.e., due to system or sample lag. Traditionally, in a method comparison, one of the measurements is assigned "reference" status, and the other "measurement." With the caveats above in mind, this can be adopted for continuous monitoring results as well. The comparison can be done in a traditional way, meaning using a Consensus Error Grid or Bland–Altman plot to give a good idea of how the two methods compare for measuring glucose. In order to make a valid comparison, it is important that the reference and measurement points line up in time.

The Bland–Altman plot (reference value vs. bias) is a good choice for method comparison as it quickly conveys the range of error observed in the data, the distribution of error, and the occurrence of outliers.

In addition to the Bland–Altman and Consensus Error Grid plots, a compilation of mean, minimum, and maximum bias, range and distribution of reference values, as well as linear regression statistics such as Pearson correlation coefficient (R), standard error ($Sy.x$), slope, and intercept are all useful for a comparison.

It is also appropriate to break data sets into regions of rising glucose, falling glucose, and relative steady state, as the bias for these regions are affected by the lag effect, and this data stratification can quickly highlight bias issues relative to glucose trending.

In addition to the method comparisons mentioned above, a measure of lag relative to blood glucose should also be made, when the data are appropriate. Appropriate data display a range of glucose values collected over a period of rising and falling glucose such as are observed in a glucose tolerance test. Lag should be measured as the time shift at which best correlation is observed between the reference and continuous data, that is, when the reference data are moved back and forth in time relative to the continuous data.

A very good way to convey a comparison is simply to plot the data against time, with a

¹LifeScan, Milpitas, California.

²LifeScan, St. Paul, Minnesota.

time-axis resolution great enough so that differences can easily be seen (i.e., don't squash all the data together).

Experimental design is an extremely important consideration for comparing different continuous monitors. The experimental designs must be comparable, producing a significant range in ambient glycemia (>100 mg/dL), as well as significant rates of glucose change, i.e., greater than ± 2 mg/dL/min, in order to adequately evaluate the effects of lag.

In the case of continuous monitoring used in hospitals (during surgery or in intensive care units) as a component of tight glycemic control, the capability to measure venous or arterial blood should be a requirement for these devices in this setting. Since there will almost certainly be a central line, obtaining a specimen is straightforward. Of course, the differences observed among capillary, venous, and arterial blood glucose concentrations, particularly after feeding, must be considered.

What are the metrics to characterize the performance of a well-performing continuous monitoring system? What are the limits of these metrics to achieve self-monitoring blood glucose (SMBG) device equivalent standard of care? Would you set a different set of performance criteria for different situations (e.g., 40–80, 80–200, and >200 mg/dL)?

The protocol proposed below involves feeding, insulin administration, and exercise in order to obtain changing glucose values. At least 50% of the values should be changing by >1 mg/dL/min. Glucose values from 50 to 400 mg/dL should be distributed throughout the range, and samples should be taken every 15 min by fingerstick for measurement on an episodic SMBG device. A portion of the sample should be measured on a reference instrument. Similarly, samples should be taken from an alternate site close to the continuous monitoring device and measured on the reference instrument. Use both fingerstick and alternate site samples to assess differences in physiological lag.

All data collection is manual. No exclusions are permitted. Data will be analyzed by determining bias versus the reference method for both the continuous and episodic devices. Data

will be stratified into hypoglycemic (≤ 70 mg/dL), euglycemic (>70 and ≤ 180 mg/dL), and hyperglycemic (>180 mg/dL). Data will also be stratified by rate of change into static (within ± 1 mg/dL/min) and changing (greater than ± 1 mg/dL/min).

The protocol has been structured around the following assumptions and constraints: A minimum of 20 users are included in the study to include enough subject to subject variability; there are 100 data points per cell (not 100 participants); and if a bin is filled with 100 data points, further data points should not be included. If cells have more than 100 data points, performance should be compared by randomly drawing 100 pairs for each bin. This will standardize most performance measures across the cells and would not weight heavily one side of the table.

The acceptance criteria here are based on the Rate Error Grid.¹ The performance criteria for glucose measurements are based upon Zones A and B of this Error Grid.

Based upon 100 observations per cell and at least 90% of observations being the acceptable region, then one could be 90% confident that statistically each cell would have more than 85% of observations meeting the acceptance criteria. Based upon 95% of all 900 observations being in their acceptable region, then one could be 90% confident that overall statistically more than 94% of observations met the acceptance criteria.

Protocol. Create two tables of the format shown in Table 1: one for fingerstick assessment and one for alternative site assessment. Fill each cell with 100 pairs of observations, with a pair being one measurement from the continuous monitoring system and one measurement from fingerstick or alternative site. In general, since continuous monitoring systems can be programmed to measure frequently, data can be acquired for difficult to fill cells (e.g., ≤ 70 mg/dL) by having the system read more frequently.

There are two measurement types available with the continuous glucose monitors, that is, rate of change computed on the basis of two consecutive measurements and average glucose reading computed on the basis of two consecutive measurements. With these two mea-

TABLE 1. REGIONS FOR DATA STRATIFICATION

Glucose/rate of change	≤ 70 mg/dL	$70 < BG \leq 180$ mg/dL	$BG > 180$ mg/dL
≤ -1 mg/dL/min			
Between -1 and $+1$ mg/dL/min			
$\geq +1$ mg/dL/min			

BG, blood glucose.

surements, two errors are available for comparison with the reference measurements. These errors are tabulated as the percentage falling within the acceptable region as defined below.

The acceptable region for each cell is defined in Table 2.

After tabulation, apply the following acceptance criteria to determine if performance is acceptable.

Acceptance criteria. Each cell has at least 90% of observations in the acceptable region, and the total of all cells has at least 95% observations in the acceptable region.

To help illustrate the reasoning behind the acceptance criteria, Table 3 gives some examples.

Whether sensitivity and specificity are sufficient parameters to evaluate the predictive value of a hypoglycemia alarm.

Both sensitivity and specificity are standard measurements of assay performance. However, they are not the most relevant measurements for the patient who wants to know

whether an alarm is to be believed. A more relevant measurement of alarm accuracy for the patient is how likely he is to be in a hypoglycemic state when the system alarms him that he is in a hypoglycemic state (positive predictive value) and how likely he is to be in a state of control when the system is not alarming him that he is in a hypoglycemic state (negative predictive value). This likelihood depends not only on the sensitivity and specificity, but also on the frequency the patient is in a hypoglycemic state (prevalence).

To set forth a couple of concrete examples, assume a system has a sensitivity of 90% (alarms 90% of the time when the patient is in a hypoglycemic state) and a specificity of 90% (does not alarm 90% of the time when the patient is not in a hypoglycemic state). Patient 1 remains out of hypoglycemia 90% of the time and is hypoglycemic only 10% of the time. When the alarm sounds, this patient is in a true hypoglycemic state only 50% of the time (half the alarms are false-positives). On the other hand, consider Patient 2 who remains out of hypoglycemia only 10% of the time and is hypoglycemic 90% of the time. This patient is in a hypoglycemic state over 98% of the time

TABLE 2. PROPOSED PERFORMANCE CRITERIA BASED ON THE CONTINUOUS ERROR GRID

Glucose/rate of change	≤ 70 mg/dL	$70 < BG \leq 180$ mg/dL	$BG > 180$ mg/dL
≤ -1 mg/dL/min	Bias $< +30$ mg/dL and rate of change error $< +1$ mg/dL/min	Bias within $\pm 30\%$ and rate of change error $< +1$ mg/dL/min	Bias $> -30\%$ and rate of change error $< +1$ mg/dL/min
Between -1 and $+1$ mg/dL/min	Bias $< +30$ mg/dL and rate of change error within ± 2 mg/dL/min	Bias within $\pm 30\%$ and rate of change error within ± 2 mg/dL/min	Bias $> -30\%$ and rate of change error within ± 2 mg/dL/min
$\geq +1$ mg/dL/min	Bias $< +30$ mg/dL and rate of change error > -2 mg/dL/min	Bias within $\pm 30\%$ and rate of change error > -2 mg/dL/min	Bias $> -30\%$ and rate of change error > -2 mg/dL/min

BG, blood glucose.

TABLE 3. EXAMPLES TO DEMONSTRATE THE ACCEPTANCE REGION

<i>Cell</i>	<i>True glucose</i>	<i>True rate of change</i>	<i>Limit on predicted glucose</i>	<i>Limit on predicted rate of change</i>
1 (upper left)	50 mg/dL	-1 mg/dL/min	<80 mg/dL	< 0 mg/dL/min
1 (upper left)	70 mg/dL	-3 mg/dL/min	<100 mg/dL	< -2 mg/dL/min
2 (left middle)	50 mg/dL	-1 mg/dL/min	<80 mg/dL	Between +1 mg/dL/min and -3 mg/dL/min
2 (left middle)	70 mg/dL	+1 mg/dL/min	<100 mg/dL	Between +3 mg/dL/min and -1 mg/dL/min
3 (left bottom)	50 mg/dL	+1 mg/dL/min	<80 mg/dL	> -1 mg/dL/min
3 (left bottom)	70 mg/dL	+3 mg/dL/min	<100 mg/dL	> 1 mg/dL/min
4 (middle top)	80 mg/dL	-1 mg/dL/min	Between 56 and 104 mg/dL	< 0 mg/dL/min
4 (middle top)	170 mg/dL	-3 mg/dL/min	Between 119 and 221 mg/dL	< -2 mg/dL/min
5 (middle)	80 mg/dL	-1 mg/dL/min	Between 56 and 104 mg/dL	Between +1 mg/dL/min and -3 mg/dL/min
5 (middle)	170 mg/dL	+1 mg/dL/min	Between 119 and 221 mg/dL	Between +3 mg/dL/min and -1 mg/dL/min
6 (middle bottom)	80 mg/dL	+1 mg/dL/min	Between 56 and 104 mg/dL	> -1 mg/dL/min
6 (middle bottom)	170 mg/dL	+3 mg/dL/min	Between 119 and 221 mg/dL	> 1 mg/dL/min
7 (right top)	200 mg/dL	-1 mg/dL/min	>140 mg/dL	< 0 mg/dL/min
7 (right top)	400 mg/dL	-3 mg/dL/min	>280 mg/dL	< -2 mg/dL/min
8 (right middle)	200 mg/dL	-1 mg/dL/min	>140 mg/dL	Between +1 mg/dL/min and -3 mg/dL/min
8 (right middle)	400 mg/dL	+1 mg/dL/min	>280 mg/dL	Between +3 mg/dL/min and -1 mg/dL/min
9 (right bottom)	200 mg/dL	+1 mg/dL/min	>140 mg/dL	> -1 mg/dL/min
9 (right bottom)	400 mg/dL	+3 mg/dL/min	>280 mg/dL	> 1 mg/dL/min

when the alarm sounds. Clearly, the two patients would have a very different view of the hypoglycemic alarm accuracy when using the same system.

In general, it is important to have accurate results in the region for which the patient spends most of his time. Hence, the false-negative rate is of greater importance than the false-positive rate for patients who spend a large proportion of time in a hypoglycemic state, since a high false-negative rate means hypoglycemia occurs frequently without an alarm. Correspondingly, for patients who spend a large proportion of time outside of hypoglycemia a false-positive rate is more important, since most alarms will be false ones. One can increase (decrease) the false-negative rate while decreasing (increasing) the false-positive rate by changing the alarm threshold. Thus it is possible to “adjust” the hypoglycemic alarm for individual patients, provided one understands the trade-off between

the false-negative and false-positive rates. The receiver operating curve (also called ROC) provides a visual representation of this trade-off.

The advantage of sensitivity, specificity, and the ROC is that they are performance parameters of the system alone, independent of the patient. The disadvantage is that these measurements are not the most relevant measurements of performance from an individual patient's point of view. More relevant measurements are prevalence, positive predictive value, and negative predictive value, as the examples above illustrate. The disadvantage is that these measurements vary from patient to patient. A complete understanding of system performance requires not only understanding the sensitivity, specificity and their trade-off in the ROC for the system, but also the prevalence, positive predictive value, and negative predictive value for a representative sample of the patient population.

How long would it be acceptable to be under 70 mg/dL (under the hypoglycemia threshold) before receiving an alarm?

This depends on the rate of change and the user's activity level. The user should never go below 50 mg/dL at any time without an alarm. For high cognition states, such as driving, the level should not go below 60 or 65 mg/dL. If glucose level is changing slowly or if the user is in a state not requiring high cognition, then 10–15 min might be acceptable. If glucose is dropping rapidly, up to 4 mg/dL/min have been observed, then less than 5 min would be acceptable before receiving an alarm. With a 70 mg/dL threshold for hypoglycemia and with a user in a sleeping state, the permissible time to be below 70 mg/dL would be $(70 - 50 \text{ mg/dL}) / 4 \text{ mg/dL/min} = 5 \text{ min}$. In the case of high cognition, $(70 - 65 \text{ mg/dL}) / 4 \text{ mg/dL/min}$ is 75 s, a little over a minute.

A related point is how far below 70 mg/dL can glucose drop before an alarm can be considered a true-positive. Even mild hypoglycemia may induce an undesirable counter regulatory hormone response and induce hypoglycemia unawareness. True-positive alarms need to be given before glucose reaches the 60 mg/dL level.

Would you accept a greater rate of false alarms if your glucose was falling rapidly at 80–90 mg/dL compared with falling slowly?

Yes, a greater rate of false alarms would be acceptable given the greater risk of this situation. Again, whatever notification was required to prevent the user from going below 50 mg/dL would be appropriate in the case of sleep or below 60 or 65 mg/dL in the case of high cognition activities.

What is an acceptable sensor response time (that is, total system lag) in various user scenarios such as exercise, rapidly rising glucose, rapidly falling glucose, medication, meals, etc.? Does the glucose concentration affect this response time?

Generally, a sensor response time of anything less than 10 min is probably acceptable for most situations. The greater the lag time is, the greater the potential effect of rapidly chang-

ing glucose in terms of creating a measurement error. This is probably an acute problem only when glucose is rapidly falling and glucose concentration is normal to low ($<110 \text{ mg/dL}$), and has been relatively steady, meaning no discernible trend has been found with the previous two or more data points. If we assume that the fastest rate of change is 4 mg/dL/min , then a lag of 10 min might mean a reporting error of 40 mg/dL (assuming an otherwise accurate measurement). Under these somewhat unusual conditions of rapid glucose change, it is possible that a subject could drop into hypoglycemia undetected.

A more likely situation is that some trending information will be available to make a prediction of hypoglycemia in the case of rapidly falling glucose. However, because this cannot be guaranteed, reducing system lag to $<5 \text{ min}$ is desirable when blood glucose is $<110 \text{ mg/dL}$.

If blood glucose is rising, or the glucose concentration is high, the response time can be relaxed without serious acute effect.

How many times per night would a caregiver or a user accept being awakened by a false alarm for hypoglycemia?

The anecdotal experience from experience with the Cygnus (Redwood City, CA) Gluco-Watch® is that more than once or twice a week is unacceptable for either the caregiver or the user. More than this and they stop using the device.

For an individual user it would probably depend on his or her past experience with nocturnal hypoglycemia and the severity of those episodes. For those who have had severe episodes, possibly one false alarm a night might be acceptable. For those who are hypoglycemia aware and have not required assistance to treat hypoglycemia, very high levels of sensitivity and specificity would be required.

How should calibrations preceding a rapid change in glucose be handled (particularly in the case of lag)?

For devices susceptible to lag, such devices should not be calibrated at a time when it is likely that glucose concentrations in blood or ISF are

likely to be changing rapidly. In order to achieve the optimal accuracy, calibrations should be performed when glucose concentrations are relatively stable. If, however, the device measures a change in glucose level prior to or following calibration, the calibration should be locked out or rejected and the user prompted to recalibrate when glucose levels have stabilized. A rejected calibration may still be valuable as a quality control check if it confirms the device is reading to within, say, a 99% confidence level. In this situation, device readings would be accepted until the recalibration. If the quality control check fails, the user should be instructed to use another device to measure glucose levels until recalibration can be accomplished. If it is possible to factory-calibrate a device, the effects of lag and SMBG error are mitigated.

Which of your patients would benefit most from a continuous monitoring system?

Intensive insulin users, patients with hypoglycemic unawareness and/or nocturnal hy-

poglycemia, patients with brittle diabetes whose control varies from day to day, and those patients for whom it has been difficult to find a therapeutic regimen that achieves targeted glycemic control (both hemoglobin A1c target and glucose excursion, i.e., postprandial, targets).

REFERENCE

1. Kovatchev BP, Gonder-Frederick LA, Cox DJ, Clarke WL: Evaluating the accuracy of continuous glucose-monitoring sensors: Continuous glucose-error grid analysis illustrated by TheraSense Freestyle Navigator data. *Diabetes Care* 2004;27:1922–1928.

Address reprint requests to:
Paul G. Hayter, M.S.
LifeScan
1000 Gibraltar Drive
Milpitas, CA 95035

E-mail: phayter@lfsus.jnj.com