Journal of Diabetes Science and Technology
Volume 7, Issue 4, July 2013
© Diabetes Technology Society

SYMPOSIUM

Performance Evaluations of Continuous Glucose Monitoring Systems: Precision Absolute Relative Deviation Is Part of the Assessment

Karin Obermaier, B.S., Günther Schmelzeisen-Redeker, Ph.D., Michael Schoemaker, Ph.D., Hans-Martin Klötzer, Ph.D., Harald Kirchsteiger, Ph.D., Heino Eikmeier, Ph.D., and Luigi del Re, Ph.D.

Abstract

Background:

Even though a Clinical and Laboratory Standards Institute proposal exists on the design of studies and performance criteria for continuous glucose monitoring (CGM) systems, it has not yet led to a consistent evaluation of different systems, as no consensus has been reached on the reference method to evaluate them or on acceptance levels. As a consequence, performance assessment of CGM systems tends to be inconclusive, and a comparison of the outcome of different studies is difficult.

Materials and Methods:

Published information and available data (as presented in this issue of *Journal of Diabetes Science and Technology* by Freckmann and coauthors) are used to assess the suitability of several frequently used methods [International Organization for Standardization, continuous glucose error grid analysis, mean absolute relative deviation (MARD), precision absolute relative deviation (PARD)] when assessing performance of CGM systems in terms of accuracy and precision.

Results:

The combined use of MARD and PARD seems to allow for better characterization of sensor performance. The use of different quantities for calibration and evaluation, e.g., capillary blood using a blood glucose (BG) meter versus venous blood using a laboratory measurement, introduces an additional error source. Using BG values measured in more or less large intervals as the only reference leads to a significant loss of information in comparison with the continuous sensor signal and possibly to an erroneous estimation of sensor performance during swings. Both can be improved using data from two identical CGM sensors worn by the same patient in parallel.

continued →

Author Affiliation: Roche Diagnostics GmbH, Mannheim, Germany

Abbreviations: (ARD) absolute relative deviation, (BG) blood glucose, (CG-EGA) continuous glucose error grid analysis, (CGM) continuous glucose monitoring, (CLSI) Clinical and Laboratory Standards Institute, (EGA) error grid analysis, (MARD) mean absolute relative deviation, (PARD) precision absolute relative deviation, (SD) standard deviation

Keywords: assessment criteria, continuous glucose monitoring, glucose sensors, mean absolute relative deviation, precision absolute relative deviation

Corresponding Author: Karin Obermaier, B.S., Roche Diagnostics GmbH, DCRDVL 6164, Sandhofer Straße 116, 68305 Mannheim, Germany; email address karin.obermaier@roche.com

Abstract cont.

Conclusions:

Evaluation of CGM performance studies should follow an identical study design, including sufficient swings in glycemia. At least a part of the study participants should wear two identical CGM sensors in parallel. All data available should be used for evaluation, both by MARD and PARD, a good PARD value being a precondition to trust a good MARD value. Results should be analyzed and presented separately for clinically different categories, e.g., hypoglycemia, exercise, or night and day.

J Diabetes Sci Technol 2013;7(4):824–832

Introduction

The performance of blood glucose (BG) meters is evaluated according to standards described in guidance papers, but no internationally accepted standard exists for continuous glucose monitoring (CGM) systems, comparable with the International Organization for Standardization (ISO) 15197 standard.[1] Although all manufacturers claim that their CGM systems are "adjunctive" devices, which are not intended to replace BG measurements for insulin dosing adjustments but to provide actual information about the current glycemia, it cannot be excluded that patients base their therapeutic decision on CGM information rather than on BG results.

The measurement performance of CGM systems is typically described by comparing their signal to BG values and quantifying the deviation and its clinical relevance, mostly using point and trend accuracy (defined with respect to the reference BG value) and algorithms that differ from company to company.

Continuous glucose monitoring systems are calibrated with BG measurement after insertion and recalibrated thereafter in regular intervals in which the CGM signal measures the interstitial glucose whereas BG systems measure capillary glucose. The glucose in these compartments differs physiologically (e.g., depending on glucose rate of change). This results in a so-called "physiological lag time" between these compartments. An additional delay (physical lag time) is introduced by the glucose measurement *per se* (glucose transport within the sensor) and the averaging algorithms built into the CGM systems to reduce the noise of the signal.

All this may explain why the procedures suggested by the existing Clinical and Laboratory Standards Institute (CLSI) guideline for evaluation of CGM systems[2] are not widely used. Indeed, several years of experts' discussions have not yet led to a "standard measurement setup" for evaluation of CGM system performance, and it is difficult to compare different systems evaluated by manufacturer-specific procedures.

Indeed, the CLSI guideline proposes procedures for the design of clinical studies to collect data describing CGM system performance, asking for both numerical and clinical evaluation. Unfortunately, this guideline limits the evaluation to paired data series between a frequently sampled BG measurement (i.e., reference) and CGM readings, even though considerable intersensor variations applied in the same subject are known.[3] To date, the recommendations by the CLSI guideline have been followed completely in only one study; results from other studies based on these recommendations are presented in this issue of *Journal of Diabetes Science and Technology*.

The lack of definition of standard assessment methods for CGM performance is a serious shortcoming considering the wide number of different measures used and the (subtle) differences in the procedures used to compute them, like in the case of the standard deviation (SD). In some cases, inconsistent reporting (e.g., indicating accuracy in percentage for some ranges and in milligrams per deciliter for others) add to the confusion.

Against this background, several authors have analyzed the pitfalls of widely used measures and suggested new methods or (complex) evaluation procedures [e.g., Kollman and coauthors,[4] Kovatchev and coauthors,[5] Wentholt and coauthors,[6] and Zisser and coauthors[7] tested the SEVEN® CGM system by Dexcom with high measurement frequency (20 min) against a laboratory method using a precision measure (precision absolute relative deviation [PARD]) on a subgroup of patients to characterize sensor precision].

With the widespread use of CGM systems, transparent and clinically meaningful criteria are needed to allow a significant comparison of different CGM systems. These criteria should be simple and straightforward to compute while describing the characteristics of CGM systems appropriately. Therefore the aim of this article is to critically evaluate different summary measures to characterize their properties and suggest a best practice.

Materials and Methods

Data

As this paper is not related to a specific clinical study or CGM system; data from a clinical study with a CGM system that is in clinical development (presented by Freckmann and coauthors[8] in this issue of *Journal of Diabetes Science and Technology*) are used as basis for a number of analyses. The clinical study was designed following the recommendations provided by the CLSI guideline POCT05-A, "Performance Metrics for Continuous Interstitial Glucose Monitoring." Glucose swings were obtained by induced hypoglycemia and induced hyperglycemia on two study days, and two identical but independent CGM systems were applied in parallel on the patients at all times. **Figure 1** shows a typical study day, where hyperglycemia was induced in area C followed by a hypoglycemic episode around lunchtime.

Data Analysis

Choice of the Reference

Most commonly, BG measurement results are used as reference data for comparisons. Those are measured preferably with a laboratory standard device in venous blood samples or alternatively with a standard test-strip-based BG meter in capillary blood samples (e.g., if one is interested in the CGM accuracy in an outpatient setting). In the latter case, the limited accuracy of BG meters should be considered in the evaluation. Relying on venous blood samples introduces an additional uncertainty because it is well-known that, postprandially, there is a difference between these and arterial (capillary) glucose levels. More important, the rate of change of BG levels in capillary or venous blood or in interstitial fluid is usually different at a given point in time, so different glucose measurement results obtained do not necessarily indicate a measurement error but may arise from a physiological difference. In other words, points paired according to their measurement time stamp may be inadequately paired from the physiological point of view. As a consequence, when the CGM system is calibrated, as mostly done by means of BG meter readings, it is questionable whether an evaluation of its accuracy against laboratory values measured in venous samples really improves the quality of the evaluation.

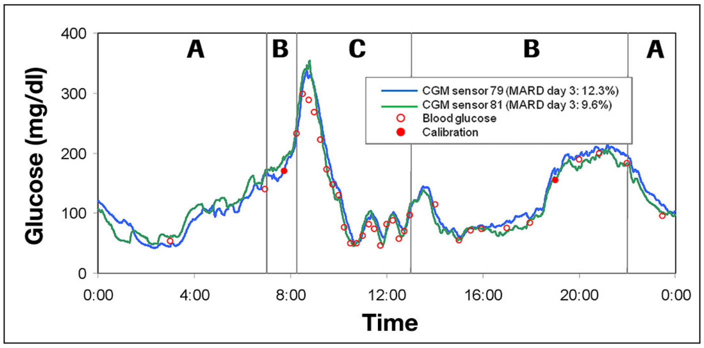

Figure 1. Glycemic profile of a given patient wearing two identical CGM systems over a period of 24 h. A, nighttime with 1–2 BG readings; B, daytime with ≥1 BG readings every hour; C, dynamic phase 1 BG reading every 15 min.

Independent from this, BG measurements will provide paired data only for a fraction of the readings of a CGM system. Deviations occurring between BG measurements in time (there might be hours in between) are not detected, and some of them, like the effect of mechanical pressure on the sensor tip in the tissue, represent specific problems of the CGM sensor technology.

Unfortunately, there is no real alternative to BG measurements for accuracy evaluation, but it is possible and important both to improve the plausibility of the accuracy estimation and to gain better insight into the performance of the

CGM system by using paired measurements with another, identical sensor worn by the same patient. This allows analysis using a much larger number of CGM data collected to assess precision.

Numerical Assessment of Accuracy

Many numerical approaches have been proposed for the assessment of accuracy; the most common one is calculation of the mean (or median) absolute relative deviation (MARD). The absolute relative deviation (ARD) is defined as

$$ARD = 100\,\frac{|y_{CGM} - y_{RBG}|}{y_{RBG}},$$

where y_{RBG} is the reference BG concentration and y_{CGM} is the glucose concentration measured at the same point in time by the CGM system. Consequently, the MARD is calculated as the mean value of individual ARDs and the median ARD as the median of individual ARDs. The latter is less dependent on outliers and tends therefore to be lower than the MARD. Unfortunately, in many publications, it is not clear if a mean or median ARD was calculated.

The MARD is easy to compute and interpret and allows summarizing the properties of a CGM system by few figures. However, the MARD allows no distinction between positive and negative errors or between systematic and random errors. As it is computed as relative variation, its value is affected by the glucose values of the study participants. The MARD strongly depends on the composition of the study cohort and the study setting, i.e., how much swings in glycemia are induced and how often and how long BG is in the hypoglycemic or hyperglycemic range. In the study of Nielsen[9] for example, the MARD of the same CGM system is significantly lower for a cohort of patients with type 2 diabetes than for those with type 1 diabetes, who are known to exhibit higher glucose variability. For these reasons, it is highly advisable to perform separate evaluations for different glucose ranges (hypoglycemic, euglycemic, and hyperglycemic ranges) and clearly mark which patient group was studied.

ISO 15197:2013[1] essentially calls for a classification of the errors of BG meters according to their relative magnitude (≤±5%, 10%, 15%, and 20%). It is easy to compute but essentially suffers from the same problems as the error grid analysis (EGA)—somehow arbitrary limits and, if separate glucose ranges are used, too many figures to allow a simple comparison.

Clinical Assessment

Numerical methods do not take into account the clinical relevance of the numbers calculated. Separate evaluation of numerical criteria in different clinical ranges allows an easier clinical evaluation, but specific methods have been developed. A frequently used assessment is the Clarke EGA used in its original version[10] or extended with analysis of the rate error [continuous glucose error grid analysis (CG-EGA)].[5] This graphical tool should connect the imprecision of the CGM system with implications on the therapy, e.g., by choosing an inappropriate insulin bolus based on a wrong BG measurement. A potential benefit of those methods is that they establish a correlation to BG that is the actual variable of interest in diabetes therapy. Additionally, the CG-EGA incorporates the trend information that can be used to compensate for lag times.

However, EGA also suffers from several drawbacks. As it is also based on a comparison between BG and CGM readings, all the limits mentioned earlier—neglect of most data and "wrong" pairing—hold true here as well. Additionally, the choice of the borders between regions is rather arbitrary, and regions A and D are adjacent. Attempts to define a better distribution—such as the consensus grid[11]—have not succeeded; however, there are new EGA versions in development to overcome certain limitations.

The CG-EGA represents a noticeable progress with respect to EGA for the CGM sensors, but it suffers from some additional drawbacks, in particular, the high number of BG measurements required to estimate the rate of change of glucose. In some studies, glucose rates of change ≥5 mg/dl/min have been observed. For such rates of change, the CG-EGA does not capture the information that a CGM user would have at the same time from the combination of value and trend.

It might be questionable whether the EGA is an appropriate means to quantify sensor performance. In the 2011 meeting of the Diabetes Technology Society,[12] it was concluded that these error grids, which were introduced in 1987, were no longer meaningful for BG meters partly because very few meters have ever failed these criteria and partly because of the seemingly arbitrary divisions between the zones. A similar statement might hold true for CGM systems, even though, as mentioned earlier, there is an ongoing effort to develop a new and improved EGA[13] that might prove useful for CGM evaluation.

In summary, while error grids yield very interesting information, they are not suitable for a conclusive performance evaluation, especially for comparisons between different CGM systems.

Precision Using Relative Accuracy between Two Sensors Running in Parallel

A logical step forward to overcome the limitations of using single BG measurements as reference is using another identical CGM reading as a second "reference." While this approach is not simple to use for determining accuracy, the absence of relative delays and the availability of large number of data that can be analyzed allow getting a complementary insight into the sensor properties (also by calculating other summary measures).

In principle, all these mentioned methods could be adapted, taking one CGM system reading instead of the BG. In particular, the MARD can be generalized to the PARD[7,14] as

$$PARD = 100\,\frac{|\,y_{CGM1} - y_{CGM2}\,|}{\text{mean}(y_{CGM1}, y_{CGM2})}\ (0.1).$$

Mean and median values of PARD are as easy to compute and interpret as MARD. Calculation of the SD is an alternative (see **Figure 2**), but the existence of several different algorithms for its computation—e.g., SD of all differences, SD of the average difference of the single experiments, robust or less robust estimators—makes the comparability difficult. Notice that using 95% percentiles of the absolute differences between the two sensors would better assess large intermittent differences between the sensors (e.g. due to dropouts of one sensor). Even with the PARD or SD, intermittent large differences are not detectable due to the large number of differences in a normal range in the data sets. However, this would add another value, and for the sake of simple comparability, this is probably not too important.

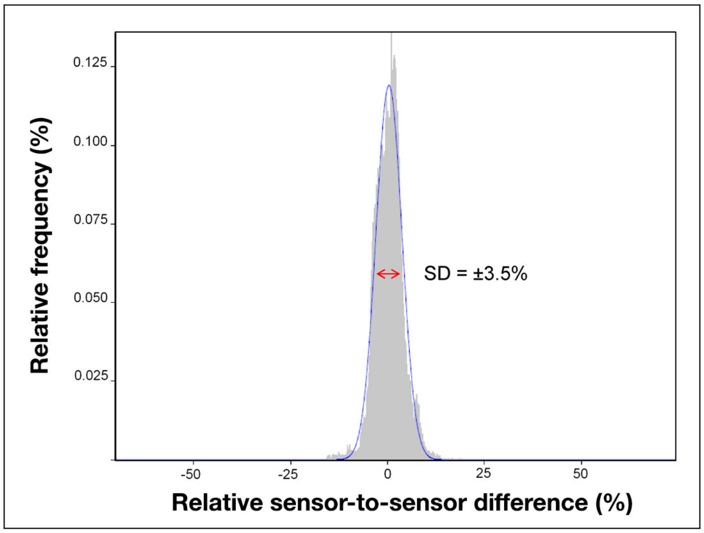

Figure 2. Interpretation of SD.

Results and Discussion

Analysis of some typical cases of CGM recordings (combined with some unusual periods) was performed (see **Figure 3**) to discuss the pros and cons of the different measures described earlier for characterization of CGM systems. One case (case 3 in **Figure 3**) is of interest, as it would lead to a wrong clinical decision depending on which measure is applied. It is important to stress that those compression effects are not inherent to the CGM technology *per se* but depend on a specific patient–tissue interaction. In other words, such effects may not appear in a predictable manner. Another CGM recording shows a parallel measurement with two identical sensors worn by the same patient (see **Figure 4**). The recording of one sensor strongly deviates from the other one in a given time interval. An analysis based on the one sensor would detect a clinically relevant issue, whereas the other one not.

It is noteworthy that such deviations are not an exception and may occur also due to technological issues. **Figure 5** shows a defined time period of a CGM study presented by Freckmann and coauthors[8] in this issue of *Journal of Diabetes Science and Technology*, in which several such cases are documented. For example, on day 3 of their study, one of the Guardian sensors deviates and reaches its saturation value; the consequence is that the CGM system displayed glucose numbers that were erroneous and too high. While the issue was detected in due time in this case and the issue was solved by a recalibration, under daily life conditions, one wonders what the consequences of such an issue would have been (especially when the signal would have been used in an artificial pancreas setup). Such a phase of deviations of the CGM signal from glycemia could have induced a potential risk for the user if he would have drawn therapeutic conclusions without checking glycemia with a capillary measurement. Incidentally, the same sensor again generates misleading results 3 days after this first phase with erroneous results. Notice also that all CGM systems exhibit certain differences, however, not to this extent. If only the results obtained with the "good" CGM system would have been used for evaluation of the system performance (or if only infrequent measurements would have been taken), the measurement quality of this system would have been overestimated.

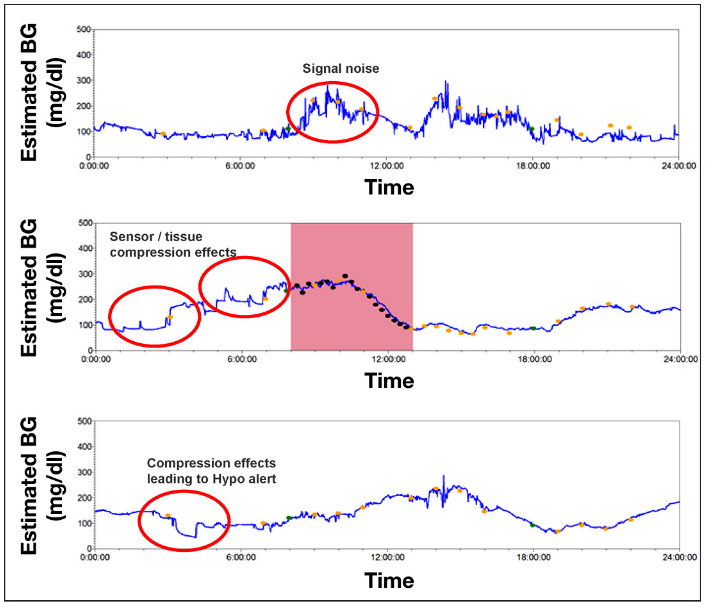

Figure 3. Anomalies seen in CGM recordings.

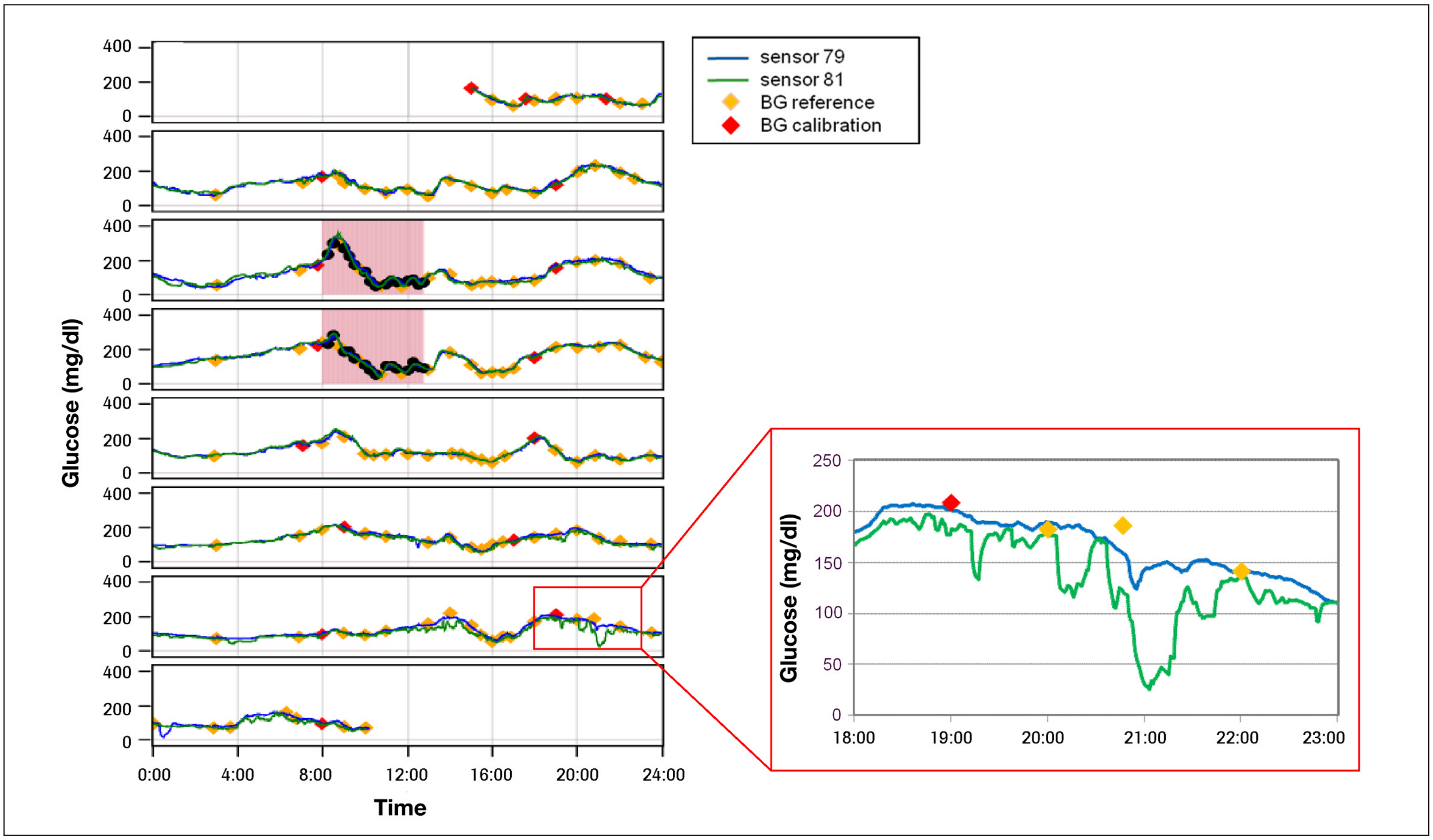

Figure 4. Differences in two CGM recordings in the same patient in parallel, one day with significant differences.

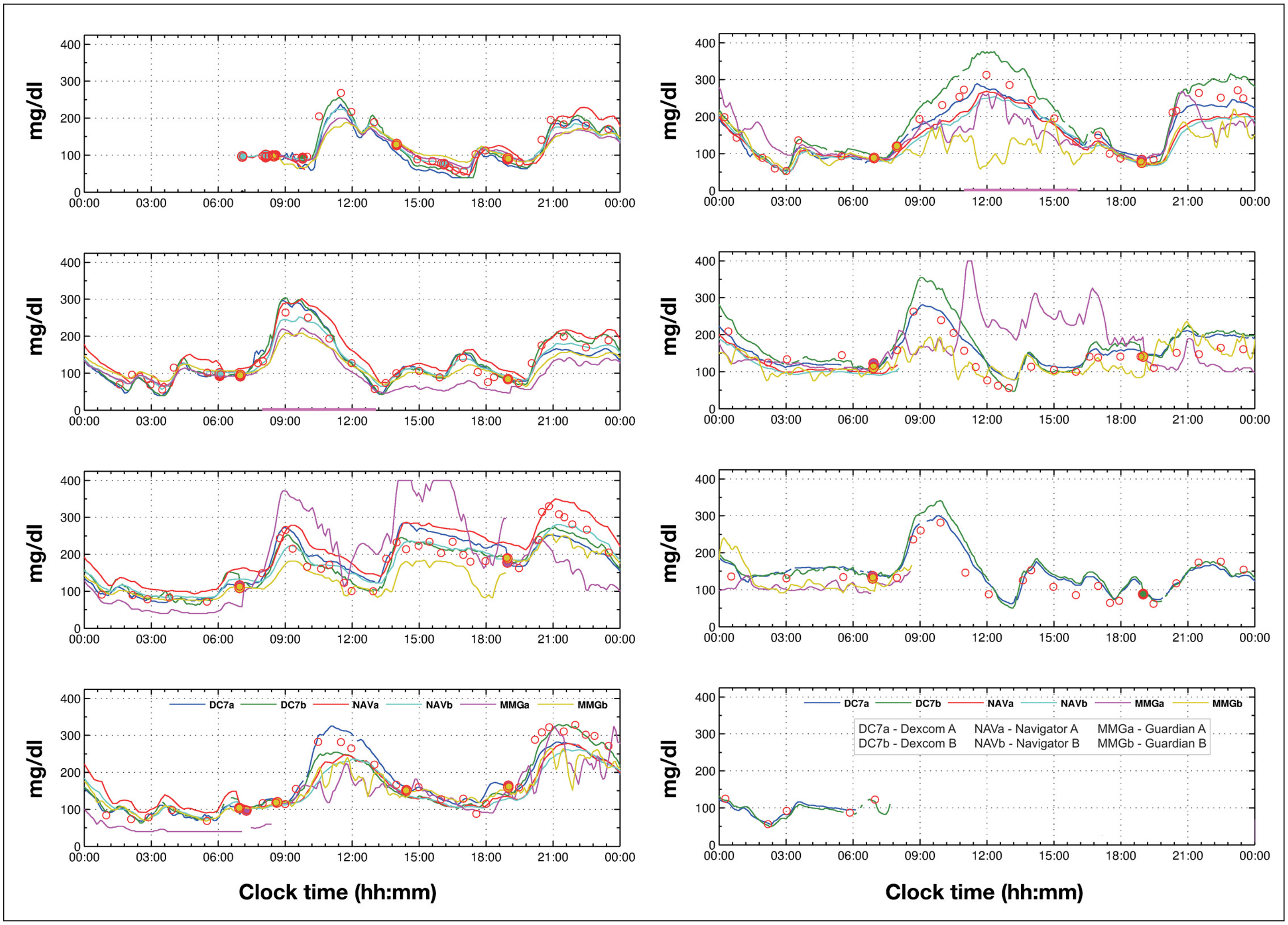

Figure 5. Glucose profiles registered with different CGM systems used in parallel in a given patient (unpublished data).

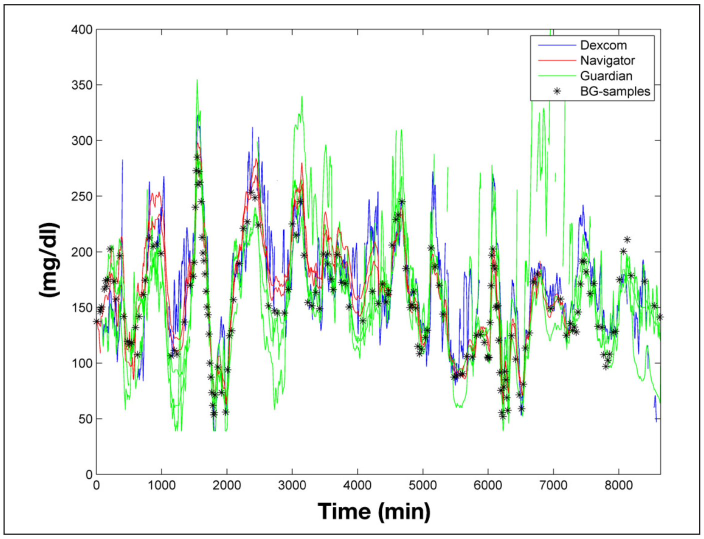

Figure 6. Raw CGM and BG measurements for a patient wearing six sensors in parallel.

To better understand how two averaged measures (MARD and PARD) detect (or not) such differences, another CGM recording from the same data[8] has been analyzed (see **Figure 6**, which shows the measurements for a Dexcom SEVEN (Dexcom Inc., San Diego, CA). **Figure 7** shows the corresponding values of ARD and PARD (and not the mean values), compared with the mean values of PARD and MARD for both sensors plotted in red using a window length of one day. The considerable deviation of one Dexcom CGM system remained undetected by MARD (computed between BG and CGM system), but it was detected by PARD.

It is important to notice that, for evaluation of a particular sensor, the two sensors elements should be redundant, meaning, for example, not be housed in the same probe. Tissue compression would affect two (not independent) sensors in one housing, and this would remain undetected by PARD. However, there is one more important key point: the number of data that can be reasonably acquired from BG. We also do not suggest using two sensors in daily use, but two sensors (on different sides) should be used for the evaluation only. In general, detection of abnormal conditions such as tissue compression can and should be done using data plausibility analysis.

The computation of the PARD value as presented earlier could be modified also to account specifically for different physiological conditions. In particular, instead of calculating a single value, one could compute a value for stable glucose conditions and one for changing glucose conditions. Since the peaks of MARD typically occur during transient conditions, this is essentially equivalent to concentrating on the PARD during phases with high MARD.

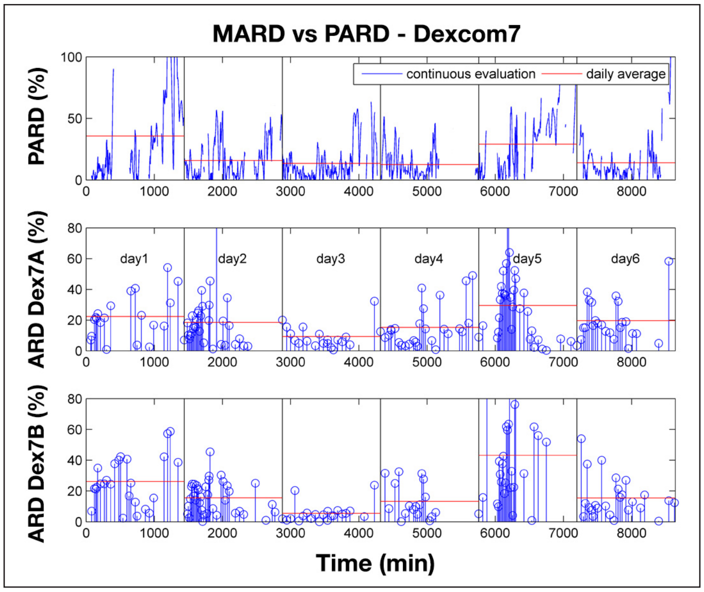

Figure 7. Continuous PARD and MARD evaluation of the measurements of **Figure 6**.

Calibration of both sensors for PARD evaluation has to be done with the same glucose reference values because the aim is to address sensor errors. The effect of erroneous calibration measurements was avoided by repeatedly measuring BG until two consecutive measures show identical values. We believe that calibration must be strictly the same, as PARD gives the difference between two sensors and this difference—if present—should reflect sensor errors and not the different calibration.

Conclusion

The aim of this article was to discuss different approaches to characterize CGM recordings (along with their pros and cons). Evaluation of CGM systems by means of infrequent BG measurements does not really reflect the nature of CGM and may lead to misleading results. On the other hand, because BG values are still regarded as the "gold standard," accuracy has to be documented in comparison with BG values. Nevertheless, the MARD should be computed separately for different glucose ranges and for specific clinically relevant conditions. The MARD and median ARD have the clear advantage of providing essential accuracy information in an easily interpretable and comparable way. However, MARD does not appropriately detect certain issues, either because of the limited number of paired data points, or because of the inherent physiological difference between the two compartments in which glucose is measured. In case of rapid changes in glycemia, the MARD may provide too high results.

In view of these shortcomings, calculation of the PARD becomes important to support the accuracy assessment by the MARD. The PARD *per se* does not convey sufficient information—not only for the lack of comparison with the BG values, but for instance, an increase of MARD during transient phases can be a good indicator of changes in the lag time of CGM systems that are not detected by PARD as the system remains reproducible (but reproducibly slower).

If possible, more evaluation criteria, such as false alarm rates, sensitivity, and specificity, should be calculated, because they provide additional information and also allow a better plausibility assessment of the MARD and PARD results. However, they should be computed only if the MARD and PARD are reasonable.

Funding:

Data presented in this study were collected during clinical studies organized by Roche Diagnostics GmbH. Preparation of this manuscript was supported by an unrestricted grant by Roche Diagnostics GmbH.

Disclosures:

Karin Obermaier, Guenther Schmelzeisen-Redeker, Michael Schoemaker, and Heino Eikmeier are employees of Roche Diagnostics.

Acknowledgments:

The authors thank Dr. Lutz Heinemann, Dr. Guido Freckmann, and Dr. Volker Lodwig for their constructive comments to the manuscript.

References:

1. International Organization for Standardization. ISO 15197:2013. *In vitro* diagnostic test systems -- requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus. *http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=54976*.
2. Clinical and Laboratory Standards Institute. Performance metrics for continuous interstitial glucose monitoring; approved guideline. *http://shopping.netsuite.com/c.1253739/site/Sample_pdf/POCT05A__sample.pdf*.
3. Castle JR, Ward WK. Amperometric glucose sensors: Sources of error and potential benefit of redundancy. J Diabetes Sci Technol. 2010;4(1):221–5.
4. Kollman C, Wilson DM, Wysocki T, Tamborlane WV, Beck RW; Diabetes Research in Children Network Study Group. Limitations of statistical measures of error in assessing the accuracy of continuous glucose sensors. Diabetes Technol Ther. 2005;7(5):665–72.
5. 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(8):1922–8.
6. Wentholt IM, Hart AA, Hoekstra JB, Devries JH. How to assess and compare the accuracy of continuous glucose monitors? Diabetes Technol Ther. 2008;10(2):57–68.
7. Zisser HC, Bailey TS, Schwartz S, Ratner RE, Wise J. Accuracy of the SEVEN continuous glucose Monitoring System: Comparison with frequently sampled venous glucose measurements. J Diabetes Sci Technol. 2009;3(5):1146–54.
8. Freckmann G, Pleus S, Link M, Zschornack E, Klötzer HM, Haug C. Performance evaluation of three continuous glucose monitoring systems: comparison of six sensors per subject in parallel. J Diabetes Sci Technol. 2013;7(4):842–53.
9. Nielsen JK, Freckmann G, Kapitza C, Ocvirk G, Koelker KH, Kamecke U, Gillen R, Amann-Zalan I, Jendrike N, Christiansen JS, Koschinsky T, Heinemann L. Glucose monitoring by microdialysis: performance in a multicentre study. Diabet Med. 2009;26(7):714–21.
10. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-Monitoring of blood glucose. Diabetes Care. 1987;10(5):622–8.
11. Parkes JL, Slatin SL, Pardo S, Ginsberg BH. A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. Diabetes Care. 2000;23(8):1143–8.
12. Walsh J, Roberts R, Bailey T. Guidelines for optimal bolus calculator settings in adults. J Diabetes Sci Technol. 2011;5(1):129–35.
13. Klonoff DC. The need for clinical accuracy guidelines for blood glucose monitors. J Diabetes Sci Technol. 2012;6(1):1–4.
14. Bailey T, Zisser H, Chang A. New features and performance of a next-generation SEVEN-day continuous glucose monitoring system with short lag time. Diabetes Technol Ther. 2009;11(12):749–55.