Assessing the Performance of Handheld Glucose Testing for Critical Care

Gerald J. Kost, M.D., Ph.D., 1 Nam K. Tran, B.S., 1 Richard F. Louie, Ph.D., 1 Nicole L. Gentile, B.S., 1 and Victor J. Abad, M.A. 2

Abstract

Background: We assessed the performance of a point-of-care (POC) glucose meter system (GMS) with multitasking test strip by using the locally-smoothed (LS) median absolute difference (MAD) curve method in conjunction with a modified Bland-Altman difference plot and superimposed International Organization for Standardization (ISO) 15197 tolerance bands. We analyzed performance for tight glycemic control (TGC).

Methods: A modified glucose oxidase enzyme with a multilayer-gold, multielectrode, four-well test strip (StatStrip®, NOVA Biomedical, Waltham, MA) was used. There was no test strip calibration code. Pragmatic comparison was done of GMS results versus paired plasma glucose measurements from chemistry analyzers in clinical laboratories. Venous samples \( n = 1,703 \) were analyzed at 35 hospitals that used 20 types of chemistry analyzers. Erroneous results were identified using the Bland-Altman plot and ISO 15197 criteria. Discrepant values were analyzed for the TGC interval of 80–110 mg/dL.

Results: The GMS met ISO 15197 guidelines; 98.6% (410 of 416) of observations were within tolerance for glucose <75 mg/dL, and for ≥75 mg/dL, 100% were within tolerance. Paired differences (handheld minus reference) averaged -2.2 (SD 9.8) mg/dL; the median was -1 (range, -96 to 45) mg/dL. LS MAD curve analysis revealed satisfactory performance below 186 mg/dL; above 186 mg/dL, the recommended error tolerance limit (5 mg/dL) was not met. No discrepant values appeared. All points fell in Clarke Error Grid zone A. Linear regression showed \( y = 1.018x - 0.716 \) mg/dL, and \( r^2 = 0.995 \).

Conclusions: LS MAD curves draw on human ability to discriminate performance visually. LS MAD curve and ISO 15197 performance were acceptable for TGC. POC and reference glucose calibration should be harmonized and standardized.

Introduction

The locally-smoothed (LS) median absolute difference (MAD) curve method 1 employs a nonparametric statistical algorithm that provides quantitative assessment of performance of a point-of-care (POC) test. Errors do not offset each other because there is no algebraic summing of positive and negative errors. For POC glucose testing, continuity of the LS MAD curve enhances simultaneous visual assessment of performance in hypo-, normo-, and hyperglycemic ranges and allows quick interpretation for glucose ranges relevant to tight glycemic control (TGC) protocols in critical care settings.

Our objectives were: (1) to use the LS MAD curve method and a modified Bland-Altman 2 difference plot with superimposed International Organization for Standardization (ISO) 15197 3 tolerance bands to evaluate POC glucose testing performance across multiple institutions; (2) to evaluate a handheld glucose meter system (GMS) that corrects for hematocrit effect and compensates for oxidizing substances; and (3) to assess suitability for a TGC interval of 80-110 mg/dL and for the critical adjacent ranges immediately outside the TGC interval where insulin infusion rates are changed to achieve the target glucose level.

Materials and Methods

Multicenter strategy

The Point-of-Care Testing Center for Teaching and Research (POCT・CTRSM) (University of California, Davis, California).
Davis, CA) acted as an independent arbitrator without remuneration. The University of California Davis Medical Center did not perform instrument comparisons or contribute a dataset. The Institutional Review Board at the University of California, Davis approved the multicenter-arbitrator strategy. Subjects included patients in critical care sites (e.g., intensive care unit, neonatal intensive care unit, operating room, emergency room, and labor and delivery) and also diabetes clinics, the nursery, and outpatient centers.

Thirty-five U.S. medical centers anonymously provided datasets obtained from parallel analysis of fresh remnant lithium or sodium heparinized venous or blood gas syringe samples (n = 1,703) using a uniform protocol, the GMS to measure whole-blood glucose, and chemistry analyzers to measure plasma glucose immediately following centrifugation (Table 1). No fingerstick samples were used. Glucose was measured in singleton. At the majority of sites, the manufacturer provided assistance with testing. Samples were processed without delay except when placed on a rocker at room temperature to allow glycolysis to produce low range glucose levels.

**GMS**

Hospitals evaluated the StatStrip™ handheld glucose meter (NOVA Biomedical, Waltham, MA), which uses a modified glucose oxidase enzyme method and multilayer-gold, multielectrode, four-well test strip. No calibration codes or lot numbers need be entered before measurement. The sample volume was 1.2 μL, and the analysis time was 6 s. StatStrip measures hemocrit by an impedance method and corrects glucose values for abnormal hematocrits. The StatStrip received clearance from the Food and Drug Administration (FDA) for use in neonatal testing and is intended for in vitro diagnostic use with capillary, venous, and arterial whole blood. It is approved for all hospital areas, including but not limited to critical care, the operating room, inpatient sites, and outpatient sites, such as diabetes clinics.

**Quality control and reference instruments**

GMS measurements were compared to parallel measurements of plasma glucose performed within a few minutes using 20 types of clinical laboratory chemistry analyzers at 35 U.S. hospitals (Table 1). Test strips consisted of approximately 20 different lots. Reference instruments were quality controlled daily according to manufacturers’ specifications as part of laboratory requirements for reporting patients’ results.

Glucose meters were operated within control according to the manufacturer’s specifications. Three control levels were provided: 46–76, 88–128, and 253–323 mg/dL; the middle

<table>
<thead>
<tr>
<th>Manufacturer (<a href="http://www">www</a>.)</th>
<th>Reference instrument</th>
<th>Sites</th>
<th>Number of observations</th>
<th>Group</th>
<th>Bias (mg/dL) [SD, P], median (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbott (abbott.com)</td>
<td>Architect</td>
<td>1</td>
<td>30</td>
<td></td>
<td>-1.47 [10.09, 0.284]</td>
</tr>
<tr>
<td></td>
<td>Areoset</td>
<td>1</td>
<td>25</td>
<td>55</td>
<td>1 (-43 to +17)</td>
</tr>
<tr>
<td>Bayer (bayer.com)</td>
<td>Advia 1650</td>
<td>2</td>
<td>51</td>
<td>51</td>
<td>-3.82 [9.42, 0.006]</td>
</tr>
<tr>
<td></td>
<td>DCx</td>
<td>2</td>
<td>60</td>
<td>656</td>
<td>1 (-96 to +45)</td>
</tr>
<tr>
<td></td>
<td>Synchron</td>
<td>1</td>
<td>200</td>
<td></td>
<td>-1 (-32 to +18)</td>
</tr>
<tr>
<td></td>
<td>cx-9</td>
<td>1</td>
<td>30</td>
<td></td>
<td>-0.64 [7.48, 0.09]</td>
</tr>
<tr>
<td>Dade Behring</td>
<td>RxL</td>
<td>4</td>
<td>270</td>
<td></td>
<td>-1 (-29 to +43)</td>
</tr>
<tr>
<td></td>
<td>Dimension</td>
<td>1</td>
<td>30</td>
<td>384</td>
<td>-1 (-29 to +43)</td>
</tr>
<tr>
<td></td>
<td>Xpand</td>
<td>1</td>
<td>84</td>
<td></td>
<td>-1 (-29 to +43)</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>Vitros Fusion</td>
<td>1</td>
<td>30</td>
<td></td>
<td>-2.46 [8.86, &lt;0.001]</td>
</tr>
<tr>
<td></td>
<td>Vitros</td>
<td>4</td>
<td>215</td>
<td></td>
<td>-2 (-65 to +16)</td>
</tr>
<tr>
<td></td>
<td>Vitros 5.1</td>
<td>1</td>
<td>30</td>
<td>497</td>
<td>-2 (-65 to +16)</td>
</tr>
<tr>
<td></td>
<td>Vitros 250</td>
<td>1</td>
<td>72</td>
<td></td>
<td>-2 (-65 to +16)</td>
</tr>
<tr>
<td></td>
<td>Vitros 950</td>
<td>1</td>
<td>150</td>
<td></td>
<td>-2 (-65 to +16)</td>
</tr>
<tr>
<td>Olympus (olympusamerica.com)</td>
<td>AU series</td>
<td>1</td>
<td>30</td>
<td>30</td>
<td>-7.03 [3.72, &lt;0.001]</td>
</tr>
<tr>
<td>Roche (rochediagnostics.com)</td>
<td>Integra</td>
<td>1</td>
<td>30</td>
<td>30</td>
<td>-7.03 [3.72, &lt;0.001]</td>
</tr>
<tr>
<td>Types of instruments</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total medical centers</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total multicenter observations</td>
<td>1,703</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean bias (mg/dL), all observations [SD, P]</td>
<td>-2.2 [9.8, &lt;0.001]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median bias (mg/dL), all observations [range]</td>
<td>-1 [-96 to +45]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
range and one other control level were checked each 24 h. Only paired observations meeting these dual POC-laboratory quality control requirements were fulfilled were analyzed. The GMS linear range is 10–600 mg/dL. Linearity validation levels include 12–24, 46–76, 88–126, 253–323, and 454–574 mg/dL.

**LS MAD curve**

In brief, local smoothing transforms discrete points in the $x$-$y$ plane into a curve that helps reveal underlying patterns. The LS MAD curve contains points $(x, y)$ where $y$ is the median of the values, $y^*$, for all original points $(x^*, y^*)$ in the range $[x-k] \leq x^* \leq [x+k]$. The bandwidth, $h$, controls the degree of smoothing.

The LS MAD curve is continuous from start to end; it started at 35 mg/dL and ended at 220 mg/dL to ensure adequate points were included in the first (20–50 mg/dL) and last (205–235 mg/dL) computational bands. We set the bandwidth, $h$, to 15 mg/dL, and therefore, $2h$, or a span of 30 mg/dL, corresponded to the TGC interval of 80–110 mg/dL.

**Clarke Error Grid**

When plotted on a Clarke Error Grid, performance is considered acceptable if $>95\%$ of observations fall within zones A and B and no or negligible points fall in zones D and E. Note that the FDA does not require Clarke Error Grid analysis for device licensing.

**Modified Bland-Altman plot**

Bland-Altman plots base interpretation of performance on differences (y-axis) in paired GMS and reference values versus means (x-axis) of pairs. We used a modified Bland-Altman plot where the x variable represents the singlet reference result. Horizontal lines show zero bias and mean difference. Visual inspection reveals overall unacceptable bias or ranges where bias appears unexpectedly large. We combined this plot with the ISO 15197 guideline.

**ISO 15197 guideline and erroneous results**

The ISO 15197 guideline states that meter measurements should be within 15 mg/dL (0.83 mmol/L) of the reference result for glucose <75 mg/dL (4.2 mmol/L) and within 20% for glucose ≥75 mg/dL (4.2 mmol/L). A GMS is within the guideline if 95% of pairs satisfy these criteria separately for each range. We define erroneous results as all points falling outside the ISO 15197 tolerance bands. The ISO 15197 guideline does not represent a standard, but currently is under consideration for such by the FDA.

**Bracket predictive value (BPV)**

Positive BPV is defined as $[TP/(TP + FP)]$ where TP (true positive) represents the number of GMS-reference pairs within the TGC bracket (80–110 mg/dL), and FP (false positive) represents the number of pairs where the GMS result is inside the bracket and the reference result is not. Negative BPV is $[TN/(TN + FN)]$, that is, the number of GMS-reference paired observations outside the bracket divided by the number of GMS results outside the bracket. TN is true negative, and FN is false negative. Two-dimensional positive BPV additionally constrains GMS TP results to within 15 mg/dL of paired reference results, chosen to be equivalent to the bandwidth, 15 mg/dL, for the LS MAD curve.

![FIG. 1. Modified Bland-Altman plot with superimposed ISO 15197 tolerance bands.](image-url)
Discrepant values

Class I and II discrepancies represent GMS measurement errors that could significantly impact the effectiveness of TGC protocols. Class I discrepancies are pairs with reference <80 mg/dL and GMS >110 mg/dL. Class II discrepancies are pairs with reference >110 mg/dL and GMS <80 mg/dL. Class I discrepancies could lead to dangerous clinical decisions worsening hypoglycemia, while Class II discrepancies could lead to aggravation of hyperglycemia.

Statistics and units

We used SPSS version 14.0 (SPSS Inc., Chicago, IL) for descriptive statistics, analysis of paired differences (meter minus reference), Kruskal-Wallis nonparametric analysis of the equality of medians among groups, and least squares linear regression. Minitab® (version 14.20, 2005; Minitab, Inc., State College, PA) was used for Ryan-Joiner analysis of the normality of paired difference distributions. Nonparametric symmetric confidence intervals were calculated for the medians of the absolute differences. We report glucose in conventional units, mg/dL, used by participant medical centers. Conversion calculations used [glucose] (in mg/dL) × 0.05551 = [glucose] (in mmol/L).

Results

Figure 1 presents the modified Bland-Altman plot with integrated ISO 15197 tolerance bands (dashed lines) for 1,703 paired observations. ISO 15197 bin populations (% n, mg/dL span, and ISO target %) were: 12.0, 203, <50, 5; 18.0, 306, 50–80, 15; 25.8, 440, >80–120, 20; 15.7, 267, >120–200, 30; 10.6, 180, >200–300, 15; 9.1, 155, >300–400, 10; and 8.9, 151, >400, 5. The mean of the paired differences (meter minus reference) was −2.2 (SD 9.8) mg/dL (P < 0.001), and the median was −1 mg/dL (range, −96 to 45 mg/dL) (Table 1).

The paired difference distribution for all 1,703 observations was not normally distributed (P < 0.01). Individual distributions for the seven brand groups listed in the left-hand column of Table 1 were not normally distributed (P = 0.05), and in one case the distribution was somewhat bimodal. Kruskal-Wallis analysis for the seven brand groups showed P < 0.01.

In Figure 1, all GMS values were within tolerance when the reference glucose was ≥75 mg/dL. For reference glucose <75 mg/dL, 98.6% (410 of 416) were inside, and six (1.4%) were outside. Represented as (x; y; bias), these six were: (50, 34; −16), (52, 34; −18), (52, 29; −23), (57, 41; −16), (57, 40; −17), and (61, 43; −18) mg/dL, which reflected results from four different types of reference instruments and two different brands. This cluster of erroneous results is located below the lower tolerance band on the left in Figure 1.

Figure 2 presents the LS MAD curve. The breakout was at 186 mg/dL (10.32 mmol/L). Positive and negative BPVs were 87.1% (TP, 365; FP, 54) and 96.9% (TN, 1,244; FN, 40), respectively; two-dimensional positive BPV was 87.1% [363/(363 + 54)]. Only two GMS and reference pairs, (91, 107) and (86, 104), in the TGC interval had bias (16 and 18, respectively) greater than 15 mg/dL. There were no Class I or Class II discrepancies. Linear regression showed y = 0.018x − 0.716 mg/dL and r² = 0.995 (Fig. 3). The ranges of reference and GMS glucose values were 16–623 and 17–600 mg/dL, respectively. All points fell within Clarke Error Grid zone A.

Discussion

For hospital glucose meters, we recommend that the LS MAD curve not exceed an error tolerance limit of 5 mg/dL.1 Ideally, the LS MAD curve should be as close as possible to the x-axis (minimal offset), indicating congruence with the hospital laboratory chemistry analyzer from hypoglycemic through hyperglycemic ranges. The GMS studied performed satisfactorily in the vicinity of the TGC interval, where the MAD was approximately 4 mg/dL. Suitable performance in the ranges immediately adjacent to the TGC interval assures accurate glucose results for critical adjustments in insulin infusion rates and will help moderate glycemic variability attributable to asymmetric stochastic measurement error.

When used to monitor patients hourly in critical care settings, glucose meters should be optimized for the relevant span of the TGC interval8–18 since a critical care team adjusts insulin therapy to maintain patients within the TGC interval. The GMS studied here generally met that criterion for 80–110 mg/dL. Other whole-blood glucose meters perform poorly in low (hypoglycemic) zones.1 They may generate er-
roneous results and discrepant values, including dangerous ones falling in Class I that could affect bedside decision-making adversely. The modified Bland-Altman plot with superimposed ISO 15197 tolerance bands (see Fig. 1) reveals deviations from the paired reference glucose in the form of asymmetrical scatter. We recommend that an ISO 15197 difference plot be used to identify erroneous results. Errors increased at high glucose levels (heteroscedasticity) but still fell within the ISO tolerance bands, which are not very demanding. BPV\textsuperscript{7} reflects whether meter results inside or outside the TGC interval reliably reflect paired reference results. BPVs were acceptable. There were no Class I or Class II discrepancies that, if present, potentially can impact patient outcomes adversely. Table 2 summarizes other performance evaluation tools and their utility.

Confounding variables and specimen sources (e.g., arterial, venous, or capillary) may influence measurement error. Fluctuations in O\textsubscript{2} pressure, CO\textsubscript{2} pressure, and pH, as well as in hematocrit, and myriad interferences,\textsuperscript{19-22} operator errors,\textsuperscript{23} environmental factors,\textsuperscript{24} and pathophysiological perturbations (e.g., low perfusion index, arterial hypotension, peripheral hypoperfusion, and generalized mottling with capillary samples\textsuperscript{25}) can affect glucose meter performance adversely. A recent editorial cautions users regarding the potential harm that may result from inaccurate bedside results.\textsuperscript{26} The FDA has warned physicians of “...the potential for life-threatening falsely elevated glucose readings in patients who have received parenteral products containing (or metabolized to) maltose or galactose, or oral xylose, and are subsequently tested using glucose dehydrogenase pyrroloquinolinequinone (GDH-PQQ) ...”.\textsuperscript{27,28} However, the GMS test strip does not use GDH-PQQ chemistry. Other investigators\textsuperscript{29-31} showed that this GMS minimized hematocrit effects.

Limitations of the present study include: (1) reference instrument types reflected hospital choice and were not equally weighted in frequency of use; (2) plasma glucose reference measurements appeared not to be harmonized because of presumed differences in manufacturer calibration; (3) aggregated results from different hospitals and geographic locations derived from heterogeneous patient populations (and sampling sites), which were subject to a variety of confounding variables; (4) ISO 15197 bin populations differed from guidelines specifications mainly by overpopulating the critical range below 120 mg/dL; (5) measure-
Critical insulin rate adjustments must be made to ensure adequate glycemic control. Physicians and nurses should scrutinize performance for specific TGC intervals. Inadequate glycemic control can result in increased mortality and morbidity. In the early 2000s, point-of-care (POC) glucose testing devices were rapidly introduced into the market, with the stated purpose of reducing delays that can compromise patient outcomes. These devices were intended for hospital use, particularly for acute care, and were to be standardized for calibration accuracy. However, discrepancies in performance evaluation have been noted. The development of a consensus-based periodic proficiency testing approach for assessment of glucose meter performance was proposed. This approach included the LS MAD curve pattern, which is designed to be applied to other POC tests to help facilitate informed bedside decision-making.

Conclusions

LS MAD curves draw on the unique human ability to recognize patterns quickly and discriminate performance visually. Judged by the LS MAD curve pattern with an error tolerance limit of 5 mg/dL, in conjunction with the ISO 15197-integrated modified Bland-Altman plot results, the performance of the GMS was acceptable for the glucose range encountered in the TGC interval.

We recommend that future studies investigate performance in specific critical care conditions, such as sepsis and burns; incorporate bedside operators, including nursing staff; and minimize delays that can compromise fresh whole-blood samples.

In collaboration with POC coordinators, critical care physicians and nurses should scrutinize performance for a specific TGC interval and the immediately adjacent ranges used for critical insulin rate adjustments. They should also consider the impact of both erroneous results and discrepant values. Discrepant glucose meter values can impair therapeutic decisions, but no discrepancies occurred. A small set of erroneous results occurred in the hypoglycemic range.

Different brands of clinical laboratory analyzers and POC glucose devices intended for hospital use, particularly for acute care, should be standardized for calibration accuracy to reduce disharmony, and, then, accuracy-based rather than consensus-based periodic proficiency testing should be used in hospitals nationwide and globally.

Bedside glucose testing in critical care demands careful attention to patient status in order to avoid adverse effects of confounding factors, such as peripheral hypotension and its impact on capillary glucose. Asymmetric stochastic measurement error may exacerbate glycemic variability.

We recommend that future licensing criteria for bedside glucose meters include the LS MAD curve pattern recognition approach for assessment of GMS performance. LS MAD curves with analyte-specific error tolerance limits also could be applied to other POC tests to help facilitate informed bedside decision-making.

Acknowledgments

The tables and figures were provided courtesy and permission of Knowledge Optimization®, Davis, CA. This study was supported by the Point-of-Care Testing Center for Teaching and Research (POCT • CTRSM), School of Medicine, University of California, Davis, and by a National Institute of Biomedical Imaging and Bioengineering Point-of-Care Technologies Center grant (G.J.K., Principal Investigator, National Institutes of Health grant number 1U54 EB007959-01). This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Biomedical Imaging and Bioengineering or the National Institutes of Health.

Author Disclosure Statement

No competing financial interests exist.

References


Table 2. Performance Evaluation Toolbox for Handheld POC Devices

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error grid Differences</td>
<td>Clarke—Clarke grid not useful when data fall 100% within constrained zone “A”</td>
</tr>
<tr>
<td>Statistical Regression</td>
<td>Student’s t test for paired differences—robust detection of non-zero mean bias, but positive and negative bias in different ranges can offset each other and mask non-equivalence</td>
</tr>
<tr>
<td>Graphical ISO standard</td>
<td>Bland-Altman plot—provides clinical perspective over measurement range</td>
</tr>
<tr>
<td>Post hoc Hybrid</td>
<td>Modified Bland-Altman plot with superimposed ISO 15197 error tolerance bands—visualizes scatter (heteroscedasticity) and reveals erroneous results that fall outside the tolerance bands (proposed FDA licensing requirement)</td>
</tr>
<tr>
<td>LS MAD curve</td>
<td>Reveals errors relevant to decision intervals (e.g., TGC, hypoglycemia, and hyperglycemia), employs analyte-specific error tolerance (i.e., LS MAD &lt;5 mg/dL for glucose), visually displays quantitative performance, includes 95% confidence bands, and quickly identifies problem zones with curve “breakouts”</td>
</tr>
</tbody>
</table>
LS MAD PERFORMANCE AND TIGHT GLUCOSE CONTROL


Address reprint requests to:
Gerald J. Kost, M.D., Ph.D.
Point-of-Care Testing Center for Teaching and Research
3455 Tupper Hall
Pathology and Laboratory Medicine
School of Medicine
University of California, Davis
Davis, CA 95616
E-mail: gjkost@ucdavis.edu