

# Clinical and Statistical Evaluation of Self-Monitoring Blood Glucose Meters

JEAN-YVES POIRIER, MD  
NADINE LE PRIEUR, MD  
LOÏC CAMPION, MD

ISABELLE GUILHEM, MD  
HUBERT ALLANNIC, MD  
DIDIER MAUGENDRE, MD, PHD

**OBJECTIVE** — Our objective was to compare statistical and clinical methods for the evaluation of five self-monitoring blood glucose (SMBG) meters.

**RESEARCH DESIGN AND METHODS** — Two successive capillary blood glucose measurements were performed, and a simultaneous laboratory venous glucose measurement was used as the reference value. Accuracy was studied by comparing each of the two successive meter values with the reference value by 1) a Spearman's correlation test, 2) a Wilcoxon's paired test, 3) the percentage of values within the 10% interval of the reference value according to the American Diabetes Association consensus statement, and 4) the error grid analysis.

**RESULTS** — The first two methods did not discriminate between the SMBG systems:  $r$  was  $>0.92$  for the five meters, and a significant difference between the meter and reference values was found for all but one meter. The two other methods allowed classification of the devices into three groups according to their accuracy: good (two meters), acceptable (two meters), and unacceptable (one meter). These two methods gave consistent results and both had a good reproducibility, because the classification was similar for the two successive measurements.

**CONCLUSIONS** — Both the Spearman's and Wilcoxon's paired tests, although commonly used, are inappropriate to evaluate SMBG systems. The percentage of SMBG values within the  $\pm 10\%$  interval and the error grid analysis are more accurate, because they consistently classified the five glucose meters tested in our study with a high degree of reproducibility.

*Diabetes Care* 21:1919–1924, 1998

Self-monitoring of blood glucose (SMBG) has become a major tool in the management of diabetes in the past decade. However, there has been some controversy over what methods should be used to analyze the accuracy of these systems. Statistical methods, including the Wilcoxon's paired test and correlation and linear regression tests, are the most commonly used, especially by the manufacturers. On the other hand, the American Diabetes Association (ADA) Consensus Conferences (1,2) recommend that the performance goal of all SMBG systems should be to achieve a total error (system plus user) of  $<10\%$  at glucose concentrations

ranging from 30 to 400 mg/dl (1.6 to 22.2 mmol/l). Finally, Clarke and colleagues (3,4) developed the error grid, a graphical display method that evaluates the clinical significance of the errors generated by the glucose meters.

The objective of our study was to compare the ability and the reproducibility of these different statistical and clinical methods to classify five SMBG systems according to their accuracy.

**RESEARCH DESIGN AND METHODS** — The study included 225 insulin-treated diabetic patients (mean age:  $48 \pm 12$  years, mean duration of diabetes

$7.5 \pm 6$  years). The exclusion criteria were renal insufficiency (serum creatinine  $>110$   $\mu\text{mol/l}$ ) and treatment with sulfonylurea, biguanide antidiabetic agents, paracetamol, aspirin, or vitamin C.

Five new glucose meters using non-wipe technology were assessed: Glucometer 4 (Ames Bayer Diagnostic), Accucheek-Easy (Boehringer-Mannheim), One Touch Basic (LifeScan), Exactech Companion (Medicines), and Suprême (Vermed). All were used properly according to the manufacturer's instructions. Two successive measurements on two different fingers were performed by nurses for each patient, who were all in the morning fasting state. The two measurements were read on one of the five meters randomly designated ( $2 \times 45$  measurements for each SMBG system). A venous blood sample was drawn immediately thereafter for laboratory measurement of plasma glucose with a CX5 analyzer (Beckman Instruments, Brea, CA) (intra-assay coefficient of variation: 2%) and simultaneous measurement of hematocrit.

## Statistical and clinical analysis

The hematocrits of the patients ranged from 35 to 49%, which is within the operative specifications of the five meters. However, because of the physiological difference between capillary whole-blood glucose and venous plasma glucose levels (5,6), the laboratory results were corrected for the hematocrit according to the following equation (7): whole blood glucose = plasma glucose  $\times [1 - (0.0024 \times \text{hematocrit})]$ . The five meters were randomly designated from I to V; their names were intentionally concealed because we chose to focus on methodological issues and not to classify the meters according to their performances.

Accuracy was assessed by comparing each of the two successive SMBG-system values with the reference value. Four methods were used: 1) the Spearman's correlation test, 2) the Wilcoxon's paired test, 3) the percentage of SMBG values within  $\pm 10\%$  of the laboratory value, and 4) the error grid analysis (3). This grid (Fig. 1) defines the x-axis as the reference blood glucose and the y-axis as the value generated by the SMBG system. The data points

From the Department of Metabolic and Endocrine Diseases, CHU Hôpital-Sud, Rennes, France.

Address correspondence and reprint requests to Dr. Jean-Yves Poirier, Department of Metabolic and Endocrine Diseases, CHU Hôpital-Sud, Bd de Bulgarie, BP 56129, 35056 Rennes Cedex, France.

Received for publication 15 April 1998 and accepted in revised form 31 July 1998.

**Abbreviations:** ADA, American Diabetes Association; SMBG, self-monitoring of blood glucose.

A table elsewhere in this issue shows conventional and Système International (SI) units and conversion factors for many substances.

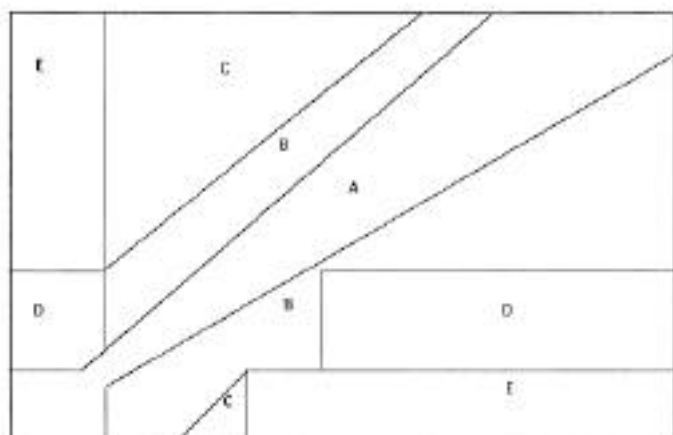


Figure 1—Error grid analysis.

obtained for each measurement fall into one of the different zones drawn on the grid. The zones indicate how adequate the therapeutic decision taken on the basis of the SMBG-system result would be compared with the decision that would have been taken on the basis of the laboratory result. In zone A, the difference between the two measurements is <20% and would lead to clinically correct treatment decisions. In the upper and lower zones B, the difference is >20%, and the therapeutic decision would be inappropriate but without any serious immediate consequence. In zone C, the difference leads to an overcorrection of a normal or subnormal glucose level (the meter displaying low or high values). Zone D represents "dangerous failure to detect and treat errors": the reference values are high or low while the meter values are within the normal range. Zone E is defined as "erroneous treatment zone": the meter-generated values are opposite to the reference values, and the corresponding treatment decisions would be unsuitable and dangerous.

Precision of the glucose meters was examined as follows: 1) the Spearman's correlation test and the Wilcoxon's paired test

between the two successive capillary blood glucose measurements and 2) 10 successive measurements from a venous blood sample were made on each meter for 4 blood glucose levels (2.4, 4.7, 11, and 20.4 mmol/l) to calculate the coefficients of variation.

Statistical analysis was performed with a Statview software program. Statistical significance was defined as  $P < 0.05$ .

RESULTS

Accuracy

The coefficient of correlation and the Wilcoxon's paired test between each of the two successive SMBG-system values and the reference value were not discriminant (Table 1). The coefficients of correlation were not different for the five meters: the lowest  $r$  coefficient was 0.94 and the highest was 0.98 for the first SMBG measurement and ranged from 0.92 to 0.99 for the second SMBG measurement. In the same way, the difference between the SMBG values and the reference value (Wilcoxon's paired test) was statistically significant for all but one meters (meter III for the second measurement) (Table 1). On the contrary, the percentage of values within  $\pm 10\%$  of

the reference value allowed classification of the five meters into three groups according to their accuracy: good, acceptable, and unacceptable for clinical use (Table 2). The first group consisted of two meters (I and II) with >60% of each of the two successive measurements within the  $\pm 10\%$  interval. In the second group (two meters: III and IV), the percentages of measurements within the  $\pm 10\%$  interval were 60 and 51%, respectively, for the first SMBG measurement and 53 and 51% for the second measurement. The third group consisted of one unacceptable meter (V), which displayed only 35% of the first SMBG measurements and 37% of the second within the  $\pm 10\%$  interval.

The results of the error grid analysis were consistent with those of the  $\pm 10\%$  interval (Table 2): >97% of the two successive measurements performed with meters I and II (good accuracy according to the  $\pm 10\%$  interval method) were in zone A of the error grid (Fig. 2); the performance of the unacceptable meter, number V, was 82% for both measurements (Fig. 4); and the two meters classified as acceptable accuracy (III and IV) displayed intermediate performances (Fig. 3). The reproducibility of the error grid was satisfactory, as the classification of the five meters was similar for the two successive measurements (Table 2).

Precision

No significant difference between the two successive capillary glucose measurements was found for any of the five meters; the two measurements were significantly correlated and the slope of the regression line was not different from 1. The coefficients of variation calculated for each of the four blood glucose levels among 10 successive measurements from a single venous blood sample were satisfactory (1–6%) except for the meter versus showing the lowest precision (coefficients of variation >10% whatever the level of blood glucose).

Table 1—Comparison between each of the two successive SMBG values and the reference method by nonparametric tests

Meter	Spearman's correlation test				Wilcoxon's paired test			
	1st measurement	P	2nd measurement	P	1st measurement	P	2nd measurement	P
I	0.98	<10 <sup>-4</sup>	0.99	<10 <sup>-4</sup>	Z = 2.21	0.027	Z = 3.86	10 <sup>-4</sup>
II	0.98	<10 <sup>-4</sup>	0.98	<10 <sup>-4</sup>	Z = 5.05	<10 <sup>-4</sup>	Z = 5.27	<10 <sup>-4</sup>
III	0.97	<10 <sup>-4</sup>	0.97	<10 <sup>-4</sup>	Z = 3.46	0.0005	Z = 1.53	NS
IV	0.94	<10 <sup>-4</sup>	0.92	<10 <sup>-4</sup>	Z = 3.62	0.0003	Z = 3.03	0.002
V	0.94	<10 <sup>-4</sup>	0.96	<10 <sup>-4</sup>	Z = 4.25	<10 <sup>-4</sup>	Z = 4.27	<10 <sup>-4</sup>

**Table 2—Comparison between the reference values and each of the two successive SMBG values with the percentage of results within the  $\pm 10\%$  interval and the error grid analysis (percentage of values within zone A of the grid)**

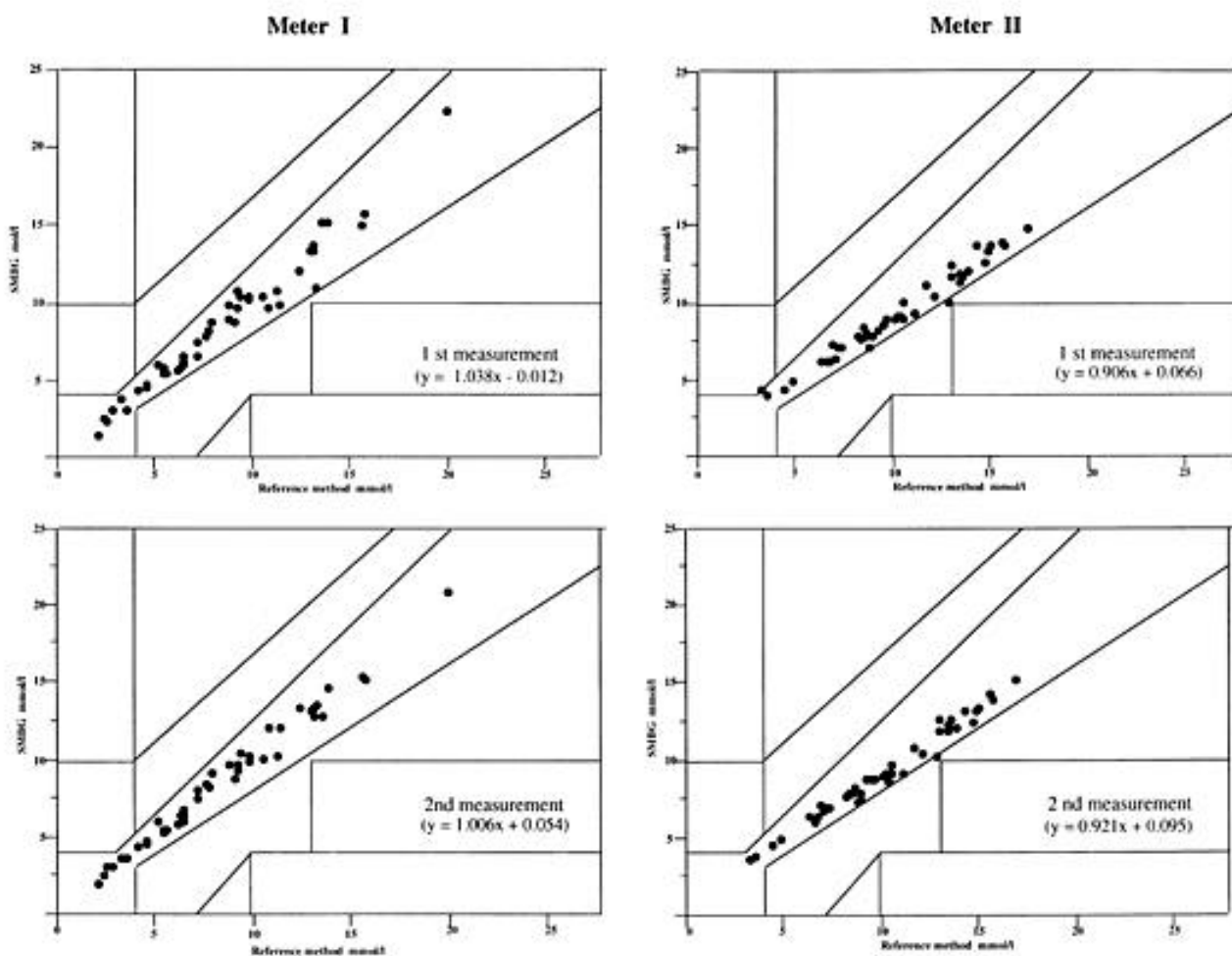
Meter	% of SMBG values within $\pm 10\%$ interval		% of SMBG values in zone A of the error grid	
	1st measurement	2nd measurement	1st measurement	2nd measurement
I	66	69	100	100
II	62	66	97	100
III	60	53	91	93
IV	51	51	87	87
V	35	37	82	82

Data are %.

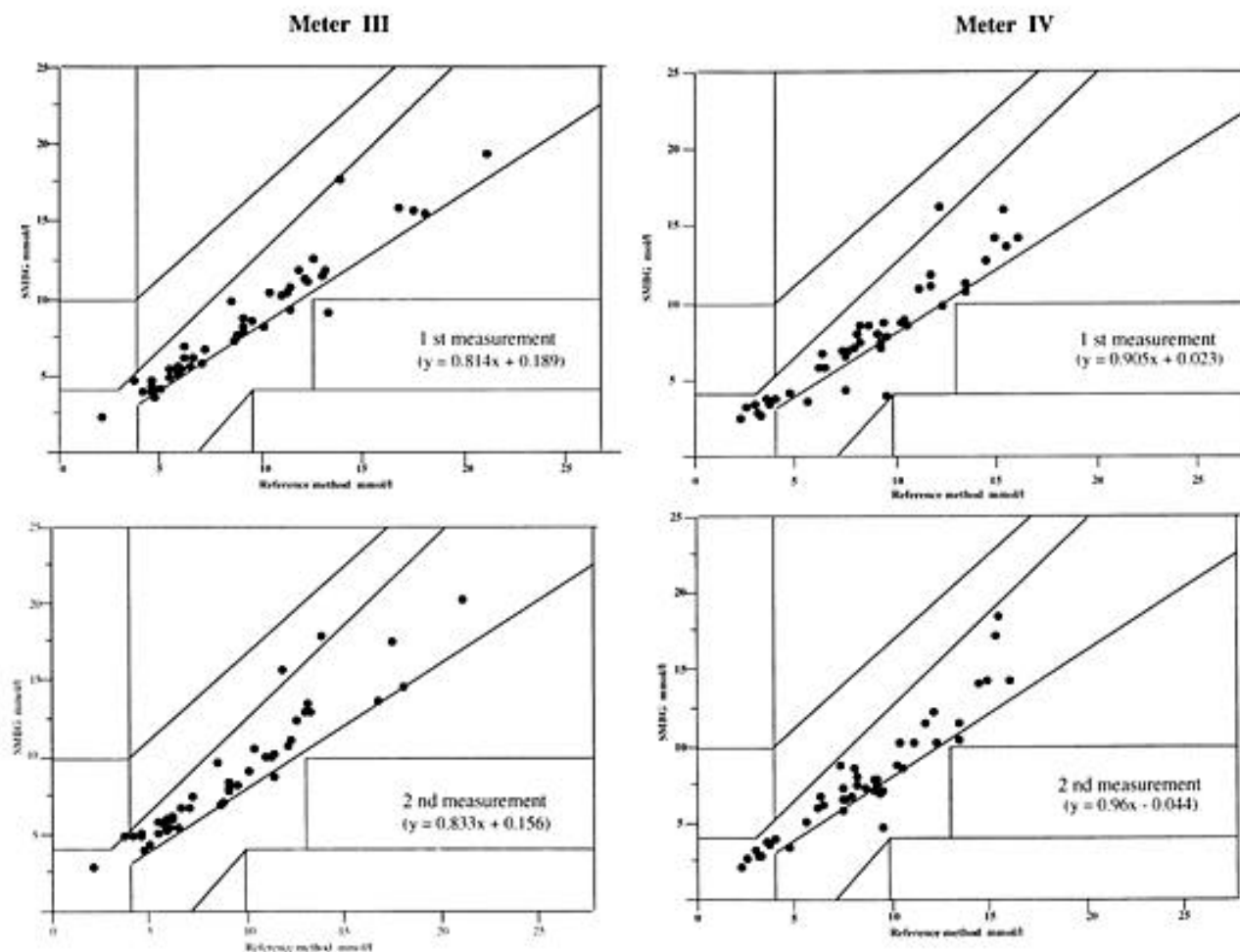
**CONCLUSIONS** — Our results illustrate the limitations of statistical methods in evaluating the accuracy of SMBG systems. This problem is crucial because an incorrect evaluation may have critical conse-

quences on the treatment decision. An additional problem is the common misconception of the statistical tests. The  $r$  coefficient, which is easily calculated on modern computers, is most widely applied

to the evaluation of SMBG systems. However, it should be stressed that most investigators and manufacturers assume that a perfect correlation indicates a small amount of error in the SMBG system. In fact, this assumption is not correct, because the  $r$  coefficient measures the extent to which two sets of data fit a linear relationship, not the consistency between data. For example, if measurements from a glucose meter differ from the reference method by a constant amount, the coefficient of correlation is 1.00 despite the error generated by the meter. In this case, the two variables are associated, but obviously the consistency equals 0. Moreover, the  $r$  value depends on the distribution of the data. Thus, a correlation can be found to be significant in a large set of data over the reference glucose range from 2 to 20 mmol/l and not significant in subsets of data restricted to low or high values (8,9). Once applied to our



**Figure 2**—Error grid analysis for the two successive SMBG values obtained with the two meters classified as having good accuracy.



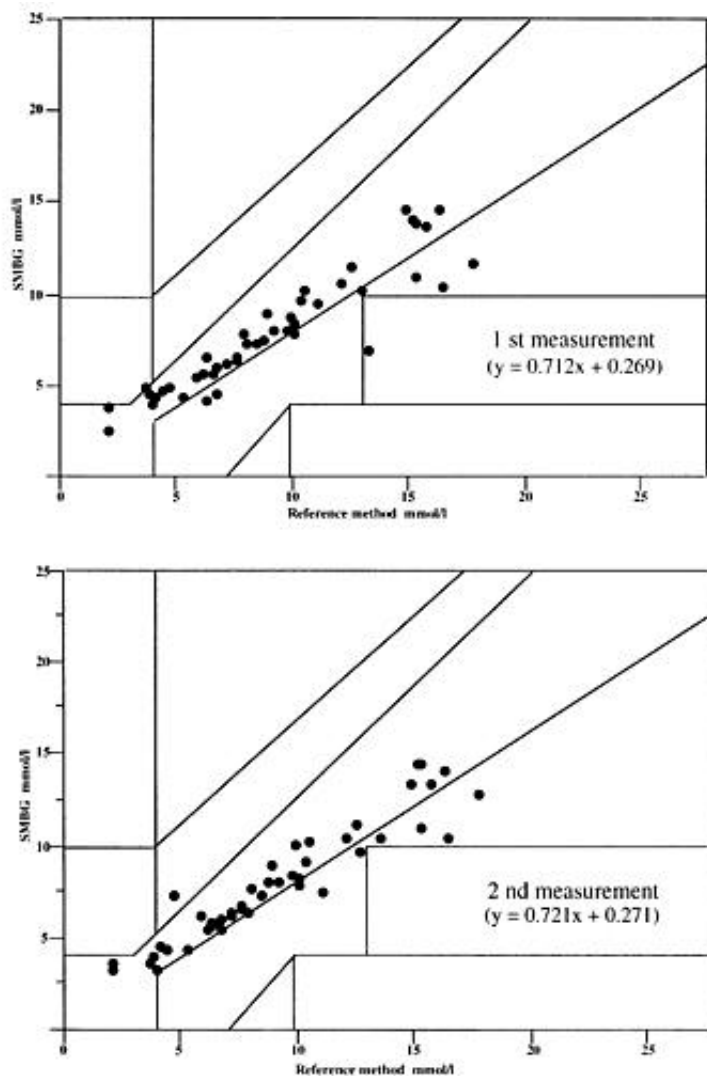
**Figure 3**—Error grid analysis for the two successive SMBG values obtained with the two meters with acceptable accuracy.

study, this method did not prove valuable to discriminate between the SMBG systems because the *r* coefficients were quite similar and above 0.92 for the five meters at both measurements. The percentage of SMBG values differing from the reference value has been proposed as an alternative test that could be more appropriate. The ADA consensus statement conferences on SMBG (1,2) have stated that SMBG measurements should be within the  $\pm 10\%$  interval of the reference values. However, none of the five meters used in our study could achieve this goal because the most accurate displayed only 66% of the values within this range at the first measurement and 69% at the second one (Table 2). This is consistent with a recent conclusion of the College of American Pathologists, who found no more than 56% of the SMBG results within 10% of the corresponding

laboratory result (10,11). Moreover, the clinical relevance of this goal is questionable because it depends on the glucose level (8). For example, a 10% deviation from a reference glucose of 16 mmol/l has no therapeutic consequence, but the same deviation from a reference glucose of 3.5 mmol/l may result in an inappropriate treatment decision.

The error grid (3,4) takes into account both the difference between the reference and the SMBG-system values and the pertinence of the treatment decision resulting from the SMBG value. The error is analyzed according to its effect on the decision that would have been made if the blood glucose had been measured with the reference method. This grid was initially developed with the objective of evaluating the capacity of insulin-dependent diabetic patients to estimate their blood glucose level from

symptoms they had learned to recognize during hypoglycemic, normoglycemic, and hyperglycemic clamp tests. After this training period, the patient-estimated value was compared with the measurement performed with a glucose meter (4). The results were disappointing, and the method was later used to evaluate the accuracy of the SMBG systems compared with the laboratory method (3). Applied to our study, the error grid allowed us to identify specific features for each device, in particular to identify both the two accurate meters and the one inaccurate meter. Note that the coefficients of correlation between the SMBG meter and the reference measurements were not different for these three meters. Moreover, the classification of the five meters according to their accuracy was similar with the error grid and the percentage of results within the  $\pm 10\%$  interval. The repro-



**Figure 4**—Error grid analysis for the inaccurate meter

ducibility of both methods was satisfactory because this classification was not different for the two successive measurements.

Despite its real advantages, the error grid analysis also has its limitations, basically concerning the definition of the lower zone B, which represents benign errors. This qualitative assessment would not be correct for some points located in this zone—for example, for SMBG values at 2.2 and 10.4 mmol/l for reference values of 8.2 and 22 mmol/l, respectively. Since 1991, the successive ADA consensus statement conferences that dealt with issues regarding the accuracy of SMBG systems have focused on this problem, stating that “effort should be made to refine the Error Grid target ranges to account for intensive treatment goals” (2).

The above methodological problems are further complicated by several issues that should be taken into account when comparing an SMBG system with a reference method. It is well established that a capillary whole-blood glucose is lower than a venous plasma equivalent (5,6). This difference depends partly on the hematocrit concentration and the fasting or the non-fasting state. The presence of reducing agents (glutathion, vitamin C, cysteine), hyperuricemia, or treatment with tolbutamide may result in an underestimation, particularly when using the glucose oxidase method. In contrast, high-dose paracetamol, aspirin, and hemolysis can cause an overestimation, particularly with the hexokinase method. These factors were controlled in our study: the capillary and

venous samples were drawn at patient admission as the routine samples for laboratory tests (lipids, urea, creatinine, etc.). In addition, the clinical and biological characteristics of the patients whose data points were outside the zone A in the error grid analysis were not different from those found in the other patients.

Other potential causes of errors have been well documented for ambulatory SMBG, namely the technical skills of the operator (particularly in the case of hypoglycemia, which can alter psychomotor ability), the device wear, and above all, the compliance to maintenance procedures. These factors were not examined in the present study because our objective was to evaluate the methodological approach of the intrinsic performances of the glucose meters independently of human factors that might alter their total reliability. For this reason, we chose to use new devices and the measurements were performed by trained nurses.

In conclusion, the common use of the coefficient of correlation for the evaluation of SMBG devices is inappropriate. As determined by the nonparametric Wilcoxon's paired test, it does not provide any realistic information for clinical use. On the contrary, the percentage of SMBG values within the  $\pm 10\%$  interval and the error grid are more appropriate, because they both allowed the same classification of the five glucose meters tested in our study. Moreover, these two methods have a good reproducibility, as the classification of the five meters was similar for the two successive measurements.

**Acknowledgments**— We thank J. Yaouanq from the Department of Epidemiology for statistical assistance and the nursing staff of the University Hospital of Rennes.

#### References

1. American Diabetes Association: Consensus statement on self-monitoring of blood glucose. *Diabetes Care* 10:95–99, 1987
2. American Diabetes Association: Self-monitoring of blood glucose. *Diabetes Care* 18:47–52, 1995
3. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl S: Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* 10:622–628, 1987
4. Cox DJ, Clarke WL, Gonder-Frederick LA, Pohl S, Hoover C, Snyder A, Zimelman L, Carter WR, Bobitt S, Pennebaker J: Accuracy of perceiving blood glucose in IDDM. *Diabetes Care* 8:529–536, 1985
5. Chmielewski SA: Advances and strategies

## Evaluation of blood glucose meters

- for glucose monitoring. *Am J Clin Pathol* 104:S59-S71, 1995
6. Maley TC, D'Orazio P: Biosensors for blood glucose: a new question of what is measured and what should be reported. *Clin Lab News* 21:12-13, 1995
  7. Frishman D, Ardito DM, Graham SM: Performance of glucose monitors. *Lab Med* 23:179-184, 1992
  8. Dedrick RF, Davis WK: What do statistics really tell us about the quality of the data from self-monitoring of blood glucose? *Diabet Med* 6:267-273, 1989
  9. Moberg E, Lundblad S, Lins PE, Adamson U: How accurate are home blood-glucose meters with special respect to the low glycemic range? *Diabetes Res Clin Pract* 19:239-243, 1993
  10. Jones BA, Howanitz PJ: Bedside glucose monitoring quality control practices: a College of American Pathologists Q-probes study of program quality control documentation, program characteristics, and accuracy performance in 544 institutions. *Arch Pathol Lab Med* 20:339-345, 1996
  11. Howanitz PJ, Jones BA: Bedside glucose monitoring: comparison of performance as studied by the College of American Pathologists Q-probes program. *Arch Pathol Lab Med* 20:333-338, 1996