Andreas Bietenbeck*, Wolf J. Geilenkeuser, Frank Klawonn, Michael Spannagl, Matthias Nauck, Astrid Petersmann, Markus A. Thaler, Christof Winter and Peter B. Lupp

External quality assessment schemes for glucose measurements in Germany: factors for successful participation, analytical performance and medical impact

<https://doi.org/10.1515/cclm-2017-1142>

Received December 7, 2017; accepted February 6, 2018; previously published online May 1, 2018

Abstract

Background: Determination of blood glucose concentration is one of the most important measurements in clinical chemistry worldwide. Analyzers in central laboratories (CL) and point-of-care tests (POCT) are both frequently used. In Germany, regular participation in external quality assessment (EQA) schemes is mandatory for laboratories performing glucose testing.

Methods: Glucose testing data from the two German EQAs “Reference Institute for Bioanalytics” (RfB) and “INSTAND – Gesellschaft zur Förderung der Qualitätssicherung in medizinischen Laboratorien” (Instand) were analyzed from 2012 to 2016. Multivariable odds ratios (OR) for the probability to reach a “good” result were calculated. Imprecision and bias were determined and clinical risk of measurement errors estimated.

Results: The device employed was the most important variable required for a “good” performance in all EQAs. Additional participation in an EQA for CL automated

analyzers improved performance in POCT EQAs. The reciprocal effect was less pronounced. New participants performed worse than experienced participants especially in CL EQAs. Imprecision was generally smaller for CL, but some POCT devices reached a comparable performance. Large lot-to-lot differences occurred in over 10% of analyzed cases. We propose the “bias budget” as a new metric to express the maximum allowable bias that still carries acceptable medical risk. Bias budgets were smallest and clinical risks of errors greatest in the low range of measurement 60–115 mg/dL (3.3–6.4 mmol/L) for most devices. **Conclusions:** EQAs help to maintain high analytical performances. They generate important data that serve as the foundation for learning and improvement in the laboratory healthcare system.

Keywords: bias budget; learning healthcare system; point-of-care testing; proficiency testing; ring trials.

List of abbreviations: CL, central laboratories; POCT, point-of-care test; EQA, external quality assessment; GO, glucose oxidase; GDH, glucose dehydrogenase; HK, hexokinase; Rili-BAEK, Guideline of the German Medical Association on Quality Assurance in Medical Laboratory; RfB, reference institute for bioanalytics; Instand, INSTAND Gesellschaft zur Förderung der Qualitätssicherung in medizinischen Laboratorien e.V.; CL-RfB, proficiency test RfB KS (“Clinical chemical analytes in serum – wet chemistry”); CL-Instand, proficiency test Instand 100 (“Clinical Chemistry – Wet Chemistry”); POCT-RfB, proficiency test RfB GL (“Glucose [wet and dry chemistry]”); POCT-Instand, proficiency test Instand 800 (“Dry Chemistry 01 – POCT: Glucose”); OR, odds ratio; CV, coefficients of variation; SEG, surveillance error grid; TGC, tight glycemic control; mCV, weighted mean of the coefficients of variation.

***Corresponding author: Dr. med. Andreas Bietenbeck**, Institut für Klinische Chemie und Pathobiochemie, Klinikum rechts der Isar der Technischen Universität München, Ismaninger Str. 22, 81675 München, Germany, Phone: +49 89 4140 4757, E-mail: andreas.bietenbeck@tum.de, <http://orcid.org/0000-0002-1228-0770>

Wolf J. Geilenkeuser: Reference Institute for Bioanalytics, Bonn, Germany

Frank Klawonn: Department of Biostatistics, Helmholtz Centre for Infection Research, Braunschweig, Germany; and Department of Computer Science, Ostfalia University of Applied Sciences, Wolfenbüttel, Germany

Michael Spannagl: Department of Hemostasis and Transfusion Medicine, Ludwig-Maximilian-University of Munich, Munich, Germany; and INSTAND e.V. Reference Laboratory, Düsseldorf, Germany

Matthias Nauck and Astrid Petersmann: Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany

Markus A. Thaler, Christof Winter and Peter B. Lupp: Institut für Klinische Chemie und Pathobiochemie, Klinikum rechts der Isar der Technischen Universität München, Munich, Germany

Introduction

The diagnosis and treatment of diabetes mellitus relies heavily on blood glucose concentration measurements. With a worldwide prevalence of diabetes mellitus close to 7%, glucose measurements are among the most frequently

performed analyses in medical laboratories. The advent of portable glucometers represents a breakthrough in the monitoring of diabetes patients. Glucometers were among the first point-of-care testing (POCT) instruments and are of major medical and economic importance [1].

Most analytical methods use one of three enzymatic reactions to quantify glucose: glucose oxidase (GO), glucose dehydrogenase (GDH) or hexokinase/glucose-6-phosphate dehydrogenase (HK). Enzymatic activity produces an electrical current or a color change proportional to glucose concentration. Isotope dilution gas chromatography mass spectrometry serves as higher-order reference procedure in reference laboratories, whereas the hexokinase method is widely accepted for routine calibration and accuracy evaluation [2]. Analytical performance goals have been derived from expert opinions or through computer simulations [3–6]. In central laboratories (CL), glycolysis is a major source of error. Centrifuging samples with minimal delay or adding sodium fluoride and a citrate buffer to the sample prevent glycolysis [7, 8]. Portable glucometers measure glucose immediately and close to the patient from capillary blood. However, their compact instrument design and the use of the less specific enzymes GO or GDH results in a greater susceptibility to interferences [9–11]. While specialized training can usually be assumed for operators in CLs, operator qualification varies for POCT measurement and represents a major medical risk [12].

External quality assessments (EQAs) are crucial to ensure continuous high quality in medical laboratories worldwide. In Germany, the “Guideline of the German Medical Association on Quality Assurance in Medical Laboratory” (Rili-BAEK) [13] requires medical laboratories performing glucose testing to pass a glucose EQA at least twice a year. Only two organizations, “INSTAND Gesellschaft zur Förderung der Qualitätssicherung in medizinischen Laboratorien e.V.” (Instand) and “Reference Institute for Bioanalytics” (RfB), are licensed to offer glucose EQAs in Germany. Therefore, results of these organizations allow a comprehensive overview of glucose measurements in Germany. Especially for POCT EQAs, a lack of commutability of the applied samples often impedes interpretation [14–16]. The reports of the EQA organizers nevertheless provide valuable information to participants [17], making EQA an important educational tool. On a higher level, data from EQAs can form the basis for improvements in laboratory medicine as a whole [18].

In this study, we analyzed glucose EQAs of Instand and RfB to identify factors affecting participant performance. We estimated the imprecision and bias of glucose measurements and assessed clinical risks of inaccurate measurements.

Materials and methods

Data

Glucose data for the years 2012–2016 were obtained from the EQA schemes RfB KS, “Clinical chemical analytes in serum – wet chemistry” (CL-RfB) and Instand 100, “Clinical Chemistry – Wet Chemistry” (CL-Instand). These schemes were designed for automated analyzers in CL where glucose measurements were part of larger panels. In addition, measurements from RfB GL, “Glucose (wet and dry chemistry)” (POCT-RfB), and Instand 800, “Dry Chemistry 01 – POCT: Glucose” (POCT-Instand), were acquired. All data were reformatted to a common structure and analyzed using the R statistical software version 3.4.0. Wherever possible, devices were matched to a common denomination across both organizations. All other devices were subsumed into the “others” device category. Full code is available at <https://github.com/acnb/Glucose-EQA>. Data are released under <https://doi.org/10.17605/OSF.IO/F3CD5>.

EQA material

Two liquid plasma samples (POCT-Instand) or two lyophilized serum samples (CL-RfB, CL-Instand, POCT-RfB) were used in each glucose EQA distribution (Supplemental Methods).

Odds ratios

We classified deviations of measurements from the target value as follows. In line with the analytical performance specifications stipulated in the German Rili-BAEK [19], deviations >15% from the target value were classified as “failed” participation of the EQA distribution (which also resulted in the denial of the EQA certificate). In line with Bukve et al. [20], participation outcomes in an EQA distribution other than “failed” were classified as “poor” if a result exceeded the target interval (target value ± 2 mg/dL [0.1 mmol/L]) by >10%, as “acceptable” for deviation of 10%–5% from the target interval, and as “good” for deviation $\leq 5\%$. The sample with the largest deviation from the target interval was used for the final classification unless the EQA organization retracted one sample (e.g. due to insufficient stability).

Odds ratios (OR) were calculated for the probability to reach a “good” result and for the probability to not reach a “failed” result. Independent variables were the device employed, experience from previous participations in the same EQA scheme and simultaneous participations in other glucose EQAs from the same organization. Missing experience data were substituted with plausible values using multiple imputation by chained equations [21] (Supplemental Methods).

Pathway analysis

Consequences of a year with or without failed participations were analyzed (Supplemental Methods).

Imprecision

Imprecision was calculated as coefficient of variation (CV) and as so-called characteristic function [22, 23].

For each sample, robust device-specific central location and scale were calculated using the Huber M-Estimator implemented in the R software package “robustbase” [24]. CVs were established by dividing the scale through the central location. For the overall CVs of devices, the weighted mean of the individual CVs was determined with the inverse of the squared standard error of the estimation as weight.

Additionally, imprecision was estimated using the characteristic function:

$$\sigma_c = \sqrt{\alpha^2 + (c\beta)^2} \quad (1)$$

The characteristic function (Eq. 1) allows specifying an absolute imprecision α in the lower measuring range close to the limit of detection. The imprecision β for higher values depends on the concentration c and resembles the traditional CV. Parameters α and β were fitted using the Nonlinear Least Squares algorithm with the inverse of the squared standard errors as weights (Supplemental Methods).

Bias

For CL EQAs, the differences between the robust central location of the measured values of a device and the reference method value were regarded as bias. Biases were not estimated in POCT EQAs to avoid distortions by possible non-commutability of samples.

Comparison of lots

In POCT-Instand, participants could voluntarily specify the lot number of their test strip. Differences between the robust central location of the results of different lots for the same device and sample were

calculated. The 95% confidence interval of these differences was determined using bootstrapping [25] (Supplemental Methods).

Bias budget

We introduce the “bias budget”, which denotes the maximum allowable bias that is still clinically acceptable for a laboratory test with a given imprecision. Clinical risk assessment was based on the surveillance error grid (SEG), a tool to assess the degree of clinical risk for diabetes patients from inaccurate blood glucose monitors [6]. For imprecision, the previously calculated CVs as well as the fitted parameters from the characteristic function were used. For each true glucose concentration up to 500 mg/dL (27.7 mmol/L), the bias had to be small enough such that for 99.7% (three standard deviations) of all measurements with the given imprecision the total error posed less than a “moderate” SEG risk. The largest bias still meeting these constraints was termed “bias budget”. Similarly, a bias budget was also calculated using insulin dosage errors during a tight glycemic control (TGC) simulation according to Karon et al. [3]. Here, 99.7% of all measurements were permitted to result in a maximum “2-category” (not “dangerous”) dosing error.

Results

POCT measurements failed four times as often in EQAs compared to CL automated analyzer measurements

Between 2012 and 2016, overall more than 1000 individual laboratories participated in each of the four EQA schemes. For schemes run by Instand, there were six distributions per year. POCT-RfB had four and CL-RfB eight distributions per year (Table 1). Two samples were supplied in

Table 1: Description of the glucose EQAs analyzed in this study.

		CL-Instand	CL-RfB	POCT-Instand	POCT-RfB
Number of distinct participants		1168	1516	1189	1017
Total number of samples		29,992	50,268	28,678	24,747
Number of distinct devices		44	39	65	54
Distributions per year		6	8	6	4
Participants per distribution	Minimum	161	555	196	567
	Mean	511	629	481	620
	Maximum	796	746	725	648
Success rate per distribution	Minimum	0.94	0.97	0.79	0.86
	Mean	0.98	0.99	0.90	0.91
	Maximum	0.99	1	0.94	0.93
Reference method value, mg/dL (mmol/L)	Minimum	77 (4.25)	76 (4.19)	48 (2.66)	66 (3.66)
	Mean	170 (9.46)	152 (8.42)	179 (9.96)	135 (7.5)
	Maximum	309 (17.15)	250 (13.87)	313 (17.37)	245 (13.6)

All EQAs were conducted in the years 2012–2016. Means were always calculated over all distributions of these years and not over individual samples. EQA for automated analyzers in central laboratories are displayed with a gray background.

Figure 1: Relative deviation from assigned value. The central 95% of the results are highlighted in red. EQA for automated analyzers in central laboratories are displayed with a gray background.

each distribution. Sixty-five and 54 unique POCT devices were employed in POCT-Instand and POCT-RfB, respectively. Across EQA organizations, nine and 11 devices were often employed in CL and POCT EQAs, respectively, and therefore were used as common device categories.

For evaluation of POCT EQAs, reference method values and device specific consensus values were both used as target values. Instand gradually increased the percentage of device subgroups evaluated according to the reference method value from 1.3% in 2012 to 40.0% in 2016 (Supplemental Figure 1). In POCT-RfB, device-specific consensus values have always been used as target values if the device subgroup was large enough.

In general, success rates were higher in the EQAs for CL automated analyzers. On average, 1%–2% of participants failed in CL schemes, compared to 9%–10% of participants in POCT schemes (Table 1). In CL schemes, the central 95% of all measurements deviated less than 10% from the assigned target values after exclusion of outliers, whereas in POCT schemes, the central 95% slightly exceeded 15% deviation from the target value (Figure 1).

The device had the highest influence on performance

We investigated the influence of different factors on the percentage of “failed”, “poor”, “acceptable” and “good”

participations (Supplemental Figures 2–7). Few devices accounted for the majority of all employed devices (Supplemental Figures 2–5). The “others” group subsumes devices that were rarely used or could not be coded otherwise. Especially in the POCT EQAs, this group had a high percentage of failed results. Similarly, we investigated the effect of additional participation in other EQAs (Supplemental Figure 6). A substantial number of laboratories participated in EQAs for both CL and POCT glucose measurements. These laboratories had a higher fraction of “good” participations than laboratories that took part in only one EQA. To assess the influence of a laboratory’s experience, participants were classified as “new” (first participation), “intermediate” (2–10 participations) and “experienced” (>10 participations) [20]. Most participations were submitted by “experienced” participants (Supplemental Figure 7). In all examined EQAs, the fraction of “good” participations was lowest for new participants.

Next, we calculated univariable OR for the probability to reach a “good” result (Supplemental Figures 8, 9). Although ORs for the same devices differed between EQAs, the employed device had the strongest influence in all EQAs.

To compare effect sizes and analyze interactions, multivariable ORs to reach a “good” result were calculated for CL automated analyzers (Figure 2). The choice of a device not in the “others” category increased the probability for a “good” result up to 5.4 times (for “Abbott” device). Large 95% confidence intervals for some devices (up to 1.9–4.5 “Roche Diagnostics” with “other” method) indicate a high degree of uncertainty, at least partially caused by a small sample size. “Experienced” participants had significantly higher odds (OR 1.5) to reach a “good” result than “new” participants. The effect of an additional participation in a POCT EQA was negligible (OR 1.02).

In POCT EQAs, the device had again the greatest influence on the probability to reach a “good” result (Figure 2). ORs ranged from 0.6 (“Bayer Vital Contour”) to 6.2 (“Roche Accu-Chek Inform II”). The effect of additional participation in a CL EQA scheme was greater (OR 1.2) than the reciprocal influence of POCT EQAs on CL EQAs. All ORs denoting past experience included 1 in their confidence interval.

To verify these results, we recalculated multivariable ORs for the likelihood not to fail (Supplemental Figure 10). The spread of ORs was larger but showed comparable tendencies. For 29,561 (out of 67,295) participations, the level of experience could not be determined with certainty as the first participation likely occurred before 2012. These participations were excluded from multivariable ORs calculations. To evaluate their contribution, missing data were imputed and multivariable ORs to reach a “good”

Figure 2: Multivariable odds ratios for factors affecting the probability to reach a “good” result in CL automated analyzer and in POCT glucose EQAs.

An additional variable modeled distribution-specific variations (not shown). Error bars represent 95% confidence intervals. Methods in squared brackets: HK, hexokinase; GO, glucose oxidase. Numbers on the left indicate participations with the respective factor.

result ratios recalculated (Supplemental Figure 11) [21]. The recalculated ORs slightly exceeded the 95% confidence interval of the ORs without imputed data for three and four devices in the CL and POCT EQA, respectively.

After “failed” EQA distributions participants switched devices more often and improved

The worse the result of the previous EQA participation, the larger was the fraction of failing participations

(Supplemental Figure 12). After having failed the previous distribution, 11% (CL-Instand), 13% (CL-RfB), 33% (POCT-Instand) and 35% (POCT-RfB) of participants failed again.

For a better understanding of ways to improve performance after failing an EQA participation, a pathway analysis was conducted (Figure 3). After a year with failed participations, laboratories left the EQA or employed a new device more often than laboratories that did not fail in the previous year. Although small, this effect could be observed in all EQAs. Of the participants who failed but continued in the EQA, 14% (CL-Instand), 15% (CL-RfB),

Figure 3: Development after failed participations.

(A) Pathway of participants after a year with and without failed participations. Number of participants are indicated above bar. EQA for CL automated analyzers are displayed with a gray background. (B) Comparison of the year after a failed participation with and without a new device. Numbers on top of bars depict total number of participations. Too few participants failed in CL EQAs for meaningful analysis.

41% (POCT-Instand) and 50% (POCT-RfB) failed again at least once in the following year. For participants who changed their failed POCT device (or the declaration of the same), the percentage of “good” results was higher and the percentage of “failed” results lower than for those who did not change in the following year.

Certain POCT devices approached the analytical performance of central laboratory instruments

To estimate imprecision, the weighted mean of the coefficients of variation (mCV) was determined for each device (Table 2). The mCV of CL automated analyzers ranged from 0.022 to 0.031 (median 0.026). Some POCT devices reached comparable imprecision. The mCVs of all POCT devices ranged from 0.025 to 0.084 (median 0.055).

The characteristic function was fitted to measured values of each device to determine changes of imprecision over the measuring range (Table 2, Figure 4, Supplemental Figure 13). Most CL devices exhibited small fitted values for parameter α denoting low absolute imprecision in the lower measuring range (overall median 1.15, IQR, 0.69–1.62). For POCT devices, values for α were higher but often exhibited wide confidence intervals (overall median 2.12, IQR 0.20–4.68) (Supplemental Table 1). Parameter β expresses a relative imprecision in the higher measuring

range. The smaller the α , the more closely the β matched the respective mCVs for most devices. We used the characteristic function and fitted parameters to calculate CVs at 80 and 300 mg/dL (4.4 and 16.8 mmol/L). In the lower measuring range, CVs were up to 1.6 (CL) and 2.4 (POCT) times larger than in the higher range.

Biases were determined for CL EQAs (Supplemental Figure 14). The same device exhibited the highest median bias in CL-Instand and in CL-RfB (Siemens Advia, 2.2% and 1.8%).

Lot-to-lot differences are a serious source of variation for POCT glucose testing

In POCT-Instand, 80 samples were measured with the same device but with at least two different test strip lots. Maximum differences between lots exceeded 5% of the assigned target value in nine samples (11%) (Figure 5). These differences were estimated with wide confidence intervals.

The smallest allowable bias, the bias budget, occurs in the lower measurement range for most devices

The bias budget, the maximum allowable bias that is still clinically acceptable for a laboratory test with a given

Table 2: Precision of POCT devices and CL automated analyzers independently determined as weighted mean CV (mCV) and as characteristic function with parameters α and β fitted to the data.

Device	Type	EQA	mCV	α	β	CV at 80 mg/dL	CV at 300 mg/dL	Ratio CV80/CV300
Abbott	CL	Instand	0.022	0.638	0.021	0.023	0.021	1.1
		RfB	0.023	1.503	0.019	0.027	0.020	1.4
Beckman Coulter AU-Series	CL	Instand	0.025	1.096	0.024	0.028	0.024	1.1
		RfB	0.023	0.583	0.023	0.024	0.023	1.0
Beckman Coulter other devices [GO]	CL	Instand	0.025	1.925	0.020	0.031	0.021	1.5
		RfB	0.031	2.193	0.025	0.037	0.026	1.4
Beckman Coulter other devices [HK]	CL	Instand	0.026	0.856	0.025	0.027	0.025	1.1
		RfB	0.030	2.557	0.023	0.039	0.024	1.6
Roche Diagnostics [GO]	CL	Instand	0.029	1.379	0.027	0.032	0.027	1.2
		RfB	0.028	1.663	0.024	0.032	0.025	1.3
Roche Diagnostics [HK]	CL	Instand	0.025	0.910	0.024	0.026	0.024	1.1
		RfB	0.025	1.109	0.023	0.027	0.024	1.1
Roche Diagnostics [others]	CL	Instand	0.028	0.151	0.031	0.031	0.031	1.0
		RfB	0.030	0.046	0.033	0.033	0.033	1.0
Siemens Advia	CL	Instand	0.028	0.413	0.028	0.028	0.028	1.0
		RfB	0.028	1.199	0.026	0.030	0.026	1.1
Siemens Dimension	CL	Instand	0.024	1.736	0.020	0.029	0.020	1.4
		RfB	0.024	1.245	0.022	0.027	0.023	1.2
Abbott Precision Xceed Pro	POCT	Instand	0.069	4.995	0.057	0.084	0.059	1.4
		RfB	0.056	3.723	0.046	0.066	0.048	1.4
Bayer Vital Contour	POCT	Instand	0.060	1.730	0.057	0.061	0.057	1.1
		RfB	0.066	0.082	0.077	0.077	0.077	1.0
Bayer Vital Contour XT	POCT	Instand	0.041	0.075	0.042	0.042	0.042	1.0
		RfB	0.042	0.053	0.047	0.047	0.047	1.0
Dytrex (Infopia) Easy Gluco	POCT	Instand	0.065	2.405	0.062	0.069	0.063	1.1
		RfB	0.081	8.435	0.057	0.120	0.063	1.9
HemoCue B-Glucose Analyzer	POCT	Instand	0.053	6.033	0.033	0.082	0.039	2.1
		RfB	0.061	3.579	0.056	0.071	0.057	1.3
Lifescan One Touch Vita	POCT	Instand	0.084	9.142	0.064	0.131	0.071	1.8
		RfB	0.063	7.434	0.033	0.099	0.042	2.4
Nova StatStrip	POCT	Instand	0.047	0.565	0.047	0.048	0.047	1.0
		RfB	0.051	2.051	0.047	0.054	0.047	1.1
Roche Accu-Chek Aviva	POCT	Instand	0.025	0.532	0.025	0.026	0.025	1.0
		RfB	0.046	2.193	0.041	0.049	0.041	1.2
Roche Accu-Chek Inform II	POCT	Instand	0.026	0.041	0.029	0.029	0.029	1.0
		RfB	0.037	1.373	0.034	0.038	0.034	1.1
Roche Accu-Chek Performa	POCT	Instand	0.028	0.087	0.029	0.029	0.029	1.0
		RfB	0.041	0.078	0.042	0.042	0.042	1.0
Roche Accu-Chek Sensor/Comfort/Inform	POCT	Instand	0.061	5.246	0.037	0.075	0.041	1.8
		RfB	0.058	3.614	0.043	0.063	0.045	1.4

The mCVs remain constant over the measuring range. The characteristic function models CVs that change depending on the measured value. CVs at 80 and 300 mg/dL are calculated using the parameters of the characteristic function. The ratio of these CVs is provided. EQA for automated analyzers in central laboratories are displayed with a gray background.

imprecision, was calculated with the previously determined mCVs and parameters from the characteristic function (Figures 6–8). Bias budgets calculated using the TGC simulation as risk assessments were slightly smaller than the respective bias budgets derived from the SEG. They were symmetric for most devices implying that a positive

bias carries nearly the same risk as a negative bias. Regardless of whether imprecision was formulated as mCV or as a characteristic function, glucose concentrations in the range 60–115 mg/dL (3.3–6.4 mmol/L) were likely to display an unacceptable bias and determined the bias budget for nearly all devices. The absolute bias budgets of CL automated

Figure 4: Relative imprecision over the measuring range modeled using the characteristic functions.

Colored ribbons represent 95% confidence interval determined by bootstrapping. EQA for CL automated analyzers are displayed with a gray background.

analyzers were largely homogeneous and ranged from 20 to 27 mg/dL (1.1–1.5 mmol/L) (median 24, IQR: 23–25 mg/dL [1.3, 1.2–1.4 mmol/L]). For two POCT devices (Dytrex (Infopia) Easy Gluco; Lifescan One Touch Vita), the imprecision alone was too high to reach an acceptable risk. Overall, bias budgets for POCT devices were smaller and differences between devices greater (median: 17, IQR: 13–21 mg/dL [0.9, 0.7–1.2 mmol/L]) than for CL devices.

Discussion

In this work, we analyzed glucose measurements of over 130,000 samples from the German EQA organizations Instand and RfB. As regular participation in EQAs is mandatory for all medical laboratories performing analysis of glucose, these data provide a comprehensive overview of glucose measurements in Germany. The high number of

Figure 5: Maximum difference between robust central locations of different lots used in the same device and for same sample. Filled dots represent differences >5%. Error bars represent 95% confidence intervals determined by bootstrapping.

Figure 6: Schematic calculation of the bias budget based on the data from EQAs and on risk assessment with the Surveillance Error Grid.

Gray overlay represents 99.7% (three standard deviations) of measurements given the imprecision. Black dotted lines represent the border between acceptable and unacceptable errors. Two blue bars represent the allowable positive or negative bias, the “bias budget”.

analyzed samples and mandatory laboratory participation constitute a strength of this study. On the other hand, especially the results of POCT EQAs have to be interpreted with caution as the lack of commutability may have contributed to different behaviors of stabilized EQA samples

as compared to patient samples. Participants’ mistakes during data entry, different encodings, different assessments and different samples are further possible sources of variations. In this work, analytical performance specifications for failing a distribution have been derived from Rili-BAEK, which is legally binding in Germany [19], but they might be determined differently in other EQAs [26]. To increase comparability with other studies, we reused existing classifications for non-failed participations [20]. As far as possible, we used more than one method to analyze the same question to increase validity. Robust statistics were employed to avoid an oversized influence of outliers.

To identify factors associated with good analytical quality, we calculated uni- and multivariable ORs. Considerable differences between ORs for different devices indicate that the device itself had the highest influence on analytical quality for POCT and CL automated analyzers. Especially for POCT, concurrent participation in other glucose EQAs and in particular for CL, a higher number of previous participations is associated with a higher chance for a “good” result in an EQA distribution. Similar observations were made in studies with other EQAs [20, 27, 28]. However, EQA studies cannot differentiate between the experience in conducting the specific EQA and the experience in conducting the actual analytical test. It therefore remains an open question if the improvement represents a real improvement in analytical quality or if this effect is merely the result of increased experience with EQAs. Nevertheless, educational guidance to healthcare professionals with only limited experience in laboratory medicine seems worthwhile [17]. Given the importance of the device, this could include an evaluation of POCT devices according to standards such as DIN EN ISO 15197, e.g. for primary health care [20, 29].

A participant who had failed is more likely to fail again. A failed participation in an EQA led to higher rates of participants leaving the EQA (and likely stop measuring glucose as participation in EQAs is mandatory). Also, more laboratories switched their device and achieved better results. Both actions likely improve overall analytical performances in Germany.

EQAs of glucose measurements from CL automated analyzers exhibited excellent agreement with reference method values. Single POCT devices approached the imprecision of CL automated analyzers. However, considerable variation among POCT devices was observed. Our data suggest that for some POCT devices, imprecision increases in the low concentration range. Similar considerable differences in performance in the low-glucose range were also observed in other studies using unaltered capillary blood samples [30]. The large fraction of lots

Figure 7: Bias budget based on the surveillance error grid for risk assessment. EQA for automated analyzers in central laboratories are displayed with a gray background.

(11%) with a difference of $>5\%$ of the target value constitute another substantial source of variation. Lot-to-lot differences are known to significantly affect measurements of native patient samples [31, 32]. Results from control material, however, may again not be readily transferable [33]. Of note, the stated imprecisions represent long-term reproducibility imprecisions. They already include short-term biases such as the lot-to-lot-variations or differences in daily calibrations [34]. In line with the guide to the expression of uncertainty in measurement, long-term biases affecting all samples should be corrected [35]. However, for some devices, a consistent bias over many EQA distributions has been found.

To evaluate the impact of imprecision on medical decision making, we propose the concept of the “bias budget” to express a maximum permissible bias that still avoids unacceptable risks from erroneous glucose measurements. The bias budget relies primarily on detailed clinical risk assessments, such as the SEG [6] or the TGC simulation [3], available for glucose measurements. Because of this direct approach, the influence of errors in the very low end of the measurement range that are large relative to the true value but small in absolute terms is limited to a clinically justifiable influence [36]. The bias budget can avoid mathematical assumptions such as a linear relationship between bias and imprecision [37], or a

Figure 8: Bias budget based on the tight glucose control simulation for risk assessment. EQA for automated analyzers in central laboratories are displayed with a gray background.

predetermined distribution of glucose values [3]. The bias budget covers biases arising in all steps of the total testing process and therefore exceeds the analytical bias determined in performance specifications.

For almost all CL and POCT devices, the most critical glucose values are in the lower measuring range from 60–115 mg/dL (3.3–6.4 mmol/L). EQA organizers should focus on this range and design their glucose schemes accordingly. The bias budget can be depleted by random bias caused by various, mostly sample-specific factors such as unimpeded glycolysis [7], abnormal hematocrit values [11] or interfering factors [9, 10]. Especially with POCT, laboratory medicine does not aim for perfect measurements but

has to balance turnaround time, costs and accuracy based on medical needs. The estimate through the bias budget could help the clinician decide which sample-specific error factors are still permissible. For example, a negative bias of 10 mg/dL (0.6 mmol/L) occurring during 1–2 h of uninhibited glycolysis is still acceptable for most CL automated analyzers. Interfering factors, e.g. from special medications, often induce biases exceeding the budget for most POCT devices. These risks need to be mitigated, e.g. by operator training when implementing a POCT testing program [38].

EQAs can ensure a continuous high analytical quality of glucose measurements. Their data must be made

available to facilitate learning in the laboratory healthcare system. To further increase the informative value of EQAs for science and quality control, commutable materials for POCT glucose EQAs or alternative designs for proficiency testing schemes are urgently needed [14, 39, 40].

Acknowledgments: The authors would like to thank Dirk Illigen for technical assistance and Evangeline Thaler for reviewing the manuscript. We would also like to thank David C. Klonoff and Michael A. Kohn for providing the individual data points of the Surveillance Error Grid.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

- Clarke SF, Foster JR. A history of blood glucose meters and their role in self-monitoring of diabetes mellitus. *Br J Biomed Sci* 2012;69:83–93.
- Andreis E, Küllmer K, Appel M. Application of the reference method isotope dilution gas chromatography mass spectrometry (ID/GC/MS) to establish metrological traceability for calibration and control of blood glucose test systems. *J Diabetes Sci Technol* 2014;8:508–15.
- Karon BS, Boyd JC, Klee GG. Glucose meter performance criteria for tight glycemic control estimated by simulation modeling. *Clin Chem* 2010;56:1091–7.
- 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;23:1143–8.
- Clarke WL. The original Clarke error grid analysis (EGA). *Diabetes Technol Ther* 2005;7:776–9.
- Klonoff DC, Lias C, Vigersky R, Clarke W, Parkes JL, Sacks DB, et al. The surveillance error grid. *J Diabetes Sci Technol* 2014;8:658–72.
- Gambino R, Piscitelli J, Ackattupathil TA, Theriault JL, Andrin RD, Sanfilippo ML, et al. Acidification of blood is superior to sodium fluoride alone as an inhibitor of glycolysis. *Clin Chem* 2009;55:1019–21.
- Bruns DE, Knowler WC. Stabilization of glucose in blood samples: why it matters. *Clin Chem* 2009;55:850–2.
- Schleis TG. Interference of maltose, icodextrin, galactose, or xylose with some blood glucose monitoring systems. *Pharmaco-therapy* 2007;27:1313–21.
- Erbach M, Freckmann G, Hinzmann R, Kulzer B, Ziegler R, Heinemann L, et al. Interferences and limitations in blood glucose self-testing: an overview of the current knowledge. *J Diabetes Sci Technol* 2016;10:1161–8.
- Ramljak S, Lock JP, Schipper C, Musholt PB, Forst T, Lyon M, et al. Hematocrit interference of blood glucose meters for patient self-measurement. *J Diabetes Sci Technol* 2013;7:179–89.
- Schifman RB, Howanitz PJ, Souers RJ. Point-of-care glucose critical values: a Q-probes study involving 50 health care facilities and 2349 critical results. *Arch Pathol Lab Med* 2016;140:119–24.
- German Medical Association. Revision of the “Guideline of the German Medical Association on Quality Assurance in Medical Laboratory Examinations–RiliBAEK”. *J Lab Med* 2015;39:26–69.
- Jacobs J, Fokkert M, Slingerland R, De Schrijver P, Van Hoovels L. A further cautionary tale for interpretation of external quality assurance results (EQA): commutability of EQA materials for point-of-care glucose meters. *Clin Chim Acta* 2016;462:146–7.
- Petersmann A, Lupp P, Michelsen A, Sonntag O, Nauck M. Gemeinsame Stellungnahme zur Situation der Bewertung von Ringversuchen für Glucose mittels Systemen für die patientennahe Sofortdiagnostik (POCT)/Joint statement on the situation of external quality control for glucose in POCT systems. *Laboratoriumsmedizin*. 2012;36:165–8.
- Wood WG. Problems and practical solutions in the external quality control of point of care devices with respect to the measurement of blood glucose. *J Diabetes Sci Technol* 2007;1:158–63.
- Stavelin A, Sandberg S. Essential aspects of external quality assurance for point-of-care testing. *Biochem Med (Zagreb)* 2017;27:81–5.
- Cerioti F. The role of external quality assessment schemes in monitoring and improving the standardization process. *Clin Chim Acta* 2014;432:77–81.
- Orth M. Are regulation-driven performance criteria still acceptable? – The German point of view. *Clin Chem Lab Med* 2015;53:893–8.
- Bukve T, Stavelin A, Sandberg S. Effect of participating in a quality improvement system over time for point-of-care C-reactive protein, glucose, and hemoglobin testing. *Clin Chem* 2016;62:1474–81.
- Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:67.
- Koch M, Magnusson B. Use of characteristic functions derived from proficiency testing data to evaluate measurement uncertainties. *Accred Qual Assur* 2012;17:399–403.
- Coucke W, Charlier C, Lambert W, Martens F, Neels H, Tytgat J, et al. Application of the characteristic function to evaluate and compare analytical variability in an external quality assessment scheme for serum ethanol. *Clin Chem* 2015;61:948–54.
- Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibián-Barrera M, et al. robustbase: Basic Robust Statistics R package version 0.92-7. 2016. <http://CRAN.R-project.org/package=robustbase>.
- Davison AC, Hinkley DV. Bootstrap methods and their applications. Cambridge: Cambridge University Press, 1997.
- Jones GRD, Albareda S, Kessler D, MacKenzie F, Mammen J, Pedersen M, et al. Analytical performance specifications for external quality assessment – definitions and descriptions. *Clin Chem Lab Med* 2017;55:949–55.
- Howerton D, Krolak JM, Manasterski A, Handsfield JH. Proficiency testing performance in US laboratories: results reported to the Centers for Medicare & Medicaid Services, 1994 through 2006. *Arch Pathol Lab Med* 2010;134:751–8.

28. Morandi PA, Deom A, Kessler D, Cohen R. Retrospective analysis of 88,429 serum and urine glucose EQA results obtained from professional laboratories and medical offices participating in surveys organized by three European EQA centers between 1996 and 2007. *Clin Chem Lab Med* 2010;48:1255–62.
29. Mahoney JJ, Ellison JM. Assessing glucose monitor performance – a standardized approach. *Diabetes Technol Ther* 2007;9:545–52.
30. Heinemann L, Zijlstra E, Pleus S, Freckmann G. Performance of blood glucose meters in the low-glucose range: current evaluations indicate that it is not sufficient from a clinical point of view. *Diabetes Care* 2015;38:e139–40.
31. Baumstark A, Pleus S, Schmid C, Link M, Haug C, Freckmann G. Lot-to-lot variability of test strips and accuracy assessment of systems for self-monitoring of blood glucose according to ISO 15197. *J Diabetes Sci Technol* 2012;6:1076–86.
32. Stavelin A, Riksheim BO, Christensen NG, Sandberg S. The importance of reagent lot registration in external quality assurance/proficiency testing schemes. *Clin Chem* 2016;62:708–15.
33. Kristensen GB, Christensen NG, Thue G, Sandberg S. Between-lot variation in external quality assessment of glucose: clinical importance and effect on participant performance evaluation. *Clin Chem* 2005;51:1632–6.
34. Theodorsson E. Uncertainty in measurement and total error: tools for coping with diagnostic uncertainty. *Clin Lab Med* 2017;37:15–34.
35. Oosterhuis WP, Theodorsson E. Total error vs. measurement uncertainty: revolution or evolution? *Clin Chem Lab Med* 2016;54:235–9.
36. Freckmann G, Schmid C, Baumstark A, Rutschmann M, Haug C, Heinemann L. Analytical performance requirements for systems for self-monitoring of blood glucose with focus on system accuracy: relevant differences among ISO 15197:2003, ISO 15197:2013, and current FDA recommendations. *J Diabetes Sci Technol* 2015;9:885–94.
37. Oosterhuis WP, Sandberg S. Proposal for the modification of the conventional model for establishing performance specifications. *Clin Chem Lab Med* 2015;53:925–37.
38. Barabas N, Bietenbeck A. Application guide: training of professional users of devices for near-patient testing. *J Lab Med* 2017;41:215–8.
39. Delatour V, Lalere B, Saint-Albin K, Peignaux M, Hattchouel J-M, Dumont G, et al. Continuous improvement of medical test reliability using reference methods and matrix-corrected target values in proficiency testing schemes: application to glucose assay. *Clin Chim Acta* 2012;413:1872–8.
40. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. *Clin Chim Acta* 2003;327:25–37.

Supplemental Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/cclm-2017-1142>).