This manuscript has been published in:

Diabetes Technology & Therapeutics

http://doi.org/10.1089/dia.2022.0331

A statistical approach for assessing the compliance of integrated continuous glucose monitoring systems with FDA accuracy requirements

Peter Stephan, MSc[1]; Manuel Eichenlaub, PhD[2]; Delia Waldenmaier, PhD[2]; Stefan Pleus, PhD[2]; Martina Rothenbühler, PhD[3]; Cornelia Haug, MD[2]; Guido Freckmann, MD[2]

1 Mannheim, Germany

2 Institut für Diabetes-Technologie, Forschungs- und Entwicklungsgesellschaft mbH an der Universität Ulm, Ulm, Germany

3 Diabetes Center Berne, Bern, Switzerland

Running title

Assessing the compliance of CGM systems with FDA accuracy requirements

Corresponding Author

Manuel Eichenlaub, PhD, Institut für Diabetes-Technologie Forschungs- und Entwicklungsgesellschaft mbH an der Universität Ulm, Lise-Meitner-Strasse 8/2, D-89081 Ulm, Germany, e-mail: manuel.eichenlaub@idt-ulm.de, phone +49-731-50-990-50

Key words

FDA iCGM requirements, confidence interval, agreement rate, bootstrapping, Wilson interval, continuous glucose monitoring, accuracy

Abstract (max. 150)

To assess the compliance of "integrated" continuous glucose monitoring (CGM) systems with U.S. Food and Drug Administration (FDA) requirements, the calculation of confidence intervals on agreement rates, i.e., the percentage of CGM measurements lying within a certain deviation of a comparator method, is stipulated. However, despite the existence of numerous approaches that could yield different results, a specific procedure for calculating confidence intervals is not described anywhere. This report therefore proposes a suitable statistical procedure to allow transparency and comparability between CGM systems. Three existing methods were applied to six datasets from different CGM performance studies. The results indicate that a bootstrap-based method that accounts for the clustered structure of CGM data is reliable and robust. We thus recommend its use for the estimation of confidence intervals of agreement rates. A software implementation of the proposed method is freely available (https://github.com/IfDTUlm/CGM_Performance_Assessment).

Introduction

In 2018, the United States Food and Drug Administration (FDA) introduced requirements for "integrated" continuous glucose monitoring (CGM) systems. In terms of point accuracy, i.e. the agreement between CGM and comparator blood glucose measurements at a given point in time, these requirements are mainly defined for agreement rates (AR) (Table 1).[1] The AR provides the percentage of CGM measurements lying within certain limits, e.g. ±20%, of their paired comparator measurements.

This report deals with one specific aspect of the FDA requirements: the fact that the AR requirements are defined for the lower bound of the corresponding one-sided 95% confidence interval (CI). The use of CIs marks a fundamental shift compared to the acceptance requirements of blood glucose monitoring systems (BGMS) and rightly acknowledges the existence of statistical uncertainty when drawing conclusions about the accuracy of CGM systems in general on the basis of individually planned clinical studies. The lower bound of the CI can be interpreted as a threshold above which we are 95% confident that the true AR lies. An issue with these requirements is, however, that a specific description of the procedure for calculating the CI is neither required by the FDA, nor provided in existing FDA approval requests.[1,6] Furthermore, discussions on possible CI calculation procedures within the context of CGM performance studies are missing so far. In contrast, numerous general statistical approaches for this procedure potentially leading to different results have been proposed in the literature.[2–5] The standardization of the statistical method is therefore essential to allow transparency and comparability of the results between CGM systems. The aim of this brief report is to review and evaluate selected approaches for CI calculation with regard to their applicability to CGM data from real clinical performance studies and to provide a statistical foundation for developers, manufacturers and researchers.

Of note is also that the ARs and their associated CIs in different glucose ranges are calculated with respect to the measurements of the CGM system. In contrast to the assumed to be more accurate comparator method, this categorization with respect to the CGM system is unusual in the field of measurement method comparison and means that the comparison of different CGM systems amongst each other is hindered. However, this aspect will not be further

discussed as the approaches for CI calculation are independent from the measurement method used for glucose range determination.

Methods

Selection of statistical approaches for confidence interval calculation

A well-known characteristic of the CI is its dependence on sample size. A larger sample size typically reduces the statistical uncertainty, which is reflected in a narrower CI, i.e., the distance between AR and associated lower bound. Another relevant factor is the variability of the measurement values. A higher variability extends the CI and can be compensated by an increase in sample size. In the case of CGM systems, there is an additional important point to consider: measurement values within one sensor correlate as they are affected, e.g. by physiological factors of the individual subject and by measurement properties of the sensors. Therefore, the distinction between sensor-to-sensor variability and variability within a sensor is essential. Data from each sensor is thereby considered as a cluster. Standard approaches that disregard the clustered structure of the data, tend to underestimate the width of the CI,[3,7] which could lead to a requirement being wrongfully met. Therefore, several methods for calculating the CI of ARs in clustered data have already been proposed in the context of clinical trials.[4,5] In this report three approaches for both clustered and unclustered data were selected and compared.

The first approach is the standard Clopper-Pearson (CP) method,[8] that neglects the clustered nature of the data. It was selected for comparison with the following two approaches to highlight the importance of using a method that takes the clustered structure of the data into account.

The second approach considered in this report is the continuity-corrected Wilson (WCC) method for clustered data. Here, the CI is calculated by applying defined statistical formulas according to specific assumptions to the CGM data. This semi-parametric method is relatively new and has been shown to outperform other semi-parametric methods for small sample sizes under a wide range of scenarios. The WCC approach accounts for the clustered structure of the data by differentiating between sensor-to-sensor variability and variability within a sensor using a standard analysis of variance (ANOVA) approach.[4,5]

The third approach uses a technique known as bootstrapping.[9,10] It allows general conclusions about the accuracy of the CGM system without relying on assumptions about the statistical

properties of the target parameter and is therefore a viable alternative to the semi-parametric methods. The bootstrapping method randomly resamples the data with replacement, thereby mimicking the data collection process and allowing the simulation of many “virtual” repetitions of the clinical study. Here it is important to note that the resampling is done with respect to the sensors, therefore preserving the clustered structure of the data. For each repetition, the AR is calculated, and its CI is then estimated from the bootstrap samples of “virtual” clinical studies. A recommended variant of the bootstrapping method for calculating CIs is the bias-corrected and accelerated bootstrapping (BCa) method.[9,11] It allows the correction of possible bias and skewness of the bootstrapping estimator. To ensure reliability and reproducibility of the CI, the number of bootstrap samples was determined to be 10,000 by repeating the entire procedure multiple times and ensuring sufficiently good reproducibility.

The results are expressed in terms of the negative width of the CI as we are concerned with the lower bound of the CI. This facilitates the comparison of results independent from the value of the AR and indicates the conservative nature of the methods, i.e. a more conservative method will lead to a smaller lower bound of the CI and thus to a more negative value for the CI width.

Data description

To evaluate the methods for CI calculation six datasets corresponding to six CGM systems obtained in four clinical CGM performance studies are used. Comparator BG measurements were collected from capillary samples using commercially available BGMS. The number of subjects with valid data per dataset varies between 23 and 48, and only data from one sensor was used per subject. Details of the used datasets are provided in the supplementary materials.

Results

The results of calculating the CI widths with the three selected approaches for the different datasets are depicted in Figure 1. More detailed results for each dataset are provided in the supplementary materials.

Discussion

The results in Figure 1 demonstrate that, except for the glucose range <70 mg/dL, the WCC and BCa approaches yield similar median CI widths while the CP method leads to narrower CIs. This confirms the expectation that the CP method underestimates the width of the CIs when applied to CGM data, because the clustered data structure is neglected. However, the difference between WCC and BCa on the one hand, and CP on the other hand, is decreased for requirement 2 (Figure 1 (B)), indicating that the cluster effect is less prominent in this case.

Comparing WCC and BCa, it can be observed that, although similar, the WCC approach yields slightly broader CI estimates in almost all cases. This confirms the results of previous works that used Monte Carlo simulations and concluded that the WCC approach tends to be overly conservative.[5]

Inspecting the results for the glucose range <70 mg/dL, especially for requirement 2, larger differences between WCC and BCa approaches become apparent. This is mainly caused by the fact that in this glucose range, individual sensors in the datasets have no or only a single data point (see the details of the used datasets in the supplementary materials), because reliably inducing multiple glucose values in the hypoglycemic range for every subject can be challenging in practice. In this case, the WCC approach is inadequate as the required ANOVA-based calculation of within- and between-sensor variability is impaired. This can be demonstrated by excluding the respective data of sensors with only one data point per range from analysis (detailed results are provided in the supplementary materials). Here, the median CI widths of the CP and BCa methods are only marginally affected, while the WCC leads to considerably smaller median CI widths and thus approaches the results of the BCa method.

Considering our goal to suggest a universally applicable method without explicit knowledge of data characteristics and the overly conservative nature of the WCC approach as well as its issue with small sample sizes, we propose to apply the BCa method to calculate the CIs of ARs in CGM accuracy studies. Furthermore, the BCa method could be easily adapted and evaluated in the context of CI estimation for other CGM accuracy parameters such as mean absolute relative difference.

In the extreme case of an AR of exactly 100% (all values are within the limit), the variability between sensors can no longer be calculated as every individual sensor has the same AR of 100%. In this case, the BCa method yields no result while the WCC approach only considers the number of sensors and disregards the number of data points per sensor. This provides a far too conservative estimate (Figure 1 (B), glucose range >180 mg/dL, datasets 5 and 6) and for this reason, these results were excluded from the median calculation. In this particular case we suggest to apply the CP method instead of the BCa method, as the difference between CP and BCa is no longer pronounced for ARs close to 100%.

Although the sample size of any single dataset examined in this report is limited, we argue that by considering six datasets with a total of 191 sensors and ~26,000 data points, the overall findings regarding the comparability and suitability of the considered methods translate to larger datasets typically used for FDA approval submission (~150 sensors with ~20,000 data points).[1,6]

Conclusions

This report evaluated different methods for calculating the CI of ARs from data collected in CGM performance studies and found that the bootstrap-based BCa approach accounting for the clustered nature of the data is most suitable. In the case of an observed AR of 100% the CP method should be applied. We thus encourage researchers and manufacturers to apply the procedure to CGM performance studies in general to benefit from its meaningfulness.

In the interest of transparency and to facilitate the use of the proposed method by manufacturers and the scientific community a software implementation in Python and R is published alongside this brief report (https://github.com/IfDTUlm/CGM_Performance_Assessment).

Acknowledgements

The authors would like to thank the Diabetes Center Berne for their financial support.

Authorship contribution

PS: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft

ME: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – review & editing

DW: Conceptualization, Methodology, Visualization, Writing – review & editing

SP: Methodology, Validation, Writing – review & editing

MR: Methodology, Validation, Writing – review & editing

CH: Writing – review & editing

GF: Conceptualization, Writing – review & editing

Conflict-of-Interest Statement

GF is general manager and medical director of the IfDT (Institut für Diabetes-Technologie Forschungs- und Entwicklungsgesellschaft mbH an der Universität Ulm, Ulm, Germany), which carries out clinical studies on the evaluation of BG meters, with CGM systems and medical devices for diabetes therapy on its own initiative and on behalf of various companies. GF/IfDT have received speakers' honoraria or consulting fees from Abbott, Ascensia, Berlin Chemie, Beurer, BOYDsense, CRF Health, Dexcom, i-SENS, Lilly, Metronom, MySugr, Novo Nordisk, Pharmasens, Roche, Sanofi, Sensile, Terumo and Ypsomed.

ME, DW, SP and CH are employees of the IfDT.

PS is an advisor to the IfDT

MR is an employee of Diabetes Center Berne.

Funding Source

This study was supported by Diabetes Center Berne, Switzerland.

References

1. U.S. Food and Drug Administration. EVALUATION OF AUTOMATIC CLASS III DESIGNATION FOR Dexcom G6 Continuous Glucose Monitoring System: Decision Summary. Published online 2018. Accessed May 24, 2022. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170088.pdf

2. Vollset SE. Confidence intervals for a binomial proportion. *Stat Med*. 1993;12(9):809-824.

3. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17(8):857-872.

4. Saha KK, Miller D, Wang S. A Comparison of Some Approximate Confidence Intervals for a Single Proportion for Clustered Binary Outcome Data. *Int J Biostat*. 2016;12(2).

5. Short MI, Cabral HJ, Weinberg JM, LaValley MP, Massaro JM. A novel confidence interval for a single proportion in the presence of clustered binary outcome data. *Stat Methods Med Res*. 2020;29(1):111-121.

6. U.S. Food and Drug Administration. 510(k) SUBSTANTIAL EQUIVALENCE DETERMINATION DECISION SUMMARY. Published online 2020. Accessed June 23, 2022. https://www.accessdata.fda.gov/cdrh_docs/reviews/K193371.pdf

7. Rao JNK, Scott AJ. A Simple Method for the Analysis of Clustered Binary Data. *Biometrics*. 1992;48(2):577-585.

8. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*. 1934;26(4):404-413.

9. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. CRC Press; 1993.

10. Rutter CM. Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Acad Radiol*. 2000;7(6):413-419.

11. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Stat Sci*. 1996;11(3):189-228.

Figures

Figure 1: Confidence interval widths calculated as the difference of the lower bound of the one-sided 95% confidence intervals to the corresponding agreement rate of FDA requirement 1 (A) and requirement (B). Compared are the results from the Clopper-Pearson (CP), continuity-corrected Wilson (WCC) and bias-corrected and accelerated bootstrapping (BCa) approach for all datasets and separated by CGM glucose ranges.

Tables

Table 1: FDA requirements defined for point accuracy.[1]

CGM glucose range	*<70 mg/dL*	*70-180 mg/dL*	*>180 mg/dL*	*Total*
*Requirement 1 for agreement rate**	>85%	>70%	>80%	>87%
Limit 1	±15 mg/dL	±15%	±15%	±20%
*Requirement 2 for agreement rate**	>98%	>99%	>99%	-
Limit 2	±40 mg/dL	±40%	±40%	
Requirement 3	No comparator value above 180 mg/dL	-	No comparator value below 70 mg/dL	-

** Lower bound of one-sided 95% confidence interval*