Supplementary Material

Table S1a. Geography of Experienced Clinicians

Location	Number	%
United States	198	60
Africa	9	3
Asia	31	9
Europe	57	17
North America (excluding US)	2	1
Oceania	21	6
South America	12	4
Total	330	100

Table S1b. Specialty of Experienced Clinicians

Specialty	Number	%
Adult endocrinologist	218	66
Pediatric endocrinologist	25	8
Other physician	13	4
Diabetes educator	74	22
Total	330	100

Table S2. Ranges of seven AGP metrics for each of the four types of subjects whose CGM tracings were studied

T1D Closed Loop (N = 56)				Percentile			
% of time	Avg	SD	Min	25th	50th	75th	Max
Very Low (<54 mg/dL; <3.0 mmol/L)	0.39%	0.65%	0.0%	0.0%	0.0%	1.0%	3.0%
Low (54 - <70 mg/d; 3.0 - <3.9 mmol/L)	1.79%	1.52%	0.0%	1.0%	1.5%	2.5%	7.0%
In Range (70 - 180 mg/dL; 3.9 – 10.0 mmol/L)	77.11%	12.39%	43.0%	67.0%	80.0%	88.0%	95.0%
High (>180-250 mg/dL; > 10.0 – 13.9 mmol/L)	15.88%	7.50%	3.0%	9.5%	15.0%	21.5%	32.0%
Very High (> 250 mg/dL; >13.9 mmol/L)	4.84%	6.58%	0.0%	1.0%	2.5%	7.0%	35.0%
Total	100.0%						
Other							
Mean Glucose (mg/dL)	146	20	120	131.5	142	158	214
Coefficient of Variation	0.32	0.06	0.21	0.28	0.31	0.36	0.47

T1D Pump (N = 56)				Percentile			
% of time	Avg	SD	Min	25th	50th	75th	Max
Very Low (<54 mg/dL; <3.0 mmol/L)	0.98%	1.30%	0.0%	0.0%	0.5%	2.0%	5.0%
Low (54 - <70 mg/d; 3.0 - <3.9 mmol/L)	3.00%	2.77%	0.0%	0.0%	2.0%	5.0%	8.0%
In Range (70 - 180 mg/dL; 3.9 – 10.0 mmol/L)	63.80%	20.19%	16.0%	47.0%	63.0%	81.5%	97.0%
High (>180-250 mg/dL; > 10.0 – 13.9 mmol/L)	22.05%	12.68%	1.0%	12.0%	21.0%	33.5%	58.0%
Very High (> 250 mg/dL; >13.9 mmol/L)	10.16%	11.43%	0.0%	1.0%	7.0%	15.0%	57.0%
Total	100.0%						
Other							
Mean Glucose (mg/dL)	159	35	94	129.5	151.5	188.5	267
Coefficient of Variation	0.34	0.07	0.18	0.31	0.33	0.37	0.51

T1D MDI (N = 56)				Percentile			
% of time	Avg	SD	Min	25th	50th	75th	Max
Very Low (<54 mg/dL; <3.0 mmol/L)	2.52%	2.86%	0.0%	0.0%	2.0%	4.0%	12.0%
Low (54 - <70 mg/d; 3.0 - <3.9 mmol/L)	3.68%	3.14%	0.0%	1.0%	3.0%	5.2%	15.0%
In Range (70 - 180 mg/dL; 3.9 – 10.0 mmol/L)	43.95%	15.94%	13.0%	31.0%	43.0%	53.0%	88.0%
High (>180-250 mg/dL; > 10.0 – 13.9 mmol/L)	26.20%	7.60%	7.0%	22.5%	26.5%	32.0%	42.0%
Very High (> 250 mg/dL; >13.9 mmol/L)	23.66%	13.95%	1.0%	13.0%	22.0%	35.2%	52.0%
Total	100.0%						
Other							
Mean Glucose (mg/dL)	188	35	110	162	185	216	251
Coefficient of Variation	0.41	0.08	0.26	0.34	0.43	0.47	0.58

T2D MDI (N = 57)				Percentile			
% of time	Avg	SD	Min	25th	50th	75th	Max
Very Low (<54 mg/dL; <3.0 mmol/L)	0.75%	1.62%	0.0%	0.0%	0.0%	1.0%	8.0%
Low (54 - <70 mg/d; 3.0 - <3.9 mmol/L)	1.58%	2.76%	0.0%	0.0%	0.0%	2.0%	14.0%
In Range (70 - 180 mg/dL; 3.9 – 10.0 mmol/L)	54.89%	20.11%	14.0%	38.0%	52.0%	72.0%	93.0%
High (>180-250 mg/dL; > 10.0 – 13.9 mmol/L)	29.33%	12.41%	4.0%	20.0%	30.0%	40.0%	51.0%
Very High (> 250 mg/dL; >13.9 mmol/L)	13.42%	10.94%	0.0%	4.0%	12.0%	21.0%	37.0%
Total	100.0%						
Other							
Mean Glucose (mg/dL)	176	29	112	153	180	201	231
Coefficient of Variation	0.32	0.09	0.20	0.27	0.30	0.37	0.62

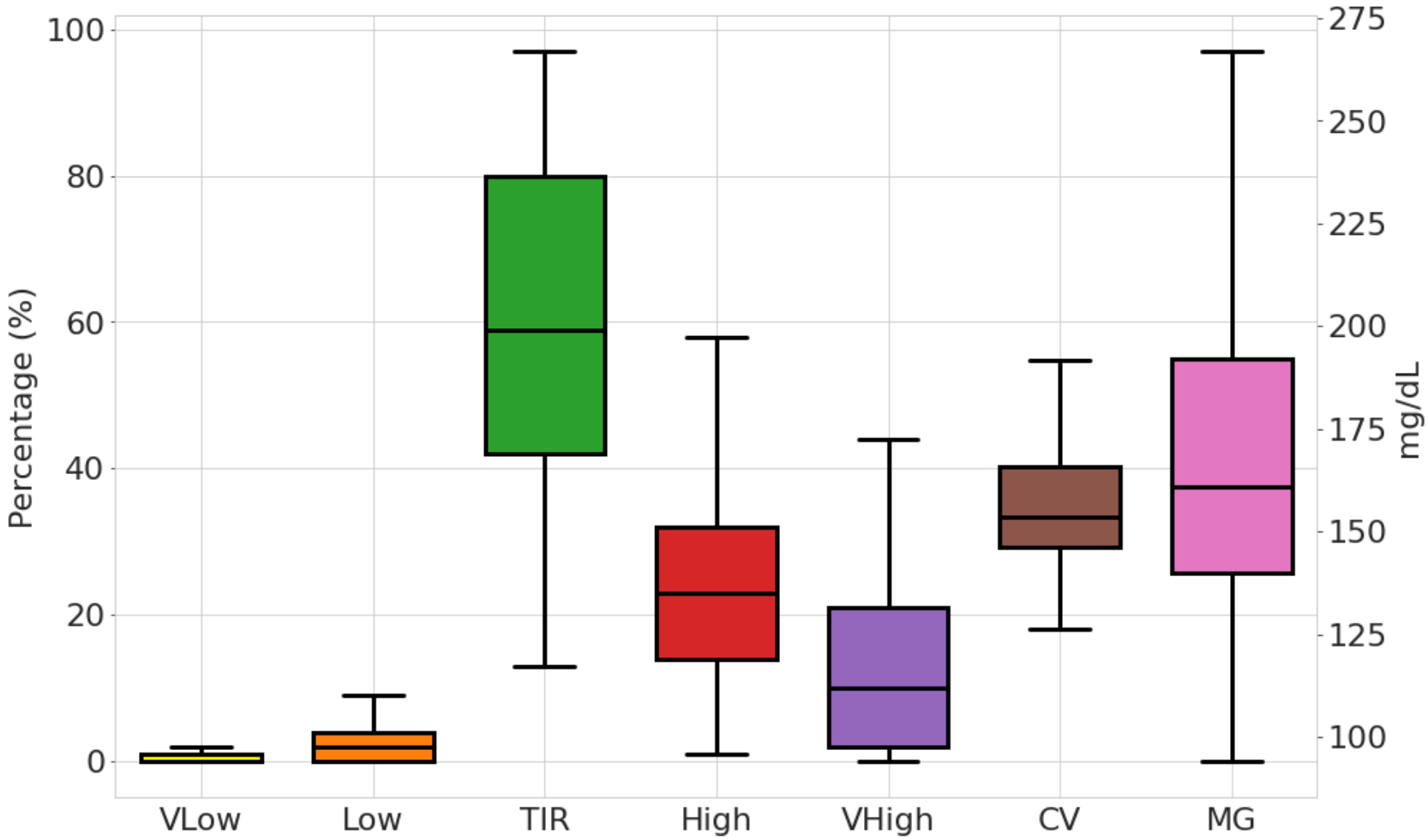

Figure S1. The 7 AGP metrics in 225 CGM tracings.
Abbreviations: VLow, very low–glucose hypoglycemia (<54 mg/dL; <3.0 mmol/L) (level 2 hypoglycemia); Low, low-glucose hypoglycemia (54-<70 mg/dL; 3.0-<3.9 mmol/L) (level 1 hypoglycemia); TIR, time in target range (70-180 mg/dL; 3.9-10.0 mmol/L); High, high-glucose hyperglycemia (>180-250 mg/dL; >10.0-13.9 mmol/L) (level 1 hyperglycemia); VHigh, very high–glucose hyperglycemia (>250 mg/dL; >13.9 mmol/L) (level 2 hyperglycemia); CV, coeffecient of variation, units on left-hand vertical axis are multiplied by 100; MG, mean glucose, units on right-hand vertical axis.

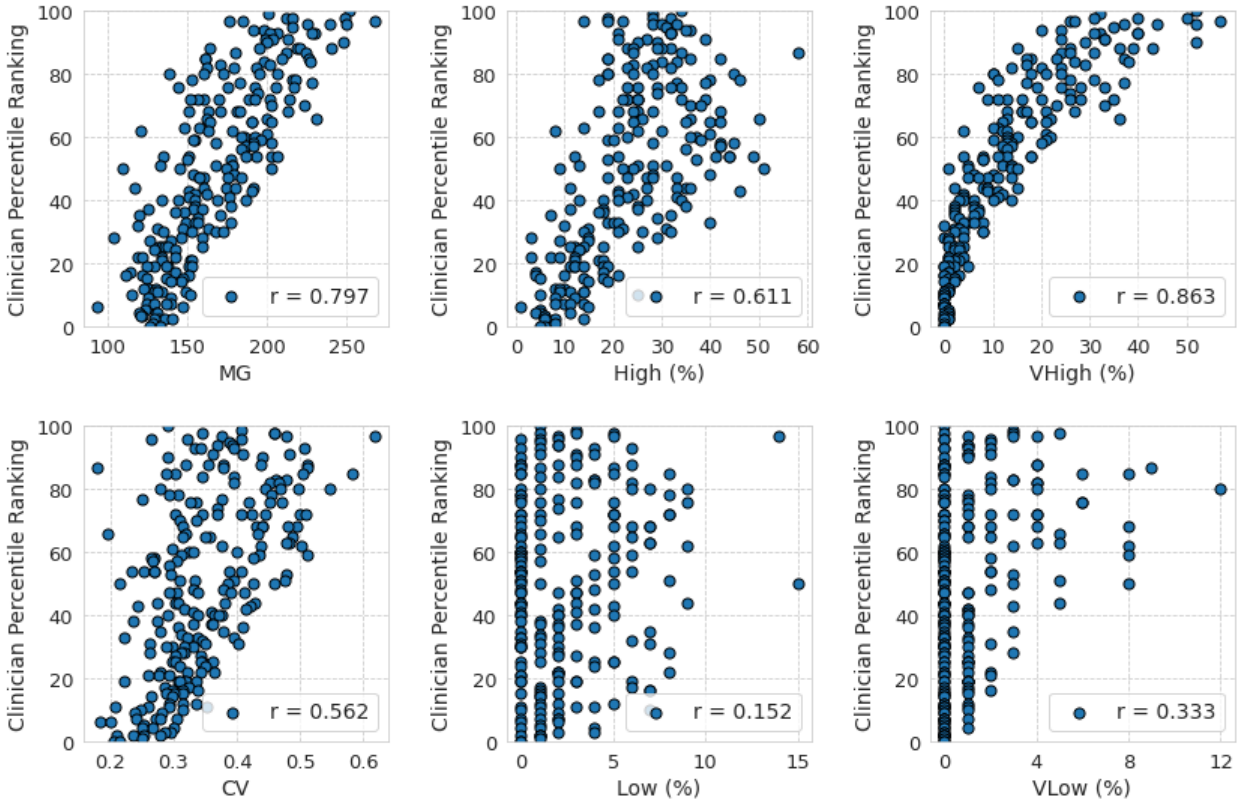

Figure S2. Correlations between six AGP metrics and clinician rankings.
The plot for the correlation between time in range (TIR) and clinician rankings is shown in Figure 3.
Abbreviations: MG = mean glucose (mg/dL); High = high-glucose hyperglycemia (>180 – 250 mg/dL; >10.0 – 13.9 mmol/L); VHigh = very high-glucose hyperglycemia (>250 mg/dL; >13.9 mmol/L); CV = coefficient of variation (standard deviation of glucose/mean glucose); Low = low-glucose hypoglycemia (54 - <70 mg/dL; 3.0 - <3.9 mmol/L); VLow = very low-glucose hypoglycemia (< 54 mg/dL; < 3.0 mmol/L)

Balanced Incomplete Block Design and Assessment of Inter-rater Agreement

In this study, the independent ratings of 330 clinicians were combined to rank a list of 225 tracings from best to worst in terms of quality of glycemia. Pre-testing showed that each clinician could only rank 5 tracings at one time, so our block size was 5. When the block size is smaller than the total number of items to be ranked, the ranking study used is an incomplete block design. For an incomplete block design to be "balanced," each possible pair of items should appear in the same number of blocks, so that any given item is compared against all the other items in the list the same number of times. We originally designed the blocks so that each pair of items would be compared against all of the other items exactly once, but we ultimately re-used the blocks so that each pair of items would be compared exactly twice.

We randomly separated the original list of 225 tracings into 5 separate sets of 45 each. Within a set of 45 tracings, there are 990 total pairs of tracings (45 × 44 /2). Our original design called for all pairs to appear exactly once. Since a block of 5 includes 10 distinct pairs (5 × 4 /2), we required 99 blocks (990/10). With 5 tracings per block that is 495 slots to be filled by 45 tracings or 11 slots per tracing. We used the classic paper by Bose (1), to create a balanced incomplete block design with 45 tracings, 99 blocks of 5, and 11 repetitions of each tracing.

The same 99 blocks were re-used so that each CGM tracing appeared in 22 groups of 5 tracings and therefore received 22 rankings from 0 (worst) to 4 (best). The tracings showed a wide range of variability. The worst tracing received 21 rankings of "0" and 1 ranking of "1". The second worst tracing received 19 rankings of "0" and 3 rankings of "1". The best tracing received 22 rankings of "4". The second-best tracing received 21 rankings of "4" and 1 ranking of "3".

There are several methods for assessing inter-rater agreement on a 5-point scale (2). A commonly used metric is the average deviation index, AD_m (3). With 5 possible ratings, an AD_m value less than 5/6 = 0.833 is considered "strong agreement". For the 225 tracings in this study, AD_m ranged from 0 to 1.2 with a mean of 0.726. Another commonly used metric for assessing the consistency of ratings is James's r_{wg} (4), which is 1 when agreement is perfect (as in the tracing that received 22 ratings of "4") and 0 when agreement is consistent with random ranking (from a discrete uniform distribution). For these 225 tracings, r_{wg} ranged from 0.064 to 1.000 with a mean of 0.558, which is considered "moderate agreement" (5). Given this study design, the mean r_{wg} will closely approximate the intraclass correlation coefficient (ICC(1,1)), which was 0.559 (6).

References

1. Bose RC. ON THE CONSTRUCTION OF BALANCED INCOMPLETE BLOCK DESIGNS. Ann Eugen. 1939 Dec;9(4):353–99.
2. O'Neill TA. An Overview of Interrater Agreement on Likert Scales for Researchers and Practitioners. Front Psychol. 2017 May 12;8:777.
3. Burke MJ, Dunlap WP. Estimating Interrater Agreement with the Average Deviation Index: A User's Guide. Organ Res Methods. 2002 Apr;5(2):159–72.
4. James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. J Appl Psychol. 1984 Feb;69(1):85–98.
5. LeBreton JM, Senter JL. Answers to 20 Questions About Interrater Reliability and Interrater Agreement. Organ Res Methods. 2008 Oct;11(4):815–52.
6. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420–8.