Reproducibility and accuracy of large language model vision APIs for carbohydrate estimation from food photographs: a four-model batch comparison with implications for automated insulin dosing

Preprint — This article has not been certified by peer review. Published at Diabettech.com, April 2026. Submitted to Diabetologia for peer review. ##
Authors

Tim Street [ORCID: 0009-0008-4417-6581]

Diabettech.com

Corresponding author: Tim Street, tim@diabettech.com

Word count: ~4,200 (excluding abstract, references, tables and figure legends)

Abstract

Aims/hypothesis

We aimed to characterise the within-image reproducibility of carbohydrate estimates from four large language model (LLM) vision APIs and to quantify the clinical risk for insulin dosing, stratifying accuracy by reference value quality.

Methods

Thirteen food photographs were each submitted 495–561 times to four LLM vision APIs (GPT-5.4, Claude Sonnet 4.6, Gemini 2.5 Pro, Gemini 3.1 Pro Preview) using an identical structured prompt adapted from the iAPS automated insulin delivery system (26,904 total queries, temperature 0.01). The primary outcome was within-image variation (coefficient of variation [CV], range, distributional normality). Secondary outcomes included accuracy against reference values for nine images, stratified by quality tier (packet label, weighed/measured, portioned, or visual estimate). Clinical risk was translated at an insulin-to-carbohydrate ratio of 1:10.

Results

Median within-image CV was 2.4% (Claude), 8.4% (GPT-5.4), 10.3% (Gemini 3.1 Pro) and 11.0% (Gemini 2.5 Pro), translating to median insulin dosing uncertainties of 0.9, 2.3, 2.9 and 4.7 U respectively. All 52 model–image distributions were non-normal (Shapiro–Wilk $p<0.05$). On the five strong-reference images (packet label or weighed), Claude achieved the lowest mean absolute error (8.7 g, 95% CI 8.6–8.9; 100% within 20 g) and 0% of queries would have caused an insulin overdose exceeding 2 U, compared with 12–37% for the other models (all pairwise $p<10^{-20}$). Claude also achieved the lowest MAE across all nine reference images (11.6 g, 86.4% within 20 g).

Conclusions/interpretation

LLM vision APIs pose two distinct clinical risks: systematic overestimation bias and stochastic within-image variability, both invisible to end users. These findings support multi-query ensemble approaches, uncertainty communication and mandatory user confirmation before LLM-derived values are used for insulin dosing.

Keywords: artificial intelligence; automated insulin delivery; carbohydrate counting; computer vision; diabetes; food recognition; hypoglycaemia; insulin dosing; large language models; reproducibility; type 1 diabetes

---

Research in context

What is already known about this subject? - LLM vision APIs are being integrated into diabetes management applications, including open-source automated insulin delivery systems, to estimate carbohydrate content from food photographs - These systems are probabilistic by design, but the implications for clinical reproducibility of their outputs have not been systematically characterised with large sample sizes - Standard accuracy reporting (mean and standard deviation) does not capture the within-image variation that determines safety for single-query clinical use

What is the key question? - How reproducible are current LLM vision APIs when asked the same question about the same food photograph hundreds of times, and how does this reproducibility relate to accuracy when measured against reference values of known quality?

What are the new findings? - Within-image CV differs by more than four-fold across four leading models (median 2.4% to 11.0%), translating to insulin dosing uncertainties of 0.9 to 4.7 U at the median - All 52 model-image distributions are non-normal; standard summary statistics inadequately describe the safety profile - On the five images with the strongest reference values (packet label or weighed), the most reproducible model (Claude Sonnet 4.6) is also the most accurate (MAE 8.7 g), with 100% of its estimates within 20 g of reference – demonstrating that consistency and accuracy can co-vary when reference quality is controlled

How might this impact on clinical practice in the foreseeable future? - Diabetes management applications integrating LLM-based food analysis should treat reproducibility as a first-class safety criterion, prefer models with documented low within-image variation, present uncertainty ranges rather than point estimates, use ensemble or multi-query approaches, and require user confirmation before any LLM-derived value is used for insulin dosing

---

Abbreviations

AID, automated insulin delivery; CI, confidence interval; CV, coefficient of variation; ICR, insulin-to-carbohydrate ratio; IQR, interquartile range; LLM, large language model; MAE, mean absolute error; MAPE, mean absolute percentage error

---

Introduction

Carbohydrate counting is the primary nutritional intervention recommended for individuals with type 1 diabetes and insulin-treated type 2 diabetes [1]. For people using insulin, errors in carbohydrate estimation translate directly into insulin dosing errors with immediate glycaemic consequences. Published studies report human carbohydrate estimation errors of 15–60% [2,3], reaching 21 g in controlled settings [4], and these errors are a leading determinant of postprandial glucose variability [5]. Mobile applications designed to assist carbohydrate counting have shown variable accuracy, with a recent randomised controlled trial demonstrating benefits that vanished post-discontinuation and 44% mean error in canteen settings [6]. The cognitive burden of manual carbohydrate counting remains a significant barrier to optimal self-management [7].

The recent integration of large language model (LLM) vision APIs into diabetes management applications offers a new approach to photo-based carbohydrate estimation. The open-source iAPS automated insulin delivery (AID) system now offers food analysis through APIs from OpenAI, Anthropic and Google [8]. Dedicated applications and general-purpose LLMs are increasingly used by people with diabetes to estimate carbohydrate content from meal photographs [9,10]. Recent studies have evaluated the accuracy of these systems: Joubert et al reported a mean absolute error of 18.0 g for ChatGPT-5 across 246 hospital meals [9], while Goncalves et al found dangerous overestimations (exceeding 20 g) in up to 38% of queries for some models [10]. The Diabetes Technology Network UK has stated that generic LLMs must never be used as autonomous advisory calculators for insulin delivery [11].

Alongside these research-oriented evaluations, a growing number of commercial applications and subscription services now offer LLM-powered carbohydrate estimation directly to people with diabetes, often with polished interfaces and confident presentation of results that may generate a level of trust disproportionate to the underlying accuracy and reproducibility of the technology. The commercial framing of these tools as reliable meal-logging aids — rather than as probabilistic estimates with significant uncertainty — creates a specific safety concern when outputs are used to inform insulin dosing decisions.

Despite this proliferation, existing studies have focused on accuracy (how close to truth on average) rather than **reproducibility** (how much the same model disagrees with itself on the same input). LLM vision APIs are fundamentally probabilistic; even at near-zero temperature settings, repeated queries of the same photograph may yield different outputs. End users see only a single point estimate per query and have no visibility into the underlying distribution. This study addresses that gap, characterising within-image reproducibility as the primary outcome across 26,904 queries to four leading LLM vision APIs, with accuracy stratified by reference value quality as a secondary outcome.

Methods

Study design

This was a prospective benchmark study conducted in April 2026 comparing four commercially available LLM vision APIs. As the study involved no human participants and used only programmatic API services, ethics approval was not required. The primary outcome (reproducibility) and secondary outcome (accuracy stratified by reference quality) were pre-specified prior to data collection.

Test images

Thirteen food photographs were captured under typical real-world dining conditions (smartphone cameras, variable lighting, naturalistic plating) (Table 1). Selection prioritised diversity across meal type (single-item, composite, restaurant-prepared), carbohydrate density and food categories known to challenge human estimation.

Reference values for accuracy analysis

For nine of the thirteen images, the author estimated the carbohydrate content using methods described in Appendix 1. Reference quality was categorised into four tiers:

- **Tier 1 (packet label):** Carbohydrate values derived from manufacturer nutrition labelling. Two images (cheese sandwich, soup with bread) used bread with labelled carbohydrate content of 20 g per slice.
- **Tier 2 (weighed/measured):** Portions directly weighed and cross-referenced with established composition data. Three images (Bakewell tart, bakery cookie, breakfast burrito).
- **Tier 3 (portioned):** Portions estimated by the author (not weighed) and combined with USDA composition data. Three images (roast dinner, chilli con carne with rice, stuffed pork loin).
- **Tier 4 (visual estimate):** Portions and composition estimated from visual inspection. One image (churros).

For the four restaurant dishes (pizza capricciosa, eggs benedict, crema catalana, paella), no reference value was established. These images were used for the primary reproducibility analysis only.

Carbohydrate values follow the EU convention with dietary fibre excluded.

Table 1. Test images, reference carbohydrate values and reference quality tier

#	Meal	Reference (g)	Range (g)	Quality tier
1	Roast beef dinner with potatoes and broccoli	47.5	45-50	3 (portioned)
2	Bakewell tart slice	40.0	37-43	2 (weighed)
3	Tomato, chorizo, butterbean soup with bread	60.0	57-63	1 (packet label)
4	Cheese sandwich on thick white bread	40.0	38-42	1 (packet label)
5	Matcha churros with pistachio ice cream	80.0	70-90	4 (visual estimate)
6	Chilli con carne on white rice	58.0	54-62	3 (portioned)
7	Large bakery cookie	47.5	45-50	2 (weighed)
8	Stuffed pork loin with bread stuffing and fried courgette	40.0	35-45	3 (portioned)
9	Breakfast burrito	34.0	32-36	2 (measured)
10	Neapolitan pizza capricciosa	–	–	Reproducibility only
11	Eggs benedict on sourdough with sweet potato	–	–	Reproducibility only
12	Crema catalana	–	–	Reproducibility only
13	Paella, chicken and green beans (whole pan)	–	–	Reproducibility only

Models evaluated

Four production-grade LLM vision APIs were evaluated (Table 2): GPT-5.4 (OpenAI), Claude Sonnet 4.6 (Anthropic), Gemini 2.5 Pro (Google) and Gemini 3.1 Pro Preview (Google). These represent the most capable generally available vision models from each of the three major providers at the time of data collection (10–12 April 2026). Prior studies have evaluated earlier-generation models (e.g., ChatGPT-4o [10], GPT-5 [9]); the models tested here represent a subsequent generation with improved multimodal capabilities. All data were collected via the respective providers' batch processing APIs.

Table 2. Models, sample sizes and data collection

Model	Provider	API	Queries (n)	Per image (n)
GPT-5.4	OpenAI	Batch API	7,279	559-561
Claude Sonnet 4.6	Anthropic	Message Batches API	6,630	510
Gemini 2.5 Pro	Google	Batch API	6,500	500
Gemini 3.1 Pro Preview	Google	Batch API	6,495	~500
Total			**26,904**	

Sample sizes differ across models because each provider's batch API has distinct operational constraints. OpenAI enforces a per-organisation enqueued-token limit that necessitated sequential sub-batch submission across multiple API keys, yielding the largest dataset. Anthropic and Google batch APIs allowed single-submission processing but with different payload size limits, resulting in slightly different final counts. All models exceed 500 independent trials per image, well above the threshold required for stable CV and distributional estimates.

Prompt design and execution

An identical structured prompt was used across all four models, adapted verbatim from the production prompt in the iAPS open-source AID application [8]. The prompt instructs the model to identify each food component, estimate per-100 g nutrition values according to USDA/FDA standards, estimate visible portion size in grams and return results in a structured JSON schema. The complete prompt text and its SHA-256 hash are provided as Supplementary Material S1.

All queries used temperature 0.01 (the lowest reliably available across all four providers). Images were resized to each provider's documented maximum dimension (1,568 pixels for Claude, 2,048 pixels for GPT-5.4 and Gemini models) and compressed to JPEG at quality 85. Each query was executed as an independent stateless API call with no conversation history; a fresh HTTP client was instantiated per request to ensure complete isolation between queries.

All four models were accessed via their respective providers' batch processing APIs. The batch APIs are functionally equivalent to the real-time APIs in terms of model behaviour: requests are processed by the same underlying models with the same parameters, temperature settings and prompt. The only differences are operational (asynchronous processing, separate quota pools, 50% cost discount). There is no documented difference in model output distributions. The choice of batch mode was made to enable the data volumes required for distributional characterisation.

Outcomes

Primary outcome (reproducibility): within-image variation in total portion carbohydrate estimates, characterised by: - Range (maximum minus minimum) across repeated queries of the same image - Interquartile range (IQR) - Coefficient of variation (CV) - P5-P95 range - Distributional normality (Shapiro-Wilk test)

Secondary outcome (accuracy): for the nine images with reference values, stratified by reference quality tier: - Mean absolute error (MAE) in grams with 95% confidence intervals - Mean absolute percentage error (MAPE) - Systematic bias (signed mean error) - Proportion of estimates within 10 g and 20 g of reference

Clinical risk translation: insulin dosing and glycaemic impact at a typical ICR of 1:10 [12] and insulin sensitivity factor (ISF) of 2.0 mmol/L per unit: - Mean insulin dosing error per model (systematic bias risk) - Proportion of individual queries that would cause insulin overdose > 2 U (clinically significant; predicted glucose drop > 4 mmol/L) and > 5 U (severe hypoglycaemia risk; predicted glucose drop > 10 mmol/L) - Worst-case single-query insulin overdose per model-image combination - Within-image insulin dosing variability (same photo, different query)

Statistical analysis

Descriptive statistics included mean, median, SD, IQR, range, skewness and excess kurtosis. CV was calculated per image per model. Normality was assessed using the Shapiro-Wilk test (with random subsampling to n=5,000 where necessary). Between-model comparisons of within-image CV used the Wilcoxon signed-rank test (paired across 13 images). Between-model comparisons of absolute error used Welch's t test (unpaired) with Cohen's d effect size. The Mann-Whitney U test was used as a non-parametric confirmatory analysis for all pairwise comparisons. Analyses were conducted in Python 3.12 using NumPy 2.2 and SciPy 1.14. All analysis code is available in the study repository.

Reproducibility statement

The complete dataset (26,904 query results), the verbatim prompt, all analysis scripts and the model identifiers are available at https://github.com/tim2000s/llm-food-benchmark-academic (private; access on request).

Results

Dataset

All 26,904 queries across four models returned successfully parsed food item data (100% success rate for all models). Token usage data was available for Claude Sonnet 4.6 (mean input 3,371 +/- 15 tokens, output 1,499 +/- 604 tokens) and GPT-5.4 (mean input 4,502 +/- 83, output 819 +/- 353); Gemini token usage was not captured.

Primary outcome: within-image variation

Within-image variation in total portion carbohydrate estimates differed substantially across the four models (Table 3, Fig. 1). Median CV was 2.4% for Claude Sonnet 4.6, 8.4% for GPT-5.4, 10.3% for Gemini 3.1 Pro Preview and 11.0% for Gemini 2.5 Pro. The Wilcoxon signed-rank test confirmed that Claude's within-image CV was significantly lower than both Gemini models (Claude vs Gemini 2.5 Pro: W=1, p=0.0005; Claude vs Gemini 3.1 Pro: W=4, p=0.0017). The difference between Claude and GPT-5.4 was not statistically significant at the image level (W=24,

p=0.15), reflecting a small number of images where GPT-5.4 achieves comparable or lower CV (e.g., crema catalana: GPT-5.4 CV 2.8% vs Claude 2.8%; chilli con carne: GPT-5.4 CV 0.5% vs Claude 2.1%).

Table 3. Within-image reproducibility metrics by model (primary outcome)

Metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
Queries (n)	6,630	7,279	6,495	6,500
Per image (n)	510	559-561	~500	500
CV – median	**2.4%**	8.4%	10.3%	11.0%
CV – mean	**5.5%**	9.3%	12.9%	17.0%
CV – max	29.2%	**19.2%**	30.0%	68.3%
Within-image range (g) – median	**8.6**	22.8	28.8	46.5
Within-image range (g) – max	135.6	165.6	162.0	428.7
IQR (g) – median	**1.7**	7.7	9.8	10.9
P5-P95 range (g) – median	**3.7**	13.9	23.6	20.7
Insulin range (U) – median	**0.9**	2.3	2.9	4.7
Insulin range (U) – max	13.6	16.6	16.2	42.9
Images with range > 2 U insulin	3/13	7/13	10/13	11/13
Images with range > 5 U insulin	1/13	3/13	4/13	6/13
Non-normal distributions (Shapiro-Wilk $p<0.05$)	13/13	13/13	13/13	13/13

Bold values indicate the best (most reproducible) performance for each metric.

The distribution of within-image CVs across the 13 images is shown in Fig. 1. Claude's CVs are tightly clustered below 5% for most images, with a single outlier (the paella, 29.2%). The Gemini models show wider spread with medians near or above 10%.

The clinical significance of these differences is illustrated by insulin dosing translation at an ICR of 1:10. Claude's median within-image insulin dosing uncertainty was less than 1 U (0.9 U), meaning that for most images, repeating the same query would change the recommended insulin dose by less than 1 unit. GPT-5.4's median was 2.3 U, approaching the threshold for clinical significance. Both Gemini models exceeded this threshold at the median (2.9 U and 4.7 U respectively), with Gemini 2.5 Pro producing insulin dosing variability exceeding 5 U for 6 of 13 images – a magnitude associated with symptomatic hypoglycaemia risk [13,14].

Distributional characteristics

All 52 model-image distributions were non-normal (Shapiro-Wilk $p<0.05$). Three distinct distributional patterns emerged:

Point-mass distributions characterised many Claude outputs. The cheese sandwich had CV=0.3% with near-zero range (1.3 g), and the bakery cookie had CV=0.8%. While this yields the highest reproducibility, the model provides no uncertainty signal when its confident estimate diverges from the true value.

Moderate-variance distributions characterised most GPT-5.4 outputs, with occasional near-deterministic behaviour (chilli con carne: CV=0.5%, range 2.6 g; crema catalana: CV=2.8%, range 2.6 g).

Heavy-tailed distributions were prominent in Gemini 2.5 Pro outputs. The stuffed pork loin image had CV=68.3% with a 119 g range across 500 queries. The paella spanned 429 g (55-484 g equivalent).

The paella image illustrates the distributional differences most clearly: Claude produced a distribution with most values between 80 and 150 g, while Gemini 2.5 Pro produced an extremely heavy-tailed distribution spanning 429 g for the same photograph. All four models showed CV > 14% on this image, suggesting an inherently ambiguous visual stimulus.

Secondary outcome: accuracy

Accuracy was analysed across all nine reference images and stratified by reference quality tier (Table 4).

Table 4. Accuracy metrics stratified by reference quality (secondary outcome)

Panel A: Strong reference (tier 1-2: packet label + weighed/measured, 5 images)

Metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
Queries (n)	2,550	2,797	2,500	2,500
MAE, g (95% CI)	**8.7 (8.6-8.9)**	17.4 (16.8-17.9)	12.5 (12.2-12.8)	16.6 (15.8-17.4)
MAPE, %	**20.1**	45.0	31.1	45.1
Mean bias, g	**-1.2**	+11.7	+1.8	+9.6
Within 10 g, %	43.0	**61.2**	35.0	52.2
Within 20 g, %	**100.0**	63.2	87.8	80.8

Panel B: All nine reference images

Metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
Queries (n)	4,590	5,037	4,500	4,500
MAE, g (95% CI)	**11.6 (11.4-11.8)**	20.2 (19.8-20.6)	13.8 (13.5-14.0)	17.7 (17.2-18.2)

Metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
MAPE, %	**22.7**	45.8	29.6	41.4
Mean bias, g	**+1.9**	+12.9	+4.8	+9.2
Within 10 g, %	**40.0**	44.7	38.0	42.2
Within 20 g, %	**86.4**	54.8	77.4	68.8

Panel C: Weaker reference (tier 3-4: portioned + visual estimate, 4 images)

Metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
Queries (n)	2,040	2,240	2,000	2,000
MAE, g (95% CI)	**15.1 (14.8-15.5)**	23.8 (23.2-24.3)	15.3 (14.8-15.8)	19.1 (18.6-19.6)
MAPE, %	**26.0**	46.8	27.6	36.8
Mean bias, g	**+5.8**	+14.4	+8.4	+8.8

Bold values indicate the best performance for each metric within each panel.

The stratification by reference quality revealed that Claude Sonnet 4.6 achieved the lowest MAE across all reference tiers: 8.7 g on strong-reference data (100% within 20 g), 15.1 g on weaker-reference data, and 11.6 g overall. This represents the best accuracy of any model tested on every subset. On strong-reference data, Claude was the only model to achieve 100% of estimates within 20 g of reference.

Claude also showed near-zero mean bias on strong-reference data (-1.2 g), while all other models showed substantial positive bias (+1.8 to +11.7 g). On weaker-reference data, all models including Claude showed positive bias, consistent with systematic overestimation of complex or restaurant-prepared meals.

All pairwise accuracy comparisons on strong-reference data were statistically significant (Welch's t test, all $p<10^{-20}$; Cohen's d 0.26-0.76), with the exception of Gemini 2.5 Pro vs GPT-5.4 ($p=0.13$, $d=-0.04$), which did not differ.

Clinical risk translation

Table 5 translates the observed carbohydrate estimation errors into insulin dosing errors and predicted glycaemic impact, assuming a typical ICR of 1:10 and ISF of 2.0 mmol/L per unit of rapid-acting insulin [6,7]. These parameters represent a typical adult with type 1 diabetes; individuals with lower insulin sensitivity (higher ISF) would experience proportionally smaller glucose excursions, and those with higher sensitivity proportionally larger ones.

Table 5. Clinical risk by model (strong-reference images, tier 1-2, n=5 images)

Risk metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
Queries (n)	2,550	2,797	2,500	2,500

Risk metric	Claude Sonnet 4.6	GPT-5.4	Gemini 3.1 Pro	Gemini 2.5 Pro
Mean insulin error (U)	**-0.1**	+1.2	+0.2	+1.0
Queries causing > 2 U overdose (%)	**0.0**	36.8	12.2	19.2
Queries causing > 5 U overdose (%)	**0.0**	0.0	0.0	11.6
Queries causing > 4 mmol/L glucose drop (%)	**0.0**	36.8	12.2	19.2
Queries causing > 10 mmol/L glucose drop (%)	**0.0**	0.0	0.0	11.6
Worst single-query overdose (U)	**1.2**	4.5	5.0	11.3
Worst single-query underdose (U)	**1.5**	1.3	2.0	1.6

Two distinct types of clinical risk emerged from these data:

Systematic bias risk. All four models showed positive mean bias (overestimation) across the full nine-image dataset (mean bias +1.9 to +12.9 g), meaning the dominant direction of error is toward insulin overdosing and hypoglycaemia. On strong-reference data, Claude had near-zero mean bias (-0.1 U), while GPT-5.4 showed the largest systematic overdose tendency (+1.2 U per query on average). A consistent +1.2 U bias on every meal throughout the day (e.g., three meals) would produce a cumulative overdose of approximately 3.6 U per day – sufficient to cause recurrent mild hypoglycaemia in many patients.

Variability risk. Even for a single photograph queried repeatedly, the insulin dose recommendation varied substantially for three of four models. At the extremes, a person using Gemini 2.5 Pro to photograph a breakfast burrito (reference 34 g, tier 2) could receive estimates ranging from 40 g to 147 g across repeated queries –

translating to insulin doses of 4.0 U to 14.7 U for the same meal. For Claude, the same photograph would yield estimates between 36 g and 44 g (3.6-4.4 U), a clinically manageable range.

The interaction between systematic bias and variability compounds the risk. GPT-5.4 on the cheese sandwich (tier 1 reference, 40 g) illustrates this: its mean estimate of 73.9 g (7.4 U) already represents a 3.4 U overdose compared with the true value (4.0 U), and its within-image variability (range 55.7 g, CV 19.2%) means individual queries could recommend anywhere from 2.9 U to 8.5 U – ranging from near-correct to dangerous overdose – from the same photograph.

The invisible nature of this risk warrants emphasis. A diabetes app user photographs their meal once and receives a single number. They have no way to know whether they have received a typical estimate, a tail-end outlier, or a value from an entirely different mode of a bimodal distribution. For Claude, the single-query experience is largely safe on strong-reference foods (worst case 1.2 U overdose). For Gemini 2.5 Pro, a single unlucky query could cause an 11.3 U overdose – a potentially life-threatening event.

Per-image case studies

Four images illustrate the consistency-accuracy relationship and the impact of food misidentification (Fig. 2):

Cheese sandwich (tier 1 reference, 40 g). This image has the strongest reference in the dataset: two slices of thick-cut white bread with packet-labelled carbohydrate content of 20 g per slice, plus cheddar cheese contributing negligible carbohydrate. Claude estimated 28.2 g with CV=0.3% (range 1.3 g) – highly consistent but systematically 12 g below reference. Three of four models converged on approximately 28 g (Claude 28.2, Gemini 2.5 Pro 27.9, Gemini 3.1 Pro 28.3), suggesting a shared systematic error possibly related to portion size estimation of the bread. GPT-5.4 estimated 73.9 g with CV=19.2% – both inaccurate and inconsistent.

Bakery cookie (tier 2 reference, 47.5 g). Claude estimated 49.4 g with CV=0.8% (MAE 1.9 g) – the best single-image performance in the dataset, virtually matching the weighed reference. Gemini 2.5 Pro estimated 47.9 g (MAE 4.8 g) with greater variability (CV 11.0%). GPT-5.4 and Gemini 3.1 Pro slightly underestimated (44.1 g and 44.0 g respectively).

Breakfast burrito (tier 2 reference, 34 g). Claude estimated 40.1 g with CV=6.6% (MAE 6.1 g), the closest to the user-measured reference. GPT-5.4 estimated 67.9 g with CV=6.3% – equally consistent but nearly double the reference value. Gemini 2.5 Pro estimated 86.8 g with CV=23.3%, the worst performance on this image.

Stuffed pork loin (tier 3 reference, 40 g). Claude estimated 44.8 g with CV=2.8% (MAE 4.8 g) – the closest to reference and with tight consistency, correctly identifying the food as meat with bread stuffing. GPT-5.4 estimated 77.4 g (MAE 37.5 g) with CV=19.0%, nearly doubling the reference. Gemini 2.5 Pro showed the highest variability on this image (CV=68.3%, range 119 g), while Gemini 3.1 Pro underestimated (27.3 g, MAE 12.7 g).

Food item identification and misidentification

Models differed in the number of food items identified per query. Claude identified the most items on average (mean 3.08 per query) compared with GPT-5.4 (2.41), Gemini 3.1 Pro (2.33) and Gemini 2.5 Pro (2.18). All models had a median of 2 items per query.

Analysis of the food item names returned across all 26,904 queries revealed that food misidentification – not just portion estimation error – was a significant contributor to carbohydrate estimation failures. Three patterns emerged:

Correct identification with variable specificity. The stuffed pork loin with bread stuffing (Image 8) was identified by Claude as "bread stuffing + grilled pork ribs" – substantively correct, with a corresponding MAE of only 4.8 g. However, GPT-5.4 and Gemini models used varying descriptions ("roasted chicken with skin + stuffing"; "glazed chicken + bread stuffing") that, while identifying the stuffing component correctly, misidentified the meat as chicken rather than pork. This meat misidentification had variable impact on carbohydrate estimates: the carbohydrate content is dominated by the bread stuffing regardless of meat type, but GPT-5.4's tendency to add spurious components (e.g., "cream sauce", "casserole") inflated its estimates to a mean of 77.4 g (MAE 37.5 g).

Related-item substitution. The Bakewell tart (Image 2) was consistently identified by Claude as a "Linzer torte" (100% of 510 queries) – a visually similar but compositionally different pastry. GPT-5.4 described it as a "jam-filled cake bar" or "jam tart slice." Only Gemini 3.1 Pro correctly identified it as a Bakewell tart (99.8%). The crema catalana (Image 12) was identified as "creme brulee" by three of four models in 100% of queries; only Gemini 3.1 Pro returned "crema catalana" (3.4% of queries). These substitutions had modest nutritional impact due to similar compositions, but illustrate systematic visual-semantic confusion.

Spurious item addition. For the pizza image (Image 10), Gemini 2.5 Pro and Gemini 3.1 Pro consistently identified a "burrata salad" from what appears to be an adjacent plate partially visible in the frame, adding spurious carbohydrate estimates from a dish that was not the intended subject. Gemini 3.1 Pro added non-existent "deli meat" to the cheese sandwich in 17.4% of queries, directly inflating carbohydrate estimates for a meal where cheese is the only filling.

Self-reported confidence calibration

The structured prompt required each model to return a confidence score (0-1) for each identified food item. All four models returned confidence values for 100% of food items. However, self-reported confidence was poorly calibrated against actual accuracy (Table 6).

Table 6. Self-reported confidence vs actual accuracy

Model	Mean confidence	% items conf > 0.9	Correlation with absolute error (r)	MAE when conf >= 0.85	MAE when conf < 0.85
	0.80	18.4%	**-0.01**	17.3 g	9.1 g

Model	Mean confidence	% items conf > 0.9	Correlation with absolute error (r)	MAE when conf >= 0.85	MAE when conf < 0.85
Claude Sonnet 4.6					
GPT-5.4	0.78	35.7%	-0.17	30.6 g	17.9 g
Gemini 3.1 Pro	0.91	75.5%	-0.11	13.8 g	12.8 g
Gemini 2.5 Pro	0.91	85.7%	-0.14	17.4 g	19.4 g

Claude's self-reported confidence had essentially no correlation with actual carbohydrate estimation accuracy (r = -0.01). Both Gemini models reported confidence above 0.9 for the majority of food items regardless of accuracy, consistent with systematic overconfidence. For Claude and GPT-5.4, high-confidence queries were paradoxically *less* accurate than low-confidence queries, suggesting an inverse calibration pattern.

At the per-image level, Claude's mean confidence showed weak variation by image difficulty (range 0.65 for the churros — its most frequently misidentified item — to 0.92 for the crema catalana), but the range was too narrow and poorly calibrated to serve as a clinically actionable uncertainty signal.

Discussion

Reproducibility as the primary safety question

This study demonstrates that within-image variation in LLM-derived carbohydrate estimates is substantial, model-dependent and invisible to end users. Across 26,904 queries to four leading LLM vision APIs, within-image CV ranged from a median of 2.4% (Claude Sonnet 4.6) to 11.0% (Gemini 2.5 Pro). Translated to insulin dosing at an ICR of 1:10, the most reproducible model (Claude) produced a median within-image insulin uncertainty of less than 1 unit, while the least reproducible (Gemini 2.5 Pro) produced a median of 4.7 units with a maximum of 42.9 units for a single photograph.

The non-normality of all 52 model-image distributions is a methodologically important finding. Standard reporting of mean +/- SD assumes approximate normality and understates the risk profile. Point-mass distributions (e.g., Claude's cheese sandwich: CV=0.3%), heavy-tailed distributions (e.g., Gemini 2.5 Pro's stuffed pork loin: CV=68.3%) and moderate-variance distributions coexist within and across models. Applications reporting a single "best estimate" without uncertainty quantification mask the actual risk.

Consistency and accuracy are not independent

The stratification of accuracy by reference quality resolved an apparent paradox in the unstratified data. Claude Sonnet 4.6 achieved the lowest MAE across all reference quality tiers: 8.7 g on strong-reference data, 15.1 g on weaker-reference data, and 11.6 g overall. On strong-reference data, Claude was the least biased (mean bias -1.2 g) and the only model to achieve 100% of estimates within 20 g of reference. The effect sizes for Claude's accuracy advantage on strong-reference data were large (Cohen's d 0.53-0.76 vs all other models).

This finding has important methodological implications. Studies evaluating LLM food analysis accuracy should explicitly report the quality of their reference values and stratify accordingly. Aggregate accuracy metrics that mix strong and weak references risk obscuring real performance differences between models.

The cheese sandwich: a case study in systematic error

The cheese sandwich deserves particular attention because it pairs the strongest reference in the dataset (packet-label bread, negligible cheese carbohydrate) with the most striking inter-model pattern. Three of four models – Claude, Gemini 2.5 Pro and Gemini 3.1 Pro – independently converged on approximately 28 g total carbohydrate for a meal with 40 g of carbohydrate from bread alone. This systematic 12 g underestimate, replicated across three independent model architectures with high consistency, suggests a shared failure mode likely related to portion size estimation of the bread slices in the photograph. Only GPT-5.4 diverged, estimating 74 g (34 g over reference) with high variability.

This case illustrates that high reproducibility does not guarantee accuracy. A diabetes app user receiving Claude's estimate of 28 g for a 40 g meal would consistently underdose insulin by approximately 1.2 units – a clinically significant but not dangerous error in the direction of hyperglycaemia rather than hypoglycaemia.

Food identification accuracy as a distinct dimension

Analysis of the food item names returned across all 26,904 queries revealed that food identification accuracy varies independently of carbohydrate estimation accuracy. Three patterns emerged. First, correct identification can coexist with inaccurate carbohydrate estimation: GPT-5.4 correctly identified the bread stuffing in the stuffed pork loin but estimated 77.4 g (vs 40 g reference) due to spurious additional components. Second, related-item substitution (e.g., Claude identifying the Bakewell tart as a Linzer torte in 100% of queries) can have modest nutritional impact when the substituted item has similar composition, but illustrates systematic visual-semantic confusion. Third, spurious item addition – such as Gemini models identifying a burrata salad from an adjacent plate in the pizza image – introduces carbohydrate estimates from foods not in the subject's meal. Goncalves et al similarly reported that food misidentification drove the largest errors in their evaluation [10].

These findings have implications for clinical deployment: applications should implement food-identification confirmation (e.g., displaying the identified items for user review) before proceeding to carbohydrate estimation. The current paradigm of silent, single-pass estimation provides no opportunity for users to catch identification errors before they propagate into insulin dosing decisions.

Clinical risk: two failure modes

The clinical risk analysis reveals two distinct failure modes that have different implications for patient safety:

Failure mode 1: systematic bias (chronic risk). All four models overestimate carbohydrate content on average, producing a systematic tendency toward insulin overdosing. On strong-reference data, GPT-5.4 averaged +1.2 U insulin error per meal. For a patient eating three LLM-assisted meals per day, this translates to a cumulative overdose of approximately 3.6 U per day – equivalent to lowering mean blood glucose by approximately 7 mmol/L over 24 hours, a magnitude that would cause recurrent hypoglycaemia in most patients. This systematic bias is in principle correctable through calibration or bias-adjustment in applications, but no current diabetes management app implements such correction.

Failure mode 2: stochastic variability (acute risk). The within-image variation creates the possibility of a single catastrophic outlier estimate. This is the more dangerous failure mode because it is unpredictable and invisible to the user. On strong-reference data, Gemini 2.5 Pro produced individual queries that would cause an 11.3 U insulin overdose for a 34 g meal – an error that could cause severe hypoglycaemia requiring third-party assistance [7,8]. Unlike systematic bias, variability cannot be calibrated away in a single-query application; it can only be mitigated by multi-query ensemble approaches or by selecting models with inherently low variability.

The models tested span the full spectrum of these failure modes. Claude Sonnet 4.6 exhibited low variability and near-zero systematic bias on strong-reference data, with 0% of queries causing > 2 U overdose. At the other extreme, GPT-5.4 combined high systematic bias (36.8% of queries causing > 2 U overdose on strong-reference data) with moderate variability, meaning the clinical risk is both chronic and unpredictable.

Critically, these risks are entirely invisible to the end user. A person with type 1 diabetes who photographs their cheese sandwich and receives “74 g carbohydrate” from GPT-5.4 has no indication that the true value is 40 g, that three other models would have said 28 g, or that repeating the same query might return a substantially different number. The single-query, single-model paradigm used by all current diabetes management apps provides no safety net against either failure mode.

Self-reported confidence is not a safety mechanism

A potential mitigation for the risks identified above would be to use the models’ self-reported confidence scores to flag uncertain estimates. However, the data demonstrate that confidence scores are poorly calibrated across all four models. Claude’s confidence had near-zero correlation with actual accuracy ($r = -0.01$), and both Gemini models reported confidence above 0.9 for the vast majority of items

regardless of whether the estimate was accurate. For Claude and GPT-5.4, high-confidence queries were paradoxically less accurate than low-confidence ones, meaning that filtering on confidence would preferentially retain the *worst* estimates.

This finding has a specific practical implication: diabetes applications cannot use LLM-reported confidence as a gatekeeper for clinical safety. The only reliable uncertainty signal in the current generation of models comes from observing the empirical spread across multiple independent queries of the same image — not from the model's self-assessment.

Implications for application design

1. **Multi-query ensemble approaches should be the default.** Querying the same image 3-5 times and presenting the median with IQR would mitigate stochastic variability risk at modest additional cost (~$0.01-0.05 per ensemble at current API pricing). This would not address systematic bias but would eliminate tail-end outlier estimates.

2. **Model selection matters more than prompt engineering.** The four-fold difference in median CV between the most and least reproducible models (2.4% vs 11.0%) dwarfs any plausible improvement from prompt optimisation. Application developers should prioritise model selection based on documented reproducibility.

3. **Bias correction is feasible but not currently implemented.** The consistent positive bias across all models suggests that a simple multiplicative or additive correction factor, calibrated against known-reference meals, could reduce systematic error. This is an engineering problem, not a fundamental limitation.

4. **User confirmation is non-negotiable.** No model tested, including the best-performing Claude Sonnet 4.6, was sufficiently accurate and reproducible to support autonomous insulin dosing without human oversight. This is consistent with the DTN-UK position that generic LLMs must never be used as autonomous advisory calculators for insulin delivery [11]. The cheese sandwich case (three models converging on 28 g for a 40 g meal) demonstrates that even high inter-model agreement does not guarantee correctness.

5. **Standard summary statistics are inadequate for safety assessment.** Reporting mean +/- SD assumes normality. All 52 model-image distributions were non-normal. Applications should present ranges or percentiles, not point estimates, and safety evaluations should report tail-risk metrics (e.g., proportion of queries causing > 2 U or > 5 U insulin error) rather than aggregate accuracy alone.

Comparison with human performance

Published human carbohydrate counting performance in adults with type 1 diabetes shows MAE of approximately 15.4 g [3], reaching 21 g in controlled settings [4], with percentage errors of 20-30% [2]. Recent LLM accuracy studies report MAE of 18.0 g for ChatGPT-5 across 246 hospital meals [9] and dangerous overestimations exceeding 20 g in up to 38% of queries [10]. On strong-reference data, Claude (MAE 8.7 g, MAPE 20.1%) and Gemini 3.1 Pro (12.5 g, 31.1%) are better than published human performance. On all nine images, Claude (MAE 11.6 g)

outperforms the human baseline, while GPT-5.4 (20.2 g) and Gemini 2.5 Pro (17.7 g) are comparable to or slightly below human performance. However, human counters do not produce the non-normal within-image distributions characteristic of probabilistic AI systems. The reproducibility issue has no direct human analogue.

Limitations

Several limitations should be noted. First, the nine reference carbohydrate values vary in quality from packet labels (tier 1) to visual estimates (tier 4). The stratified analysis addresses this explicitly, but accuracy conclusions for the weaker-reference images should be interpreted with appropriate caution. Second, sample sizes differ across models (500-560 per image) due to different batch API operational constraints; all exceed the minimum for stable distributional estimates but direct between-model comparison of tail statistics should account for this. Third, the 13-image test set is geographically biased toward UK/European cuisine and may not generalise. Fourth, all queries used temperature 0.01, representing the best case for reproducibility; production deployments with higher temperatures would experience greater variation. Fifth, a single prompt was tested, though it is a production prompt deployed in a diabetes management application. Sixth, the underlying models will be updated and reproducibility characteristics may change; the complete dataset and analysis code are publicly available to enable longitudinal re-testing.

Recommendations

Based on these findings, we recommend that diabetes management applications integrating LLM-based food analysis:

1. **Treat reproducibility as a first-class safety criterion** alongside accuracy when selecting models for clinical integration.
2. **No autonomous insulin dosing.** LLM-derived carbohydrate values should not drive insulin doses without explicit user confirmation.
3. **Display ranges, not point estimates.** Applications should query the model multiple times per image and present the median and interquartile range.
4. **Prefer models with documented low within-image variation** for any single-query operation.
5. **Report reference value quality** in evaluation studies, and stratify accuracy analyses accordingly.
6. **Exercise particular caution with composite and restaurant meals**, where intra-image variation is largest across all models.

Conclusions

LLM vision APIs pose two distinct clinical risks when used for insulin dosing from food photographs: systematic overestimation bias that could cause recurrent hypoglycaemia (mean overdose +0.2 to +1.2 U per meal on strong-reference data for the three overestimating models), and stochastic within-image variability that could cause acute severe hypoglycaemia from a single outlier estimate (worst-case single-query overdose up to 11.3 U). These risks are entirely invisible to end users who receive only a single point estimate per query. Within-image CV ranged from 2.4% (Claude Sonnet 4.6) to 11.0% (Gemini 2.5 Pro), and all 52 model-image distributions were non-normal. On strong-reference data, 0% of Claude's queries would have caused a clinically significant insulin overdose (> 2 U), compared with

12-37% for the other three models. On all nine images, Claude achieved the lowest MAE (11.6 g, 86.4% within 20 g). Diabetes management applications integrating LLM food analysis should implement multi-query ensemble approaches as the default, require user confirmation before any LLM-derived value is used for insulin dosing, and treat reproducibility as a first-class safety criterion alongside accuracy.

Data availability

The complete dataset (26,904 query results), the verbatim prompt, all analysis scripts and the model identifiers used are available at https://github.com/tim2000s/llm-food-benchmark-academic (private; access on request).

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. API costs were self-funded by the author.

Authors' relationships and activities

The author has used open-source automated insulin delivery systems since 2016, has contributed code to AID systems in the oref algorithm family, and maintains an active fork of AndroidAPS. The author has no affiliation with the iAPS project. The benchmark prompt used in this work was adapted from the iAPS open-source codebase as a representative example of a production LLM food-analysis prompt currently deployed in a diabetes management application. The author declares no other relationships or activities that could appear to have influenced the submitted work.

Contribution statement

TS conceived the study, designed the methodology, developed the benchmark infrastructure and analysis software, conducted all experiments, performed the statistical analysis, and wrote the manuscript. TS is the guarantor of this work and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Acknowledgements

The author acknowledges the use of Claude Code (Anthropic) for assistance with development of the benchmark testing infrastructure, batch API management tooling, statistical analysis pipeline and manuscript preparation. All scientific analysis, interpretation of results and conclusions are the sole responsibility of the author.

Supplementary Material

S1. Verbatim prompt text used for all model queries, adapted from the iAPS open-source automated insulin delivery system, with SHA-256 hash for reproducibility verification.

S2. Per-image distributional statistics for all 52 model-image combinations.

S3. Full per-image accuracy breakdown including reference quality tier, mean estimate, SD, CV, MAE and reference range for each model.

S4. Food item identification analysis: most frequent food item names returned by each model for each image, with misidentification rates.

Appendix 1. Reference carbohydrate value methodology

#	Meal	Portion determination	Composition source	Reference (g)	Quality tier
1	Roast beef dinner	Author estimate: 5 medium-small potatoes (~200 g), 80 g broccoli, 120 g beef	USDA FoodData Central	47.5	3
2	Bakewell tart slice	Weighed slice (~90 g)	Commercial composition (~47 g/100 g)	40.0	2
3	Soup with bread	2 slices bread (packet: 20 g/ slice), soup (home-made, measured: 20 g)	Manufacturer label + author measurement	60.0	1
4	Cheese sandwich	2 slices bread (packet: 20 g/ slice), cheddar (~40 g, negligible carbs)	Manufacturer label	40.0	1
5	Matcha churros + ice cream	Visual estimate: ~150 g churros, ~70 g ice cream	Typical commercial composition	80.0	4
6	Chilli con	Portioned: 350 g chilli, 120 g		58.0	3

#	Meal	Portion determination	Composition source	Reference (g)	Quality tier
	carne + rice	rice (not weighed)	USDA FoodData Central		
7	Bakery cookie	Weighed cookie (~85 g)	Commercial composition (~58 g/100 g)	47.5	2
8	Stuffed pork loin + stuffing + courgette	Portioned: ~200 g stuffed pork, ~100 g extra stuffing, ~80 g courgette	USDA FoodData Central	40.0	3
9	Breakfast burrito	Direct ingredient summation from individual component weights and labels	Ingredient labels	34.0	2

Quality tier definitions: - **Tier 1 (packet label):** Primary carbohydrate source has manufacturer-labelled nutrition data. Highest confidence. - **Tier 2 (weighed/ measured):** Portion directly weighed or ingredients individually measured. High confidence. - **Tier 3 (portioned):** Portions estimated by the author (not weighed), combined with USDA composition data. Moderate confidence; uncertainty dominated by portion size estimation. - **Tier 4 (visual estimate):** Restaurant dish or meal where neither portion nor composition could be directly measured. Lowest confidence; accuracy results for these images should be interpreted accordingly.

Carbohydrate values follow the EU convention (dietary fibre excluded). All composition values were cross-referenced against USDA FoodData Central, with secondary verification against UK Composition of Foods Integrated Dataset (CoFID) for region-specific items.

Figure legends

Fig. 1 Within-image coefficient of variation (CV) for total portion carbohydrate estimates across the four models. Each point represents one of the 13 test images (500+ independent queries per point). Boxes show the median and interquartile range; whiskers extend to 1.5 x IQR. The dashed line indicates 10% CV. Claude Sonnet 4.6 shows markedly lower and more tightly clustered CVs than the other three models (median 2.4% vs 8.4-11.0%).

Fig. 2 Distribution of total portion carbohydrate estimates for four case-study images, selected to illustrate key findings. Each box plot summarises 500+ independent queries per model. Boxes show the median and interquartile range;

whiskers extend to 1.5 x IQR; outliers shown as individual points. Red dashed lines indicate reference values with quality tier. (a) Cheese sandwich (tier 1: packet label, 40 g): three models converge on ~28 g (consistently wrong); GPT-5.4 estimates ~74 g with high variability. (b) Bakery cookie (tier 2: weighed, 47.5 g): Claude near-perfectly matches reference with minimal spread. (c) Breakfast burrito (tier 2: measured, 34 g): Claude closest to reference; GPT-5.4 and Gemini 2.5 Pro substantially overestimate. (d) Stuffed pork loin (tier 3: portioned, 40 g): Claude correctly identifies the food and estimates accurately (mean 44.8 g, MAE 4.8 g); GPT-5.4 substantially overestimates (mean 77.4 g); Gemini 2.5 Pro shows extreme variability (CV 68.3%).

References

[1] Evert AB, Dennison M, Gardner CD et al (2019) Nutrition therapy for adults with diabetes or prediabetes: a consensus report. Diabetes Care 42(5):731-754. https://doi.org/10.2337/dci19-0014

[2] Bell KJ, Barclay AW, Petocz P, Colagiuri S, Brand-Miller JC (2014) Efficacy of carbohydrate counting in type 1 diabetes: a systematic review and meta-analysis. Lancet Diabetes Endocrinol 2(2):133-140. https://doi.org/10.1016/S2213-8587(13)70160-0

[3] Brazeau AS, Mircescu H, Desjardins K et al (2013) Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes. Diabetes Res Clin Pract 99(1):19-23. https://doi.org/10.1016/j.diabres.2012.10.024

[4] Baumgartner M, Kuhn C, Nakas CT, Herzig D, Bally L (2024) Carbohydrate estimation accuracy of two commercially available smartphone applications vs estimation by individuals with type 1 diabetes. J Diabetes Sci Technol. https://doi.org/10.1177/19322968241264744

[5] Smart CE, Ross K, Edge JA, King BR, McElduff P, Collins CE (2010) Can children with type 1 diabetes and their caregivers estimate the carbohydrate content of meals and snacks? Diabet Med 27(3):348-353. https://doi.org/10.1111/j.1464-5491.2010.02945.x

[6] Piazza CD et al (2025) App-based automated meal analysis in adults with T1D using AID: a randomised controlled trial. eClinicalMedicine 78. https://pmc.ncbi.nlm.nih.gov/articles/PMC12538901/

[7] Housni A, Katz A, Bergeron LJ et al (2025) Bridging the gap in carbohydrate counting with a mobile app: needs assessment survey. J Med Internet Res 27:e63278. https://doi.org/10.2196/63278

[8] iAPS Project (2026) iAPS: open-source automated insulin delivery system. Available from https://github.com/Artificial-Pancreas/iAPS. Accessed 11 April 2026

[9] Joubert M, Dreves B, Arnould T et al (2026) Comparative accuracy of smartphone apps and a generative AI tool for carbohydrate counting: an independent bicentric study. Diabetes Obes Metab. https://doi.org/10.1111/dom.70396

[10] Goncalves S, Coelho C, Pretre L et al (2025) Chat, Gemini and Claude at the dinner table: assessing general-purpose AI tools for carbohydrate counting in type 1 diabetes. Diabetes Res Clin Pract 113031. https://doi.org/10.1016/j.diabres.2025.113031

[11] Diabetes Technology Network UK (DTN-UK) (2026) Policy statement on the use of large language models in diabetes care

[12] Walsh J, Roberts R (2024) Pumping insulin, 6th edn. Torrey Pines Press, San Diego

[13] Cryer PE (2016) Hypoglycaemia in diabetes: pathophysiology, prevalence, and prevention, 3rd edn. American Diabetes Association, Alexandria

[14] International Hypoglycaemia Study Group (2017) Glucose concentrations of less than 3.0 mmol/L (54 mg/dL) should be reported in clinical trials: a joint position statement of the American Diabetes Association and the European Association for the Study of Diabetes. Diabetologia 60(1):3-6. https://doi.org/10.1007/s00125-016-4146-6