Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation of an Audio-Based Monitoring System for Detecting Retching and Vomiting in Gastroparesis: An Observational Cohort Study

Yishu Gong
yishu.gong@takeda.com
Takeda Development Center Americas, Inc.
Jasmin Imsirovic
Takeda Development Center Americas, Inc.
Brian Tracey
Takeda Development Center Americas, Inc.
Nicholas Stylopoulos
Takeda Development Center Americas, Inc.
Yanwei Zhang
Takeda Development Center Americas, Inc.
Muna Tahir
Takeda Development Center Americas, Inc.
Jing Yuan
Takeda Development Center Americas, Inc.
Xuefang Wu
Takeda Pharmaceuticals U.S.A., Inc
Lisa Graf
Redenlab Ltd
Thayabaran Kathiresan
Redenlab Ltd
Adam Vogel
Redenlab Ltd
David Levinthal
University of Pittsburgh School of Medicine

Article

Keywords: Gastroparesis, audio-based detection, vomit, retch, patient-reported outcomes

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9180913/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. Yishu Gong, Jasmin Imsirovic, Brian Tracey, Nicholas Stylopoulos, Yanwei Zhang, Muna Tahir, Jing Yuan, and Xuefang Wu are employees of Takeda Pharmaceutical Company Limited. Adam P. Vogel, Lisa Graf, and Thayabaran Kathiresan are employees of Redenlab. David J. Levinthal has served as an advisor for Takeda. The authors declare no other competing interests.

Abstract

This study aimed to evaluate a novel, audio-based digital monitoring system for detecting retching and vomiting (expulsion) events in a noninterventional, longitudinal study involving 25 participants with gastroparesis (GP). Participants used an audio-based system comprising an iOS device (iPhone) and Watch in their home environments for approximately 14 days. Retching and expulsion events were quantified using human-annotated counts from audio recordings and daily electronic patient-reported outcome (ePRO) counts. Device compliance was defined as the percentage of expected recording time successfully captured over the 14-day study period, calculated as recorded hours divided by the total expected hours (24 × 14). Compliance was high overall, with a mean rate of 83.4% for the iOS device and 69.4% for the Watch. Event timing analyses from the audio devices revealed elevated retching and expulsion activity in the morning, with symptoms persisting during nighttime hours. Strong correlations ($r > 0.85$) were observed between event counts from the Watch and iOS devices, although the agreement of device counts with ePRO counts was lower ($r < 0.40$). While ePRO-based expulsion counts showed a stronger within-participant correlation with patient global impression of frequency/severity (PGI-F/PGI-S) measures, device-based counts were better at distinguishing between participants across different PGI-F/PGI-S levels. Weekly averaging of event counts improved consistency across both audio-based and ePRO measures relative to daily event counts. These preliminary findings suggest that audio-based technologies offer an objective, sensitive, and scalable approach to detecting vomiting and retching in episodic disorders, with strong potential to support remote patient monitoring, real-world evidence generation, and earlier proof-of-concept evaluation in clinical trials for gastroparesis and other episodic vomiting disorders.

Introduction

Gastroparesis (GP) is a chronic disorder characterized by delayed gastric emptying in the absence of mechanical obstruction [1]. Patients with GP commonly experience symptoms such as nausea, vomiting and retching [2], which can significantly impair quality of life [3]. However, these episodic symptoms vary in frequency and severity over time, posing challenges for accurate monitoring and effective management. Traditional methods for assessing vomiting and retching rely heavily on patient self-report via daily diaries [4] or structured questionnaires with longer recall periods [5]. Although these tools provide valuable insights into patients' subjective experiences, they are prone to recall and reporting biases [6], presenting significant limitations for evaluating treatment effects in clinical trials and for monitoring clinical outcomes in both healthcare and home settings. Therefore, there is a critical need for objective, continuous, and reliable methods to capture vomiting-related symptom burden in patients with GP.

Emerging digital health technologies, particularly those leveraging passive sensing and artificial intelligence, present promising opportunities for objective, low-burden, and reliable symptom monitoring in ambulatory settings. Among these, audio-based monitoring systems have demonstrated feasibility in detecting respiratory and vocal biomarkers across various domains [7]. However, their application to gastrointestinal symptoms of retching and vomiting remains largely unexplored. Given the distinct acoustic signatures of retching and vomiting events, we hypothesized that audio analysis could provide a non-invasive, real-time solution for objectively monitoring vomiting patterns.

Here, we present an exploratory evaluation of an audio-based system designed to detect and quantify retching and vomiting (expulsion) events in a non-interventional pilot study involving participants with GP. This study aimed to differentiate auditory profiles of "retch" events (i.e., those without expelled material) from "expulsion" events (i.e., those involving the ejection of material). In addition, this study sought to evaluate the feasibility of audio-based symptom monitoring, with device compliance defined as the percentage of expected recording time successfully captured over the 14-day study period (calculated as recorded hours divided by the total expected hours [24 × 14]). The study also aimed to assess the reliability and validity of retch and expulsion event counts derived from audio recordings and to characterize the temporal patterns of these events in individuals with GP.

Methods

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Core Human Factors Institutional Review Board (Core Human Factors IRB; IORG0007854; IRB Registration #IRB00009432) via expedited review. All participants provided written informed consent prior to participation using the IRB-approved consent form. The study was conducted in accordance with applicable ethical guidelines for research involving human participants and the principles of the Declaration of Helsinki. Clinical trial number: not applicable.

Participants

A total of 25 participants with GP were recruited for this study through private clinical sites specializing in the treatment of GP patients that are not associated with any academic groups. Adult patients with medically-refractory GP were included in the study if they self-reported moderate-to-severe symptom severity and vomiting at least one day per week for the past 4 weeks. Patients were excluded if they were hospitalized or pregnant at the time of consent.

Study Design and Setting

A non-interventional, fully decentralized 14-day study was conducted to quantify the number of retching and expulsion events in participants with chronic, recurring, and episodic vomiting associated with GP. Upon enrollment, participants provided informed consent and were shipped a study kit that included two study watches (needed to swap for 14 days of continuous recording) and one iOS device (iPhone). Participants received training via teleconference upon receipt of the study kit and were given the opportunity to ask questions to ensure proper use of the devices. All data were collected remotely from the participants' home settings, allowing for real-world monitoring of symptoms.

Data Collection

Audio data were recorded using two types of devices across 14 days. Participants were instructed to wear an audio recording **Watch** (Esonic, SR-600 Voice Activated Audio Recording Watch) throughout the 14-day period, with guidance to pause recording when they left the house or had other privacy concerns. Additionally, participants were provided an iOS device with a customized audio recording application (Redenlab®). The iOS device was positioned in a location where participants typically experience symptoms, such as a bathroom, and was set to remain plugged in and recording continuously throughout the study period. Audio recordings were securely uploaded to a cloud-based system. For the iOS device, data was transferred in real time to a secure cloud location, while watch data was uploaded at the end of the study once the device was returned to the team. After data aggregation, all audio files were manually annotated by trained human listeners skilled in evaluating audio data to identify and label retch and expulsion events (Redenlab Ltd., Melbourne, Australia). Raters recorded the start and stop times of all retch and expulsion events, as well as related sounds such as spits, coughs, throat clearing, and interference that occurred close in time to retching or expulsion. To optimize the manual annotation process and ensure privacy, an automated pre-screening step was performed using Redenlab's Voice Activity Detection module to remove silent segments and exclude speech. To ensure quality and consistency of annotations, a structured training and quality assurance framework was implemented. Raters received training using individual audio events sourced from publicly available repositories. This training familiarized raters with the range of signal characteristics encountered in the study and standardized their annotation approach. All annotations were completed using a signal-based alignment and labelling tool, which provided objective temporal markers for audio events. To safeguard quality, every automated annotation underwent human review by trained listeners, who systematically verified and corrected outputs, as necessary. This double-checking process ensured that errors introduced by automated tools were minimized. To further strengthen reliability, the annotation summaries were cross-referenced with contemporaneous patient notes documenting daily events (described below), allowing for contextual validation of the audio labels. Specifically, participants completed a web-based questionnaire during each day of the study period to assess:

1. **Electronic Patient-Reported Outcome (ePRO) Retch/Expulsion Counts**: A gated question ("During the past 24 hours, did you vomit or retch?" Yes/No) was first posed to participants. Participants who answered "yes" were then asked to report the number of times they vomited/retched separately, with vomiting defined as the expulsion of food or liquid, and retching defined as dry heaves without expulsion. Participants were instructed to record each instance of vomiting or retching based on the physical act itself, rather than the number of trips to the bathroom.
2. **Patient Global Impression of Severity (PGI-S)**: The PGI-S is a single-item patient-reported outcome (PRO) assessing the severity of vomiting over the past 24 hours. Participants rated severity on a 5-point scale as follows: None (0), Mild (1), Moderate (2), Severe (3), and Very Severe (4).
3. **Patient Global Impression of Frequency (PGI-F)**: The PGI-F is a single-item PRO assessing the frequency of vomiting over the past 24 hours. Responses were rated on a 5-point scale as follows: None (0), Rarely (1), Sometimes (2), Often (3), and Very Often (4). Note that the PGI-F was added as an amendment to the study protocol after the first five participants had completed data collection; as such, it is not available for all participants.

All patient-reported items (ePRO retch/expulsion counts, PGI-S, and PGI-F) were administered as part of a single, daily web-based session. If a participant did not complete the session, data for that day were considered missing for all three measures.

Data Analysis

Daily data were reviewed for completeness, duplicate entries, and outliers. No imputation was performed. To assess measurement stability, weekly event counts for retching (R), expulsion (E), and combined retching and expulsion (R+E) from the Watch and iOS devices were calculated by summing all events across all recording hours and normalizing to a 7-day period. Events collected via ePRO were similarly summed over all 14 days and normalized to a 7-day period. Beyond raw counts, event bouts were defined as clusters of combined retching and expulsion events occurring in close temporal proximity to avoid splitting of physiologically continuous symptoms, as illustrated in Figure 1. Specifically, if the time between the end of one event and the start of the next was 10 seconds or less, the events were grouped into a single bout (see Supplementary Materials for details on the threshold selection approach). The computed daily and weekly event endpoints, as well as event bout endpoints, are summarized in Table 1. Event

bout endpoints are only defined for data collected via audio devices, given that ePRO lacks a precise timestamp of events to allow for similar analyses.

Table 1. Summary of tabulated endpoints

Measure	Data Source	Time period	Tabulated endpoints (names)
Device compliance (average percentage of valid recording hours in 24 hours over 14 days)	iOS, Watch	Daily	Compliance
Number of retch events	iOS, Watch, ePRO	Daily	R (Daily) from iOS, Watch, and ePRO
Number of expulsion events	iOS, Watch, ePRO	Daily	E (Daily) from iOS, Watch, and ePRO
Number of combined retch and expulsion events	iOS, Watch, ePRO	Daily	R+E (Daily) from iOS, Watch, and ePRO
Aggregated weekly retch events	iOS, Watch, ePRO	Weekly	R (Weekly) from iOS, Watch, and ePRO
Aggregated weekly expulsion events	iOS, Watch, ePRO	Weekly	E (Weekly) from iOS, Watch, and ePRO
Aggregated weekly combined retch and expulsion events	IOS, Watch, ePRO	Weekly	R+E (Weekly) from iOS, Watch, and ePRO
Number of combined retch and expulsion bouts	iOS, Watch	Daily	R+E Bout (Daily) from iOS and Watch
Patient global impression of frequency of vomiting (expulsion)	ePRO	Daily	PGI-F
Patient global impression of severity of vomiting (expulsion)	ePRO	Daily	PGI-S

All statistical tests were two-tailed with a significance level of $\alpha = 0.05$. The following analyses were performed:

Patient characteristics: Basic clinical and demographic characteristics of the participants were recorded, including age, gender, and duration of disease.

Device Compliance: Participant compliance in capturing audio data was assessed separately for the Watch and iOS device. Compliance was calculated as the number of valid recording hours per day divided by 24 and averaged over the 14-day study. To standardize event rates, all count endpoints collected via audio devices were adjusted to a 24-hour day using the following formula:

Adjusted count/day = round((count / recorded hours) × 24).

To ensure data quality and minimize bias from incomplete wear or technical issues, only days with more than 8 hours of valid recording were included in the analysis.

Descriptive statistics Daily and weekly retch, expulsion, and combined event counts (by source) were summarized using means, standard deviations (SD), medians, and the 10^{th}–90^{th} percentiles. Similar descriptive statistics were calculated for combined R+E bouts.

Time-of-day effects: Temporal patterns of audio-derived daily events were evaluated by analyzing timestamp data for retch and expulsion events. A bootstrap approach was utilized to minimize the impact of outlier subjects. In each of 2,000 iterations, 25 participants were randomly sampled with replacement. For every 3-hour time bin, the median and 10th–90th percentile range of the event proportions across all iterations were computed, providing an overview of the distribution of event timing throughout the day.

Association between audio-based endpoints and patient-reported outcomes: The agreement between daily event counts collected by audio devices (Watch and iOS device) and daily counts reported by patients (ePRO) was assessed using repeated measure correlations (rmcorr) developed by Bakdash and Marusich [8]. Further, within-subject associations between daily event counts (collected via ePRO, Watch, and iOS devices) and corresponding daily PGI-S/PGI-F ratings was assessed using repeated measure correlations. Differences in daily event counts collected via ePRO, Watch, and iOS devices across varying frequency and severity levels, as determined by the PGI-F and PGI-S scales, were analyzed using Friedman's test, followed by Dunn's tests for post hoc pairwise comparisons. Similar statistical analyses were conducted for bout-based endpoints.

Outcome reliability and effect of averaging period: Test–retest reliability was assessed using intraclass correlation coefficients (ICCs), derived from a two-way mixed-effects ANOVA model for absolute agreement [9–10] and computed using the *iccCounts* package in R [11]. Daily event counts from Day 1 (defined individually for each participant as the first day with available data) vs Day 8, Day 1 vs Day 14, and Day 7 vs Day 14 were compared in the primary analysis. Reliability was also evaluated by comparing total event counts from the first 7 days (week 1) with those from the final 7 days (week 2) of the study.

Results

Participant Characteristics

This decentralized, observational study included 25 adult participants with nausea-vomiting gastroparesis. Participants were predominantly female (n=21 [84%]) with a mean age of 47.5 ± 12.4 years. Information on disease duration was available for 17 participants, who had a median disease duration of 5.15 years (interquartile range: ±3.44 years).

Device Compliance

During the recruitment period, 28 GP participants were contacted. Of these, none declined participation due to privacy concerns around home monitoring requirements; however, n=1 participant experienced symptoms that were too sparse and sporadic, n=1 participant had no wireless internet access, and n=1 participant had issues with the nondisclosure agreement.

Compliance was higher, on average, for the iOS device (iPhone) (83.4%, n=23) compared to the Watch (69.4%, n= 24). One participant was excluded from the Watch analyses and two were excluded from the iOS analyses due to technical setup issues; however, no participant experienced issues with both devices simultaneously (see individual participants' device compliance in Supplementary Figure 1).

Descriptive Statistics: Event Counts

Over the course of the study, the Watch captured a total of 870 retch events and 275 expulsion events, while the iOS device recorded 796 retch events and 227 expulsion events. Daily and weekly event counts for retching (R), expulsion (E), and combined events (R+E), collected from multiple sources across all available days, are summarized in Table 2. A consistent pattern emerged across all timeframes, with the Watch capturing the highest mean number of events of each type, followed by the iOS device and then the ePRO. The audio-based devices demonstrated that retching events were typically 2–4 times more frequent than expulsion events, a pattern similarly observed in ePRO counts.

Table 2. Daily and weekly retch, expulsion and combined retch and expulsion events by data source for GP participants over entire study period

Source	R (Daily)		E (Daily)		R + E (Daily)	
	Mean (SD)	10^{th} - 90^{th} percentiles	Mean (SD)	10^{th} - 90^{th} percentiles	Mean (SD)	10^{th} - 90^{th} percentiles
Watch	3.97 (10.69)	0 - 12	1.26 (6.39)	0 - 3	5.23 (14.98)	0 - 17
iOS	3.05 (9.25)	0 - 8	0.86 (3.61)	0 - 2	3.92 (11.4)	0 - 11
ePRO	1.68 (2.11)	0 - 4	0.72 (1.42)	0 - 2	2.4 (2.95)	0 - 6
Source	R (Weekly)		E (Weekly)		E (Weekly)	
	Mean (SD)	10^{th} - 90^{th} percentiles	Mean (SD)	10^{th} - 90^{th} percentiles	Mean (SD)	10^{th} - 90^{th} percentiles
Watch	24.67 (41.63)	0 - 79.5	9.88 (30.6)	0 - 17	34.58 (60.29)	0 - 94.9
iOS	20 (47.53)	0 - 39.8	5.61 (11.99)	0 - 15.4	25.66 (54.03)	0 - 52.4
ePRO	11.5 (10.61)	2 - 23.6	4.94 (4.4)	1 - 10.1	16.44 (13.22)	3.9 - 34

Bout events (combined R+E) collected via the Watch and iOS device are shown in Table 3. Using a 10-second gap, determined through cross-validation to define a bout (see Supplementary Methods), we found that bouts typically consist of 1–8 events per bout (10^{th} -90^{th} percentiles), and last 0.81–25.15 seconds (10^{th} -90^{th} percentiles). As with individual events, bout counts were consistently higher on the Watch compared to the iOS device.

Table 3. Descriptive statistics for bout events (combined R+E) by data source for GP participants over entire study period

Bout summary	Number of events per bout		Length of bout (sec)		Percent (%) time spent in events	
	Mean (SD)	10^{th} - 90^{th} percentiles	Mean (SD)	10^{th} - 90^{th} percentiles	Mean (SD)	10^{th} - 90^{th} percentiles
Watch	3.54 (2.90)	1-8	10.08 (10.79)	0.81-25.15	65.23 (26.19)	31.68 - 100.00
iOS	3.51 (3.25)	1-8	9.54 (10.02)	1.00-25.01	65.99 (26.67)	33.76 – 100.00

Figure 2 illustrates the distribution of daily retch and expulsion counts by data source, with individual subject trends depicted via connecting lines, highlighting the considerable variability both within and across patients over days.

Figure 3 presents stacked bar charts visualizing daily item-level responses and longitudinal variability in the PGI-S and PGI-F scores for GP participants. For both PGI-S and PGI-F, the distribution remained relatively consistent throughout the study period.

Descriptive Statistics: Time of Day Effects on Event Counts

Figure 4 summarizes the median distribution of events and the 10^{th} - 90^{th} percentiles from the resampling procedure. An increase in the retch and expulsion events was observed during the morning hours. Additionally, retch and expulsion events persisted throughout the night at levels comparable to those observed during the daytime.

Association Between Audio-Based Endpoints and Patient Questionnaires

Human-annotated event counts vs. patient-reported event counts

The two audio devices demonstrated strong agreement with each other. Specifically, the correlation between the Watch and iOS devices was 0.86 for retching, 0.91 for expulsions, and 0.90 for combined R+E events. In contrast, correlations between device-based and ePRO event counts were modest. For retching events, the correlation of ePRO counts with the iOS device and the Watch was 0.26 (95% CI: 0.14–0.37) and 0.40 (95% CI: 0.29–0.50), respectively. For expulsion events, both iOS and Watch showed similar correlations with ePRO at 0.48 (iOS: 95% CI: 0.41–0.54; Watch: 95% CI: 0.41–0.55). For combined R+E events, the correlation of ePRO counts with the iOS device and Watch was 0.31 (95% CI: 0.22–0.40) and 0.39 (95% CI: 0.31–0.47), respectively.

When using bout-based combined event counts, correlations with ePRO improved. The correlation between ePRO and the iOS device was 0.58 (95% CI: 0.49–0.66), while the correlation between ePRO and the Watch was 0.64 (95% CI: 0.56–0.71).

Human-annotated audio and ePRO event counts vs. PGI-S and PGI-F

We hypothesized that daily expulsion events would show stronger correlations with the PGI scales than retching events, as the PGI items specifically reference vomiting. We also anticipated that ePRO daily event counts would show stronger correlations with PGI-F/PGI-S scores than device event counts would, given that both are based on patient self-report. As summarized in Supplementary Table 1, expulsion counts from ePRO showed the highest correlations with the patient global impression items (PGI-F: r = 0.65 and PGI-S: r =0.65), followed by device-based counts (r ≈ 0.31–0.36). Correlations of retch counts with the patient global impression items were weaker across all sources (r ≈ 0.16–0.36). As expected, combined R+E raw counts showed moderate correlations with the PGI-F/PGI-S, with ePRO again outperforming both audio devices. Likewise, R+E bout counts from Watch and iOS device offered similar correlation strength to R+E raw counts.

Bar plots illustrating R (daily), E (daily), and R+E (daily) counts against the corresponding PGI-F and PGI-S reported on the same day are presented in Figure 5. For retching, ePRO event counts remained consistent across PGI-F levels, with underreporting of event counts observed for patients who rated themselves as having vomiting “Very Often” or “Often” on the PGI-F. In contrast, both the Watch and iOS device showed a relatively monotonic increase in retching counts with increasing PGI-F levels. For expulsion events, low counts across all sources were observed within the “None” PGI-F category. Expulsion counts from the Watch, iOS device and ePRO increased somewhat progressively with higher PGI-F levels, though variability was high due to day-to-day fluctuations. Since participants experienced more frequent retching than expulsion events, the combined R+E (daily) plot closely mirrored the trend observed for retching.

Figure 6 presents bar plots of the R+E bout counts (daily) from Watch and iOS device against their reported PGI-F and PGI-S scores on the same day. Notably, the bout-based counts still showed a monotonic and clearer increase with rising PGIF/PGIS levels compared to the ePRO-based counts.

The various device and patient-reported event counts failed to reliably distinguish between adjacent PGI-F groups (e.g., “Often” vs. “Very Often”) and PGI-S groups (e.g., “Severe” vs. “Very Severe”) based on pairwise comparisons. To address this problem, PGI-F responses were collapsed from five categories into three categories as follows: None, Rarely/Sometimes, and Often/Very Often.

Similarly, PGI-S responses were grouped into None, Mild/Moderate, and Severe/Very Severe. As shown in Supplementary Table 2, ePRO counts had limited ability to differentiate between these combined categories, whereas the Watch and iOS devices showed clearer distinctions, with expulsion counts increasing consistently with increasing frequency and severity.

Outcome Reliability and Effect of Averaging Period

Supplementary Table 3 summarizes the intraclass correlation coefficients (ICCs) for daily retching, expulsion, and combined (R+E) counts across three day-pair comparisons: Day 1 vs. Day 8, Day 1 vs. Day 14, and Day 7 vs. Day 14. Reliability varied across devices and event types. For the Watch, ICCs for R, E and R+E counts ranged from 0.37 to 0.80, while the iOS device ranged from 0.22 to 0.67. ePRO counts showed more stable but moderate reliability, with ICCs ranging between 0.40 and 0.56 across different day-pair comparisons.

Table 4 presents weekly ICCs (Week 1 vs. Week 2), revealing generally higher consistency over time. The Watch showed moderate reliability for retching and combined R+E events (ICCs of 0.53 and 0.50, respectively), but lower reliability for expulsion events (0.26). The iOS device also demonstrated stronger reliability for retching and R+E events (ICCs of 0.79 and 0.69), compared to expulsion events (0.36). ePRO had the most consistent performance, with ICCs of 0.76 to 0.90 across all event types. These results are visualized in Figure 7.

Bout-based R+E counts did not consistently improve reliability. Daily ICCs ranged from 0.25 to 0.72 for the Watch and iOS, depending on the day pair. Weekly ICCs remained moderate, with the Watch at 0.58 and the iOS at 0.61.

Table 4. Test-retest reliability of weekly retch, expulsion, and combined retch and expulsion events in GP participants

Source	Measure	N	Comparison	ICC
Watch	R (Weekly)	24	Week 1 vs Week 2	0.53 (0.00, 0.81)
	E (Weekly)	24	Week 1 vs Week 2	0.26 (-0.16, 0.44)
	R + E (Weekly)	24	Week 1 vs Week 2	0.50 (-0.17, 0.67)
	R + E bout (Weekly)	24	Week 1 vs Week 2	0.42 (0.05, 0.69)
iOS device	R (Weekly)	23	Week 1 vs Week 2	0.79 (0.30, 0.95)
	E (Weekly)	23	Week 1 vs Week 2	0.36 (-0.31, 0.79)
	R + E (Weekly)	23	Week 1 vs Week 2	0.69 (0.12, 0.92)
	R + E bout (Weekly)	23	Week 1 vs Week 2	0.47 (0.06, 0.74)
ePRO	R (Weekly)	25	Week 1 vs Week 2	0.90 (0.74, 0.96)
	E (Weekly)	25	Week 1 vs Week 2	0.76 (0.45, 0.91)
	R + E (Weekly)	25	Week 1 vs Week 2	0.89 (0.75, 0.96)

Discussion

This study provides foundational evidence supporting the use of audio-based devices for objectively monitoring retching and expulsion events in GP, a technology which may be applied to other disorders characterized by episodic vomiting. Our findings establish that audio-based monitoring is both feasible and effective, with objective data collected from devices complementing and building on the information gathered currently using PRO tools.

A key finding of this study was the high compliance and willingness of participants to be audio-recorded, indicating that audio-based symptom tracking is acceptable within this population. Notably, compliance was higher with the iOS device, likely due to its passive nature—once activated, it required minimal participant involvement. Audio monitoring devices remain relatively uncommon in clinical studies, but prior research has reported median daily wear times exceeding 70% of recording hours [12] and most participants maintaining device use for at least 10 of 14 days [13], supporting their feasibility. In the present study, although the audio-based device fell just below the commonly cited 70% compliance threshold [12], overall adherence on both devices was sufficient to support future clinical use with a more structured study design. Although recording times on the iOS device were longer, the Watch captured more events, likely due to its mobility and ability to travel with participants. This suggests that mobile wearables, as opposed to fixed recording systems, may be better suited for tracking symptoms in real-world environments.

This study further highlights the complementary roles of ePRO and audio-based measures in clinical research. The ePRO counts demonstrated stronger within-person correlations with patient global impression items (PGI-S and PGI-F), indicating greater sensitivity to perceived intra-individual changes over time. In contrast, audio-based measures showed stronger discriminative

performance in distinguishing severity and frequency levels based on the global impression items, despite modest within-person correlations. In addition, audio devices are limited to capturing specific symptom events and cannot assess other important patient experiences, such as nausea. A multiple-endpoint strategy incorporating both methods may thus provide a more holistic, patient-centered understanding of clinical benefit, capturing distinct treatment effects aligned with regulatory expectations.

The data collected in this study also provided new insights into the symptoms of gastroparesis. For example, recordings from the Watch and iOS device suggest that retching may be a more common symptom of GP than previously recognized. This challenges traditional views of the disease and highlights the need for further research into the impact of frequent retching on quality of life. Analysis of time-of-day patterns further revealed frequent nighttime retching, which may contribute to sleep disruption—an area often overlooked in traditional symptom diaries. Morning symptoms were also higher than other times of the day, potentially related to morning movement or eating. The mid-day dip in symptoms data should, however, be interpreted with caution, as subjects may be more likely to be outside their home during this time, away from the iOS device and with the Watch recording paused. Future studies should confirm and further explore these time-of-day patterns, by potentially incorporating physical movement tracking and monitoring of mealtimes to determine links to GP symptoms.

There was substantial symptom variation at the daily level, which likely represents real day-to-day changes rather than measurement error. As such, this reflects the dynamic nature of GP symptoms. However, weekly aggregation of the same data (i.e., week-to-week test-retest reliability) marginally improved repeatability. Due to the length of the observation period, longer aggregation intervals (e.g., over 2 weeks or more) could not be assessed in this study but are expected to further stabilize the measures; future studies could test this hypothesis. Similar findings are seen in studies of chronic cough, where high day-to-day variability leads to a need for longer observation periods [14-15]. Less variability was observed in the daily ePRO data compared to audio-based metrics, perhaps reflecting an implicit averaging in participants' recall of events.

This work explored the definition and utility of "retch and/or expulsion bouts, inspired by similar investigations in the cough literature [16]. As described in Supplemental Methods, we developed a data-driven strategy that allows for the identification of bouts that are both sufficiently frequent and of an appropriate length, seeking to capture disease burden while avoiding both excessive fragmentation and oversimplification of the event data. We found that bout-level audio data showed stronger correlations with ePRO data than raw count-level audio data, suggesting that participants may naturally perceive and report their symptoms in discrete bouts. Additionally, bouts allow computation of measures of bout severity (for example number of events per bout or bout length). However, aggregating individual events into bouts did not improve test-retest reliability.

Our work has important limitations that can be addressed in future studies. Our use of human-annotated audio introduced privacy concerns, limiting recordings to the home environment, which likely resulted in undercounting of events and challenges in interpreting the time-of-day findings. Additionally, our reliance on human scoring poses logistical challenges for scaling to larger populations. Moreover, placebo effects can be sizeable in self-reported gastrointestinal measures [17]. Objective, passively captured measures, such as audio-based symptom tracking may reduce placebo effects by reducing expectancy-driven responses inherent in subjective measures [18]. While this could not be evaluated in the non-interventional pilot study presented here, future interventional trials in GP could assess whether the use of objective monitoring of the symptoms could be valuable in settings where placebo effects are a concern.

The development of a wrist-worn, fully automated system could enable privacy-preserving monitoring both inside and outside the home and across larger populations. Although full automation is technically challenging given that retch and expulsion sounds constitute only a small fraction of a day's recorded acoustic data, we anticipate that advancements in large audio datasets and AI will make such systems feasible. Additionally, modern smart watches capture physiological signals (e.g., acceleration, heart-rate data) which could enhance AI algorithms with additional inputs.

In conclusion, this study demonstrates that an audio-based system for detecting retch and expulsion events is feasible, reliable, and valid in patients with gastroparesis. These results highlight the potential of audio-recording technologies as objective and sensitive tools for symptom monitoring in gastrointestinal disorders characterized by episodic vomiting. By enabling scalable, remote, and real-time data capture, such systems could advance clinical research, support real-world evidence generation, and facilitate earlier proof-of-concept assessments in clinical trials for gastroparesis and other episodic vomiting disorders.

Declarations

Author Contributions

YG designed the methodology, performed data curation, formal analysis, and visualization, and wrote the original draft. JI conducted formal analysis, contributed to data acquisition and visualization, and reviewed and edited the methodology. BT contributed to conceptualization, formal analysis, and manuscript review and editing. NS provided conceptualization, clinical investigation, and resources, and reviewed and edited the manuscript. YZ contributed to conceptualization, methodology, statistical analysis, supervision, and manuscript review and editing. MT provided supervision, methodology, statistical analysis, and

manuscript review and editing. JY contributed to methodology, statistical analysis, and manuscript review and editing. AV contributed to data acquisition, conceptualization, methodology, supervision, and manuscript review and editing. LG performed data acquisition, coordinated the project, and reviewed and edited the manuscript. TK contributed to data acquisition, methodology, and manuscript review and editing. XW contributed to conceptualization, project coordination, resources, and manuscript review and editing. DL provided conceptualization, clinical expertise, supervision, and manuscript review and editing. All authors read and approved the final manuscript.

Funding Statement

This work was supported by Takeda Pharmaceutical Company Limited. The funder had a role in study design, data collection, analysis, interpretation, and the decision to publish. No specific grant number.

Competing Interests

Yishu Gong, Jasmin Imsirovic, Brian Tracey, Nicholas Stylopoulos, Yanwei Zhang, Muna Tahir, Jing Yuan, and Xuefang Wu are employees of Takeda Pharmaceutical Company Limited. Adam P. Vogel, Lisa Graf, and Thayabaran Kathiresan are employees of Redenlab. David J. Levinthal has served as an advisor for Takeda.

The authors declare no other competing interests.

Data Availability

Individual event–level datasets are not publicly available due to patient privacy and confidentiality requirements. Individual and de-identified summary-level data may be requested, and each request will be reviewed for compliance with privacy and data-use restrictions.

Code Availability

Custom Python code used for feature extraction and R code for statistical analysis is available from the authors upon request.

References

1. Camilleri, M., Chedid, V., Ford, A.C., Haruma, K., Horowitz, M., Jones, K.L., Low, P.A., Park, S.-Y., Parkman, H.P. & Stanghellini, V. Gastroparesis. Nat Rev Dis Primers 4, 41 (2018). https://doi.org/10.1038/s41572-018-0041-3
2. Reddivari, A.K.R. & Mehta, P. Gastroparesis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Available at: https://www.ncbi.nlm.nih.gov/books/NBK551528/ [Accessed July 8, 2025].
3. Reddivari, A.K.R. & Mehta, P. Factors that contribute to the impairment of quality of life in gastroparesis. [Journal name placeholder]. PMC8310540. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8310540/ [Accessed July 8, 2025].
4. Revicki, D.A., Lavoie, S., Speck, R.M., Puelles, J., Kuo, B., Camilleri, M., Almansa, C. & Parkman, H.P. The content validity of the ANMS GCSI DD in patients with idiopathic or diabetic gastroparesis. J Patient Rep Outcomes 2, 61 (2018). https://doi.org/10.1186/s41687-018-0081-2
5. Rentz, A.M., Kahrilas, P., Stanghellini, V., Tack, J., Talley, N.J., de la Loge, C., Trudeau, E., Dubois, D.A. & Revicki, D.A. Development and psychometric evaluation of the patient assessment of upper gastrointestinal symptom severity index (PAGI-SYM) in patients with upper gastrointestinal disorders. Qual Life Res 13, 1737–1749 (2004). https://doi.org/10.1007/s11136-004-9567-x
6. Zini, M.L.L. & Banfi, G. A narrative literature review of bias in collecting patient-reported outcomes measures (PROMs). Int J Environ Res Public Health 18, 12445 (2021). https://doi.org/10.3390/ijerph182312445
7. Gong, Y., Yu, J. & Glass, J. VocalSound: A dataset for improving human vocal sounds recognition. ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 151–155 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746828
8. Bakdash, J.Z. & Marusich, L.R. Repeated measures correlation. Front Psychol 8, 456 (2017). https://doi.org/10.3389/fpsyg.2017.00456
9. Hays, R.D. & Revicki, D.A. Reliability and validity (including responsiveness). In: Fayers, P.M. & Hays, R.D. (eds.) Assessing quality of life in clinical trials: Methods and practice. 2nd ed. Oxford: Oxford University Press, 25–39 (2005).
10. Hays, R.D. & Revicki, D.A. Reliability and validity (including responsiveness). In: Fayers, P.M. & Hays, R.D. (eds.) Assessing quality of life in clinical trials: Methods and practice. 2nd ed. Oxford: Oxford University Press, 25–39 (2005).
11. Carrasco JL. *iccCounts: An R Package to Estimate the Intraclass Correlation Coefficient for Assessing Agreement with Count Data*. *The R Journal*, 14(2), 230–244. https://doi.org/10.32614/RJ-2022-034 (2022)
12. Zhang, Y. et al. Wearable compliance in real-world monitoring studies. NPJ Digit. Med. 8, 117 (2025).

13. Ding, X. et al. Device usage consistency in clinical monitoring studies. Digit. Health 10, 205520762411560 (2024). https://doi.org/10.1016/j.jcm.2016.02.012
14. Chung, K.F., Chaccour, C., Jover, L., Galvosas, M., Song, W.-J., Rudd, M. & Small, P. Longitudinal Cough Frequency Monitoring in Persistent Coughers: Daily Variability and Predictability. Lung 202, 561–568 (2024). https://doi.org/10.1007/s00408-024-00734-x
15. Morice, A., Alge, M., Kuett, L., Hart, S., Rigby, A. & Elkayam, D. Cough frequency has a high daily variation in patients with chronic cough. ERJ Open Res 11, 00670–2024 (2025). https://doi.org/10.1183/23120541.00670-2024
16. Kelsall, A., Decalmer, S., Webster, D., Brown, N., McGuinness, K., Woodcock, A. & Smith, J. How to quantify coughing: correlations with quality of life in chronic cough. Eur Respir J 32, 175–179 (2008). https://doi.org/10.1183/09031936.00101307
17. Elsenbruch, S. & Enck, P. Placebo effects and their determinants in gastrointestinal disorders. Nat Rev Gastroenterol Hepatol 12, 472–485 (2015). https://doi.org/10.1038/nrgastro.2015.117
18. Abelson, M.B. & Dewey-Mattia, D. The placebo effect and other confounders. Rev Ophthalmol Online. Published 11 June 2010. Available from: https://www.reviewofophthalmology.com/article/the-placebo-effect-and-other-confounders

Figures

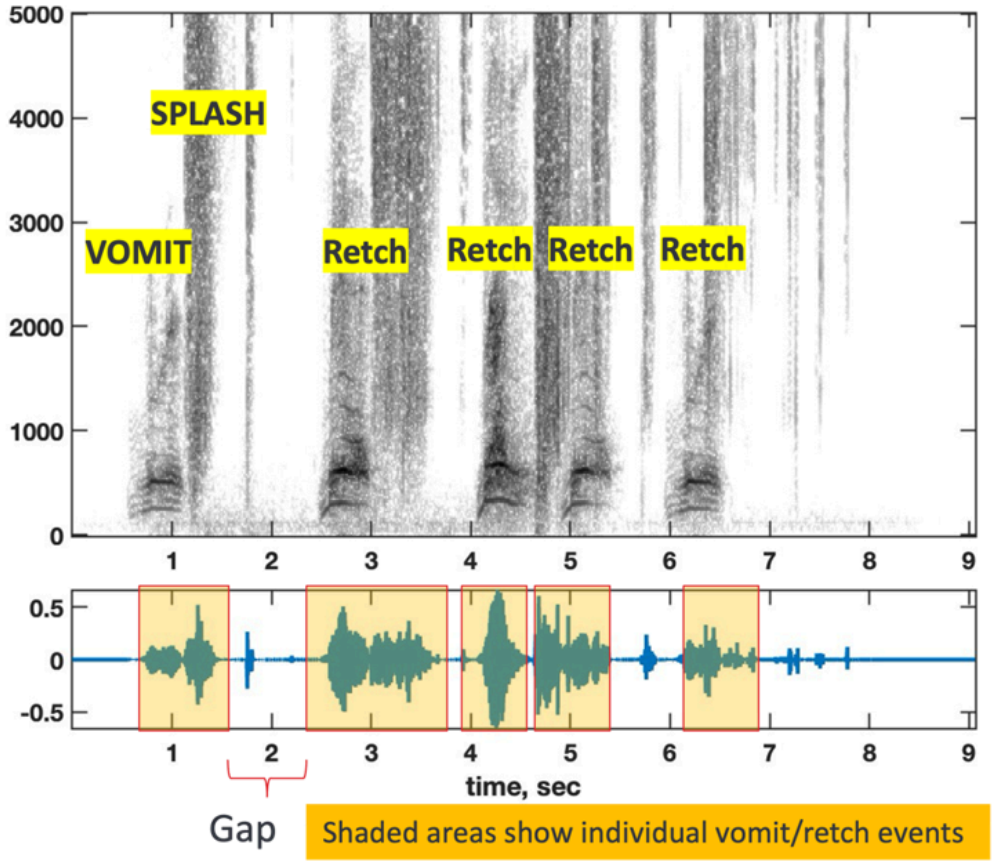


Figure 1

Demonstration of a retch and expulsion bout.

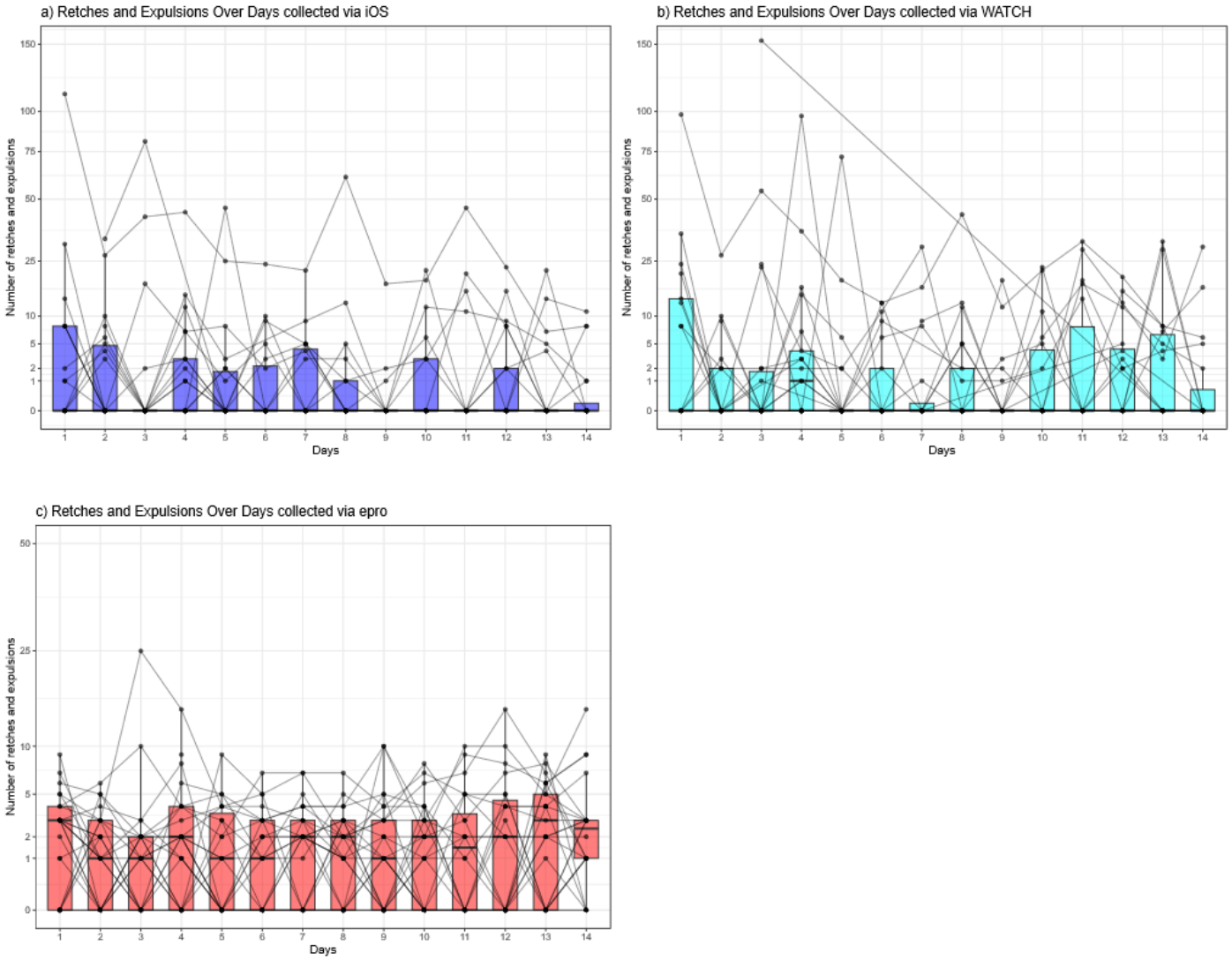


Figure 2

Boxplot of the distribution of retch and expulsion (R+E) counts over study days by data source.

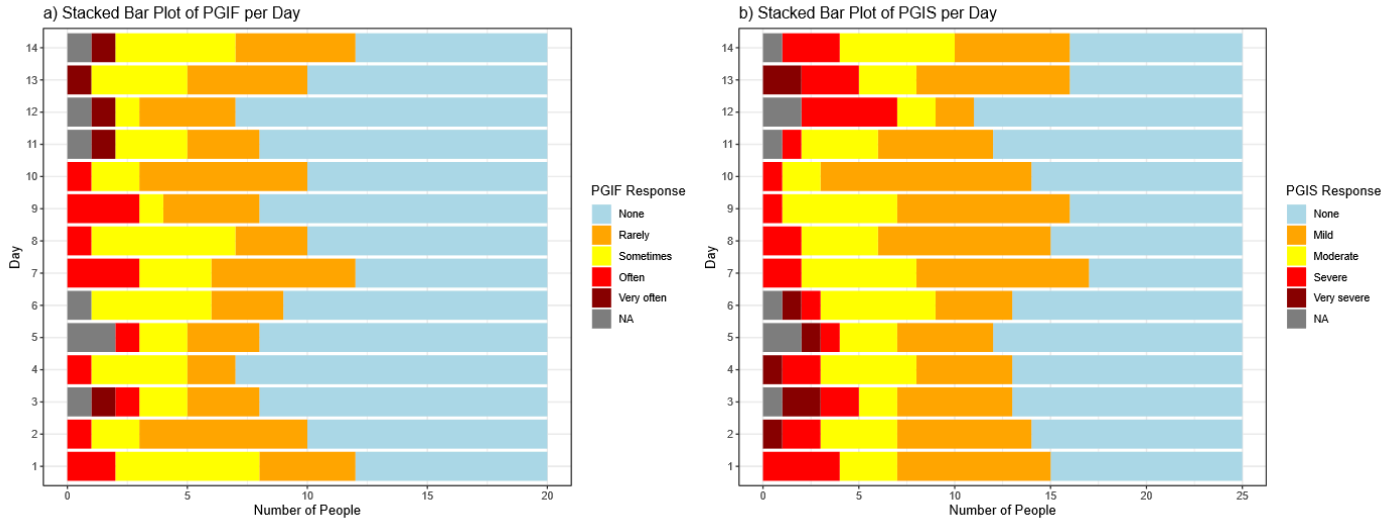


Figure 3

Stacked bar plots of PGI-F (left)/PGI-S(right).

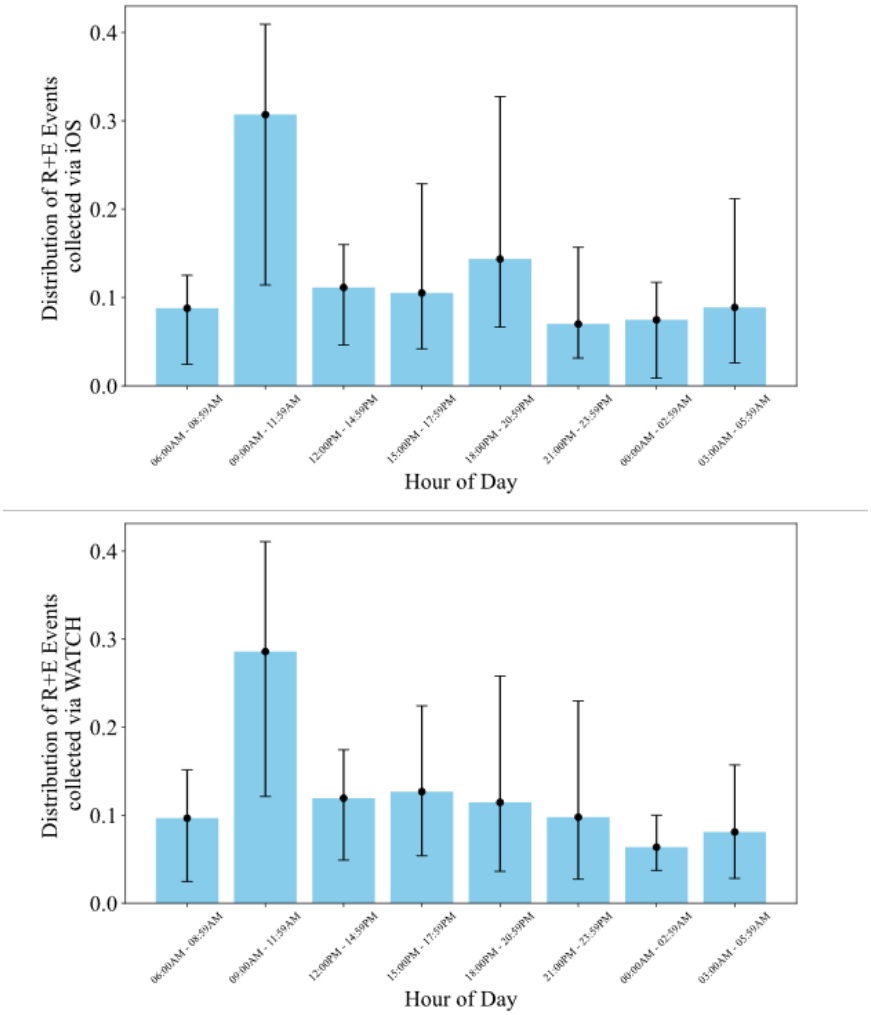


Figure 4

Distribution of retch and expulsion (R+E) events across time of day, presented in 3-hour intervals. Events collected via iOS device (top). Events collected via Watch (bottom). Error bars represent 95% confidence intervals.

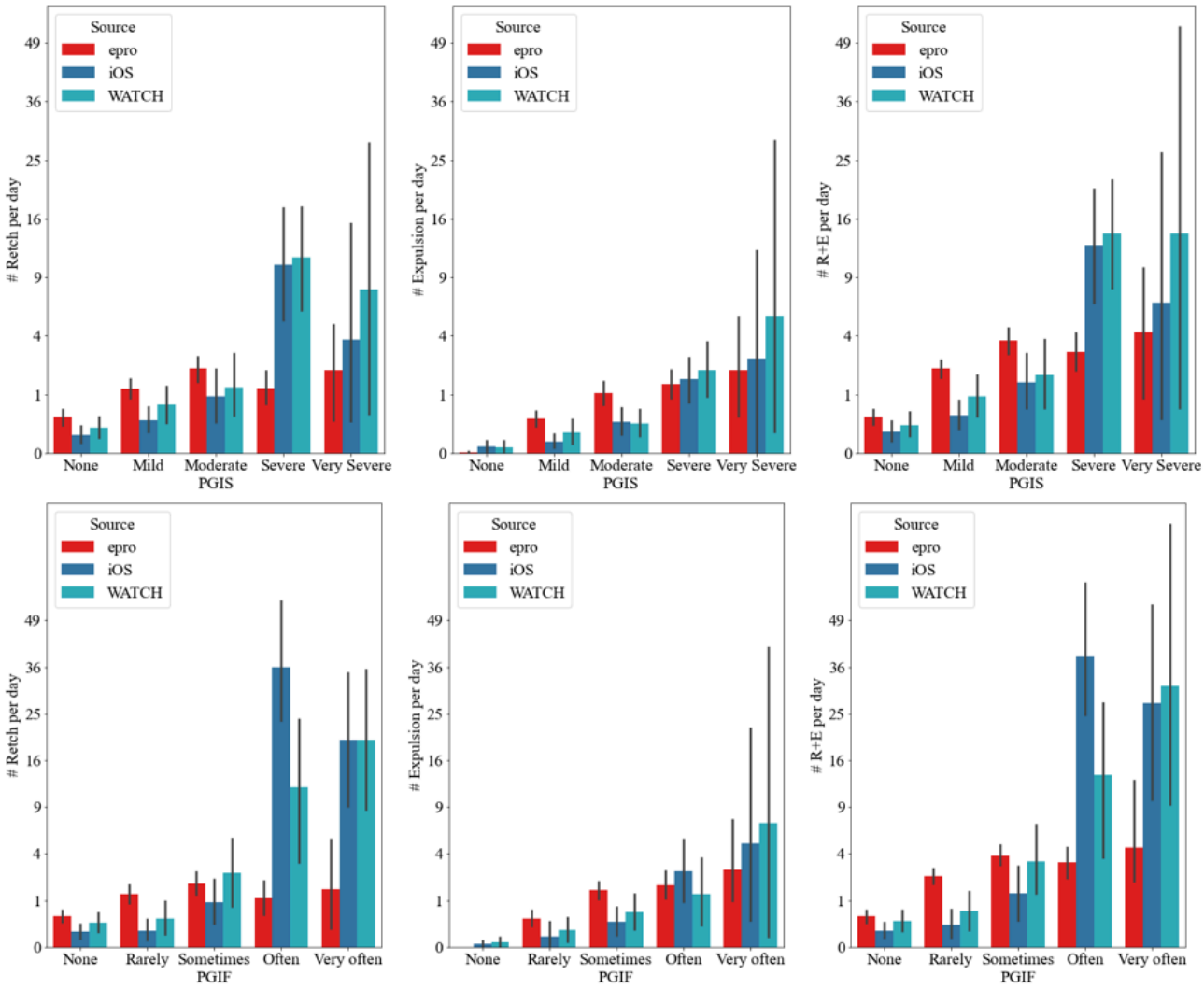


Figure 5

bar plots of daily retching (R; left), expulsion (E; middle), and combined retching and expulsion (R+E; left) events across varying levels of PGI-F (top panel) and PGI-S (lower panel). Error bars represent 95% confidence intervals.

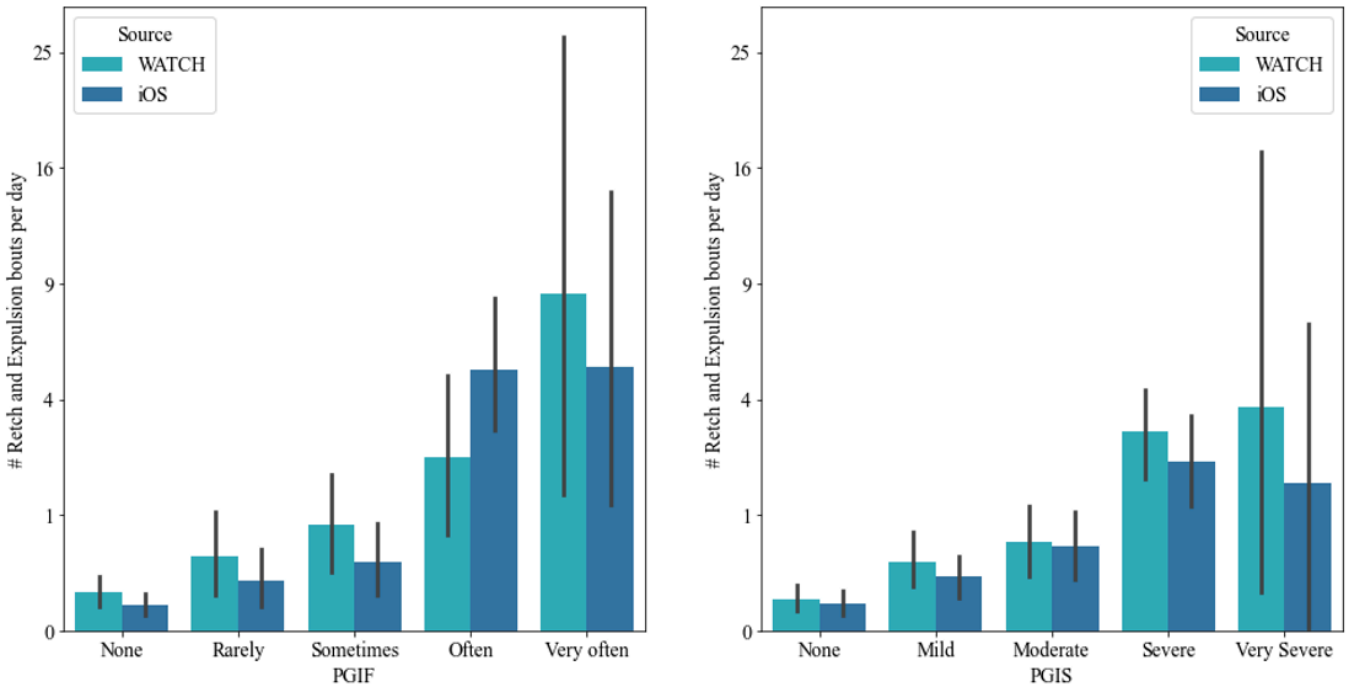


Figure 6

bar plots of daily R+E bout daily for Watch and iOS device across varying levels of PGI-F (left) and PGI-S (right). Error bars represent 95% confidence intervals.

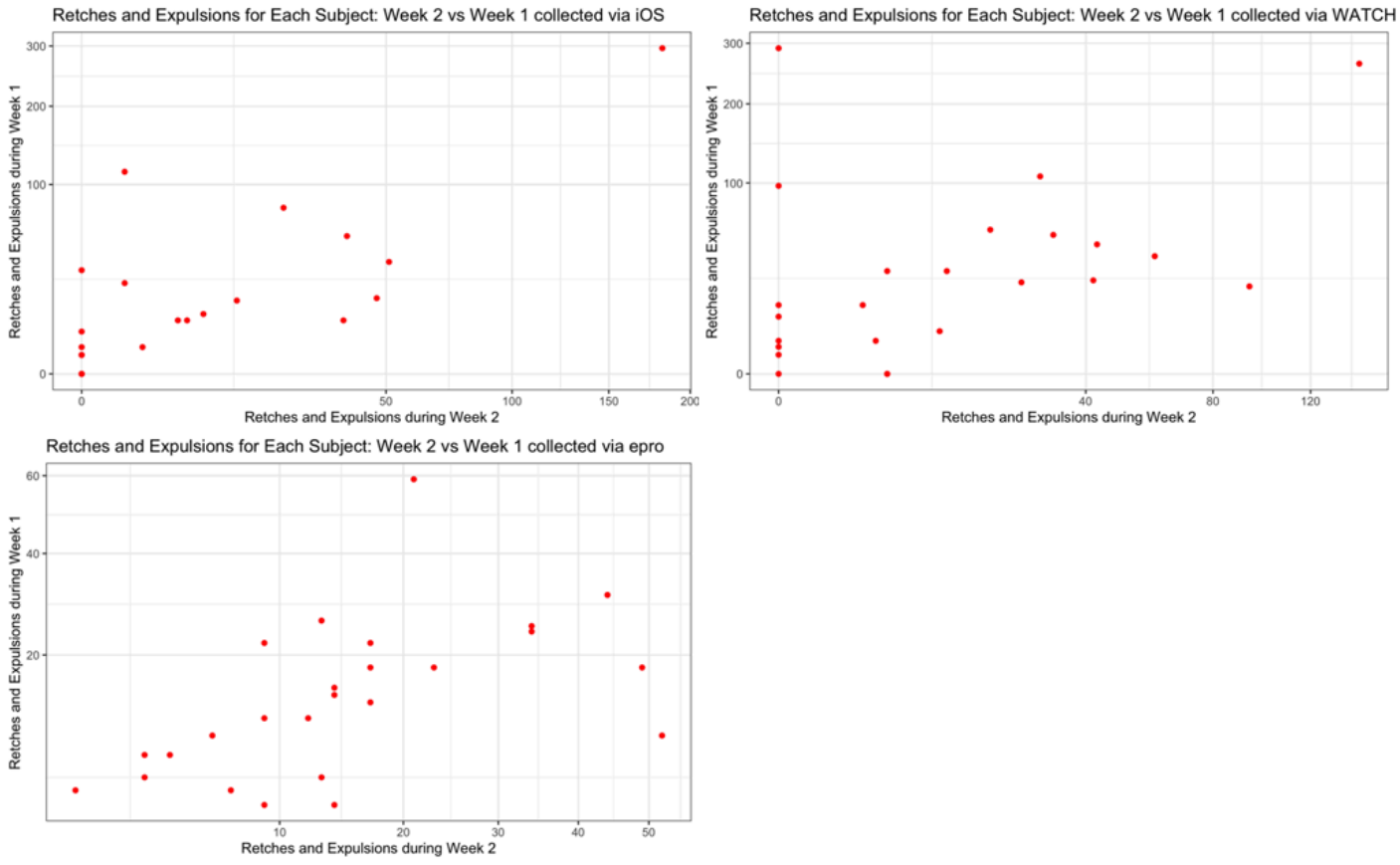


Figure 7

Scatter Plots of Weekly aggregated number of retching and expulsion (R+E) events for each subject in week 1 vs week 2 in GP patients.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GPpapersupplementary.docx