Original Paper

Agreement of Step Counting Between a Fitness Tracker and a Medical Device in Patients With Ulcerative Colitis: Prospective Observational Study

Thomas Wenske[1,2], MSc; Ovidiu Popa[1], PhD; Sandra Plachta-Danielzik[2,3], PhD; Oliver Ebenhöh[1], PhD; Bernd Bokemeyer[2,4,5], MD

[1]Quantitative and Theoretical Biology, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, North Rhine-Westphalia, Germany

[2]Competence Network IBD, Kiel, Schleswig-Holsteins, Germany

[3]Institute of Human Nutrition and Food Science, Faculty of Agricultural and Nutritional Sciences,, Christian-Albrechts-Universität zu Kiel, Kiel, Schleswig-Holstein, Germany

[4]Klinik für Innere Medizin I, University Hospital Schleswig-Holstein, Kiel, Schleswig-Holstein, Germany

[5]Interdisciplinary Crohn Colitis Centre Minden, Minden, North Rhine-Westphalia, Germany

Corresponding Author:

Thomas Wenske, MSc
Quantitative and Theoretical Biology, Faculty of Mathematics and Natural Sciences
Heinrich Heine University Düsseldorf
Universitätsstr 1
Düsseldorf, North Rhine-Westphalia 40225
Germany
Phone: 49 211 81-10171
Email: thwen100@hhu.de

Abstract

Background: Fitness trackers have become important instruments for health monitoring and long-term management of chronic conditions. In individuals with ulcerative colitis (UC), initial research indicates that these devices may effectively capture health-related parameters pertinent to disease monitoring and management. However, the agreement of widely available commercial devices, such as the Garmin vívosmart 4 (GV4), with medically certified sensors in clinical populations remains uncertain.

Objective: This study aimed to assess the agreement between GV4 step counts and those obtained from a clinically validated motion sensor, SENS motion, in patients with UC under real-world conditions.

Methods: Data were obtained from the FilgoColitis study, a prospective, observational, multicenter study conducted in Germany involving patients with UC who had recently initiated filgotinib therapy. Participants wore both devices concurrently for several days following each study visit. Data availability was described for both devices. Step counts were aggregated into 15-minute, hourly, daily, and patient-level mean intervals. For descriptive analyses, all 15-minute intervals were considered, regardless of whether any steps were recorded. In contrast, inferential statistical analyses included only those intervals where at least one device recorded 1 or more steps. The agreement between the devices was assessed using the concordance correlation coefficient, Bland-Altman analysis, and two 1-sided tests (TOST) for equivalence testing.

Results: Data from 40 patients were analyzed, yielding 37,517 paired 15-minute intervals. The concordance correlation coefficient, based on active intervals, increased with higher aggregation levels: 0.87 for 15-minute intervals, 0.93 for hourly intervals, 0.96 for daily intervals, and 0.93 for patient-level mean step counts. Step count distributions were comparable across devices, with the highest agreement at the daily and patient-level aggregations. Bland-Altman analysis showed a mean difference of 1.7 steps between the devices, with limits of agreement from –40.1 to +43.6 steps. In the exploratory post hoc two 1-sided test analysis, the 90% CI for the mean difference (–3.96 to 7.41 steps) lay within a margin of ±8 steps. Discrepancies were more pronounced at shorter aggregation intervals.

Conclusions: Step counts from the GV4 showed high agreement with a medically certified motion sensor in patients with UC when aggregated at the daily or patient level. In clinical research and digital health applications, the GV4 step count data should therefore be used as daily aggregates. Short-interval measurements should be interpreted with caution, as technical

variability and user behavior may introduce noise. Future studies should investigate the integration of wearables capturing additional health parameters into digital health interventions and evaluate their long-term use for telemedical monitoring and novel end point assessments in clinical trials.

Trial Registration: German Clinical Trials Register DRKS00027327; https://www.drks.de/DRKS00027327

JMIR Form Res 2026;10:e106910; doi: 10.2196/106910

Keywords: ulcerative colitis; wearable devices; fitness tracker; Garmin vívosmart 4; SENS motion; step count agreement

Introduction

Background

Chronic diseases, including ulcerative colitis (UC), a type of inflammatory bowel disease (IBD), significantly impact patients' quality of life. Despite advancements in therapeutic options, there is an ongoing need for innovative strategies to enhance patient care and facilitate the documentation of disease activity in daily life. For individuals with IBD, telemonitoring systems and mobile health applications have demonstrated promising results, such as reduced emergency department visits and hospitalizations, fewer outpatient visits, and earlier detection of disease flares [1-5]. Digital health technologies, particularly wearable devices, are increasingly used for disease monitoring in chronic conditions. Advanced wearables can capture physiological parameters, including step count, heart rate variability, and resting heart rate, which may reflect disease-related changes [6]. Furthermore, step counting is a fundamental feature of most wearable devices and has emerged as an accessible marker for disease monitoring. This is particularly relevant given the well-established positive effects of physical activity (PA) on health and its potential to prevent multiple diseases [7-11].

Consumer-grade wearables, such as the Garmin vívosmart 4 (GV4), offer several advantages over specialized medical devices, such as SENS motion (SM) [12-14]. They are affordable, widely available, and socially accepted products. Indeed, 32% of the German population uses such devices in their daily routines [15]. The use of devices worn regardless of disease status can improve adherence and reduce psychosocial barriers to health monitoring. However, measurement quality varies considerably across manufacturers and device types [13,16-21], and few independent validations exist, particularly in patients with UC under real-world or controlled conditions. Therefore, evaluating the step-count accuracy of the GV4 against a European Conformity (CE)–certified reference system is a crucial prerequisite before using these devices for long-term monitoring and clinical research into UC disease activity and quality of life.

Objective

The primary objective of this study was to evaluate the agreement in step measurements between the commercial GV4 fitness tracker and the CE-certified SM accelerometer system [12-14] in patients with UC under real-world conditions. The agreement was primarily evaluated at the per-patient average level using the concordance correlation coefficient (CCC) proposed by Lin, Bland-Altman analysis, and equivalence testing to address the considerable variability in the number of available measurement intervals across patients. Furthermore, CCC was calculated at various aggregation levels (15-minute, hourly, and daily intervals) to characterize concordance across different temporal resolutions.

Methods

Study Design

Data from the GV4 and SM were gathered as part of the FilgoColitis study, a multicenter, noninterventional, prospective observational study aimed at collecting longitudinal data on a new therapeutic regimen involving a JAK (Janus kinase) inhibitor for patients with active UC, including the option for treatment-switching. This study enrolled patients with UC who initiated filgotinib therapy in Germany [22]. After the induction phase (baseline through week 10), follow-up visits were scheduled approximately every 6 months. A unique feature of the study was the use of 2 devices to monitor PA and other wearable data: a commercial fitness tracker (GV4) and a medically certified leg sensor (SM).

Patients received the GV4 at the start of the study, allowing them to wear it throughout the study period, regardless of scheduled visits. It primarily monitored general activities and basic health metrics. In contrast, the SM, a medical-grade device, was attached to the leg during each visit and worn for about 7 days to collect medically focused activity data. Researchers accessed SM data through a secure online portal. Both devices could be connected, and their data were transferred to the study center using the MyTARGET (CED Service GmbH) app, a dedicated smartphone application for patients with IBD, which also enabled trial participants to submit study visit data directly through the platform. Through the app, participants accessed a secure study area and connected their wearable devices via Bluetooth to transfer activity data directly to the study server. Further details of the app and its development have been described previously [22]. In this setup, data were transmitted directly to the study server in Germany, bypassing the international Garmin server. To minimize behavioral effects, patients were shown only basic device parameters, such as step count or heart rate, and did not receive additional summaries or feedback from either device.

As FilgoColitis is a noninterventional, observational study, the use of wearables was optional for participants. No specific requirements or protocols governed the use of wearables. The study centers reminded patients at each visit to wear the watch, particularly during the 7 days following the visit,

to ensure that wearable data could supplement clinical data, focusing on clinical remission status and predicting disease course using wearables.

Devices

The SM is a validated activity sensor (medical device class I) [12-14]. It comprises a compact, lightweight triaxial accelerometer attached directly to the patient's thigh. This sensor was primarily developed for application in clinical trials or health care environments and has been used in numerous studies [23-29]. Beyond step counting, the system distinguishes among various activity patterns, including rest, standing, walking, running, high-intensity exercise, and cycling. In this study, data were collected at 15-minute intervals.

The GV4 is a commercially available fitness tracker that incorporates multiple sensors, such as an optical heart rate sensor, accelerometer, and pulse oximeter to measure oxygen saturation. In addition to step counting, the device provides data on heart rate, heart rate variability (an indicator of stress), calorie consumption, sleep phases (light, deep, and rapid eye movement), maximal oxygen uptake estimation, and body energy level ("body battery"). The recording frequency is parameter dependent. For instance, the heart rate is recorded several times per minute, whereas sleep data are aggregated by phase.

Ethical Considerations

The present data were collected as part of a clinical trial approved by the Ethics Committee of the Medical Faculty of the University of Kiel (D 609/21). The study was registered in the German Clinical Trials Register (DRKS00027327) and conducted in accordance with § 67 of the German Medicinal Products Act (AMG-Arzneimittelgesetz) [30].

Prior to enrollment, all participants provided written informed consent, which explicitly covered the collection of motion data and the comparative analysis of the technical devices used. Patients were informed of their right to withdraw from the study at any time without consequences for their further medical care.

Data Processing

Garmin data were continuously uploaded to the server and compiled into individual patient files. Timestamps were converted into a standardized format, and stress and heart rate data were summarized in 1-minute intervals. The dataset was then filtered to remove time segments without valid measurements, defined as intervals in which neither heart rate nor stress was recorded, and no valid cumulative daily step totals could be verified, ensuring that only valid wearable-derived data were retained. The data were subsequently aggregated to 15-minute intervals and retained only if the full interval contained valid data.

Data from the SM were captured during active sensor periods and uploaded to the server once the device was removed. As the SM provided data in preaggregated 15-minute intervals, only complete intervals were available for download. Individual files were generated for each patient-visit combination in which activity was recorded. Intervals with incomplete or missing data were excluded from the data processing.

For both devices, each patient was assigned a unique study ID to facilitate pseudonymized data linkage across devices. To allow a direct comparison, a dataset was constructed by identifying all 15-minute intervals for which both the SM and GV4 provided complete, synchronous data. All time periods in which the SM identified cycling as a movement pattern were excluded because the GV4 may inaccurately record steps during cycling, potentially inflating step counts. Each timestamp included the last visit date, visit identification, and the number of days between the last visit date and the timestamp.

For inferential statistical analyses, a subset was extracted from the full dataset, including only those intervals in which at least one device recorded greater than or equal to 1 step. This approach excluded periods of complete physical inactivity, thereby reducing significant zero-inflation in the data and prevented artificial inflation of the overall agreement. The discrepancy between the GV4 and the SM step measurements was calculated within this active subset.

The dataset was aggregated at multiple levels: 15-minute intervals, hourly and daily data, and the total mean value per patient. To ensure data reliability, a systematic outlier analysis was conducted on the differences in step count measurements between the 2 devices. The modified z score method, as described by Iglewicz and Hoaglin [31], was applied for this purpose.

Statistical Analysis

A descriptive analysis was first conducted to summarize participant characteristics and available wearable data. Unless otherwise specified, all inferential and agreement analyses were based on aggregated patient-level data, with 1 value per patient, to prevent undue influence from individuals with a higher number of measurements.

Descriptive statistics, including the mean and SD, were calculated for step counts from each device and for the differences between them. Visualizations included a comparison of step counts and differences across devices, as well as scatter and density plots to illustrate the relationship between SM and Garmin step counts.

Agreement between step count measurements from the 2 devices was assessed using the CCC [32]. The CCC ranges from –1 to +1, with values closer to +1 indicating stronger positive agreement. Unlike the simple correlation coefficient, the CCC accounts for both the linear relationship (precision) and systematic deviation (bias) from the line of perfect agreement (y = x). In addition to the overall CCC across aggregation levels, individual CCC values were calculated per patient to identify potential outliers. To determine whether individual CCC values were influenced by the number of available observations, a Spearman correlation analysis was conducted between patient-level CCC values and the number of recorded intervals. A lag analysis was conducted

to evaluate whether temporal discordance between devices contributed to reduced agreement in individual patients.

A Bland-Altman analysis [33] with 95% limits of agreement was performed. The Bland-Altman plot displays the differences in step counts vs the mean step count of both devices. The limits of agreement define the range within which 95% of the differences between devices are expected to lie, assuming normally distributed differences. The mean difference indicates the systematic bias between devices. A narrower range between the limits indicates greater measurement accuracy and improved interchangeability between devices.

An exploratory equivalence analysis was performed using the two one-sided tests (TOST) procedure [34]. In the absence of a clinically established equivalence threshold, no margin was prespecified. We therefore report the 90% CI of the mean difference and, descriptively, the narrowest margin within which this CI lay. Because this margin was derived from the data, the analysis does not constitute a confirmatory test of equivalence.

All statistical analyses were then carried out using R software (version 4.3; R Foundation for Statistical Computing).

Results

Baseline Characteristics

Among the 196 patients in the FilgoColitis study, 89 (45.4%) selected GV4 and 67 (34.2%) selected SM. Of these, 40 (20.4%) patients used both devices simultaneously.

This subgroup of patients wearing both GV4 and SM had a median age of 40 years and a median BMI of 25.95 kg/m². Male participants comprised 75% (30/40) of the cohort. Half of the patients (20/40) were nonsmokers. The median disease duration was 7.0 years. Based on the Montreal Classification, 62.5% (25/40) of the patients presented with extensive colitis (extensity E3) and 72.5% (29/40) exhibited a moderately active or severe disease course (severity S2+S3). Extraintestinal manifestations were observed in 25% (10/40) of the patients, and comorbidities were present in 40% (16/40) of the patients (Table 1).

Table 1. Baseline demographic and clinical characteristics.

Characteristics	Overall
Number of patients (n)	40
Age (y), median (IQR)	40.00 (32.75-58.00)
Sex (male), n (%)	30 (75)
BMI, median (IQR)	25.95 (23.58-27.20)
Smoker, n (%)	
Nonsmoker	20 (50)
Current	6 (15)
Former	14 (35)
Disease duration in years, median (IQR)	7.00 (3.00-10.70)
Extraintestinal manifestations, n (%)	10 (25)
Hospitalization, past 12 months, n (%)	7 (17.5)
History of comorbidities, n (%)	16 (40)
Disease extent (Montreal Classification), n (%)	
E1: ulcerative proctitis	4 (10)
E2: left-sided UC[a]	11 (27.5)
E3: extensive UC	25 (62.5)
Disease severity (Montreal Classification), n (%)	
S0: remission	1 (2.5)
S1: mild symptoms	10 (25)
S2: moderate symptoms	26 (65)
S3: severe symptoms	3 (7.5)

[a]UC: ulcerative colitis.

Step Counts Activity

Data availability differed between the devices (Table 2). Patients contributed GV4 data in a median of 2 (IQR 1-3) visit periods and SM data in a median of 3 (IQR 3-5) visit periods. Per visit period, GV4 data were available on a median of 24 (IQR 8-51) days, with a median daily data availability of 73% (IQR 53%-83%). In line with the study design, the SM was worn for a shorter period after each visit (median 9, IQR 8-11 days) but with higher daily data availability (median 91%, IQR 88%-94%). Individual data

availability over time is shown in Figures S1 and S2 in Multimedia Appendix 1.

Table 2. Data availability of the Garmin vívosmart 4 (GV4) and SENS motion (SM) per patient and visit period[a].

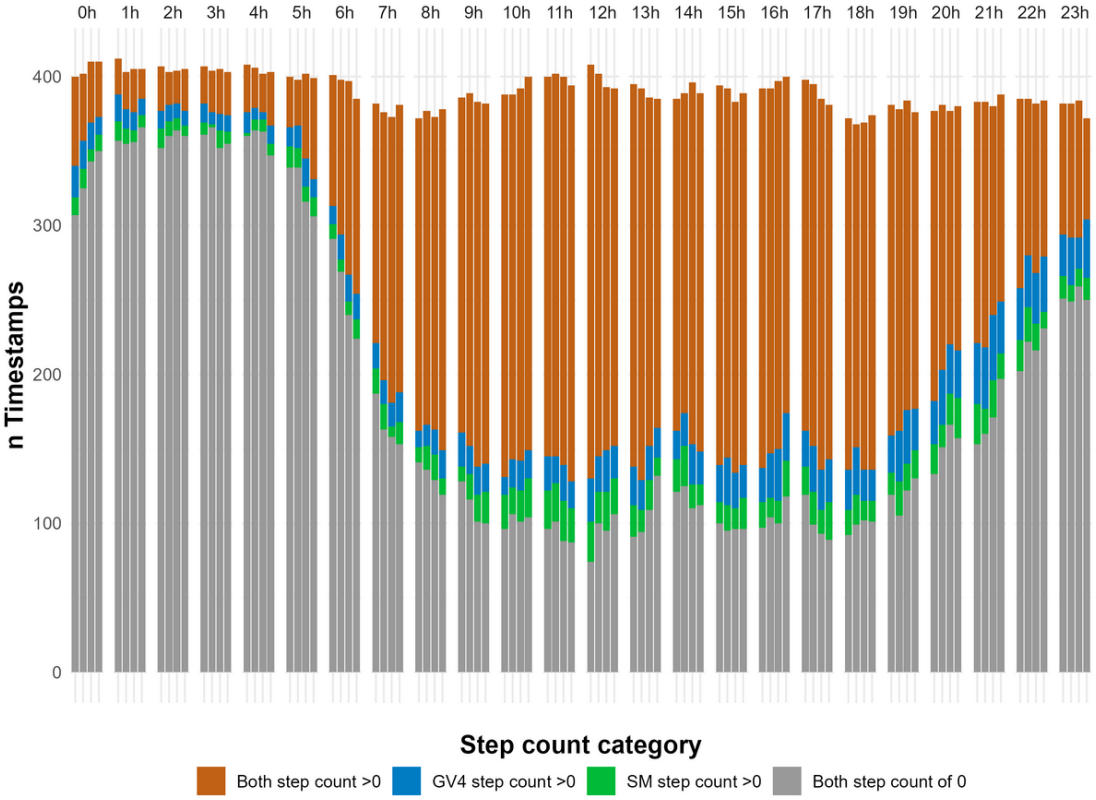
Measure	GV4	SM
Patients, n	40	40
Visits with any data per patient, median (IQR)	2 (1-3)	3 (3-5)
Days with any data per visit period, median (IQR)	24 (8-51)	9 (8-11)
Daily data availability (%), median (IQR)	73 (53-83)	91 (88-94)

[a]A visit period was defined as the time from 1 study visit to the next. Daily data availability refers to the proportion of minutes per day with valid data.

During the study period, 37,517 fifteen-minute intervals were recorded for 40 patients. Outlier analysis using the modified *z* score revealed no extreme systematic outliers. Therefore, no additional data points were excluded in this step. Of the 37,517 intervals, 48.8% (n=18,310) showed no step activity on either device, whereas 51.2% (n=19,207) included step counts from at least one device. Among the intervals with step counts, 3656 (9.8% of all intervals) were classified as discrepant, indicating that only 1 device recorded steps during these periods.

Figure 1 shows the temporal distribution of concordant and discordant step count recordings over a 24-hour period. The proportion of intervals with concurrent step detection peaked during the morning and afternoon hours (9 AM to 5 PM), reaching approximately 60% per hour. During nighttime hours (midnight to 6 AM), minimal step activity was recorded, with 80% to 90% of the intervals showing no steps. Device-specific discrepancies followed a time-dependent pattern: the GV4 recorded steps without a corresponding SM measurement in 5.7% (2128/37,517) of all intervals, while the SM recorded steps without a GV4 in 4.1% (1528/37,517) of intervals. Discrepancy frequency increased from low levels during nighttime hours to a peak in the evening.

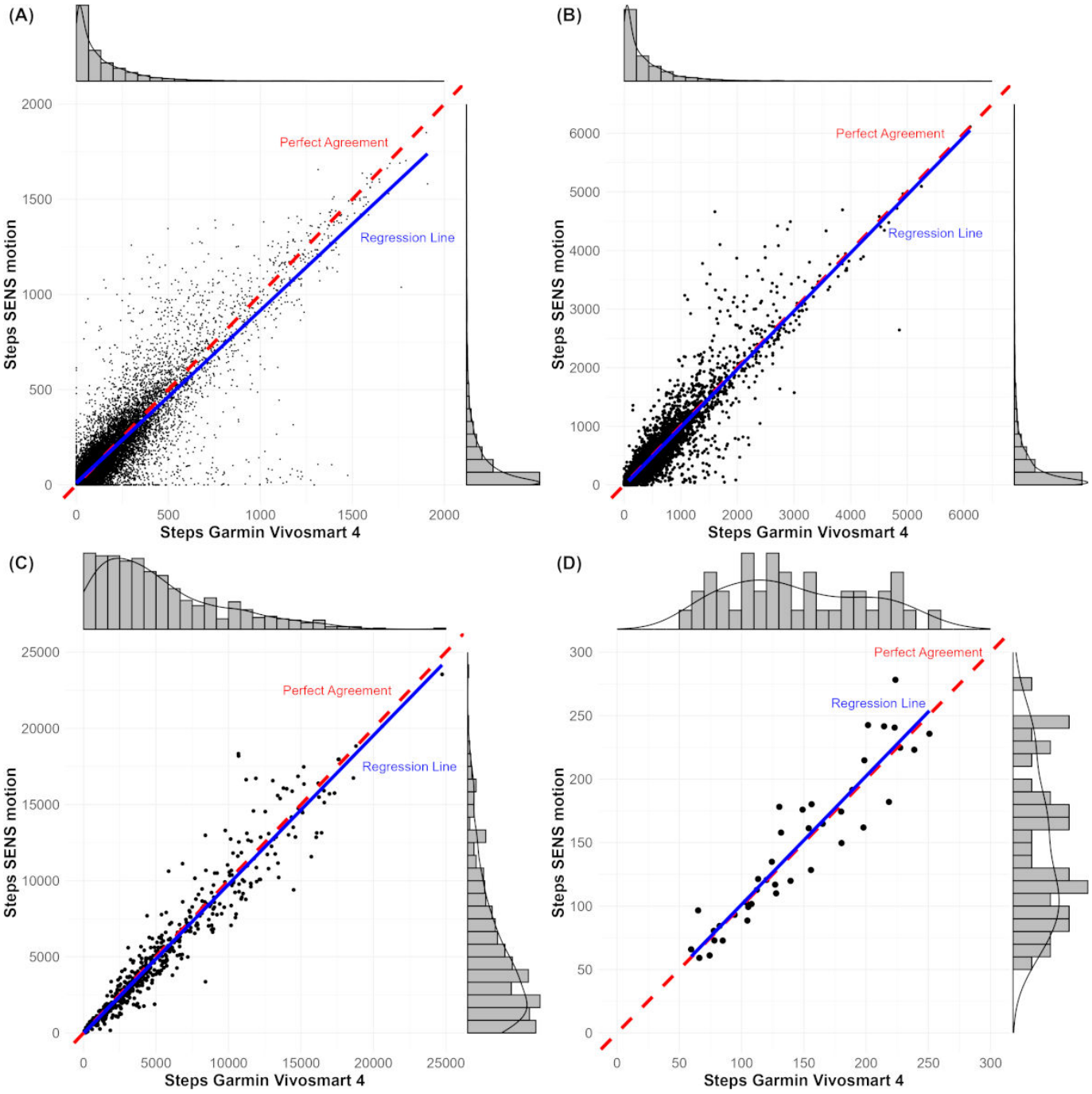
Figure 1. Hour-of-day distribution of step activity across 15-minute intervals. Facets represent hours of the day (0-23). Bar heights indicate the number of 15-minute intervals for each hour (exact values are provided in Table S1 in Multimedia Appendix 1).



Device Agreement

Figure 2 illustrates the agreement in step counts between the SM and GV4 across the 4 aggregation levels: 15-minute, hourly, daily, and patient averages. Marginal density graphs display data distributions.

Figure 2. Scatterplots comparing step counts from Garmin vívosmart 4 (GV4) and SENS motion at 4 aggregation levels. Panels represent step count comparisons at (A) 15-minute, (B) hourly, (C) daily, and (D) per-patient average intervals. The red line shows perfect agreement (y=x), while the blue line depicts linear regression.



At shorter aggregation intervals (15-minute and hourly), right-shifted distributions were observed when the SM reported higher step counts. Clustering at lower values reflects inactivity and intermittent activity. Linear regression revealed strong agreement and minimal bias across all aggregation levels, with slopes approaching unity: 0.92 for 15-minute, 0.99 for hourly, 0.98 for daily, and 1.01 for patient-level mean data. The mean step differences were –4 for 15-minute intervals, –11 for hourly intervals, –141 for daily intervals, and +2 for the patient-level mean. Aggregation improved the agreement between devices, daily aggregation reduced variability, and patient-level averages closely aligned with the identity line.

Figure 2A highlights cases in which step count differences reached up to 1400 steps per interval between the devices.

The agreement between step counts was quantified using the CCC (Table 3). Strong agreement was observed at all aggregation levels. The CCC was 0.90 for 15-minute intervals, 0.93 for hourly aggregation, and reached 0.96 at the daily level. At the patient level, mean values demonstrated strong agreement (CCC=0.93, 95% CI 0.87-0.96). The lowest agreement was observed when intervals with 0 steps were excluded (CCC=0.87).

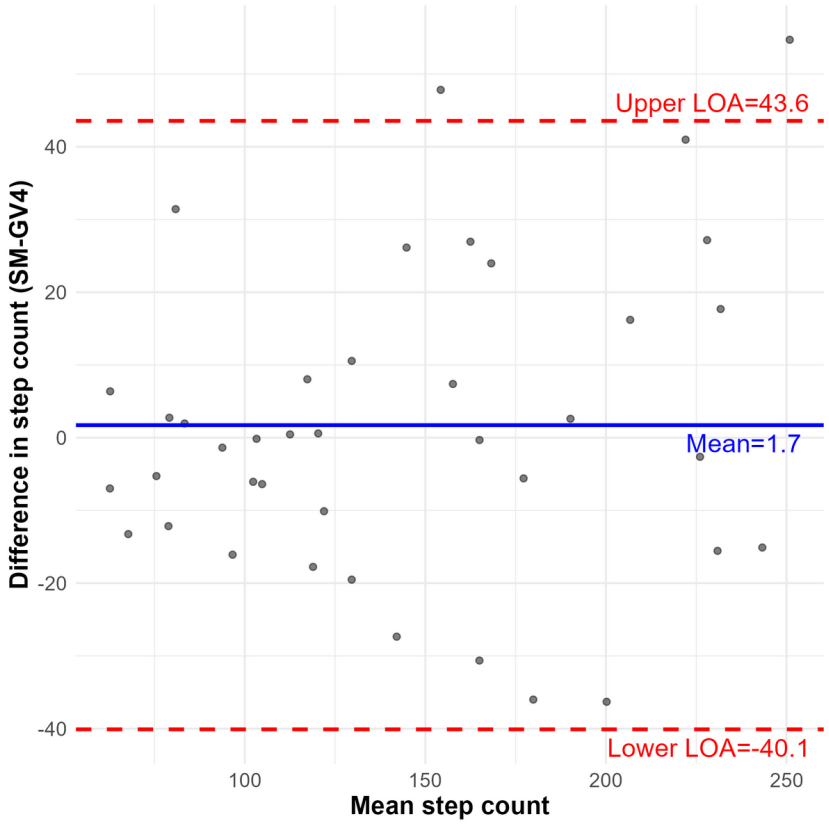
Table 3. Concordance correlation coefficient and corresponding 95% CIs between the SENS motion and Garmin vívosmart 4 (GV4) step counts over 15-minute, hourly, daily, and mean data intervals.

Dataset	CCC (95% CI)
15 min	0.90 (0.90-0.90)
15 min without 0	0.87 (0.87-0.88)
Hourly	0.93 (0.92-0.93)
Daily	0.96 (0.95-0.97)
Mean per ID	0.93 (0.87-0.96)

No statistically significant correlation was found between the CCC and the number of available intervals (r=0.04, P=.79). Patient-level analysis showed that 3 patients had CCC values below 0.70 (Table S2 in Multimedia Appendix 1). Figure S3 illustrates that the largest measurement discrepancies occurred primarily in the 2 patients with the lowest individual CCC values. A 20-minute temporal offset in the GV4 data for one patient (the lowest CCC value; Table S2 in Multimedia Appendix 1) was identified via a lag analysis. Correcting this offset, the individual CCC increased from 0.29 to 0.89 (Figure S4 in Multimedia Appendix 1).

Bland-Altman analysis was performed to assess systematic bias and limits of agreement between the GV4 and SM step count measurements at the patient level. The Bland-Altman plot (Figure 3) displays the difference between the 2 devices relative to their average. A positive difference indicates a higher step count from the SM, whereas a negative difference reflects more steps recorded by the GV4. The analysis identified a small systematic bias of +1.7 steps, which is negligible from a clinical perspective, with limits of agreement ranging from –40.1 to +43.6 steps.

Figure 3. Bland-Altman plot comparing mean step counts between Garmin vívosmart 4 (GV4) and SENS motion (SM) at the patient-level average. The plot shows the difference in mean step counts (SM minus GV4) against the average step count for each patient. The solid blue line indicates the mean difference (+1.7 steps), while dashed lines in red represent the 95% limits of agreement (LoA: –40.1 to +43.6 steps).



In the exploratory equivalence analysis, the 90% CI of the mean patient-level difference (–3.96 to 7.41) lay within ±8 steps (TOST lower: P=.003; TOST upper: P=.04). As the margin was derived post hoc, this result should be regarded as descriptive rather than as confirmatory evidence of equivalence.

Discussion

Principal Findings

The agreement between the step count measurements from the GV4 and SM was evaluated in a real-world clinical setting using 3 analytical approaches: CCCs, Bland-Altman analysis, and equivalence testing. The results demonstrated

consistent strong agreement between the 2 devices, with daily and patient-level aggregations yielding reliable estimates of PA. In contrast, the agreement at 15-minute intervals suggests that these devices are less appropriate for precise, real-time feedback over short durations and are better suited for monitoring changes in PA patterns over extended periods. Notably, these robust agreement metrics were obtained after excluding mutual time intervals with step counts of 0. Addressing the 48.8% zero-inflation in the raw data was necessary to avoid artificially inflating the overall agreement, thereby ensuring that the calculated concordance accurately reflected active physical movement rather than synchronized sedentary or sleep periods.

Linear regression slopes approaching unity across all aggregation levels (range 0.92-1.01) indicated minimal proportional bias between the 2 devices, demonstrating that the step counts scaled consistently throughout the measurement range. However, interpreting these results requires consideration of the absence of a universally accepted classification framework for CCC values. Altman [35] classified all aggregation levels in this study as demonstrating strong agreement (CCC>0.81). In contrast, McBride [36] recommended a more stringent threshold, suggesting that CCC values below 0.90 do not indicate good agreement for laboratory methods. Viscarra Rossel and Hicks [37] proposed an intermediate scale, with values above 0.90 reflecting excellent agreement and values between 0.80 and 0.90 indicating substantial agreement. Given these differing frameworks, using multiple analytical approaches, such as the Bland-Altman analysis to visualize systematic bias and equivalence testing to assess practical similarity, offers a more comprehensive assessment than relying on CCC values alone, consistent with current recommendations for validating consumer wearable step counters [38]. However, the equivalence analysis should not be interpreted as evidence of equivalence, but rather as exploratory supplementary information due to the post hoc margin.

The greater measurement discrepancy observed during daytime hours may result from several factors: (1) differences in sensor positioning (wrist-worn GV4 compared to thigh-mounted SM) [39,40], (2) algorithmic variations in step definition, (3) patient compliance with device wearing, and (4) technical synchronization errors.

Wrist-worn devices may register arm movements as steps during stationary activities, whereas the SM may detect low-intensity movements, such as slow walking or shuffling, that fall below the detection threshold of the wrist device. In addition to sensor-level differences, some discrepancies likely reflect real-world use issues inherent to free-living study conditions, such as improper GV4 wear or temporary nonwear periods. Because strict adherence to the study protocol was not required, controlling these variations was difficult. Although extreme deviations were rare, step count differences of up to 1400 steps within a single 15-minute interval were observed (Figure 2A), supporting the conclusion that isolated technical or compliance-related errors, rather than systematic sensor differences, were the primary contributors to measurement discordance in this dataset. Regardless of their cause, these short-interval discrepancies limit the use of the GV4 for interventions that require real-time behavioral responses, such as prompts triggered by activity within individual 15-minute windows. Conversely, in the context of IBD, the primary clinical application of step count data is less focused on real-time feedback and more on the longitudinal monitoring of activity patterns. It has been demonstrated that PA is reduced during periods of active disease [41]. Furthermore, changes in wearable-derived activity and physiological data have been associated with disease flares [6]. The efficacy of such applications is contingent upon the consistency of changes in daily or weekly step counts within individual patients. In this study, substantial agreement was observed between the devices (CCC=0.96 for daily data). For this purpose, short-interval discrepancies are therefore of limited relevance, provided that analyses are based on daily aggregates. Should real-time applications be considered for patients with IBD in the future, the performance of consumer devices at short intervals would necessitate specific evaluation.

The 2 patients with the lowest CCC values exemplified these challenges. The temporal offset observed in 1 patient most likely reflects a synchronization error, potentially due to incorrect device time, transmission error, or temporary malfunction rather than a genuine measurement discrepancy. The near absence of SM recordings in the second patient likely indicates either a wear compliance issue or a sensor malfunction, both of which are consistent with the difficulties in ensuring consistent device use in free-living settings. The lag analysis demonstrates the sensitivity of agreement metrics to timestamp alignment. In multidevice setups, each device maintains its own clock. Timestamps can be stored in local time or UTC (Coordinated Universal Time), adjusted for daylight saving time at different moments, or shifted when a device synchronizes with a smartphone after a period offline. In the context of decentralized clinical trials and telemonitoring, the integration of data from multiple devices occurs without the need for supervised data collection. Consequently, such discrepancies are likely to go unnoticed by patients and study personnel, potentially leading to the silent corruption of paired analyses. Consequently, the alignment of timestamps should not be delegated to post hoc analyses but rather implemented automatically on the server side. One potential solution involves the storage of all timestamps with the server time during each synchronization.

Placing these results within the context of the existing literature requires an overview of comparable studies on the accuracy of fitness trackers, particularly the GV4, in diseases other than UC. Wearables have been increasingly studied in IBD, with a recent scoping review of 37 studies identifying PA and step counts as the most frequently reported parameters [41]. However, studies examining the agreement between specific consumer devices against medical-grade reference sensors in IBD remain scarce. Few studies have specifically addressed this aspect of the GV4. In a controlled study involving patients with Parkinson disease, Bianchini et al [19] reported that the GV4 provides reliable values for healthy individuals at normal walking speeds but loses accuracy with

slower gaits or when worn on the more severely affected side. Fokkema et al [20] evaluated the step count accuracy of 10 commercially available devices at various walking speeds in a laboratory treadmill setting. The GV4 demonstrated high accuracy and good test-retest reliability at slow and average walking speeds; however, its accuracy declined considerably at fast walking speeds. Other research on the GV4 has primarily focused on sleep-tracking features [42-44].

Fuller et al [45] conducted a systematic review of commercial wearables and found that Garmin devices generally demonstrated adequate accuracy in step measurement within controlled laboratory environments, although they tended to slightly underestimate the number of steps. In direct brand comparisons, Apple and Samsung devices exhibited lower measurement variability and higher overall precision, whereas Garmin ranked in the middle range for accuracy. Similarly, a study comparing the consumer-grade Fitbit Charge 4 with a medical-grade step counter in patients with chronic obstructive pulmonary disease under free-living conditions reported good agreement for daily step counts (ICC=0.79-0.85), though the consumer device tended to overestimate steps [46], a pattern broadly consistent with the present findings and highlighting the relevance of examining device agreement under free-living conditions in chronic disease populations. To date, the agreement between step counts from the GV4 and SM has not been examined, a gap addressed by the present study.

Strengths and Limitations

This study had several notable strengths. First, the SM served as a CE-certified, medical-grade reference device, offering a validated comparison standard for the consumer-grade GV4. Second, all data were transmitted directly to a secure server in Germany, ensuring data integrity without third-party processing or modification. This approach enabled a transparent and unbiased examination of the raw step-count data from both devices. Third, the real-world setting, in which patients wore devices during normal daily activities over extended periods, provided evidence of device performance beyond controlled laboratory conditions.

This study has several limitations. Because the data were not collected under a controlled movement protocol, the actual number of steps taken by the patients could not be determined, and different walking speeds were not investigated. Wearable devices have been shown to exhibit greater measurement variability under free-living conditions than in controlled settings [45,47], which may have further contributed to the observed discrepancies. Additional metrics from the GV4 were not considered. Data availability was lower for the GV4 than for the SM. Patients contributed GV4 data in fewer visit periods and with lower daily data availability and GV4 use declined over the study period. This was likely exacerbated by technical connectivity issues that could only be tested with a limited number of operating systems or preferred their own wearable devices. Furthermore, although the sample size was comparable to that of similar studies [18,48-50], the study population consisted of patients with UC, who may be less likely to consistently wear activity trackers than younger, health-conscious individuals. The fact that the devices were provided specifically for this study, rather than being personally owned, may have further reduced wear compliance.

Conclusions

Overall, these results highlight the potential of commercial wearables, such as the GV4, to provide a cost-effective and scalable approach for continuous, free-living, step-count-based PA monitoring in patients with UC. Unlike specialized medical sensors, these devices are already prevalent and socially accepted, which may facilitate their use in everyday life. In this study, however, data availability declined over time, possibly because the devices were provided for the study rather than being personally owned.

It is important to note that measurement agreement decreased at shorter time intervals. Step count data from GV4 should therefore preferably be used as daily aggregates, and researchers should account for this when selecting aggregation levels in future study designs. Combining wearable sensor data with clinical outcomes in future studies could help define novel end points and potentially enhance treatment monitoring and therapeutic effectiveness.

Acknowledgments

The authors thank the clinical investigators and study teams for their dedicated work. The authors also wish to acknowledge the participants for their essential contributions to this project.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading, editing, and translation. The GenAI tools used were DeepL and Claude Opus 5.5 (Anthropic). All AI-generated content, including text suggestions and code, was reviewed, verified, and edited by the authors. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by the pharmaceutical company Alfasigma S.p.A. The funder was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The data underlying this article cannot be shared publicly because of the privacy of the individuals who participated in the study. The data will be shared upon reasonable request from the corresponding author.

Authors' Contributions

BB and SP-D contributed to the study protocol design. TW conducted the statistical analysis. TW, OP, SP-D, OE, and BB interpreted the data. TW wrote the first draft of this manuscript. All authors approved the final version submitted for publication.

Conflicts of Interest

BB has received consulting fees from AbbVie, Merck Sharp & Dohme, Shire, Ferring, Hospira, Takeda, AlphaSigma, Movetis, Shield Therapeutics, Pfizer, Biogen, Janssen, Hexal, Lilly, and Celgene, outside the submitted work; financial support for lectures and teaching from AbbVie, Ferring, AlphaSigma, Merck Sharp & Dohme, Merckle, Falk, HLR, Shield Therapeutics, Pfizer, Celltrion, Takeda, Lilly, Janssen, and Mundipharma, outside the submitted work; and grant and financial support for research from AbbVie, Pfizer, Lilly, Janssen, AlphaSigma, Takeda, and Ferring, outside the submitted work. All other authors declare no conflicts of interest.

Multimedia Appendix 1

Additional analyses of step-count agreement and device data availability.
[DOCX File (Microsoft Word File), 1302 KB-Multimedia Appendix 1]

References

1. Del Hoyo JD, Nos P, Faubel R, et al. Adaptation of TECCU app based on patients´ perceptions for the telemonitoring of inflammatory bowel disease: a qualitative study using focus groups. Int J Environ Res Public Health. Mar 13, 2020;17(6):1871. [doi: 10.3390/ijerph17061871] [Medline: 32183103]
2. Zand A, Nguyen A, Reynolds C, Khandadash A, Esrailian E, Hommes D. Patient experience and satisfaction with an e-health care management application for inflammatory bowel diseases. Int J Environ Res Public Health. Nov 9, 2021;18(22):11747. [doi: 10.3390/ijerph182211747] [Medline: 34831502]
3. van Deen WK, Spiro A, Burak Ozbay A, et al. The impact of value-based healthcare for inflammatory bowel diseases on healthcare utilization: a pilot study. Eur J Gastroenterol Hepatol. Mar 2017;29(3):331-337. [doi: 10.1097/MEG.0000000000000782] [Medline: 27926663]
4. Zhen J, Marshall JK, Nguyen GC, Atreja A, Narula N. Impact of digital health monitoring in the management of inflammatory bowel disease. J Med Syst. Jan 15, 2021;45(2):23. [doi: 10.1007/s10916-021-01706-x] [Medline: 33449213]
5. Burton AM, Perler BK. Mobile health technologies in inflammatory bowel disease: a narrative review. BMC Gastroenterol. Aug 18, 2025;25(1):595. [doi: 10.1186/s12876-025-04189-z] [Medline: 40826022]
6. Hirten RP, Danieletto M, Sanchez-Mayor M, et al. Physiological data collected from wearable devices identify and predict inflammatory bowel disease flares. Gastroenterology. May 2025;168(5):939-951. [doi: 10.1053/j.gastro.2024.12.024] [Medline: 39826619]
7. Warburton DER, Nicol CW, Bredin SSD. Health benefits of physical activity: the evidence. CMAJ. Mar 14, 2006;174(6):801-809. [doi: 10.1503/cmaj.051351] [Medline: 16534088]
8. Simpson RJ, Kunz H, Agha N, Graff R. Exercise and the regulation of immune functions. Prog Mol Biol Transl Sci. 2015;135:355-380. [doi: 10.1016/bs.pmbts.2015.08.001] [Medline: 26477922]
9. Durstine JL, Gordon B, Wang Z, Luo X. Chronic disease and the link to physical activity. J Sport Health Sci. Mar 2013;2(1):3-11. [doi: 10.1016/j.jshs.2012.07.009]
10. Nunan D, Mahtani KR, Roberts N, Heneghan C. Physical activity for the prevention and treatment of major chronic disease: an overview of systematic reviews. Syst Rev. Jul 10, 2013;2:56. [doi: 10.1186/2046-4053-2-56] [Medline: 23837523]
11. Lundqvist S, Cider Å, Larsson MEH, Hagberg L, Björk MP, Börjesson M. The effects of a 5-year physical activity on prescription (PAP) intervention in patients with metabolic risk factors. PLoS One. 2022;17(10):e0276868. [doi: 10.1371/journal.pone.0276868] [Medline: 36315564]
12. Ghaffari A, Rahbek O, Lauritsen REK, Kappel A, Kold S, Rasmussen J. Criterion validity of linear accelerations measured with low-sampling-frequency accelerometers during overground walking in elderly patients with knee osteoarthritis. Sensors (Basel). Jul 15, 2022;22(14):5289. [doi: 10.3390/s22145289] [Medline: 35890969]
13. Pedersen BS, Kristensen MT, Josefsen CO, Lykkegaard KL, Jønsson LR, Pedersen MM. Validation of two activity monitors in slow and fast walking hospitalized patients. Rehabil Res Pract. 2022:9230081. [doi: 10.1155/2022/9230081] [Medline: 35615755]

14. Milther C, Winther L, Stahlhut M, et al. Validation of an accelerometer system for measuring physical activity and sedentary behavior in healthy children and adolescents. Eur J Pediatr. Aug 2023;182(8):3639-3647. [doi: 10.1007/s00431-023-05014-z] [Medline: 37258775]
15. Zukunft der Consumer Technology [Article in German]. bitkom. URL: https://www.bitkom.org/zukunft-consumer-technology [Accessed 2025-05-13]
16. Wendel N, Macpherson CE, Webber K, et al. Accuracy of activity trackers in Parkinson disease: should we prescribe them? Phys Ther. Aug 1, 2018;98(8):705-714. [doi: 10.1093/ptj/pzy054] [Medline: 29718452]
17. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. Jan 2008;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
18. Degroote L, De Bourdeaudhuij I, Verloigne M, Poppe L, Crombez G. The accuracy of smart devices for measuring physical activity in daily life: validation study. JMIR mHealth uHealth. Dec 13, 2018;6(12):e10972. [doi: 10.2196/10972] [Medline: 30545810]
19. Bianchini E, Caliò B, Alborghetti M, et al. Step-counting accuracy of a commercial smartwatch in mild-to-moderate PD patients and effect of spatiotemporal gait parameters, laterality of symptoms, pharmacological state, and clinical variables. Sensors (Basel). Dec 25, 2022;23(1):214. [doi: 10.3390/s23010214] [Medline: 36616812]
20. Fokkema T, Kooiman TJM, Krijnen WP, van der Schans CP, de Groot M. Reliability and validity of ten consumer activity trackers depend on walking speed. Med Sci Sports Exerc. Apr 2017;49(4):793-800. [doi: 10.1249/MSS.0000000000001146] [Medline: 28319983]
21. Svarre FR, Jensen MM, Nielsen J, Villumsen M. The validity of activity trackers is affected by walking speed: the criterion validity of Garmin Vivosmart® HR and StepWatch™ 3 for measuring steps at various walking speeds under controlled conditions. PeerJ. 2020;8:e9381. [doi: 10.7717/peerj.9381] [Medline: 32742766]
22. Plachta-Danielzik S, Grasskemper L, Schmidt K, Schreiber S, Bokemeyer B. Health status, quality of life, psychosocial well-being, and wearables data of patients with active ulcerative colitis receiving filgotinib therapy (FilgoColitis study): protocol for a real-world observational study. JMIR Res Protoc. May 8, 2023;12(1):e42574. [doi: 10.2196/42574] [Medline: 37155235]
23. Bartholdy C, Skou ST, Bliddal H, Henriksen M. Changes in physical inactivity during supervised educational and exercise therapy in patients with knee osteoarthritis: a prospective cohort study. Knee. Dec 2020;27(6):1848-1856. [doi: 10.1016/j.knee.2020.09.007] [Medline: 33197825]
24. Kjeldsen LL, Læssøe U, Bendix JM, Maimburg RD. Maternal positions in childbirth - a cohort study of labouring women's movements and body positions the last 24 hours before birth. Sex Reprod Healthc. Mar 2025;43:101059. [doi: 10.1016/j.srhc.2024.101059] [Medline: 39729687]
25. Skov CD, Holsgaard-Larsen A, Kock Wiil U, Lindberg-Larsen M, Varnum C, Jensen CM. The patient's perspective on rehabilitation with wireless accelerometers, activity tracking and motivational feedback following knee replacement: a qualitative study prior to a randomised controlled trial (KneeActivity). Int J Med Inform. Dec 2024;192:105624. [doi: 10.1016/j.ijmedinf.2024.105624] [Medline: 39265242]
26. Søndergaard A, Gregersen M, Wilms M, et al. Inflammatory markers, somatic complaints, use of medication and health care in 11-year-old children at familial high risk of schizophrenia or bipolar disorder compared with population-based controls. The Danish High Risk and Resilience Study - via 11. Nord J Psychiatry. Aug 2024;78(6):507-517. [doi: 10.1080/08039488.2024.2369145] [Medline: 38923920]
27. Oldroyd A, Dixon W, Chinoy H, Howells K. Patient insights on living with idiopathic inflammatory myopathy and the limitations of disease activity measurement methods - a qualitative study. BMC Rheumatol. 2020;4(1):47. [doi: 10.1186/s41927-020-00146-3] [Medline: 32974608]
28. Crowley P, Ikeda E, Islam SMS, et al. The surveillance of physical activity, sedentary behavior, and sleep: protocol for the development and feasibility evaluation of a novel measurement system. JMIR Res Protoc. Jun 6, 2022;11(6):e35697. [doi: 10.2196/35697] [Medline: 35666571]
29. Dall CH, Andersen H, Povlsen TM, Henriksen M. Evaluation of a technology assisted physical activity intervention among hospitalised patients: a randomised study. Eur J Intern Med. Nov 2019;69:50-56. [doi: 10.1016/j.ejim.2019.08.019] [Medline: 31494019]
30. Gesetz über den verkehr mit arzneimitteln (arzneimittelgesetz - AMG) § 67 allgemeine anzeigepflicht. Bundesministeriumder Justiz undfür Verbraucherschutz, Bundesamtfür Justiz. URL: https://www.gesetze-im-internet.de/amg_1976/__67.html [Accessed 2026-11-08]
31. Iglewicz B, Hoaglin D. How to Detect and Handle Outliers. ASQC Quality Press; 1993. ISBN: 9780873892476
32. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. Mar 1989;45(1):255-268. [doi: 10.2307/2532051] [Medline: 2720055]

33. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. Feb 8, 1986;1(8476):307-310. [doi: 10.1016/S0140-6736(86)90837-8] [Medline: 2868172]
34. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. Dec 1987;15(6):657-680. [doi: 10.1007/BF01068419] [Medline: 3450848]
35. Altman DG. Practical Statistics for Medical Research. Chapman and Hall; 1991. [doi: 10.1201/9780429258589]
36. McBride GB. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. National Institute of Water and Atmospheric Research (NIWA); 2005. URL: https://www.medcalc.org/download/pdf/McBride2005.pdf [Accessed 2024-11-25]
37. Viscarra Rossel RA, Hicks WS. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. Eur J Soil Sci. May 2015;66(3):438-450. [doi: 10.1111/ejss.12237]
38. Johnston W, Judice PB, Molina García P, et al. Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network. Br J Sports Med. Jul 2021;55(14):780-793. [doi: 10.1136/bjsports-2020-103147] [Medline: 33361276]
39. Tudor-Locke C, Barreira TV, Schuna JM. Comparison of step outputs for waist and wrist accelerometer attachment sites. Med Sci Sports Exerc. Apr 2015;47(4):839-842. [doi: 10.1249/MSS.0000000000000476] [Medline: 25121517]
40. Montoye AHK, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA. Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior. AIMS Public Health. 2016;3(2):298-312. [doi: 10.3934/publichealth.2016.2.298] [Medline: 29546164]
41. Pathak P, Choudhury A, Marikanty A, Jena A, Sebastian S, Sharma V. Wearable technologies in inflammatory bowel disease: a scoping review. Clin Gastroenterol Hepatol. Mar 2026;24(3):646-666. [doi: 10.1016/j.cgh.2025.06.013] [Medline: 40582389]
42. Mouritzen NJ, Larsen LH, Lauritzen MH, Kjær TW. Assessing the performance of a commercial multisensory sleep tracker. PLoS One. 2020;15(12):e0243214. [doi: 10.1371/journal.pone.0243214] [Medline: 33306678]
43. Stone JD, Rentz LE, Forsey J, et al. Evaluations of commercial sleep technologies for objective monitoring during routine sleeping conditions. Nat Sci Sleep. 2020;12:821-842. [doi: 10.2147/NSS.S270705] [Medline: 33149712]
44. Schyvens AM, Van Oost NC, Aerts JM, et al. Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP versus polysomnography: systematic review. JMIR mHealth uHealth. Mar 27, 2024;12(1):e52192. [doi: 10.2196/52192] [Medline: 38557808]
45. Fuller D, Colwell E, Low J, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR mHealth uHealth. Sep 8, 2020;8(9):e18694. [doi: 10.2196/18694] [Medline: 32897239]
46. Hermans F, Arents E, Blondeel A, et al. Validity of a consumer-based wearable to measure clinical parameters in patients with chronic obstructive pulmonary disease and healthy controls: observational study. JMIR mHealth uHealth. Nov 6, 2024;12:e56027. [doi: 10.2196/56027] [Medline: 39504450]
47. Adamakis M. Validity of wearable monitors and smartphone applications for measuring steps in semi-structured and free-living settings. Technologies. 2023;11(1):29. [doi: 10.3390/technologies11010029]
48. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act. Dec 18, 2015;12(1):159. [doi: 10.1186/s12966-015-0314-1] [Medline: 26684758]
49. Wahl Y, Düking P, Droszez A, Wahl P, Mester J. Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. Front Physiol. 2017;8:725. [doi: 10.3389/fphys.2017.00725] [Medline: 29018355]
50. Reid RER, Insogna JA, Carver TE, et al. Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. J Sci Med Sport. Jun 2017;20(6):578-582. [doi: 10.1016/j.jsams.2016.10.015] [Medline: 27887786]

Abbreviations

CCC: concordance correlation coefficient
CE: European Conformity
GV4: Garmin vívosmart 4
IBD: inflammatory bowel disease
JAK: Janus kinase
PA: physical activity
SM: SENS motion
TOST: two 1-sided tests
UC: ulcerative colitis

UTC: Coordinated Universal Time

Edited by Javad Sarvestan; peer-reviewed by Adrián Marco Moyano; submitted 13.Jul.2026; initial comments to author 21.Sep.2026; final revised version received 11.Oct.2026; accepted 26.Oct.2026; published 26.Nov.2026

Please cite as:
Wenske T, Popa O, Plachta-Danielzik S, Ebenhöh O, Bokemeyer B
Agreement of Step Counting Between a Fitness Tracker and a Medical Device in Patients With Ulcerative Colitis: Prospective Observational Study
JMIR Form Res 2026;10:e106910
URL: https://formative.jmir.org/2026/1/e106910
doi: 10.2196/106910

© Thomas Wenske, Ovidiu Popa, Sandra Plachta-Danielzik, Oliver Ebenhöh, Bernd Bokemeyer. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.