Can fitness apps work long term? A 24-month quasiexperiment of 516 818 Canadian fitness app users

Lisa Nguyen ,[1] Guy Faulkner ,[2] Carolyn Taylor,[3] Marc S Mitchell [1]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bjsports-2025-109901).

[1]School of Kinesiology, University of Western Ontario, London, Ontario, Canada
[2]School of Kinesiology, The University of British Columbia, Vancouver, British Columbia, Canada
[3]Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada

Correspondence to
Dr Marc S Mitchell;
marc.mitchell@uwo.ca

Accepted 16 August 2025

ABSTRACT

Objective To examine whether a multicomponent commercial fitness app with very small ('micro') financial incentives (FI) increased population-level device-assessed physical activity (PA) over 2 years. The secondary objective was to explore the influence of select covariates on longitudinal effects.

Methods This 24 month pre–post quasiexperiment was conducted in Ontario, Canada's largest province (December 2016–June 2019). Following a 1-to-2 week baseline period, users earned micro-FIs ($0.04 CAD/day) for achieving daily step goals. Multiple linear regression models estimated changes in weekly mean daily step count from baseline to key timepoints (eg, 24 months). To address the secondary objective, separate models were developed for each level of the selected covariates (eg, start season, baseline PA).

Results The sample included 516 818 users (% female: 62.83; age (SD): 33.46 (12.65) years). Half were 'low' active at baseline (<5000 daily steps; 47.15%). Overall, daily step counts were greater than baseline at all key timepoints (eg, 242 steps/day at 24 months; p<0.001). Users from earlier start seasons and longer FI exposure exhibited larger differences from baseline (eg, 758 steps/day at 24 months; p<0.001). Differences were also more pronounced among 'low' active users (eg, 1986 steps/day at 24 months; p<0.001). Substantial daily step count reductions were observed among 'very high' active users (≥10 000 daily steps; eg, −3969 steps/day at 24 months; p<0.001).

Conclusion Modest PA increases of about 250 steps per day were sustained over 2 years. For important subgroups (ie, earlier start seasons, 'low' active) increases approached or surpassed 1000 steps/day—a level indicative of clinical significance. Substantial daily step count reductions among higher active users were also observed.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Fitness apps stimulate physical activity in the short and medium term (12 months or less) with evidence lacking regarding long-term effects (>12 months).
⇒ It is unclear how contextual factors (ie, population and intervention characteristics) influence long-term fitness app effects.

WHAT THIS STUDY ADDS

⇒ A multicomponent commercial fitness app intervention with financial incentives can sustain population-level device-assessed physical activity increases over 2 years.
⇒ While *average* daily step count increases did not meet the 1000 step/day threshold level indicative of clinical significance, more app users exhibited clinically significant improvements (40%) than reductions (25%).
⇒ Population and intervention characteristics positively (eg, financial incentives) and negatively (eg, physical activity level) influenced long-term effects.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings may spur academic-industry research partnerships, as well as strengthen quasiexperimental fitness app research designs.
⇒ Fitness app features tested here may inform better fitness app designs in the future (eg, limited-time only financial incentives tied to adaptive daily step goals).
⇒ Illustrating the long-term potential of fitness app intervention may promote their 'prescription' in chronic disease prevention and management settings.

INTRODUCTION

Despite there being over 100 000 commercial fitness apps in the major app stores,[1] very little is known regarding their long-term (ie, >12 months) effects. Systematic reviews of RCTs have uncovered no studies to-date examining fitness app effects beyond 1 year.[2 3] An umbrella review of systematic reviews recently called for more primary research to address this gap.[4] This is important for several reasons. First, the limited research to-date suggests short-term and medium-term physical activity (PA) increases may wane beyond a year.[5] Second, many health benefits associated with PA accrue over longer periods (eg, weight loss maintenance).[6] Third, longer evaluations are more likely to identify PA habituation in the face of common barriers (eg, transitions between astronomical seasons or major life events like a job change or pregnancy).[7] Fourth, knowing more about long-term fitness app effects may be valuable for app publishers relying on annual paid subscriptions from satisfied customers.[8] Finally, users spend around US$4 billion each year on fitness app subscriptions, in-app purchases and premium features, and ought to have access to evidence regarding long-term effects.[9]

Fundamental to the design of impactful fitness apps in the real world is an understanding of their long-term effects. Conducting longitudinal RCTs,

To cite: Nguyen L, Faulkner G, Taylor C, *et al*. *Br J Sports Med* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bjsports-2025-109901

however, in fast-paced digital environments can be challenging (eg, retention challenges, software development costs).[10] Reliance on traditional RCT methods may be limiting fitness app benefits.[11] Robust quasiexperiments incorporating strategies for enhancing internal validity (eg, counterfactual comparisons, subgroup analyses) may shed light on this question of long-term effectiveness, while the RCT evidence base grows.[12 13] This includes gaining greater insight into the contextual factors (ie, population (eg, geographic location) and intervention (eg, behaviour change techniques, replicable intervention components) characteristics)[14] potentially influencing long-term effects. The purpose of this quasiexperimental study, therefore, is to examine whether a multicomponent commercial fitness app can increase population-level and device-assessed PA over 2 years. The secondary objective is to explore the influence of select covariates on longitudinal effects (eg, start season, baseline PA).

METHODS

This study was approved by Western University's Human Research Ethics Board (online supplemental appendix file A) and follows Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines. Adverse events (eg, exercise-related injury) were not monitored during this study.

Study setting

The *Carrot Rewards* (*Carrot*) app was a multicomponent commercial fitness app with very small ('micro') financial incentives (FIs). It was created by Carrot Insights as part of a public–private collaboration with the Public Health Agency of Canada and Canada's provincial/territorial Ministries of Health.[15] *Carrot* was free-to-download on the Apple iTunes and Google Play app stores in Canada's most populous province, Ontario, starting in the Summer of 2016 (ie, 'soft' launch to restricted audience). The app formally launched in Ontario on 9 February 2017 (ie, 'full' launch). Insufficient government funding ultimately led to the app's discontinuation on 19 June 2019 (see online supplemental appendix file B for more on study setting).[16]

Study design

A longitudinal pre–post quasiexperimental open trial study design was used to assess long-term fitness app effects (ie, up to 24 months). Participants downloaded the app on different days between December 2016 and December 2018 with data collection continuing until the day before app discontinuation (ie, 18 June 2019). On download users could initiate *Carrot*'s cornerstone feature, 'Steps'. Before 'Steps' started in earnest, users entered into a 1-to-2 week baseline or 'preintervention' period (ie, no personalised daily step goals, no PA incentives). During the baseline, users were instructed daily to 'wear their device' (ie, smartphone or Fitbit) as much as possible. At least five out of 14 days (before 26 July 2017) and three out of 7 days (on or after 26 July 2017) with valid step counts (ie, 1000–40 000 steps/day)[17] were required to generate a baseline step count (the 'counterfactual'). While these criteria were established by *Carrot* staff and not the study authors, they align with the minimum number of days needed for valid weekly mean daily step count estimates (ie, three).[18] Thereafter, weekly mean daily step counts were calculated for each valid study week (ie, four or more valid days). 'Postintervention' refers to study weeks following baseline.

For economic reasons outside the researchers' control, *Carrot* withdrew daily PA rewards ('deimplementation') near the end of the study period (ie, December 2018). Notably, users from earlier start seasons (ie, Winter 2016/2017 to Fall 2017) received full exposure to 'Steps', including daily PA rewards, for at least a year. Users from later start seasons received full 'Steps' exposure for less than a year (ie, Winter 2017/2018 to Fall 2018). Since daily PA reward withdrawal took effect in December 2018, the Winter 2017/2018 start season was considered a transition cohort, or 'washout' period, during interpretation. While this withdrawal was driven by economic necessity rather than for hypothesis testing, it introduced the programme variance (ie, natural experiment) needed for an embedded internal validity enhancement (see online supplemental appendix file B for more study design detail).

Intervention

Carrot was theoretically grounded in behavioural economics and self-determination theory.[19 20] While behavioural economics, an offshoot of traditional economics complemented by insights from psychology, describes how incentives exploit 'present bias' to *stimulate* behaviours,[21] self-determination theory, a global theory of human motivation, focuses on the extent to which behaviours are controlled by external agents (eg, physicians) or contingencies (eg, FIs) and can be *sustained*.[22] A full intervention description is in online supplemental appendix file C.

Outcome

The primary study outcome was weekly mean daily step count. Step count data were collected from the HealthKit app (Apple) on iOS or Google Fit (Google LLC) on the Android OS. With a single app-open, *Carrot* recorded step count data from the previous 7 days. Users tracked their steps using built-in iPhone (ie, 5S or higher) and Android smartphone (eg, HTC) accelerometers as well as Fitbit trackers. Step count data collected using smartphones can be influenced by a number of factors including an app's algorithm and users' carrying habits.[18] Recent validation studies have found iPhone and Android device step counting apps, as well as those for Fitbit trackers, are accurate in laboratory settings.[23 24] In free-living conditions, where smartphones may not always be carried consistently, step counts may not be as accurate.[25] If carrying time is optimised, however, with daily reminders to carry devices as much as possible (ie, as in the baseline here) and with multicomponent fitness app intervention (ie, as in the intervention here), for example, it is suggested smartphones can accurately assess step counts.[18 26]

Covariates

Age and gender were self-reported and race/ethnicity was not available. Start season levels were defined using Eastern Standard Time astrological start dates and times.[27] App engagement levels were defined using the proportion of total possible study weeks with step count data retrieved (rare: <25%; limited: 26%–50%; occasional: 51%–75%; regular: 76%–100%). The first digit of Ontario's 3-digit forward sortation area post code was used to denote participants' geographic location.[28] Finally, baseline PA level was defined using established daily step count thresholds (low: <5000; medium: 5000–7499; high: 7500–9999; very high: ≥10 000; see online supplemental appendix table D for more covariate detail).[29]

Analyses

Statistical analysis was conducted using R V.4.4.0 (24 April 2024)–'Puppy Cup'. Participants with valid baseline step counts and at least one other valid study week from study

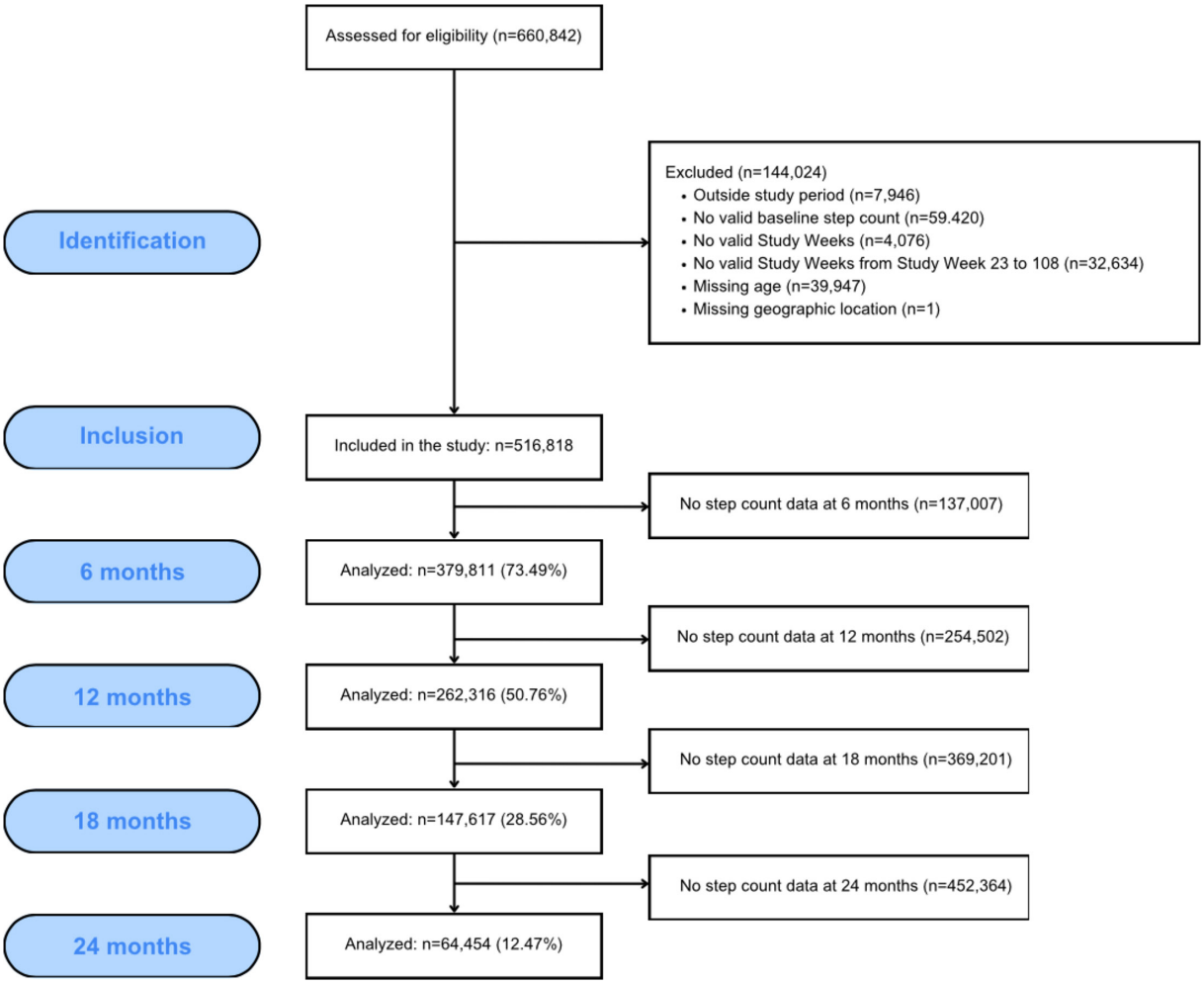

Figure 1 Study flowchart. Study period = 21 December 2016 to 18 June 2019. Valid baseline step count ≥5 days during 14-day baseline period with step count from 1000 to 40 000 steps/day, or ≥3 days during 7 day baseline period. Valid study week ≥4 days with 1000–40 000 steps/day. Geographic location = based on first digit of forward sortation area.

week 23 to 108 were included in analyses. Participants missing covariate information were excluded. To address the primary study objective, an adjusted multiple linear regression model was fit for the total sample to examine *Carrot* impact over 24 months. Users' weekly mean daily step count was the outcome modelled with fixed effects for week (109-level), gender (3-level), geographic location (5-level), app engagement (4-level) and start season (8-level) categorical variables, and with age and baseline step count continuous variables included as main effects (online supplemental appendix file E). Multiple linear regression model covariate coefficients were calculated to show covariates' upward or downward impact on weekly mean daily step count estimates.

Weekly mean daily step count averages around key timepoints were then calculated using data from 8 week windows (ie, 4 weeks before and after users' 6-, 12-, 18- and 24 month key timepoints; study weeks 23–30, 49–56, 75–82 and 101–108, respectively). Non-standardised post hoc contrasts estimated the difference from baseline for each key timepoint. Weekly mean daily step count comparisons between the 12 and 24 month timepoints and baseline were of particular interest given seasonal PA fluctuations in Canada (eg, to allow for same season comparisons, year-over-year).[30] Variances associated with the differences were also calculated (ie, 95% CIs and SEs adjusted for clustering (repeated measures) in the data using sandwich estimation). Standardised weekly mean daily step count differences (ie, Cohen's d) were then calculated, with ≥0.0, ≥ 0.2, ≥ 0.5, ≥ 0.8, ≥ 1.0 representing very small, small, medium, large and very large effect sizes, respectively.[31]

Statistical significance was assessed using a two-tailed z test with a threshold of significance of $p<0.05$. Numbers of users with 1000 or more steps/day difference from baseline at 12 and 24 months were calculated. While any PA increase may be beneficial,[32] 1000 steps/day is an often-cited level indicative of clinical significance (eg, incident cardiovascular disease).[33]

To address the secondary study objective, separate and adjusted multiple linear regression models were also developed to explore the influence of covariate levels (ie, start season, baseline PA, app engagement and geographic location levels) on longitudinal effects as previous literature suggests these may influence intervention effects.[20 34–36] As with the main model, non-standardised post hoc contrasts estimated the difference from baseline for each key timepoint (ie, 12 and 24 months). Covariate level comparisons were made by examining these non-standardised post hoc estimates and 95% CIs (ie, overlapping 95% CIs estimates were considered not different; see online supplemental appendix file F for additional secondary analysis detail).

Table 1 Sample characteristics, by app engagement level

	Rare	Limited	Occasional	Regular	Total
Sample size (n, %)	96 561 (18.68)	91 806 (17.78)	94 749 (18.33)	233 702 (45.22)	516 818
Age, years (mean, SD)	32.60 (12.65)	32.78 (12.50)	33.05 (12.35)	34.01 (12.82)	33.46 (12.65)
Gender (n, % female)	61 295 (63.48)	58 162 (63.35)	60 725 (64.09)	144 553 (61.85)	324 735 (62.83)
Start season (n, %)					
Spring	36 021 (37.30)	32 370 (35.26)	34 109 (36.00)	80 017 (34.24)	182 517 (35.32)
Summer	28 309 (29.32)	27 091 (29.51)	26 213 (27.67)	68 484 (29.30)	150 097 (29.04)
Fall	16 046 (16.62)	15 818 (17.23)	15 700 (16.57)	40 974 (17.53)	88 538 (17.13)
Winter	16 185 (16.76)	16 527 (18.00)	18 727 (19.76)	44 227 (18.92)	95 666 (18.51)
Geographic location (n, %)					
Central Ontario	39 531 (40.94)	37 309 (40.63)	38 394 (40.52)	95 527 (40.88)	210 761 (40.78)
Eastern Ontario	13 506 (13.99)	12 858 (14.01)	13 266 (14.00)	30 253 (12.95)	69 883 (13.52)
Metropolitan Toronto	22 078 (22.86)	22 261 (24.25)	24 148 (25.49)	67 854 (29.03)	136 341 (26.38)
Northern Ontario	4650 (4.82)	4047 (4.41)	3870 (4.08)	7669 (3.28)	20 236 (3.92)
Southwestern Ontario	16 796 (17.39)	15 331 (16.70)	15 071 (15.91)	32 399 (13.86)	79 597 (15.40)
Baseline physical activity (n, %)					
Low	43 710 (45.27)	42 008 (45.76)	44 231 (46.68)	113 710 (48.66)	243 659 (47.15)
Medium	21 805 (22.58)	21 354 (23.26)	22 413 (23.66)	56 854 (24.33)	122 426 (23.69)
High	15 161 (15.70)	14 412 (15.70)	14 387 (15.18)	33 899 (14.51)	77 859 (15.07)
Very high	15 885 (16.45)	14 032 (15.28)	13 718 (14.48)	29 239 (12.51)	72 874 (14.10)
Steps/day, baseline mean (SD)	6261 (3913)	6168 (3782)	6069 (3702)	5877 (3558)	6035 (3706)

App engagement level = based on proportion weeks user had step count data (rare: <25%; limited: 26%–50%; occasional: 51%–75%; regular: 76%–100%). Start season = season of baseline step count set date based on astrological calendar. Baseline physical activity = based on baseline weekly mean daily step count (low: <5000, medium: 5000–7499, high: 7500–9999, very high: 10 000+). Geographic location = based on first digit of forward sortation area.

Patient and public involvement

No patients or members of the public were involved in study planning, design, analysis or interpretation. Members of the public were involved in data collection through their intervention participation.

Equity, diversity and inclusion statement

All Ontarians downloading *Carrot* were assessed for eligibility. Compared with Ontarians in general, *Carrot* users in the province were younger, more likely to be born in Canada and more likely to identify as women. They were also more likely to self-report lower income and poorer mental health. The impact of age, gender, geographic location and baseline PA was explored. The author group of this study consisted of members of different gender identities and race/ethnicities from a single, high-income country.

RESULTS

Sample characteristics

The total analytic sample consisted of 516 818 participants (figure 1 and table 1). Baseline mean daily step count was 6035 (SD 3706) with nearly half of users categorised as 'low' active at baseline (47.15%). Twelve-month study retention ranged between 47.85% and 68.09% (ie, Winter 2016/2017 to Spring 2018), while 24-month retention was 46.95% and 38.20% for Winter 2016/2017 and Spring 2017, respectively (see online supplemental appendix table G for retention by start season).

Primary analyses

Overall, very small increases in weekly mean daily step count from baseline were observed at all key timepoints (table 2). A 464-step/day increase was found approximately 12 months after baseline, with a 242-step/day increase observed around the 24 month mark. The observed increases from baseline at 6 months were largely maintained at 12 and 18 months. Only at 24 months was a smaller daily step count increase observed. Multiple linear regression model covariate estimates are in online supplemental appendix table H (eg, for males, estimates in figure 2 shift up 285 steps/day). Daily step count ebb-and-flow attributable to the impact of seasons on PA is also observed in figure 2. Users starting to use the app during Winter 2016/2017 (figure 2a), for example, experienced weekly mean daily step count drops near study weeks 52 and 104—during the colder Canadian Winters of 2017/2018 and 2018/2019, respectively. Finally, 106 726 users (40.69%) increased their daily step count by at least 1000 per day at 12 months (65 157 users, or 24.84%,

Table 2 Weekly mean daily step count difference from baseline (study week 0), by key evaluation time point

	Sample size	Estimate	SE	Lower CI	Upper CI	Z value	P value	SMD
6 months	379 811	468	4.77	459	477	98.13	0.00E+00	0.159
12 months	262 316	464	5.58	453	475	83.21	0.00E+00	0.162
18 months	147 617	416	7.29	401	430	57.04	0.00E+00	0.148
24 months	64 454	242	11.00	221	264	22.03	1.58+E−107	0.087

Covariate reference levels set for the model are as follows: week (0), gender (female), location (Central Ontario), engagement (rare), start season (Fall 2017), age (33.36 years), baseline average daily step count (6036).
SMD, standardised mean difference (Cohen's d).

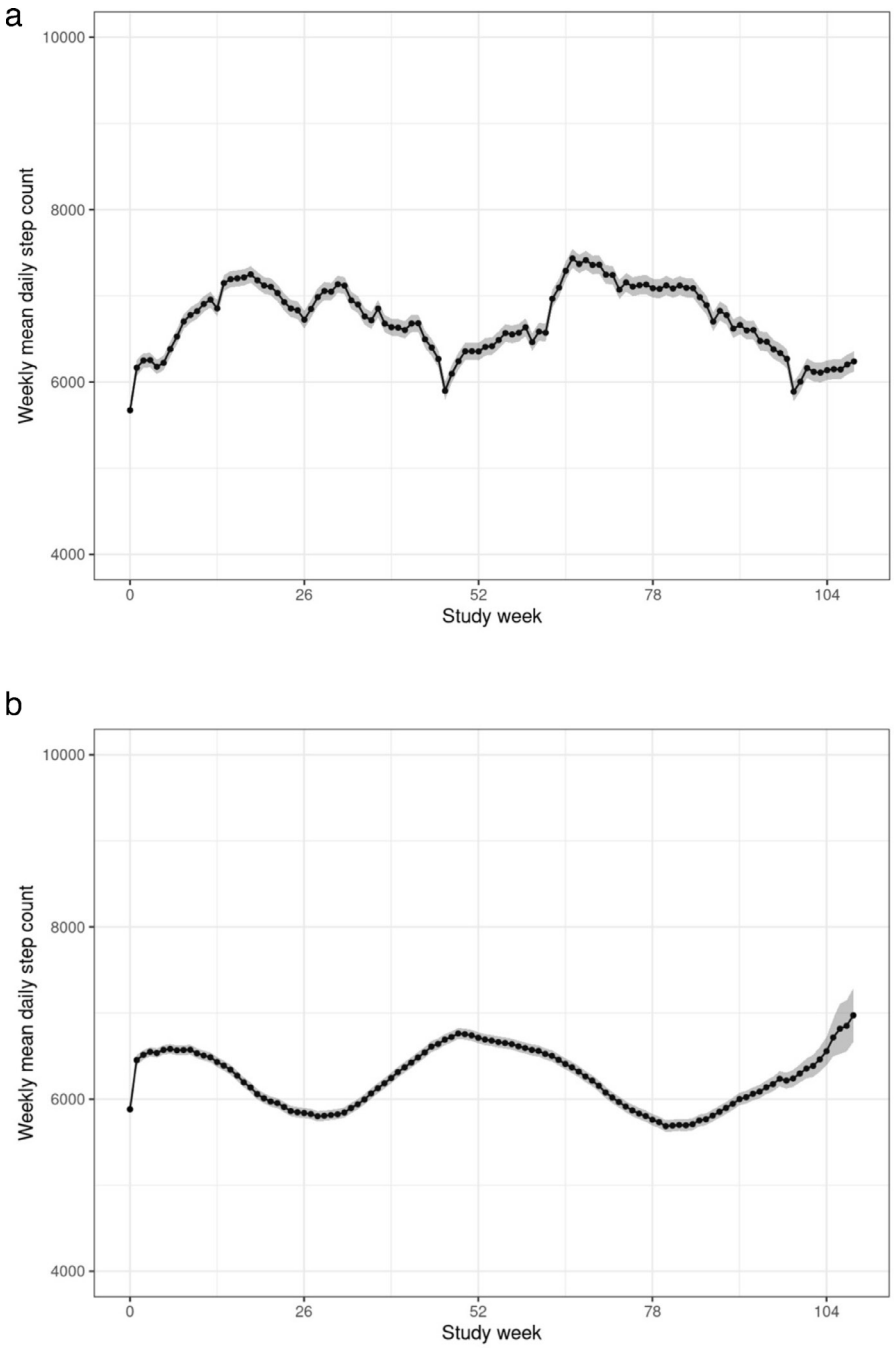

Figure 2 Predicted weekly mean daily step count with 95% CIs for each study week (0–108) for (a) Winter 2016/2017 and (b) Summer 2017 start seasons. Most common (categorical) covariate levels and mean value (continuous) set for the Winter 2016/2017 and Summer 2017 start season plots as follows: gender (female), location (Central Ontario), engagement (regular), age (33.36 years), baseline average daily step count (6036). The plot shifts upwards or downwards when adjusting the covariate values (ie, the overall pattern remains the same). The model coefficient table in online supplemental appendix table H can be used to determine how the plots shift.

decreased by that much). Additionally, 24 937 users (38.69%) experienced a 1000 daily step increase (and 17 208 users, or 26.70%, a decrease) at 24 months.

Secondary analyses

Separate and adjusted multiple linear regression models were also developed to examine the influence of start season, baseline PA, app engagement and geographic location levels on longitudinal effects (tables 3 and 4). Participants starting to use the app during the earlier start seasons (ie, with at least 12 months of daily PA reward exposure) experienced small weekly mean daily step count increases from baseline to 12 months (vs very small increases for the later start seasons (less than 12 months of daily PA reward exposure); eg, Spring 2017 = 596 steps/day vs Spring 2018 = 80 steps/day). As well, among earlier start seasons, increases from baseline diminished at 24 months (vs 12 months; eg, Summer 2017 dropped by about 10%). Large increases in weekly mean daily step count at the 12- and 24-month marks were observed among 'low' active users (ie, 49.63% of 24-month analytic sample; 31 991/64 454). On the other hand, users with 'very high' baseline PA levels exhibited large-to-very large decreases, though fewer users were in this subgroup (ie, 11.79%

Table 3 Weekly mean daily step count difference from baseline (study week 0) at 12 months, by covariate level

Covariate	Sample size	Estimate	SE	Lower CI	Upper CI	Z value	P value	SMD
Start season								
Winter 2016/2017	20 827	726	19.11	689	764	38.01	0.00E+00	0.263
Spring 2017	58 810	595	11.34	573	618	52.50	0.00E+00	0.217
Summer 2017	57 755	825	12.42	801	849	66.42	0.00E+00	0.276
Fall 2017	38 728	605	14.63	576	634	41.36	0.00E+00	0.210
Winter 2017/2018	36 026	378	14.90	349	407	25.38	4.52E−14	0.134
Spring 2018	42 486	80	13.80	53	107	5.81	6.16E−09	0.028
Summer 2018	7289	150	52.61	47	253	2.86	4.29E−03	0.033
Fall 2018	395	183	144.95	−101	467	1.26	0.20744374	0.063
App engagement level								
Rare	2175	1041	83.73	876	1205	12.43	1.87E−35	0.266
Limited	14 884	691	26.40	639	742	26.17	6.04E−151	0.214
Occasional	52 300	567	13.29	541	593	42.69	0.00E+00	0.187
Regular	192 957	473	6.74	460	487	70.24	0.00E+00	0.160
Geographic location								
Central Ontario	106 285	479	8.69	462	496	55.07	0.00E+00	0.169
Eastern Ontario	35 254	487	15.29	457	517	31.86	9.37E−22	0.170
Metropolitan Toronto	72 594	417	10.54	397	438	39.62	0.00E+00	0.147
Northern Ontario	9599	515	30.15	456	574	17.07	2.56E−65	0.174
Southwestern Ontario	38 584	482	14.86	453	512	32.46	3.88E−23	0.165
Baseline physical activity level								
Low	126 809	1968	6.63	1955	1981	296.84	0.00E+00	0.833
Medium	63 081	334	9.20	316	352	36.25	8.81E−29	0.144
High	38 602	−828	13.12	−854	−803	−63.16	0	−0.321
Very high	33 824	−3077	19.55	−3115	−3038	−157.39	0	−0.856

Covariate levels set for the model are as follows: week (0), gender (female), location (Central Ontario), engagement (rare), start season (Fall 2017), age (33.36 years), baseline average daily step count (6036). App engagement level = proportion of total possible weeks with step count data retrieved (rare: <25%; limited: 26%–50%; occasional: 51%–75%; regular: 76%–100%). Baseline PA level = low <5000; medium 5000–7499; high 7500–9999; very high≥10 000 steps/day.
SMD, standardised mean difference (Cohen's d).

of 24-month analytic sample; 7601/64 454). All app engagement levels showed very small-to-small increases in weekly mean daily step count at all key timepoints. A modest decline in PA at 12 months, but not 24 months, was also observed with increasing engagement. Finally, very small PA increases were noted across geographic locations. Metropolitan Toronto, the densest region, exhibited the smallest increases from baseline.

DISCUSSION

This study adds new evidence to the literature regarding long-term fitness app effects. Overall, very small weekly mean daily step count increases from baseline were observed at all key timepoints (ie, approximately 250–450 steps/day). Users from earlier start seasons with longer (ie, 12+ months) daily PA reward exposure experienced small increases from baseline at 12 months (ie, 600–825 steps/day). These were mostly maintained at 24 months (ie, 375–750 steps/day). None of these *average* daily step count increases met the 1000 step/day threshold level indicative of clinical significance. Roughly 40% of users at 12 and 24 months, however, increased their daily step count by 1000 or more. A smaller proportion, about 25%, decreased by at least that much. It appears, then, that modest population-level and device-assessed PA increases were sustained over 2 years with more users exhibiting clinically significant improvements than reductions.

Similar literature

A 12-month pre–post quasiexperiment was conducted in 2020 with *Carrot* users from two smaller Canadian provinces (n=39 113).[35] Similar to what was found here, typical app users experienced a 449 step/day increase from baseline (vs the 'last two recorded weeks' of app use). Study limitations, however, including a short recruitment window (ie, 13 June to 10 July 2016), unknown time-of-year of participants' 'last two recorded weeks', no within-province geographic consideration and no data collection beyond a year limit conclusion strength. Evidence reviews also suggest fitness apps produce very small-to-medium effects in the short-to-medium term (eg, 911 steps/day in the umbrella review by Singh *et al* for interventions *less than 12 months*).[2–4] This is twice the effect calculated here (ie, 468 and 460 steps/day at 6 and 12 months, respectively). It is only slightly greater, however, than what was observed for earlier start seasons with daily PA reward exposure for at least a year (ie, approximately 600–825 steps/day at 12 months). These evidence reviews also suggest greater fitness app effects when interventions include goals and planning behaviour change techniques, mention behaviour change theory, contain gamification elements (eg, points, progress bars), incorporate personalisation, target step counts (vs other PA behaviours) and have higher retention.[2–4] Each of these characterise the *Carrot* intervention as well. As mentioned, no RCT has examined fitness app effects beyond a year, and to the best of our knowledge, Kamada *et al*[37] is the only quasiexperimental study to do so.[37] They investigated the effects of a gamified fitness app among 20 052 Japanese baseball fans over 22 months. They found users' daily step count increased by 574 at 3 months (vs matched controls). This was maintained until 9 months. Afterwards, however, PA improvements were not significantly different from controls (ie, from 10

Table 4 Weekly mean daily step count difference from baseline (study week 0) at 24 months, by covariate level

Covariate	Sample size	Estimate	SE	Lower CI	Upper CI	Z value	P value	SMD
Start season								
Winter 2016/2017	14 359	486	24.25	438	533	20.01	4.33E−89	0.167
Spring 2017	35 804	375	15.58	344	405	24.06	7.09E−13	0.127
Summer 2017	11 030	758	43.53	672	843	17.40	7.69E−68	0.166
Fall 2017	931	1127	95.778	940	1315	11.771	5.47E−32	0.386
Winter 2017/2018	747	1651	102.724	1450	1853	16.073	3.91E−58	0.588
Spring 2018	884	1156	85.015	989	1323	13.598	4.09E-42	0.457
Summer 2018	487	603	121.412	365	841	4.965	6.86E−07	0.225
Fall 2018	212	24	184.227	−237	485	0.672	0.50169475	0.009
App engagement level								
Rare	489	541	178.01	192	890	3.04	2.39E−03	0.137
Limited	2633	272	64.93	145	399	4.19	2.79E−05	0.082
Occasional	7139	241	37.70	167	314	6.38	1.77E−10	0.076
Regular	54 193	256	11.94	232	279	21.41	1.03E−10	0.092
Geographic location								
Central Ontario	26 418	278	17.07	245	312	16.29	1.26E−59	0.100
Eastern Ontario	8382	316	30.55	251	371	10.17	2.82E−24	0.111
Metropolitan Toronto	17 797	143	20.41	104	184	7.03	2.08E−12	0.053
Northern Ontario	2407	249	61.06	129	369	4.08	4.54E−05	0.083
Southwestern Ontario	9450	264	30.09	206	323	8.79	1.51E−18	0.090
Baseline physical activity level								
Low	31 991	1986	13.67	1959	2013	145.28	0.00E+00	0.812
Medium	15 497	85	19.76	47	124	4.32	1.55E−05	0.035
High	9365	−1318	29.45	−1376	−1261	−44.76	0	−0.463
Very high	7601	−3969	43.34	−4054	−3884	−91.57	0	−1.050

Covariate levels set for the model are as follows: week (0), gender (female), location (Central Ontario), engagement (rare), start season (Fall 2017), age (33.36 years), baseline average daily step count (6036). App engagement level = proportion of total possible weeks with step count data retrieved (rare: <25%; limited: 26%–50%; occasional: 51%–75%; regular: 76%–100%). Baseline PA level = low <5000; medium 5000–7499; high 7500–9999; very high ≥10 000 steps/day.
SMD, standardised mean difference (Cohen's d).

to 22 months). Reasons given to explain this trajectory towards non-significance include limited longitudinal data and waning intervention engagement. It has been suggested that if meaningful engagement (eg, one or two app opens per month) can be sustained long term (eg, 12+ months) with evidence-based app features, then longitudinal effects are possible.[2–4] The fairly high *Carrot* retention rates reported previously[38] as well as here (eg, about 40% at 24 months) lend support to this hypothesis.

Secondary findings

First, 'deimplementation' analyses by start season suggest daily PA rewards were associated with PA improvements in the first 12 months of app intervention (ie, Spring 2017 vs Spring 2018 in table 3). An alternative explanation may be that those installing the app earlier ('early adopters') may have been primed for behaviour change (eg, already in the 'preparation' stage of change).[39] Intervention in the first two seasons (ie, Winter 2016/2017 and Spring 2017) appeared no more effective than the next two (ie, Summer 2017 and Fall 2017; table 3), though, somewhat abating this possibility. Furthermore, daily PA rewards provided for a year, then removed, mostly sustained PA increases when less costly supports remained in place (ie, weekly PA rewards). This is an important finding from intervention sustainability/scalability stand-points and is similar to what has been previously reported by shorter duration studies regarding partial or complete FI withdrawal (eg, PA reductions of about 25%).[40]

Second, *Carrot* was used primarily by 'low' active Ontarians at higher health risk—the app's target population.[41] Positive effects were also most pronounced among 'low' active users (ie, 1986 steps/day at 24 months). Conversely, substantial PA reductions were observed among higher active users (eg, −3969 steps/day 24 months for 'very high', respectively). The PA increases among lower active users could have been due in part to limited internalised motivation (eg, 'I do not enjoy walking') and thus greater responsiveness to external FI contingency, consistent with self-determination theory.[42] On the other hand, self-determination theory suggests that external rewards can undermine, or 'crowd out', individuals' intrinsic motives—especially when intrinsic motivation is high to begin with (eg, among the higher active)—and harm future behaviours.[43] Alternatively, regression to the mean could explain why lower active users exhibited PA increases, while reductions were observed among the higher active.[44] A few factors, however, minimise the likelihood that the observed PA improvements and reductions were due to statistical regression, including (a) longer (ie, up to 14 days), seasonally distributed and thus more stable baseline counterfactual estimates, (b) covariate inclusion in regression models to balance distribution across key timepoints/covariate levels and reduce confounding, (c) separate models fit for each baseline PA level used in post hoc change estimates (with study week as a categorical variable to capture non-linear fluctuations over time), (d) a second counterfactual ('deimplementation') mimicking an interrupted time series design to isolate the app's true effect and (e) consistency of findings across key timepoints (eg, 'low' active user increases of 1967 and 1986 step/day at 12 and 24 months, respectively) and covariate levels (eg, similar early/later start season effect sizes).[45] Finally, the PA reductions observed among the higher active could also be explained by the

Hawthorn effect (ie, leading to baseline PA overestimation)[46] or lower app engagement around key timepoints (ie, fewer steps tracked and/or taken owing to lower intervention need among higher active and/or loss of interest due to unrealistically high daily step goals of up to 15 000 steps/day).

Third, slightly smaller PA improvements were noted with greater engagement at 12, but not 24, months. These inverse and neutral dose–response relationships are unlike the positive ones previously reported by shorter duration studies.[47] One reason for this may be that psychological mechanisms influencing behaviour *maintenance* begin to shift over time (eg, more self-determination leading to greater PA)[39] with users not needing to interact with fitness apps as often (eg, progress towards daily goals becomes intuitive, visualised feedback not as critical). In other words, 'treatment fidelity' becomes less important.[48] Another explanation may be that less engaged users only opened the app during more physically active study weeks (eg, to earn rewards). In this case, step counts at key timepoints could have been overestimated. To the best of our knowledge, the dose–response relationships observed here are novel and warrant replication.

Implications

Given persistent global physical inactivity rates, these findings may be encouraging for fitness app users, healthcare providers, app publishing companies, researchers, governments and large organisations (eg, health insurers) looking to promote healthier, more active living through fitness app intervention. *Carrot* appears to have promoted population-level PA in part because it leveraged concepts from behavioural economics (eg, 'present bias' with instantaneous rewards) and self-determination theory (eg, intrinsic motivation fostered with self-efficacy promoting daily step goals). As delivered, though, the microincentives proved too costly for *Carrot*'s government partners to absorb longterm. More sustainable FI models are needed (eg, limited-time only, self-funded deposit contract or lottery-based FIs).[49] Incorporating promising artificial intelligence-led app features (eg, large language model conversational agents ('chatbots'), machine learning-driven step goals leveraging richness of data) may support long-term engagement and effectiveness while keeping FI costs low. The absence of psychological outcome assessment (eg, Behavioural Regulation in Exercise Questionnaire)[50] in this study prohibited better understanding of negative effects among higher active users and should be prioritised by researchers moving forward as well.

Limitations

First, the lack of a control group limits inferences of causality in this quasiexperiment striving to balance internal with external validity and facilitate real-world impact. To address this limitation, a 'preintervention' period was identified (ie, the baseline) to allow a counterfactual comparison. A marked PA increase between study weeks 0 and 1 (figure 2) suggests an immediate intervention effect compared with the underlying baseline trend. The naturally occurring 'deimplementation' of daily PA rewards provided a second counterfactual comparison (ie, early vs later start seasons at 12 months). Several other study design (eg, theory grounded intervention, extended data collection), data analysis (eg, adjusting for baseline values, separate (sensitivity) analyses by covariate level) and interpretation (eg, effect sizes and variances reported, comparisons to related work) phase strategies strengthen causal inference as well.[12 13] Second, with 47.85% of users retained at 12 months, and 38.20% at 24 months, attrition bias may limit conclusion strengths. Results may not extend to those ceasing app use. Third, measurement (eg, smartphone app step counting accuracy may vary in free-living contexts) and history (ie, secular trends (eg, weather) unrelated to the intervention) bias may also limit conclusion strengths. Extended data collection (ie, 2 years) at multiple timepoints (ie, four key timepoints) within smaller critical windows (ie, 8 week windows at 12 and 24 months) approaches an interrupted time series design and adds support for interpretations.[12 13] Fourth, for a small proportion of users (ie, around 3%) the baseline step count set date (ie, used to denote start season) was later than the date steps were first recorded, likely due to a user uninstalling, then reinstalling, the app. Start season may be mislabelled for these users. Finally, while a robust PA 'maintenance' definition has yet to be agreed on, an operational definition of more than a year was reasonable (ie, the transtheoretical model suggests maintenance occurs between 6 months and 5 years).[39]

CONCLUSION

This study suggests that very small but sustained population-level PA increases are possible with fitness app intervention. The influential role of contextual factors (eg, baseline PA level, geographic location, app engagement level and microincentives) may inform more impactful interventions in the future.

Acknowledgements The *Carrot Rewards* initiative was made possible in part through funding from the Public Health Agency of Canada. The views expressed here do not necessarily represent the views of the Agency. As well, this research was presented at the 2025 Society of Behavioural Medicine's 46th Annual Meeting and Scientific Sessions.

Contributors Authors contributed to the concept and design (LN, GF, CT and MSM), acquisition of the data (MSM), statistical analysis (LN, CT and MSM), interpretation of the data (LN, GF, CT and MSM) and drafting the manuscript (LN, GF, CT and MSM). All authors contributed to critical revision of the manuscript for important intellectual content. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Dr Mitchell, as the corresponding author, is responsible for the overall content as a guarantor. He accepts full responsibility for the completed work and the conduct of the study, had access to the data and made the decision to publish.

Funding The Public Health Agency of Canada and Government of Ontario.

Competing interests MSM was awarded a Partnership Development Grant by the Social Sciences Research Council of Canada in 2023 to examine engagement with a deposit contract-based app called WayBetter, and received consulting fees in 2024–2025 from The Program (makers of the Wellnify.ai health and wellness app).

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Institutional Board: Western University Health Sciences Research Ethics Board, Project ID #122932. Informed consent was not required as this research relied exclusively on the secondary use of non-identifiable data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs
Lisa Nguyen http://orcid.org/0009-0002-3217-7735

Guy Faulkner http://orcid.org/0000-0001-8898-2536
Marc S Mitchell http://orcid.org/0000-0002-2772-9614

REFERENCES

1 Grand View Research. mHealth apps market size, share & trends analysis report by type, and segment forecasts, 2019 - 2026. 2019. Available: https://www.grandviewresearch.com/industry-%20analysis/mhealth-app-market
2 Laranjo L, Ding D, Heleno B, *et al*. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. *Br J Sports Med* 2021;55:422–32.
3 Mazeas A, Duclos M, Pereira B, *et al*. Evaluating the Effectiveness of Gamification on Physical Activity: Systematic Review and Meta-analysis of Randomized Controlled Trials. *J Med Internet Res* 2022;24:e26779.
4 Singh B, Ahmed M, Staiano AE, *et al*. A systematic umbrella review and meta-meta-analysis of eHealth and mHealth interventions for improving lifestyle behaviours. *NPJ Digit Med* 2024;7:179.
5 Chaudhry UAR, Wahlich C, Fortescue R, *et al*. The effects of step-count monitoring interventions on physical activity: systematic review and meta-analysis of community-based randomised controlled trials in adults. *Int J Behav Nutr Phys Act* 2020;17:129.
6 2018 physical activity guidelines advisory committee scientific report. 2018. Available: https://odphp.health.gov/sites/default/files/2019-09/PAG_Advisory_Committee_Report.pdf
7 Gropper H, John JM, Sudeck G, *et al*. The impact of life events and transitions on physical activity: A scoping review. *PLoS One* 2020;15:e0234794.
8 Grand View Research. Fitness app market size & trends. 2024. Available: https://www.grandviewresearch.com/industry-analysis/fitness-app-market
9 Statista. Fitness apps worldwide. 2024. Available: https://www.statista.com/outlook/hmo/digital-health/digital-fitness-well-being/health-wellness-coaching/fitness-apps/worldwide
10 Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health* 2004;94:400–5.
11 West SG, Duan N, Pequegnat W, *et al*. Alternatives to the randomized controlled trial. *Am J Public Health* 2008;98:1359–66.
12 Handley MA, Lyles CR, McCulloch C, *et al*. Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research. *Annu Rev Public Health* 2018;39:5–25.
13 Ruissen GR. Establishing causal inferences from experimental and observational data: A critical review and primer for sport and exercise psychology. *Sport Exerc Perform Psychol* 2025;14:7–23.
14 Michie S, Richardson M, Johnston M, *et al*. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med* 2013;46:81–95.
15 Government of Canada. National healthy living platform: "carrot rewards" targets lifestyle improvements. 2015. Available: https://www.canada.ca/en/news/archive/2015/07/national-healthy-living-platform-carrot-rewards-targets- lifestyle-improvements.html
16 Marotta S. Ottawa-backed carrot rewards app shutting down after failing to find a buyer. 2019. Available: https://www.theglobeandmail.com/business/article-ottawa-backed-carrot-rewards-app-shutting-down-after-failing-to-find-a/
17 Bassett DR Jr, Wyatt HR, Thompson H, *et al*. Pedometer-measured physical activity and health behaviors in U.S. adults. *Med Sci Sports Exerc* 2010;42:1819–25.
18 Yao J, Tan CS, Lim N, *et al*. Number of daily measurements needed to estimate habitual step count levels using wrist-worn trackers and smartphones in 212,048 adults. *Sci Rep* 2021;11:9633.
19 Mitchell M, Faulkner G. A "nudge" at all? The jury is still out on financial health incentives. *Healthc Pap* 2012;12:31–6.
20 Mitchell M, White L, Lau E, *et al*. Evaluating the Carrot Rewards App, a Population-Level Incentive-Based Intervention Promoting Step Counts Across Two Canadian Provinces: Quasi-Experimental Study. *JMIR Mhealth Uhealth* 2018;6:e178.
21 Camerer CF, Loewenstein G. Behavioral economics: past, present, future. Princeton, NJ: Princeton University Press, 2003.
22 Deci E, Ryan R. Handbook of self-determination research. Rochester, NY: University of Rochester Press, 2002.
23 Adamakis M. Criterion Validity of iOS and Android Applications to Measure Steps and Distance in Adults. *Technologies (Basel)* 2021;9:55.
24 Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act* 2015;12:159.
25 Adamakis M. Validity of Wearable Monitors and Smartphone Applications for Measuring Steps in Semi-Structured and Free-Living Settings. *Technologies (Basel)* 2023;11:29.
26 Duncan MJ, Wunderlich K, Zhao Y, *et al*. Walk this way: validity evidence of iphone health application step count in laboratory and free-living conditions. *J Sports Sci* 2018;36:1695–704.
27 Bohlender D. When do the seasons start? 2023. Available: https://nrc.canada.ca /en/certifications-evaluations-standards/canadas-official-time/3-when-do-seasons-start
28 Canada Post. Addressing guidelines. 2022. Available: https://www.canadapost-postescanada.ca/cpc/en/support/articles/addressing-guidelines/postal-codes.page
29 Tudor-Locke C, Craig CL, Thyfault JP, *et al*. A step-defined sedentary lifestyle index: <5000 steps/day. *Appl Physiol Nutr Metab* 2013;38:100–14.
30 Merchant AT, Dehghan M, Akhtar-Danesh N. Seasonal variation in leisure-time physical activity among Canadians. *Can J Public Health* 2007;98:203–8.
31 Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
32 Hall KS, Hyde ET, Bassett DR, *et al*. Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. *Int J Behav Nutr Phys Act* 2020;17:78.
33 Stens NA, Bakker EA, Mañas A, *et al*. Relationship of Daily Step Counts to All-Cause Mortality and Cardiovascular Events. *J Am Coll Cardiol* 2023;82:1483–94.
34 Faubel R, Garriga A, Sempere-Rubio N, *et al*. Impact of seasonality on physical activity: a systematic review. *Eur J Public Health* 2022;32.
35 Mitchell M, Lau E, White L, *et al*. Commercial app use linked with sustained physical activity in two Canadian provinces: a 12-month quasi-experimental study. *Int J Behav Nutr Phys Act* 2020;17:24.
36 Schoeppe S, Alley S, Van Lippevelde W, *et al*. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. *Int J Behav Nutr Phys Act* 2016;13:127.
37 Kamada M, Hayashi H, Shiba K, *et al*. Large-Scale Fandom-based Gamification Intervention to Increase Physical Activity: A Quasi-experimental Study. *Med Sci Sports Exerc* 2022;54:181–8.
38 Lau EY, Mitchell MS, Faulkner G. Long-term usage of a commercial mHealth app: A "multiple-lives" perspective. *Front Public Health* 2022;10:914433.
39 Rhodes RE, Sui W. Physical Activity Maintenance: A Critical Narrative Review and Directions for Future Research. *Front Psychol* 2021;12:725671.
40 Spilsbury S, Wilk P, Taylor C, *et al*. Reduction of Financial Health Incentives and Changes in Physical Activity. *JAMA Netw Open* 2023;6:e2342663.
41 Arim R, Schellenberg G. An assessment of non-probabilistic online survey data: comparing the carrot rewards mobile app survey to the Canadian community health survey. 2019. Available: https://www150.statcan.gc.ca/n1/en/pub/11-633-x/11-633-x2019002-eng.pdf?st=x0OqLORg
42 Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol* 2000;55:68–78.
43 Deci EL, Koestner R, Ryan RM. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol Bull* 1999;125:627–68.
44 Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 1976;104:493–8.
45 Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005;34:215–20.
46 McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol* 2014;67:267–77.
47 Romeo A, Edney S, Plotnikoff R, *et al*. Can Smartphone Apps Increase Physical Activity? Systematic Review and Meta-Analysis. *J Med Internet Res* 2019;21:e12053.
48 Bellg AJ, Borrelli B, Resnick B, *et al*. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychol* 2004;23:443–51.
49 Salmani B, Prapavessis H, Vanderloo LM, *et al*. Financial incentives for physical activity in adults: Systematic review and meta-analysis update. *Prev Med* 2025;192:108237.
50 Markland D, Tobin V. A Modification to the Behavioural Regulation in Exercise Questionnaire to Include an Assessment of Amotivation. *J Sport Exerc Psychol* 2004;26:191–6.