



OPEN

## Quantifying and correcting bias due to outcome dependent self-reported weights in longitudinal study of weight loss interventions

Jiayi Tong<sup>1</sup>, Rui Duan<sup>2</sup>, Ruowang Li<sup>3</sup>, Chongliang Luo<sup>4</sup>, Jason H. Moore<sup>3</sup>, Jingsan Zhu<sup>5</sup>, Gary D. Foster<sup>6,7</sup>, Kevin G. Volpp<sup>8</sup>, William S. Yancy Jr.<sup>9</sup>, Pamela A. Shaw<sup>1,10</sup> & Yong Chen<sup>1</sup>✉

In response to the escalating global obesity crisis and its associated health and financial burdens, this paper presents a novel methodology for analyzing longitudinal weight loss data and assessing the effectiveness of financial incentives. Drawing from the Keep It Off trial—a three-arm randomized controlled study with 189 participants—we examined the potential impact of financial incentives on weight loss maintenance. Given that some participants choose not to weigh themselves because of small weight change or weight gains, which is a common phenomenon in many weight-loss studies, traditional methods, for example, the Generalized Estimating Equations (GEE) method tends to overestimate the effect size due to the assumption that data are missing completely at random. To address this challenge, we proposed a framework which can identify evidence of missing not at random and conduct bias correction using the estimating equation derived from pairwise composite likelihood. By analyzing the Keep It Off data, we found that the data in this trial are most likely characterized by non-random missingness. Notably, we also found that the enrollment time (i.e., duration time) would be positively associated with the weight loss maintenance after adjusting for the baseline participant characteristics (e.g., age, sex). Moreover, the lottery-based intervention was found to be more effective in weight loss maintenance compared with the direct payment intervention, though the difference was non-statistically significant. This framework's significance extends beyond weight loss research, offering a semi-parametric approach to assess missing data mechanisms and robustly explore associations between exposures (e.g., financial incentives) and key outcomes (e.g., weight loss maintenance). In essence, the proposed methodology provides a powerful toolkit for analyzing real-world longitudinal data, particularly in scenarios with data missing not at random, enriching comprehension of intricate dataset dynamics.

Obesity is a common epidemic disease across the whole world. During the past decades, there has been a dramatic increase in obesity. As of 2021, more than 70% of adults ages 20 and over are overweight or obese in the United States<sup>1–6</sup>. Increasing the incidence of diabetes, heart disease, hypertension, stroke, certain cancers, and early mortality<sup>5–7</sup>, obesity threatens people's health and becomes the most significant factor to illness<sup>8,9</sup>. What's

<sup>1</sup>Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. <sup>2</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA. <sup>3</sup>Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA. <sup>4</sup>Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis, St. Louis, MO 63110, USA. <sup>5</sup>Center for Health Incentives and Behavioral Economics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. <sup>6</sup>WW International, New York, NY 10010, USA. <sup>7</sup>Center for Weight and eating Disorders, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. <sup>8</sup>Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. <sup>9</sup>Department of Medicine, Duke University, Durham, NC 27705, USA. <sup>10</sup>Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, USA. ✉email: ychen123@upenn.edu

more, obesity also brings a burden to health care resources and raises medical expenditures. Facing this phenomenon, it is critically needed to deploy strategies that can effectively reduce body weight and maintain weight loss.

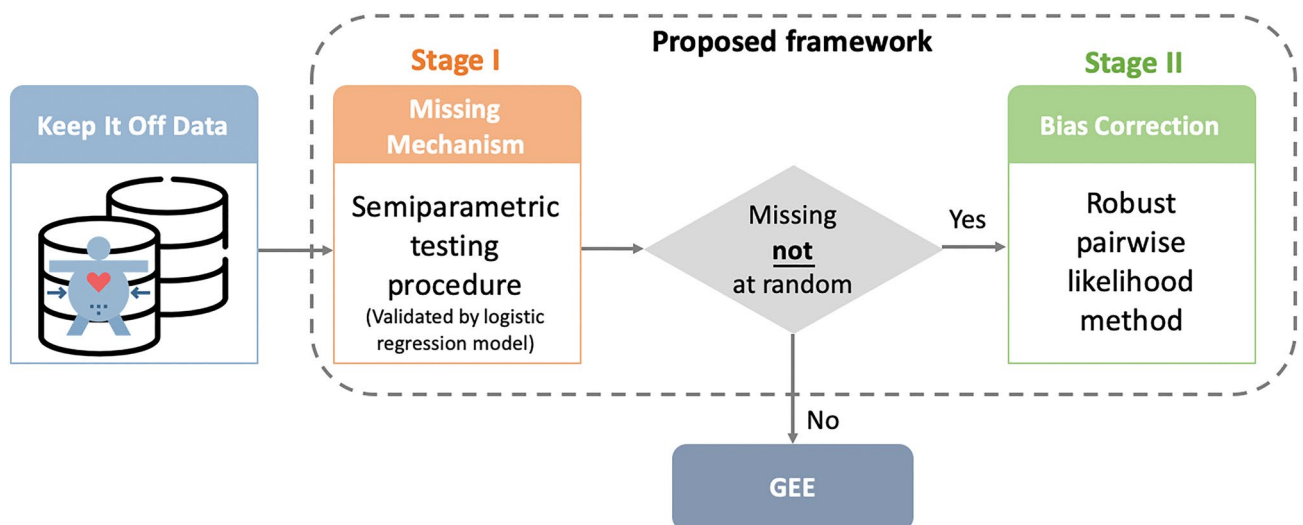
There are some strategies for achieving weight loss that have been successfully identified; however, for people with obesity, it is more challenging to maintain long-term weight loss<sup>10–12</sup>. Based on a few existing studies, it is very common to observe weight regain after initial weight loss<sup>7,12</sup>. Factors including reduced resting metabolic rate, unregulated behavioral processes, difficulty adhering to the diet, lack competing rewards and reinforcement are all possible contributors to the difficulty in maintaining weight loss<sup>13</sup>. External motivation such as financial incentives can be helpful and effective at helping people achieve initial weight loss compared with the standard approaches<sup>14</sup>. Financial incentives have been shown to be effective in initial weight loss<sup>14–17</sup>. To investigate and examine the long-term effects of financial incentives on weight-loss maintenance, a study called Keep It Off was conducted.

Keep It Off study is a three-arm randomized controlled trial with two phases after initial weight loss. Participants were randomized into three groups and given different financial incentives; the three arms were (1) a lottery-based incentive, (2) a direct payment incentive and (3) a control with no financial incentive. The daily weight of participants was measured through a wireless scale at home and for those on a financial incentive arm, receipt of the incentives was reliant on attaining the goal weight. Keep It Off study provides real-world longitudinal data from 189 participants to study the relative effectiveness of lottery-based and traditional direct payment incentives on weight loss maintenance. For studies with longitudinal data, missing data are often unavoidable, and more so when participants are relied upon to provide the data themselves, whether self-reported via phone or a website, or uploaded via wireless devices as in the Keep It Off Study. Missing data can significantly threaten the results, leading to harmful effects on the validity of conclusions and decision-making.

To analyze longitudinal data, mixed-effect model<sup>18</sup> and GEE (Generalized Estimating Equations)<sup>19</sup> methods are commonly used methods as the standard procedure. Among these two methods, GEE is robust to the misspecification of the correlation structure of the response. Additionally, it relaxes the distribution assumption on the data and can obtain consistent estimates of the population average effect. With these features, the GEE method has been widely used in biomedical studies for longitudinal data<sup>20–23</sup>. On the other hand, the GEE method is less robust to non-randomly missing longitudinal data. By assuming the observation times are predefined and are the same across subjects, the GEE method provides valid results when the observations are missing completely at random (MCAR). MCAR refers to the case that the missing data are independent of both observed and unobserved variables. MCAR is the simplest case in missing data problems, but it rarely happens in practice. The GEE method can implement the inverse probability weighting approach to handle missing data even when the data are missing at random (MAR)<sup>24</sup>. MAR refers to when data missingness depends on observed variables alone but not the unobserved ones.

Due to the nature of the design using self-weight measurements in the Keep It Off study, there is a chance that some participants chose not to weight themselves because of potentially disappointing results, such as small weight changes or weight gains, which is a common phenomenon in many weight-loss studies and self-reported outcome studies. The missingness in this longitudinal data most likely falls into the category of missing not at random (MNAR), which refers to the case where the missingness is allowed to depend on the variables that are missing. When the data are MNAR, the GEE method will yield biased estimates because GEE has a strong assumption of independence between the observation time (i.e., self-reporting process) and the outcome of interest (i.e., weight)<sup>25–31</sup>.

Therefore, to tackle the challenges of the informative reporting process in the real-world longitudinal data, we proposed a framework of methods with two stages in this paper (Fig. 1). In Stage I, a semiparametric testing approach<sup>32</sup> was utilized to quantify the evidence of MNAR due to the self-weighting mechanism. There existed evidence indicating that the data were MNAR in the testing and validation procedures. As the data were missing



**Figure 1.** The proposed framework for the Keep It Off data analysis.

not at random, the observation process was challenging to model. Thus, in Stage II, we used a pairwise likelihood method<sup>26</sup>, which does not require modeling the self-reporting process, to evaluate the impacts of the financial incentive on weight loss maintenance. With the proposed framework, we found that the enrollment time (i.e., days from the first day of enrollment to the weighing day, duration time) was associated with the weight loss maintenance after adjusting for the baseline participant characteristics (e.g., age, sex). There was some evidence that the participants in the groups with financial incentives (i.e., lottery group and direct payment group) would maintain weight loss better compared to the control group over time and that the lottery-based incentive was more likely to be effective for weight loss maintenance compared with the direct payment incentive.

## Results

### Keep It Off data

Keep It Off study is a three-arm, unblinded randomized controlled trial (RCT). The participants were recruited from WeightWatchers (WW), which is a global weight management program with over 4 million members and is empirically validated (ClinicalTrials.gov Identifier: NCT00702455)<sup>33–35</sup>. The participants aged 30–80 years old who were in stable health and had a body mass index (BMI) of 30–45 kg/m<sup>2</sup> before joining WW and had lost at least 11 lb before the start of the Keep It Off study to be eligible to enroll. Based on the inclusion criteria, the total number of randomized participants was 191. The participants' baseline characteristics at the beginning of the Keep It Off study are summarized in Table 1.

The Keep It Off study has two phases: an intervention phase (Phase I) and a follow-up phase (Phase II). Each phase lasted 6 months. In Phase I, the participants were randomized to get one of three interventions, including the control intervention which is daily weigh-ins and report without any incentive (referred to as control group), control intervention plus a traditional direct payment incentive (referred to as direct payment group), and control intervention plus a lottery-based incentive (referred to as lottery group). The participants who achieved their weight goals would get the lottery-based incentive or the direct payment incentive in the corresponding arm. In the study, the daily weights were collected through an Internet-enabled scale, which allowed wireless transmission to the database of weights measured daily at home. At the end of Phase I (i.e., at month 6), an *in-person milestone weigh-in* was required for all participants. This milestone weigh-in was aiming to examine if the participants reached/maintained the target weight.

In Phase II, all participants were observed without any intervention for an additional 6 months but were asked to continue weighing themselves daily as part of the ongoing study protocol. During the whole study, one of the participants became pregnant and one was diagnosed with lymphoma. Based on the inclusion criteria, these two participants are excluded and the final sample size of the participants for analysis is 189. The design and planned analysis of the Keep It Off study is detailed in the protocol paper by Shaw et al.<sup>7</sup>.

With the aim of comparing effective financial incentives in the weight-loss maintenance across three groups, we first examined the patterns of reporting days as a follow-up analysis in addition to the primary analyses reported by Yancy et al.<sup>13</sup>. We compared the percentages of report days in one week for the three treatment groups for 12 months (i.e., 52 weeks), including both Phases I and II. Figure 2 showed the report days patterns across the treatment groups in the Keep It Off study<sup>13</sup>. A decreasing trend was observed through the whole study period and a big drop occurred around week 26, which was the end of Phase I (i.e., end of financial incentives). In particular, the participants weighed themselves at home and reported the weights approximately 90% of days in the first week, 75% of the days in week 10, and 55% of the days in week 26. There is a lack of evidence suggesting a difference across three groups in the patterns over time<sup>13</sup>.

In addition to the weekly report days pattern, we also present the reporting pattern of participants' daily weights during Phase I in the control group (Fig. 3, left), direct payment group (Fig. 3, middle), and lottery group (Fig. 3, right). Each row was composed of the daily weights of a single participant for the first 6 months (i.e., Phase I). The purple cells were the reported daily weight, and the grey ones represented the missing values. The participants were ordered vertically from the least to the greatest percentage of missing values.

The patterns in Fig. 3 showed that the missing daily weights across all the three groups did not follow a regular pattern. Some participants reported almost every day in Phase I, but others only reported for a few days. The individual missing percentages ranged from 0.5 to 99.5%. Among the intervention groups, the control group had the most missing data (32.2%) in the first 6 months, which was larger than the missing percentages in the directly payment group (29.3%) and the lottery group (26.2%) (Table 2). This suggested that the participants with financial incentives might be more likely to maintain daily weighing and reporting.

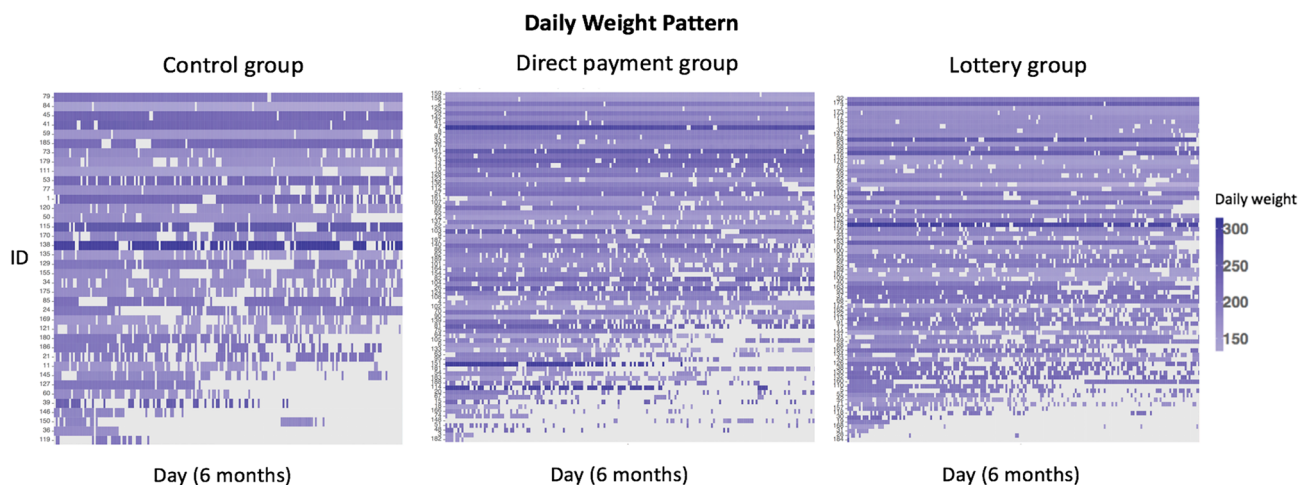
Furthermore, based on the patterns in both Figs. 2 and 3, we hypothesized that the data were missing not at random. Some participants seemed to choose not to report their weights because of small weight changes or weight gains, which is a common phenomenon in many weight-loss studies and self-reported studies. Due to the use of self-weighing process in the Keep It Off study, the missing data problem is not negligible in this real-world

Characteristic	Total (n = 191)	Control (n = 39)	Direct payment (n = 75)	Lottery (n = 77)
Age, mean (SD)	49.0 (10.5)	50.1 (10.0)	48.6 (11.0)	48.8 (10.3)
Female, n (%)	175 (91.6)	36 (92.3)	69 (92.0)	70 (90.9)
Baseline BMI (kg/m <sup>2</sup> ), mean (SD)	36.7 (4.2)	36.8 (4.3)	36.7 (4.1)	36.6 (4.4)
Baseline weight (lb), mean (SD)	198.9 (32.9)	201.0 (33.1)	200.0 (34.6)	196.7 (31.5)

**Table 1.** Available baseline characteristics of the study participants.



**Figure 2.** Percentage of report days in 1 week. The red line represents the lottery group; the green one represents the group with direct payment; the blue line is the control group.



**Figure 3.** Daily weight patterns, missing percentage, and 6-month milestone weight change in the first 6 months for three groups (i.e., control group, direct payment group, lottery group). Cells in purple are the daily weights. The first column in red and green is the heat bar to show the percentage of missing daily weight. The cells in grey represent missing values.

	Total (n = 191)	Control (n = 39)	Direct payment (n = 75)	Lottery (n = 77)
Phase I	28.6%	32.2%	29.3%	26.2%

**Table 2.** Summary of missing daily weights percentages for Phase I.

longitudinal data. Therefore, to address the issue in missing data and investigate the effectiveness of financial incentives, we proposed a novel framework to study the missing mechanism and conduct bias correction using a robust and novel pairwise likelihood method.

### Data analysis results

We utilized our proposed framework, which composed of two stages: in Stage I, we applied a semiparametric testing approach to investigate the missing data mechanism; in Stage II, we conducted the bias correction with the estimating equation derived from pairwise composite likelihood, to analyze the Keep It Off data. No standardization was performed on the variables, except for centralizing age and BMI.

In Stage I, we first applied the semiparametric testing procedure to test the missingness mechanism of the data. The test statistic is 9.12 (i.e.,  $T = 9.12$ ). Compared to a chi-squared distribution with 3 degrees of freedom (i.e.,  $m = 2$  with two covariates: age and sex), the p-value equals 0.02. It indicates that there exists evidence showing the daily self-reported longitudinal data in the Keep It Off study were most likely to be *missing not at random* (MNAR). A discussion of the validation of the testing procedure using additional data from the Keep it Off trial is provided in Supplementary Appendix B. As there was evidence that the data were missing not random, we then conducted Stage II of the proposed framework, by applying the pairwise likelihood method. The outcome is the difference between the baseline weight and the daily self-reported weight (i.e., outcome = daily weight – baseline weight). The covariates included in the model were age at enrollment (centralized), sex, baseline BMI (centralized), time since enrollment (i.e., enrollment time, duration time), and the interaction between the time since enrollment and group indicator.

To estimate the effect sizes (i.e., regression coefficient) of the covariates, we applied both GEE method by using *gee* R package<sup>36</sup> and the proposed pairwise method for comparison. The effect sizes with corresponding standard errors and the p-values of both methods are presented in Table 3. The control group was treated as the reference group in the model. The effect size of the time since enrollment on the weight change is  $-0.582$  (p-value = 0.048). There exists evidence showing that the longer duration time a participant remained in the study, the higher the weight loss (i.e., the duration time was positively associated with weight loss maintenance) after adjusting for the usage of the baseline participant characteristics. The lottery-based intervention (effect size =  $-0.616$  in Table 3) was more likely to be effective in weight loss compared with the direct payment group (effect size =  $-0.512$  in Table 3). The participants in the groups with financial incentives (i.e., lottery group and direct payment group) have weak evidence of maintaining weight loss better compared to the control group over time. Although the results were not statistically significant, the direction and relative magnitudes of the effect estimates are consistent with the findings from the main analysis of the Keep It Off study by Yancy Jr et al.<sup>13</sup> For comparison, we also presented the marginal covariate effects estimated by the GEE method in the table. Much larger negative effects were found by the GEE method, which is likely to be overestimated due to the ignorance of the informative self-report process. The standard errors (se) estimated through the GEE method were notably larger compared to those derived from the pairwise likelihood method. This discrepancy could be attributed to the high intra-class correlation present in the data, which leads to reduced effective sample size. In the context of the weight-loss data, each patient represents a distinct cluster, and the weights associated with a single patient are naturally highly correlated.

### Discussion

Here we report framework to analyze the real-world self-reported longitudinal data in the Keep It Off study. The Keep If Off study is a three-arm randomized controlled trial (RCT). The aim was to examine if participants' weight loss maintenance can be improved by financial incentives. For the self-measured weight-loss data, the missing data problem is unavoidable. Thus, in Stage I of the proposed framework, we utilize a semi-parametric testing approach to investigate the missing mechanism of the Keep It Off data. The results showed that that the missingness of the data is most likely to be missing not at random.

For Stage II, we apply a pairwise likelihood method to evaluate the impacts of financial incentives on weight-loss maintenance. Using the conditioning technique, the pairwise likelihood method provide robust estimation of the effect of financial incentives. Without imposing parametric models on the self-reporting process, the pairwise likelihood method avoids the potential bias inherent in the GEE method under MNAR. Through the

	GEE			Pairwise likelihood		
	Effect size	SE	p-value	Effect size	SE	p-value
Sex	-2.744	1.705	0.108	-0.246	0.234	0.293
Baseline BMI	-0.042	0.121	0.725	-0.002	0.018	0.908
Age	0.087	0.047	0.060	0.013	0.007	0.050
Time since enrollment	-7.653	2.227	0.001	-0.582	0.295	0.048
Lottery group* time	-4.725	2.689	0.079	-0.616	0.342	0.072
Direct payment group* time	-4.164	2.698	0.123	-0.512	0.345	0.138

**Table 3.** Summary of the data analysis for the associations between weight loss and the covariates. \*For simplicity, time = time since enrollment; effect size refers to regression coefficient.

proposed framework, we show that there is a statistically significant correlation between duration time in study and weight loss. In particular, the longer the duration time, the greater the weight loss obtains.

Additionally, there is weak evidence showing that the participants in both groups with financial incentives (i.e., lottery group and direct payment group) maintain weight loss better compared to the control group over time respectively. Specifically, the lottery-based group have weak evidence of being more effective in weight loss compared with the direct payment group with control group as reference. The results are in directional agreement with those from Yancy Jr et al. for the first 6-month weight loss but not statistically significant.

As demonstrated by this study, the proposed framework is a novel framework to analyze real-world longitudinal data. In particular, the proposed framework provides a test in Stage I to examine the missing mechanism and a robust pairwise likelihood method in Stage II to investigate the association between the exposure (e.g., financial incentives) and outcome of interest (e.g., weight loss maintenance).

The proposed framework has its limitations. In Stage II, the pairwise construction of likelihood comes with the price of higher computational cost, as the algorithm involves computation of likelihood constructed by all pairs of patients within a site. To alleviate this limitation, we implemented an algorithm with R calling C, which is about 50 times faster than using the R programming language alone. The code is available on Github (<https://github.com/Pencil/Keep-It-Off-Study>). Secondly, the results we got were directionally consistent with Yancy Jr et al., but there was no evidence indicating the statistical significance. In summary, the proposed framework has broad applicability to other research topics where data are missing at random or completely at random, especially when the observation time process is challenging to model.

## Methods

In this section, we introduced the proposed framework to analyze the real-world self-reporting data in the Keep It Off study (ClinicalTrials.gov Identifier: NCT00702455)<sup>35</sup>. The framework is composed of two stages: in Stage I, we applied a semiparametric testing approach to investigate the missing data mechanism; in Stage II, we conducted the bias correction with the estimating equation derived from pairwise composite likelihood.

### Stage I: testing missing data mechanism

When analyzing the real-world healthcare data, especially the longitudinal self-reported data, the missing data problem is inevitable. In the Keep It Off study, since the participants self-weighed, there existed a large chance that participants at times chose not to use the wireless scales because of expected small weight change, failure to lose weight, or gaining weight from baseline. Under this scenario, the missingness was most likely to fall into the class of missing not at random (MNAR). Analyzing MNAR data is more challenging compared to analyzing missing at random (MAR) or missing completely at random (MCAR) data. Thus, to ensure the validity of the data analysis of Keep It Off study, we examined the missing mechanism by utilizing a semiparametric testing approach<sup>32</sup>.

Suppose we have a response variable  $Y$  (i.e., daily weight), covariates  $X$  (i.e., age, sex), and a generalized linear model, where the conditional distribution of  $Y$  given covariates  $X$  belongs to the exponential dispersion family:

$$p(y | x; \theta) = \exp \left\{ \frac{y\eta - b(\eta)}{\lambda} + c(y; \lambda) \right\}, h\{\mu(\eta)\} = \alpha + x^T \beta_1 \quad (1)$$

where  $\mu = E(Y | X)$ , is related to covariates through a link function  $h(\cdot)$ ,  $\eta$  is the natural parameter,  $\lambda$  is the dispersion parameter,  $\mu(\eta) = b'(\eta)$  by the property of the exponential family and  $\beta = (\alpha, \beta_1)^T$  are regression coefficients of interest.

We proposed to identify two parameter estimators of the regression coefficient. Both estimators are valid when the data is missing at random, while only one of them is valid otherwise. The first estimator of  $\beta$  is denoted as  $\tilde{\beta}$ , which is obtained by solving the estimating equation derived from the likelihood function using the probability density function (Eq. (1)) of exponential family. The second estimator is denoted as  $\hat{\beta}$ , which is obtained using a semiparametric pseudo-likelihood method<sup>37</sup>.

Our test of missingness is based on the discrepancy between the two estimators of  $\beta$ . The test statistics,  $T$ , can be written as:

$$T = n(\tilde{\beta} - \hat{\beta})^T \widehat{W}^{-1} (\tilde{\beta} - \hat{\beta}) \quad (2)$$

where  $n$  is the sample size and  $\widehat{W}$  is the weighting matrix which can be estimated through the influence functions of estimators  $\tilde{\beta}$  and  $\hat{\beta}$ . The details of derivation of two estimators and the weighting matrix are provided in the Supplementary Appendix A. As  $n \rightarrow \infty$ , the test statistic  $T$  converges weakly to  $\chi^2$  with  $m + 1$  degree of freedom, where  $m$  is the dimension of covariate  $X$ . It is suggested that the data is more likely to be missing not at random when the test statistic  $T$  takes large values.

### Stage II: analyzing missing not at random data

For the Keep It Off study, the data are most likely to be missing not at random as we have shown. Thus, alternative methods are critically needed for the outcome-dependent longitudinal data. An innovative *pairwise likelihood method* was proposed in 2015<sup>26</sup>. The pairwise conditional scheme was utilized to form a composite conditional likelihood. This method provides an alternative estimating procedure for the investigation of the marginal covariate effects on the repeated measure in longitudinal weight loss data. The novelty of this method is that it does not require modeling the self-reported time process, which is challenging when the data is missing not at random. This key feature of this pairwise likelihood method brings robustness to the statistical inference and corrects the potential bias induced by the outcome-dependent weight loss data.

Suppose there are two observations from a pair of independent participants: the  $j$ -th observation of participant  $i$  and the  $j'$ -th observation of participant  $i'$ . Let  $y_{ij}$  denote daily weight response at the  $j$ -th time point of participant  $i$  and  $y_{i'j'}$  denote the daily weight response at the  $j'$ -th time point of participant  $i'$ . For one of the observations ( $y_{ij}$ ), the proportional density function<sup>26,38</sup> given the vector of covariates  $\mathbf{x}_{ij}$  can be written as:

$$f(y_{ij} | \mathbf{x}_{ij}; \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} y_{ij}) g(y_{ij})}{\int \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij} y_{ij}) dG(y_{ij})} \tag{3}$$

where  $g(\cdot)$  is the distribution of the response variable and  $G(\cdot)$  is the cumulative distribution, which represent the probability structure of the observation time process. To focus on the estimation of parameter of interest  $\boldsymbol{\beta}$ , the specification of the distribution function is not necessary. Thus, the nuisance distribution function can be cancelled out through a conditioning technique. The conditional density of the responses at  $j$ -th time point of participant  $i$  and at the  $j'$ -th time point of participant  $i'$ , ( $y_{ij}, y_{i'j'}$ ), given the order statistics  $y^{(1)} = \min(y_{ij}, y_{i'j'})$  and  $y^{(2)} = \max(y_{ij}, y_{i'j'})$ , can be calculated as:

$$f(y_{ij}, y_{i'j'} | y^{(1)}, y^{(2)}, \mathbf{x}_{ij}, \mathbf{x}_{i'j'}) = \frac{f(y_{ij} | \mathbf{x}_{ij}; \boldsymbol{\beta}) f(y_{i'j'} | \mathbf{x}_{i'j'}; \boldsymbol{\beta})}{f(y_{ij} | \mathbf{x}_{ij}; \boldsymbol{\beta}) f(y_{i'j'} | \mathbf{x}_{i'j'}; \boldsymbol{\beta}) + f(y_{ij} | \mathbf{x}_{i'j'}; \boldsymbol{\beta}) f(y_{i'j'} | \mathbf{x}_{ij}; \boldsymbol{\beta})} \tag{4}$$

where  $\mathbf{x}$  is the vector of the covariates (e.g., baseline weight, BMI, age, time, etc.), and  $\boldsymbol{\beta}$  is the vector parameters of interest. Through conditioning on the order statistics, the probability structures of the observation time process (i.e.,  $g(y_{ij})$ ) were eliminated in the density function. For this procedure, the specification of the observation time process is not required.

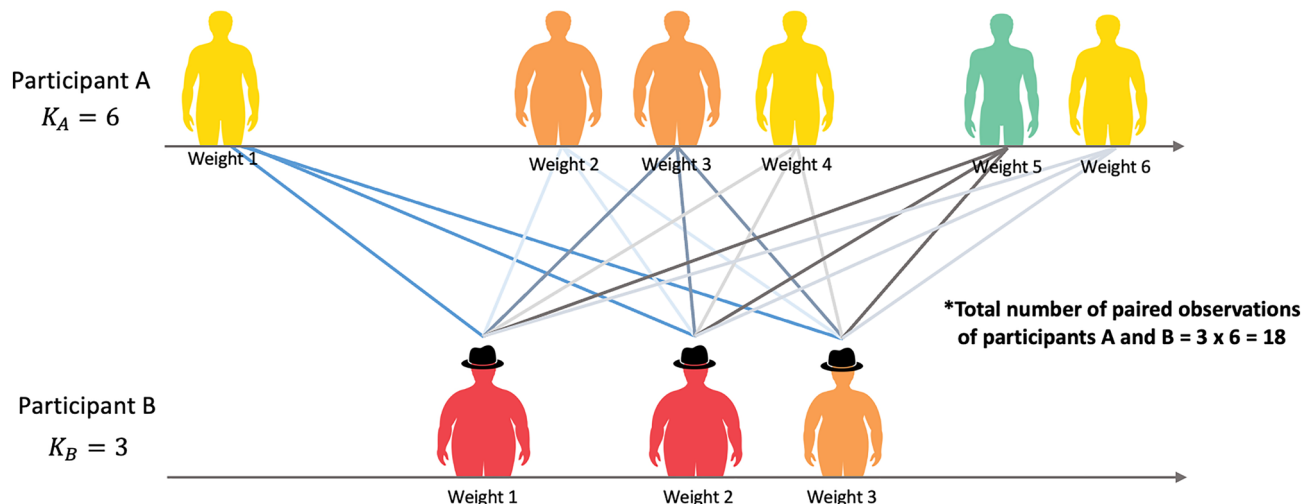
For each possible pair of participants, we can calculate the above conditional density. By multiplying these densities together, the following pairwise likelihood function for all observations can be obtained:

$$L(\boldsymbol{\beta}) = \prod_{i < i'} \prod_{j=1}^{K_i} \prod_{j'=1}^{K_{i'}} \left[ 1 + \frac{f(y_{ij} | \mathbf{x}_{i'j'}; \boldsymbol{\beta}) f(y_{i'j'} | \mathbf{x}_{ij}; \boldsymbol{\beta})}{f(y_{ij} | \mathbf{x}_{ij}; \boldsymbol{\beta}) f(y_{i'j'} | \mathbf{x}_{i'j'}; \boldsymbol{\beta})} \right]^{-1} \tag{5}$$

where  $K_i$  is the number of total observations of participant  $i$  and  $K_{i'}$  is the number of total observations of participant  $i'$ . With this pairwise conditioning approach, the missing-not-at-random mechanism, i.e. the observation of  $y$  depending on  $x$ ,  $g(y_{ij})$ , is canceled in Eq. (5). In other words, the missingness pattern of the dependent variable does not affect the estimation of the parameter of interest. The final pairwise estimator is  $\hat{\boldsymbol{\beta}} = \text{argmax} L(\boldsymbol{\beta})$ , which is the marginal covariate effects on the outcome.

The performance of this novel pairwise likelihood method in analyzing longitudinal data has been validated through extensive simulation studies and empirical data application in Chen et al.<sup>26</sup>. The key assumption of this method is that the probability of observing the response variable can be expressed as a product of functions involving the response variable and covariates, which is Eq. (2.2) in Chen et al.<sup>26</sup>. Their simulation studies indicate a degree of robustness in cases where this assumption does not meet. In the context of model misspecification, Luo and Tsai<sup>38</sup> emphasize the robustness of the proportional density ratio. This robustness underscores the method's stability and flexibility, ensuring its reliability in real-world scenarios where model assumptions might not hold. By incorporating the pairwise conditioning technique, our method maintains the effectiveness in analyzing the weight loss data.

In Fig. 4, we graphically illustrated the pairwise idea by using a pair of participants as an example. Participant A has 3 observations and participant B has 6 observations over time. The total number of pairs of observations



**Figure 4.** Illustration of the pairwise likelihood method idea in the proposed framework Stage II by using a pair of participants as example.  $K$  is the total number of observations for each participant.

between participants A and B is 18. With this idea, the pairs of observations of both participants were multiplied together to get the likelihood function as Eq. (5).

All methods were carried out in accordance with relevant guidelines and regulations. We confirmed all experimental protocols were approved. The use of human subjects was approved by the University of Pennsylvania Institute Review Board (IRB) under the protocol number 816917. Waivers of Informed Consent were granted by the Institutional Review Board of the University of Pennsylvania.

### Data availability

The data underlying this article cannot be shared publicly due to patient privacy concerns. The data will be shared on reasonable request to the corresponding author.

### Code availability

The R code is available at <https://github.com/Pencil/Keep-It-Off-Study>. The analyses were conducted using R version 4.2.1.

Received: 31 August 2022; Accepted: 31 August 2023

Published online: 04 November 2023

### References

1. FastStats—Overweight Prevalence. <https://www.cdc.gov/nchs/fastats/obesity-overweight.htm> (accessed 24 Jul 2021).
2. Adult Obesity Prevalence Maps | Overweight & Obesity | CDC. <https://www.cdc.gov/obesity/data/prevalence-maps.html> (accessed 27 Aug 2023).
3. Adult Obesity Facts | Overweight & Obesity | CDC. <https://www.cdc.gov/obesity/data/adult.html> (accessed 27 Aug 2023).
4. Ward, Z. J. *et al.* Association of body mass index with health care expenditures in the United States by age and sex. *PLoS ONE* <https://doi.org/10.1371/JOURNAL.PONE.0247307> (2022).
5. Flegal, K. M. *et al.* Prevalence and trends in obesity among US adults, 1999–2000. *JAMA* **288**, 1723–1727 (2002).
6. Ogden, C. L. *et al.* Prevalence of overweight and obesity in the United States, 1999–2004. *JAMA* **295**, 1549–1555 (2006).
7. Shaw, P. A. *et al.* The design and conduct of Keep It Off: An online randomized trial of financial incentives for weight-loss maintenance. *Clin. Trials* **14**, 29–36. <https://doi.org/10.1177/1740774516669679> (2017).
8. Adult Obesity Causes & Consequences | Overweight & Obesity | CDC. <https://www.cdc.gov/obesity/adult/causes.html> (accessed 22 Mar 2022).
9. Delahanty, L. M. Weight loss in the prevention and treatment of diabetes. *Prev. Med. (Baltim.)* **104**, 120–123. <https://doi.org/10.1016/j.ypmed.2017.07.022> (2017).
10. Anderson, J. W. *et al.* Long-term weight-loss maintenance: A meta-analysis of US studies. *Am. J. Clin. Nutr.* **74**, 579–584 (2001).
11. Dombrowski, S. U. *et al.* Long term maintenance of weight loss with non-surgical interventions in obese adults: Systematic review and meta-analyses of randomised controlled trials. *BMJ* **348**, 2646 (2014).
12. Franz, M. J. *et al.* Weight-loss outcomes: A systematic review and meta-analysis of weight-loss clinical trials with a minimum 1-year follow-up. *J. Am. Diet. Assoc.* **107**, 1755–1767 (2007).
13. Yancy, W. S. Jr. *et al.* Financial incentive strategies for maintenance of weight loss: Results from an internet-based randomized controlled trial. *Nutr. Diabetes* **8**, 1–11 (2018).
14. Volpp, K. G. *et al.* Financial incentive-based approaches for weight loss: A randomized trial. *JAMA* **300**, 2631–2637 (2008).
15. John, L. K. *et al.* Financial incentives for extended weight loss: A randomized, controlled trial. *J. Gen. Intern. Med.* **26**, 621–626 (2011).
16. Kullgren, J. T. *et al.* Individual-versus group-based financial incentives for weight loss: A randomized, controlled trial. *Ann. Intern. Med.* **158**, 505–514 (2013).
17. Glanz, K. *et al.* Effect of financial incentives and environmental strategies on weight loss in the healthy weigh study: A randomized clinical trial. *JAMA Netw. Open.* <https://doi.org/10.1001/JAMANETWORKOPEN.2021.24132> (2021).
18. Diggle, P., Diggle, P. J., Heagerty, P. *et al.* *Analysis of Longitudinal Data* (Oxford University Press, 2002).
19. Liang, K. Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22. <https://doi.org/10.1093/biomet/73.1.13> (1986).
20. Lee, J.-H. *et al.* The use of GEE for analyzing longitudinal binomial data: A primer using data from a tobacco intervention. *Addict. Behav.* **32**, 187–193 (2007).
21. Rochon, J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Stat. Med.* **17**, 1643–1658 (1998).
22. Lord, D. & Persaud, B. N. Accident prediction models with and without trend: Application of the generalized estimating equations procedure. *Transp. Res. Rec.* **1717**, 102–108 (2000).
23. Kubrak, C. *et al.* Clinical determinants of weight loss in patients receiving radiation and chemoirradiation for head and neck cancer: A prospective longitudinal view. *Head Neck* **35**, 695–703 (2013).
24. Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Stat. Assoc.* **94**, 1096–1120 (1999).
25. Lin, H., Scharfstein, D. O. & Rosenheck, R. A. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 791–813 (2004).
26. Chen, Y., Ning, J. & Cai, C. Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics* **16**, 727–739. <https://doi.org/10.1093/biostatistics/kxv008> (2015).
27. Robins, J. M., Rotnitzky, A. & Zhao, L. P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* **90**, 106–121 (1995).
28. Rotnitzky, A. & Robins, J. M. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–820 (1995).
29. Rotnitzky, A., Robins, J. M. & Scharfstein, D. O. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Stat. Assoc.* **93**, 1321–1339 (1998).
30. French, B. & Heagerty, P. J. Marginal mark regression analysis of recurrent marked point process data. *Biometrics* **65**, 415–422 (2009).
31. Cai, Y., Huang, J., Ning, J., Lee, M. L., Rosner, B., Chen, Y. Two-sample test for correlated data under outcome-dependent sampling with an application to self-reported weight loss data. *Stat. med.* **38**(25), 4999–5009 (2019).
32. Duan R, Liang CJ, Shaw PA, Tang CY, Chen Y. Testing the missing at random assumption in generalized linear models in the presence of instrumental variables. *Scandinavian Journal of Statistics*. 2023.



33. Ahern, A. L. *et al.* Extended and standard duration weight-loss programme referrals for adults in primary care (WRAP): A randomised controlled trial. *Lancet* **389**, 2214–2225. [https://doi.org/10.1016/S0140-6736\(17\)30647-5](https://doi.org/10.1016/S0140-6736(17)30647-5) (2017).
34. Gudzone, K. A. *et al.* Efficacy of commercial weight loss programs: An updated systematic review. *Ann. Intern. Med.* **162**, 501. <https://doi.org/10.7326/M14-2238> (2015).
35. Keep It Off: A Weight Loss Maintenance Study—Full Text View—ClinicalTrials.gov. <https://classic.clinicaltrials.gov/ct2/show/NCT00702455> (accessed 27 Aug 2023).
36. CRAN—Package gee. <https://cran.r-project.org/web/packages/gee/index.html> (accessed 27 Aug 2023).
37. Zhao, J. & Shao, J. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Am. Stat. Assoc.* **110**, 1577–1590 (2015).
38. Luo, X. & Tsai, W. Y. A proportional likelihood ratio model. *Biometrika* **99**, 211–222 (2012).

## Acknowledgements

This work was supported in part by National Institutes of Health (R21AI167418, 1R01LM014344, 1R01AG077820, R01LM012607, R01AI130460, R01AG073435, R56AG074604, R01LM013519, R56AG069880, U01TR003709, RF1G077820, R01GM148494). This work was supported partially through Patient-Centered Outcomes Research Institute (PCORI) Project Program Awards (ME-2019C3-18315 and ME-2018C3-14899). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

## Author contributions

J.T., R.D., and Y.C. designed methods; J.Z. and P.S. provided the Keep It Off dataset for data analysis; J.T. conducted data analysis of Keep If Off data; J.T., R.D., P.S. and Y.C. interpreted the results and provided instructive comments; J.T., R.D., R.L., C.L., and Y.C. drafted the main manuscript. All authors have approved the manuscript.

## Competing interests

GDF is an employee and shareholder at WeighWatchers.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-023-41853-4>.

**Correspondence** and requests for materials should be addressed to Y.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023, corrected publication 2023