Supplementary Materials: Adjusting for adherence in randomized trials when adherence is measured as a continuous variable: an application to the Lipid Research Clinics Coronary Primary Prevention Trial

Kerollos Nashat Wanis, Arin L. Madenci, Miguel A. Hernán, Eleanor J. Murray

---

1. Supplementary materials

This supplementary document expands on the methods described in the main text for our analysis of the Lipid Research Clinics Coronary Primary Prevention Trial (LRCCPT). In section 2, we discuss how missed visits were handled. In section 3 we describe how covariates were measured. In section 4 we describe how the primary outcome, coronary heart disease, was defined and measured. In section 5, we elaborate on how adherence was modelled. In section 6 we provide more technical details about the inverse probability weights used to adjust for confounding in our analysis. Lastly, in section 7 we describe the parametric g-formula and how it was implemented in our analysis.

2. Missed visits

Participants in the LRCCPPT missed occasional visits over the seven-year follow-up period. Because each individual was given an emergency supply of study drug, their level of adherence might not have changed after a missed visit. In our analysis, we assumed that participants would maintain their level of adherence up to one year from their last clinic visit, and therefore carried forward their last observed adherence value to the subsequent five expected visits. Individuals who missed more than 1 year of clinic visits were censored on the expected date of their sixth visit. Similarly, when individuals attended clinics but did not have their adherence measured, perhaps because they forgot to return their unused medication packets for counting, we assumed that their adherence level was the same as its previously measured value for up to one year (i.e. we carried forward

values for up to 5 visits). Sensitivity analyses in which we carried forward values for 3 or 4 visits did not result in materially different estimates (see Appendix table 2 and 3).

3. Covariates

Data on important covariates were collected by the lipid research clinics at the pre-randomization visits and at scheduled bi-monthly follow-up visits. 'Two-month visits,' 'six-month visits,' and 'annual' follow-up visits differed in terms of the variables collected. 'Two-month visits' which occurred on the two, four, eight, and ten-month follow-up each year were shorter visits at which lipid data (i.e. total cholesterol, triglycerides, HDL cholesterol, and LDL cholesterol) was measured with bloodwork, a cardiovascular system review was performed, drug toxicity and side-effect information was ascertained, and the diet instructions were reinforced. Any reported chest pain, hospitalization, or other medical problem was investigated and recorded.

'Six-month visits,' which occurred at the sixth month of follow-up each year, involved collection of data on all of the same covariates as the 'two-month visits' plus additional bloodwork (i.e., creatinine, bilirubin, glucose, uric acid, hematocrit, white blood cell count), a one-day dietary recall, a partial medical history, a cardiovascular and respiratory physical examination, and a resting electrocardiogram. 'Annual visits,' occurring at the twelfth month of follow-up each year, further included a complete clinical history and physical examination, additional bloodwork (i.e., potassium, sodium, phosphate, calcium, carotene, vitamin K, iron, iron binding, thyroxin, total protein, albumin, and globulin), an exercise ECG, and urinalysis. Adherence to the study medication was calculated and recorded at all visits.

The baseline and post-randomization variables collected by the lipid research clinics were used to estimate the stabilized weights for IP weighting (and conditional densities for the g-formula, as described in detail below) used in our analyses in order to adjust for time-varying confounding. As described in greater detail in the following sections, we included variables whose values did not change over follow-up (baseline covariates: educational status, age at baseline, physical activity level at work at baseline, race, and baseline risk strata) in models for estimation of the numerator of the stabilized weights. The baseline risk strata were composed, by the LRCPPT investigators, using combinations of the following three binary prognostic factors: LDL cholesterol $\geq$215 mg/dl (the 97.5^{th} percentile of U.S males aged 35-59), S-T segment depression during exercise, and a logistic risk

function estimated using diastolic blood pressure, age, and number of cigarettes smoked at pre-randomization visits one and two. We included both baseline and post-randomization covariates in models for estimation of the denominator of the stabilized weights. The post-randomization variables included were systolic blood pressure, diastolic blood pressure, body mass index, gastrointestinal symptom (i.e., nausea, vomiting, diarrhea) indicators, lipid laboratory values, non-lipid laboratory values (i.e., sodium, phosphate, potassium, thyroxin, iron, total bilirubin, direct bilirubin, alkaline phosphatase, creatinine, glucose, albumin, calcium, and white blood cell count), dietary recall values (i.e., calories, protein, fat, total carbohydrates), clinical history (i.e., amount of cigarettes smoked daily, history of a vascular event since the last visit, history of any vascular event since the start of follow-up, angina assessed using the Rose Questionnaire, and total number of aspirin tablets taken in the past week), physical activity level outside of work, exercise electrocardiogram abnormality with ST segment >10 V/sec, and a summary indicator variable summarizing performance during exercise. Because blood lipids were the most frequently measured covariates, our models were adjusted for flexible functions (restricted cubic splines with knots at the 5^{th}, 35^{th}, 65^{th}, and 95^{th} percentiles) of the average total cholesterol, LDL, HDL, and triglycerides in the past 6 visits.

4. Primary outcome

The primary outcome of the LRCCPPT was definite atherosclerotic coronary heart disease death or non-fatal myocardial infarction. Definite atherosclerotic coronary heart disease death was defined by a death certificate with a consistent underlying or immediate cause and either: i) pre-terminal hospitalization with definite or suspected myocardial infarction, or ii) a history of definite angina, or suspected or definite myocardial infarction and no more likely cause of death ascribed. Sudden and unexpected deaths were also ascribed to coronary heart disease if they occurred within one hour of characteristic symptoms in patients with no other potentially lethal acute or chronic co-morbidity and not confined to home, hospital, or other institution because of illness in the 24-hour period prior to death. The diagnosis of definite non-fatal myocardial infarction required one of the following: i) diagnostic electrocardiogram, ii) characteristic ischemic chest pain and diagnostic cardiac enzymes, or iii) characteristic ischemic chest pain and equivocal cardiac enzymes and equivocal electrocardiogram.

The research clinics were alerted to possible events by the individual participants, their family, or their personal physician when a hospitalization occurred

or symptoms arose. Other events were ascertained at regular follow-up visits, or by directly contacting participants, their families, or their physician when a participant failed to attend a scheduled clinic appointment. Each potential event was verified by a Lipid Research Clinic cardiologist and other physicians.

5. Adherence models

We used the parameter estimates from a discrete-time hazards (pooled logistic) model to estimate the 7-year cumulative incidence of definite atherosclerotic coronary heart disease death or non-fatal myocardial infarction for individuals in the placebo arm of the LRCCPPT under different levels of exposure.

Let t be the visit number, with visit five, the visit at which randomization occurred, defined as t=0. A_t is the value of adherence between time t and $t+1$, Y_t is an indicator for definite atherosclerotic coronary heart disease death or non-fatal myocardial infarction between time t and $t+1$, C_t is an indicator for loss to follow-up between time t and $t+1$, and Z indicates the randomization arm (i.e. $Z=0$ indicates randomization to placebo).

In order to estimate the cumulative incidence of the primary outcome among participants randomized to placebo, we fit the following weighted discrete-hazards model to all person-visits:

$$\text{logit}\Big(\hat{\Pr}(Y_{t+1}=1\Big|\bar{A}_t, V, \bar{C}_{t+1}=0, \bar{Y}_t=\bar{0}, Z=0)\Big) = \theta_{0,t} + \theta_1^T V + \theta_2 f(\bar{A}_t) + \theta_{3,t} f(\bar{A}_t)$$

where the intercept term $\theta_{0,t}$, is allowed to vary according to a flexible functional form of time and θ_2 may be a vector of parameters. Because adherence was first measured at visit 6, the analysis was necessarily restricted to those who attended visit 6. However, a sensitivity analysis which restricted to individuals who attended visit 7 (a greater proportion of individuals) did not result in materially different estimates (see Appendix table 1).

We used a dose-response function, $f(\bar{A}_t)$, which allowed the effect of average adherence since the beginning of the study to vary over time:

$$\theta_2 f(\bar{A}_t) = \theta_2^T h\Big(\frac{1}{t+1}\sum_{k=0}^{t} A_k\Big),$$

$$\theta_{3,t} f(\bar{A}_t) = \theta_3^T g(t+1) h\Big(\frac{1}{t+1}\sum_{k=0}^{t} A_k\Big).$$

We also considered a dose-response function, $f^*(\bar{A}_t)$, using average adherence prior to the most recent visit plus adherence at the most recent visit:

$$\theta_2 f^*(\bar{A}_t) = \theta_{2,1}^T h\left(\frac{1}{\max\{t,1\}}\sum_{k=0}^{t-1} A_k\right) + \theta_{2,2}^T h\left(A_t\right)$$

$$\begin{aligned}\theta_{3,t} f^*(\bar{A}_t) = \theta_{3,1}^T h\left(\frac{1}{\max\{t,1\}}\sum_{k=0}^{t-1} A_k\right) + \theta_{3,2}^T g(t+1) h\left(\frac{1}{\max\{t,1\}}\sum_{k=0}^{t-1} A_k\right) \\ + \theta_{3,3}^T h\left(A_t\right) + \theta_{3,4}^T g(t+1) h\left(A_t\right)\end{aligned}$$

and a dose-response function, $f^\dagger(\bar{A}_t)$, using average adherence up to the last year plus average adherence during the last year:

$$\theta_2 f^\dagger(\bar{A}_t) = \theta_{2,1}^T h\left(\frac{1}{\max\{t-5,1\}}\sum_{k=0}^{t-6} A_k\right) + \theta_{2,2}^T h\left(\frac{1}{\min\{t+1,6\}}\sum_{k=t-5,k\geq 0}^{t} A_k\right)$$

$$\begin{aligned}\theta_{3,t} f^\dagger(\bar{A}_t) = \theta_{3,1}^T h\left(\frac{1}{\max\{t-5,1\}}\sum_{k=0}^{t-6} A_k\right) + \theta_{3,2}^T g(t+1) h\left(\frac{1}{\max\{t-5,1\}}\sum_{k=0}^{t-6} A_k\right) \\ + \theta_{3,3}^T h\left(\frac{1}{\min\{t+1,6\}}\sum_{k=t-5,k\geq 0}^{t} A_k\right) + \theta_{3,4}^T g(t+1) h\left(\frac{1}{\min\{t+1,6\}}\sum_{k=t-5,k\geq 0}^{t} A_k\right)\end{aligned}$$

A flexible restricted cubic spline function of adherence, $h(\cdot)$ (knots at 5$^{\text{th}}$, 35$^{\text{th}}$, 65$^{\text{th}}$, and 95$^{\text{th}}$ percentiles), and time, $g(\cdot)$ (knots at visits 6, 12, 18, 30, 42, and 48), was used in all models.

6. Inverse probability weights

The time-varying stabilized weights for each patient at each time, $t+1$, are defined as:

$$SW_{t+1} = SW_t^A SW_t^M SW_{t+1}^C$$

where

$$SW_t^A = \prod_{k=0}^{t} \frac{f(A_k|\bar{A}_{k-1}, V, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0)}{f(A_k|\bar{A}_{k-1}, V, \bar{L}_{k-1}, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0)}$$

$$SW_t^M = \prod_{k=0}^{t} \frac{f(M_k|\bar{A}_{k-1}, V, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0)}{f(M_k|\bar{A}_{k-1}, V, \bar{L}_{k-1}, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0)}$$

$$SW_{t+1}^C = \prod_{k=1}^{t+1} \frac{p(C_k = 0|\bar{A}_{k-1}, V, \bar{C}_{k-1} = \bar{Y}_{k-1} = \bar{0}, Z = 0)}{p(C_k = 0|\bar{A}_{k-1}, V, \bar{L}_k, \bar{C}_{k-1} = \bar{Y}_{k-1} = \bar{0}, Z = 0)}$$

and V is a vector of covariates measured only at baseline (educational status, age at baseline, physical activity level at work at baseline, race, and baseline risk strata), L_t is a vector of time-varying covariates, and M_t is an indicator of adherence measurement at time t (e.g., $M_1 = 1$ reflects that a participant attended visit seven and had their value of adherence between visit six and visit seven recorded at that time). Overbars denote history of the variable.

To estimate $f(A_k|\bar{A}_{k-1}, V, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0)$, we modelled adherence using a truncated normal distribution with constant variance, and a truncated normal distribution with heteroskedasticity. We estimated the conditional mean using a pooled linear model:

$$\hat{\mathbb{E}}\Big[A_k|\bar{A}_{k-1}, V, M_k = 1, \bar{C}_k = \bar{Y}_k = Z = 0\Big] = \delta_{0,k} + \delta_1^T g(\bar{A}_{k-1}) + \delta_2^T V.$$

Similarly, for $f(A_k|\bar{A}_{k-1}, V, \bar{L}_{k-1}, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0)$, we estimated the conditional mean using the analogous pooled linear model:

$$\hat{\mathbb{E}}\Big[A_k|\bar{A}_{k-1}, V, \bar{L}_{k-1}, M_k = 1, \bar{C}_k = \bar{Y}_k = \bar{0}, Z = 0\Big] = \psi_{0,k} + \psi_1 g(\bar{A}_{k-1}) + \psi_2^T V + \psi_3^T \bar{L}_{k-1}$$

where $g(.)$ denotes a flexible restricted cubic spline function of previous adherence.

The estimated conditional mean and estimated variance were used to estimate the conditional density using a truncated normal distribution. To allow for heteroskedasticity, we also modelled the conditional variance using a pooled loglinear model with each observation's squared residual used as the outcome. In order to avoid extreme weights which resulted from modelling the conditional variance using all baseline and post-baseline covariates, we modelled the conditional variance using a parsimonious model which included only baseline covariates.

We also considered a hurdle gamma distribution and hurdle beta distribution for the numerator and denominator of the weights. For the hurdle gamma distribution, we fit a pooled binary logistic regression model for the probability of zero

adherence (vs. non-zero adherence), and a pooled gamma regression model with an inverse link for the mean of the non-zero adherence observations. For the hurdle beta distribution, we scaled adherence, which ranged from 0% to 130% in the data because participants were given an emergency supply of medication beyond their prescribed dose, to the range [0, 1]. We then fit a pooled multinomial logistic regression for the probability of zero, one, or non-zero/non-one scaled adherence, and a beta regression model with a logit link for the mean of the non-zero/non-one adherence. These models included the same covariates as the pooled linear models described previously. For both the beta and gamma distributions, we considered both a constant dispersion and non-constant dispersion assumption. To relax the assumption of constant dispersion, we modelled the dispersion as a function of i.) all baseline and post-baseline covariates or ii.) the baseline covariates only.

Pooled logistic models were fit to estimate the numerator and denominator of SW_t^M and SW_{t+1}^C. SW_t^A was set to 1 for person-visits without measured adherence (i.e., those for which adherence was carried forward from prior visits). We truncated the estimated weights, SW_t at the 99.9$^{\text{th}}$ percentile to protect against potential model misspecification and near violations of positivity.

Because the covariates measured only at baseline, V, were included in the numerator of the stabilized inverse probability weights, these variables were also included in the weighted outcome model.

7. Parametric g-formula

In general, for an outcome Y, intervention $\bar{A}_k = \bar{a}_k^*$, time-varying covariates L, time-fixed baseline covariates V, and visit times $k \in \{0, \ldots, t\}$, the g-formula can be written as:

$$\sum_{V}\sum_{\bar{l}_k}\sum_{k=0}^{K}\Pr\Big(Y_{k+1}=1\Big|V=v,\bar{L}_k=\bar{l}_k,\bar{A}_k=\bar{a}_k^*,\bar{C}_{k+1}=\bar{Y}_k=\bar{0}\Big)$$
$$\prod_{j=0}^{k}\Bigg\{f\Big(l_j,a_j^*\Big|v,\bar{l}_{j-1},\bar{a}_{j-1}^*,\bar{C}_j=\bar{Y}_{j-1}=\bar{0}\Big)$$
$$\Pr\Big(Y_j=0\Big|V=v,\bar{L}_{j-1}=\bar{l}_{j-1},\bar{A}_{j-1}=\bar{a}_{j-1}^*,\bar{C}_j=\bar{Y}_{j-1}=\bar{0}\Big)\Bigg\}f(V). \tag{1}$$

For more details, see Robins [1], Young et al [2], and Lin et al [3]. In the presence of one or more continuous covariates (as occurs in the observed data), this non-parametric estimation is not feasible. However, the densities in Equation (1) can

be estimated under parametric assumptions and then Monte Carlo simulation can be used to approximate the sum of interest over all covariate histories. In our analysis, this was accomplished by the procedure outlined in the following sections, which was implemented with the gfoRmula package in R [3].

7.1. Parametric assumptions

Again, let $\bar{L}_k$ be a vector of the j covariates and their histories at time k, for each $k = 0, \ldots, K$.

$$\begin{aligned} f\Big(l_j, a_j^* \Big| \bar{l}_{j-1}, \bar{a}_{j-1}^*, v, \bar{C}_j = \bar{Y}_{j-1} = \bar{0}\Big) = \\ f\Big(l_{j,k} \Big| l_{j-1,k}, l_{j-2,k}, \ldots, l_{1,k}, \bar{l}_{k-1}, \bar{a}_{k-1}^*, v, \bar{Y}_k = \bar{C}_{k+1} = 0\Big) \\ f\Big(l_{j-1,k} \Big| l_{j-2,k}, l_{j-2,k}, \ldots, l_{1,k}, \bar{l}_{k-1}, \bar{a}_{k-1}^*, v, \bar{Y}_k = \bar{C}_{k+1} = 0\Big) \\ \ldots \\ f\Big(l_{1,k} \Big| \bar{l}_{k-1}, \bar{a}_{k-1}^*, v, \bar{Y}_k = \bar{C}_{k+1} = 0\Big) \end{aligned} \tag{2}$$

such that $f\Big(l_{1,k} \Big| l_{j-1,k}, l_{j-2,k}, \ldots, l_{1,k}, \bar{l}_{k-1}, \bar{a}_{k-1}^*, v, \bar{Y}_k = \bar{C}_{k+1} = 0\Big) = 0$ when $\sum_{s=1}^{6} l_{1,k-s} = 0$ for $k \geq 6$, where $L_{1,k}$ is an indicator random variable for missed visit. This condition implies that a participant who has had 5 missed visits, must have an unmissed next visit.

7.1.1. Covariate models

Using the interval (discrete time) data structure, we regressed each covariate on a function of treatment history and covariate history at times k and $k-1$, conditional on survival and uncensored status. For covariate history at times k and $k-1$, we assumed the following topological order:

$$V, L_{1,k}, A_k, L_{j,k}, C_{k+1}, L_{m,k}, Y_{k+1}$$

for $L_{1,k}$ as an indicator for missed visit at visit k, $j \in \{2, \ldots, J\}$ for $J-1$ time-varying covariates, and $m \in \{J+1, M\}$ for $M-J$ non-outcome events that could be measured at times other than the visit. The order of elements in the covariate vector L_j was taken to be an arbitrary permutation.

The expected values of binary covariates were estimated using binary logistic regression. The expected values of continuous covariates were estimated by using linear regression. The expected values of categorical covariates were estimated using multinomial logistic regression. All covariate regressions were conditioned on prior survival and uncensored status, history as described above, and restricted cubic spline (knots at 5th, 35th, 65th, and 95th percentiles) function of time.

7.1.2. Outcome model

A discrete time hazards outcome model was fit, conditional on covariate history, treatment history, prior survival, and prior uncensored status.

$$\text{logit}\Big(\hat{\Pr}(Y_t = 1|\bar{A}_{t-1}, \bar{L}_{t-1}, V, \bar{C}_t = \bar{Y}_{t-1} = \bar{0}\Big) = \beta_{0,t} + \beta_1^T V + \beta_{2,t}^T \bar{L}_{t-1} + \beta_{3,t}^T f(\bar{A}_{t-1}) \quad (3)$$

in which, analogous to the weighted discrete time hazards model described above, the intercept may vary with time and $f(\bar{A}_{t-1})$ is a dose-response function (average adherence since the beginning of the study), the effect of which was permitted to vary over time.

7.2. Monte Carlo simulation

From the observed data, 10,000 participants were sampled with replacement. For each of these resampled observations, using parameter estimates from the covariate models described in Section 7.1.1, the covariate $L_{m,k}$ was drawn from the density functions estimated above, based on previously drawn covariates at times k and $k-1$, baseline covariates, V, and assigned treatment, conditional on prior survival and uncencored status, for $k \in \{0, \ldots, K\}$ visits and $m \in \{1, \ldots, M\}$ covariates. Under the visit process assumptions described above, if $L_{1,k} = 1$, then $L_{\setminus 1,k} = L_{\setminus 1,k-1}$.

The static treatment regimes under comparison were $\bar{a} = 1$ (i.e., 100% adherence) and $\bar{a} = 0$ (i.e., 0% adherence).

7.3. Expected counterfactual risk

To compute the risk under treatment regime $\bar{a}_k$, we estimated the g-formula as the following sample average in the pseudopopulation

$$\hat{\mathbb{E}}\Bigg\{ \sum_{k=0}^{K} \hat{\Pr}\Big(Y_{k+1} = 1 \Big| V = v, \bar{L}_k = \bar{l}_k, \bar{A}_k = \bar{a}_k, \bar{Y}_k = \bar{C}_{k+1} = \bar{0}\Big) \prod_{s=0}^{k} \Big[1 - \hat{\Pr}\Big(Y_s = 1 \Big| V = v, \bar{L}_{s-1} = \bar{l}_{s-1}, \bar{A}_{s-1} = \bar{a}^*_{s-1}, \bar{Y}_{s-1} = \bar{C}_s = 0\Big)\Big] \Bigg\} \quad (4)$$

based on the coefficients from the regression model fit in Equation (3).

Appendix Table 1: Estimated risk differences at 7 years (95% confidence interval) comparing 100% adherers to placebo at each visit versus 0% adherers at each visit in the Lipid Research Clinics Coronary Primary Prevention Trial in a sensitivity analysis which set the baseline visit at visit seven.

Dose-response model	Unadjusted	Adjusted for baseline covariates	Adjusted for baseline and post-baseline covariates with adherence modelled using a truncated normal distribution assuming constant variance (homoscedastic)	Adjusted for baseline and post-baseline covariates with adherence modelled using a truncated normal distribution with modeled variance (heteroscedastic)	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a gamma distribution assuming constant dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a gamma distribution with modeled dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a beta distribution assuming constant dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a beta distribution with modeled dispersion
Average adherence over entire follow-up	-8.3% (-18.7, -0.3)	-7.9% (-19.5, -0.5)	1.1% (-16.6, 8.4)	-0.1% (-17.0, 7.0)	0.5% (-15.1, 7.4)	0.7% (-21.3, 7.6)	2.9% (-14.0, 10.5)	0.7% (-15.6, 7.7)
Average adherence in the last visit plus average adherence prior to the last visit	-9.1% (-20.5, -0.8)	-8.9% (-21.3, -1.2)	-0.2% (-17.7, 6.6)	-1.4% (-19.8, 4.6)	-0.5% (-17.5, 5.9)	-0.5% (-22.0, 6.0)	2.1% (-14.5, 10.2)	-0.6% (-18.2, 5.5)
Average adherence in the last 6 visits plus average	-7.3% (-18.7, 1.0)	-7.2% (-19.8, 1.0)	2.1% (-14.4, 8.5)	0.7% (-18.2, 6.7)	1.3% (-15.7, 7.5)	0.9% (-22.9, 7.2)	3.2% (-14.5, 10.2)	0.7% (-17.2, 7.3)

Appendix Table 2: Estimated risk differences at 7 years (95% confidence interval) comparing 100% adherers to placebo at each visit versus 0% adherers at each visit in the Lipid Research Clinics Coronary Primary Prevention Trial in a sensitivity analysis which carried forward the adherence and covariate values from an individual's most recently attended visit for up to three consecutive visits.

Dose-response model	Unadjusted	Adjusted for baseline covariates	Adjusted for baseline and post-baseline covariates with adherence modelled using a truncated normal distribution assuming constant variance (homoscedastic)	Adjusted for baseline and post-baseline covariates with adherence modelled using a truncated normal distribution with modeled variance (heteroscedastic)	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a gamma distribution assuming constant dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a gamma distribution with modeled dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a beta distribution assuming constant dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a beta distribution with modeled dispersion
Average adherence over entire follow-up	-7.1% (-21.9, 2.3)	-5.7% (-18.9, 2.9)	2.0% (-10.1, 7.2)	0.0% (-16.9, 6.0)	1.0% (-26.1, 5.9)	0.7% (-31.8, 7.7)	3.2% (-18.7, 8.3)	0.8% (-25.9, 6.2)
Average adherence in the last visit plus average adherence prior to the last visit	-7.9% (-24.1, 1.1)	-6.2% (-20.0, 1.9)	1.2% (-12.7, 5.4)	-0.8% (-19.3, 4.2)	0.2% (-27.8, 4.8)	-0.5% (-35.8, 6.4)	2.4% (-34.0, 6.6)	-0.3% (-29.5, 4.8)
Average adherence in the last 6 visits plus average	-7.8% (-23.7, 2.2)	-5.9% (-21.0, 2.8)	1.7% (-13.1, 6.0)	-0.5% (-21.7, 5.1)	0.7% (-33.3, 5.5)	0.0% (-40.1, 7.9)	2.3% (-40.4, 7.1)	0.1% (-30.9, 5.4)

Appendix Table 3: Estimated risk differences at 7-years (95% confidence interval) comparing 100% adherers to placebo at each visit versus 0% adherers at each visit in the Lipid Research Clinics Coronary Primary Prevention Trial in a sensitivity analysis which carried forward the adherence and covariate values from an individual's most recently attended visit for up to four consecutive visits.

Dose-response model	Unadjusted	Adjusted for baseline covariates	Adjusted for baseline and post-baseline covariates with adherence modelled using a truncated normal distribution assuming constant variance (homoscedastic)	Adjusted for baseline and post-baseline covariates with adherence modelled using a truncated normal distribution with modeled variance (heteroscedastic)	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a gamma distribution assuming constant dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a gamma distribution with modeled dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a beta distribution assuming constant dispersion	Adjusted for baseline and post-baseline covariates with adherence modelled using a hurdle model with a beta distribution with modeled dispersion
Average adherence over entire follow-up	-7.3% (-21.8, 1.7)	-5.9% (-18.8, 2.3)	1.8% (-12.7, 7.4)	-0.3% (-18.6, 5.9)	0.0% (-27.7, 6.4)	-0.4% (-33.6, 7.9)	2.6% (-18.7, 8.8)	-0.1% (-26.9, 6.0)
Average adherence in the last visit plus average adherence prior to the last visit	-7.7% (-22.9, 1.2)	-6.2% (-20.1, 1.6)	0.7% (-15.3, 5.5)	-1.5% (-21.0, 4.1)	-0.9% (-30.1, 5.0)	-1.7% (-38.4, 5.0)	1.5% (-22.9, 6.7)	-1.5% (-30.4, 4.3)
Average adherence in the last 6 visits plus average	-6.3% (-22.1, 3.3)	-4.8% (-18.6, 3.8)	1.8% (-13.4, 6.4)	-0.4% (-20.4, 4.9)	0.1% (-28.9, 5.6)	-0.8% (-37.2, 6.0)	2.5% (-25.0, 7.3)	-0.6% (-31.2, 5.2)

References

[1] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical modelling. 1986;7(9-12):1393–1512.

[2] Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. Statistics in biosciences. 2011;3(1):119.

[3] Lin V, McGrath S, Zhang Z, Petito LC, Logan RW, Hernán MA, et al. gfoRmula: An R package for estimating effects of general time-varying treatment interventions via the parametric g-formula. arXiv preprint arXiv:190807072. 2019;.