Supplement 2: Multiple Imputation and Marginal Structural Model

MULTIPLE IMPUTATION
We conducted the multiple imputation procedures only as sensitivity analyses as recommended by Groenwold RH et al (Am J Epidemiol 2012;175(3):210-7). We imputed the end-point values for 544 participants (252 dropouts and 292 participants without available measurement of blood glucose during follow-up). We assumed that our missing data were missing at random (MAR), in the sense that the probability that data were missing did not depend on unobserved data but might depend on observed data. Under the assumption of MAR, the missing-data values do not contain any additional information beyond given observed data about the missing-data mechanism. An even more stringent assumption of missing completely at random (MCAR) would be tenable in the 292 participants without glucose measurements because only organizational or administrative problems, related to the characteristics of health centers but not to the clinical characteristics of participants, were the cause of missing glucose values during follow-up. Regarding the 252 participants who were dropouts, the probability of dropout depended from several baseline variables and this allowed us to hold the MAR assumption and to include these variables as predictors in the imputation procedure.

We imputed both the extra length of follow-up (to replace the missing information for time of follow-up among participants lost to follow-up) and the occurrence of events among both dropouts and participants lacking information on glucose.

We used the multivariate normal regression procedure and the logistic regression procedure for length of follow-up and event, respectively, in Stata 12.1 (mi estimate command). The imputation methods available in Stata obtain imputations by simulating from a Bayesian posterior predictive distribution of missing data (or its approximation) under the conventional (or chosen) prior distribution. When a pattern of missing values is arbitrary, iterative methods are used by Stata to fill in missing values. The multivariate normal method uses multivariate normal data augmentation to impute missing values of continuous imputation variables. An alternative method is imputation using chained equations (ICE), also known as imputation using fully conditional specifications and as sequential regression multivariate imputation in the literature. The ICE method uses a Gibbs-like algorithm to impute multiple variables sequentially using univariate fully conditional specifications. Despite a lack of rigorous theoretical justification, the flexibility of ICE has made it one of the most popular choices used in practice. As a secondary analysis, we repeated our estimates using ICE instead of multivariate normal estimation, and the results were similar.

We imputed the outcome, not the independent covariates (which only had <1% missing values), and we used also the allocated dietary groups as predictors. Therefore, predictors for the imputations were sex, age, smoking, baseline dyslipidemia, baseline hypertension, baseline adherence to Mediterranean diet, leisure-time physical activity, educational level, total energy intake, alcohol consumption, body mass index, waist circumference, and group allocation.

We run 20 sets of random imputations. The estimates of multivariate-adjusted relative risks using multiple imputation averaged across 20 sets of data (taking into account both between-set and within-set variances) according to the algorithm implemented in Stata 12.1 for Poisson regression were consistent with the findings of the primary analysis, with relative risks estimates of 0.70 (0.52-0.94) for the MedDiet+EVOO and 0.82 (0.62-1.09) for the MedDiet+nuts when we only imputed the outcome for dropouts (participants without contact for 2 years or longer). These estimates were 0.73 (0.55-0.98) and 0.83 (0.63-1.11) when we used chained equations (ice method).

When we imputed the outcome (diabetes) to all participants without contact for 2 years or longer plus to those who lacked repeated measurements of glucose control, the estimates were 0.74 (0.55-0.99) and 0.86 (0.56-1.31), respectively. These estimates were 0.74 (0.55-0.99) and 0.83 (0.63-1.11) when we used chained equations (ice method).

MARGINAL STRUCTURAL MODEL
We used the marginal structural model to provide the results of an alternative technique to analyze the data as if it were from an observational study rather than a randomized, controlled trial.

The conditional probabilities of being allocated to each of the three groups (MedDiet+EVOO, MedDiet+nuts or control diet) were estimated using three logistic regression models with the dichotomous dependent variable of the respective actual assigned group and age, sex, educational level, body mass index, waist circumference, baseline adherence to MedDiet and total energy intake as independent variables.

The respective 3 marginal probabilities corresponded to the fraction of participants allocated to each of the 3 groups. The inverse probability weights were built as the ratio of marginal to conditional probabilities. We fitted a Poisson regression model weighted by the inverse probability weights and using the sandwich-type estimate (robust) for the variance.

We believe that further expansions of these analyses to also address the issue of incomplete adherence to the assigned diets with repeated measurements of adherence will exceed the aims of this report.