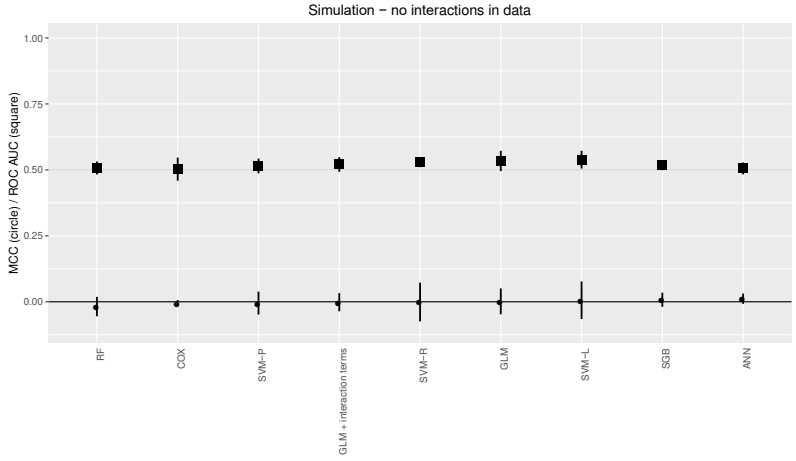
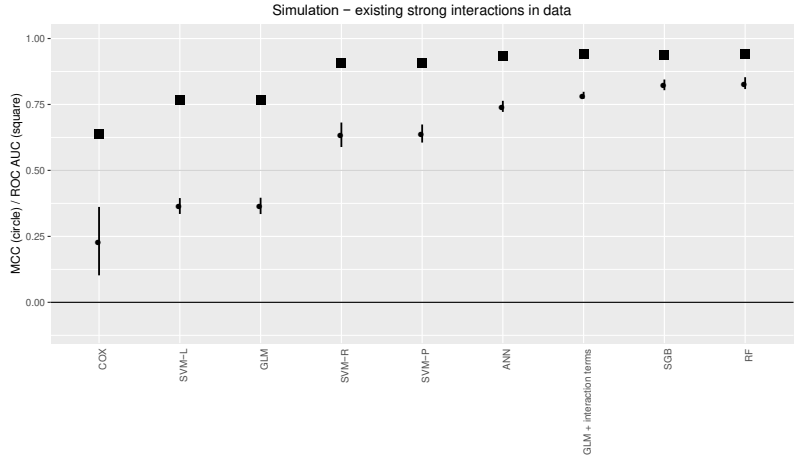
Supplemental Text 3. Simulation study on the ability of algorithms to detect effect modification.

This simulation study was performed to show whether machine learning algorithms are able to detect latent effect modifications (i.e. statistical interactions) in a predictive model without fitting explicit interaction terms. The authors generated an artificial dataset that resembles the original DPP dataset in terms of size, variable means, standard deviations and variable correlations. Input features were restricted to those in Model 8 (age, sex, self-reported ethnicity, fasting glucose, HbA1c, treatment arm, T2D family history, gestational diabetes mellitus history, BP medication use, waist, BMI, systolic blood pressure, postprandial glucose, insulinogenic index, inverse fasting insulin). The outcome T2D (0/1) and the related censored time-to-event variable were generated randomly, without any pre-set correlations to any other variables in the dataset. In a subsequent step, strong statistical interactions were generated in the dataset by artificially setting the outcome to 1 (incident T2D) above a certain BMI value (>31) in the placebo control arm only, above a certain age value (>55) in the lifestyle intervention arm only and above a certain insulinogenic index value (>8) in the metformin arm only. In addition to the statistical and machine learning algorithms used in the paper, an extra method was added, termed *GLM + interaction terms*. This method was a logistic regression model using the same covariates as the *GLM* method, with additional "BMI×treatment arm", "age×treatment arm" and "insulinogenic index×treatment arm" multiplicative interaction terms. The same five-fold nested cross-validation framework was utilized as in the main analysis.



The first figure shows the discriminative utilities (mean MCC by • , mean ROC AUC by ■, SDs denoted by vertical lines) in the dataset where the outcome has no statistical relationship with the features used for prediction. As it is apparent from the figure, MCC values are close to 0 and ROC AUC values are close to 0.5, representing no discriminative utility and no difference is apparent between the algorithms.

The second figure shows the discriminative utilities in the dataset with the outcome set to be correlated with age, BMI and insulinogenic index in the three distinct treatment arms. The *GLM*, *COX* and *SVM linear* models perform poorly, as without explicit interaction terms, the effects of age, BMI and IFI in the three treatment arms are not detected. However, by fitting explicit interaction terms, the discriminative utilities dramatically improve, as shown by the MCC and ROC AUC values for the *GLM + interaction terms* method. Notably, all the rest of the machine algorithms used in this manuscript are able to "pick up" on the interaction signal without fitting explicit interaction terms – *SVM radial, SVM polynomial, ANN, SGB* and *RF* perform comparably to the *GLM + interaction terms* method.



This simulation experiment shows the potential of the *SVM radial, SVM polynomial, ANN, SGB* and *RF* methods to outperform *GLM* and *COX* when latent interactions are present, without fitting explicit interaction terms.