Editor Comments:  
1) Line 173: For MESA cohort, HbA1c was measured but was not used as a criteria for DM. Any reason why?

**Notes**: Because we start follow-up at exam 1 or 2 depending on when first meet criteria for prediabetes, do not want different definitions of diabetes applied. Mike

Can authors clarify if there were preventive programs to prevent diabetes in MESA cohort?

**Notes**: There were none. Mike

The results are a bit difficult to read with references to supplementary materials. Can the authors reorganized to provide clarity for readers who are not statistics savvy.

**Notes**: [come back to, no good answer right now] Mike

Line 276, 278: what is preventive optimal intervention?

**Notes**: Reword to optimal... Mike

What is the potential clinical impact of the study findings? Is the intent for the study findings to serve as information to inform patients of the 3-year risk of diabetes when they enter preventive programs?

**Notes**: Not about “when” someone is starting but (1) identifying current risk for diabetes and (2) about how starting a prevention program could reduce their risk (magnitude) and (3) which preventive program would be optimal for that person. Contrast programs.

Figures: not clear. Need to resubmit.  
**Notes**: DPIs? Byron

**Reviewer Comments:**  
Reviewer 1: In this study, Byron Jaeger and coworkers, developed and validated a prediction model of the risk of diabetes. The Diabetes Prevention Program (DPP; n=2640) cohort was used for discovery while the Multi-Ethnic Study of Atherosclerosis (MESA; N=2104) was used to externally validate the model. Sex, glycated hemoglobin, fasting plasma glucose (FPG), body mass index (BMI), triglycerides, and intervention (and pairwise interactions between intervention and age, FPG, and BMI) were included in the standard and in the individualized intervention model. The mean C-statistic was 0.71 (95% CI: 0.68, 0.74) in DPP and 0.86 (95% CI: 0.83, 0.88) in MESA. Some (not many) of the performance metrics Model performance metrics were similar across different race/ethnicity groups and tended to be higher in the individualized vs. the standard model. The study adds some new knowledge on the subject. The data are well interpreted and discussed. The paper is well written. I have the following suggestions

Methods  
Please add also data on the integrated discrimination improvement (IDI; both as absolute value and % relative IDI) when comparing the individualized vs. the standard model. This will help understand if and how the two models are different (I am not sure this is the case).  
**Notes**: Add to summary table. Byron

We have added the IDI to Table X of the revised manuscript. To compute IDI for a risk prediction model, we used the survIDINRI package. The routines in this package allow for IDI to be computed at specific times (i.e., 3 years after baseline), but do not currently support confidence limits for the % relative IDI.

Results  
1.The aim is to predict incident diabetes which is defined by FBG and/or Hba1C levels during follow-up. Obviously, both FBG and Hba1C at baseline are in the model. Dos the model perform better than FBG or Hba1C considered individually and in combination? Please make these comparisons and show the data.  
**Notes**: Additional analyses, will report C-stats in response and in paper. Byron and Mike (paper)

C-statistics (95% confidence interval) were .653 (.629, .689) and .610 (.576, .643) for FPG and Hba1c, respectively, for internal validation. In external validation, the respective C-statistics were .772 (.682, .862) and .671 (.551, .791). Noting that the combination of FPG and Hba1c consistently obtained higher C-statistic and index of prediction accuracy compared to the individual variables, we replicated our main analysis with an additional model that was restricted to these two predictors (Table R1)

2. The model performs clearly better (significantly better) in MESA than DPP (higher C-statistic and no overlap of 95% CI) and this should be clearly stated, rather than saying the metrics are similar (also add C statistic value for MESA in the text at page 11, lanes 249-250). I am sure the Authors agree that a model performing significantly better in the validation sample as compared to the discovery cohort is unusual if not surprising. One would like to understand if and which of the differences between the two cohorts (age, different proportion of race/ethnicity, blood glucose, insulin resistance, beta cell function, BMI or educational attainment) explain this difference.  
**Notes**: Large C-statistic when outcome is (more) rare and easier to discriminate low and high risk groups Byron.

Thank you for raising this question. We agree that the C-statistic is higher in the validation data. This occurrence is attributed to the incidence of diabetes being rarer in the validation data (MESA) versus the derivation data (DPP). The C-statistic measures how well a model discriminates between those who experience the outcome and those who do not. It can be interpreted as the probability that a randomly chosen positive case (event) has a higher predicted risk score than a randomly chosen negative case (no event). Consider a scenario with many negative and few positive cases**.** If only a small fraction of observations have the event, the vast majority are non-events. A model can correctly identify almost all non-events (high specificity) while still flagging only a few individuals as high risk. In such a scenario, most event/non-event pairs will appear “concordant” (the event case has a higher score) simply because there are so many non-event cases to compare against. Each true event that the model ranks above a host of nonevents boosts the concordance probability. In other words, with many negatives, each event ranked near the top yields many favorable comparisons in the AUC calculation. In case the reviewer would like to verify this claim, we have included R code that creates a predictor X and outcome with the same association but different rarity and as a result different C-statistics.

library(tidyverse)

set.seed(123)

n <- 10000

data\_common <- tibble(time = round(runif(n) \* 100),

x = runif(n, min = 0.5, max = 1),

status = rbinom(n, size = 1, prob = x))

data\_rare <- tibble(time = round(runif(n) \* 100),

x = runif(n, min = 0, max = 0.5),

status = rbinom(n, size = 1, prob = x))

# outcome is more common => lower C of 0.562

summary(coxph(Surv(time, status) ~ x, data = data\_common))$concordance

# outcome is more rare => Higher C of 0.656

summary(coxph(Surv(time, status) ~ x, data = data\_rare))$concordance

Discussion  
1. Overall, the Discussion is too long and somehow repetitive. It could easily be shortened.

**Notes**: Cut down on discussion, see next comment for some material Mike

2. Please, tone down about the better metrics in the individualized vs. the standard model. Data are not robust in this sense (C statistics is not different, NRI is not statistically significant).  
**Notes**: Cut material Mike  
  
  
Reviewer 2: The authors have developed and validated a diabetes risk prediction model that incorporates individualized preventive intervention effects among racially diverse populations. Overall, this is a well-conducted and valuable study. However, I have several comments that may help improve the quality of the manuscript:  
  
(1) Line 108: There is a considerable variation in participants' ages, which may influence their responses to the intervention. Different age groups might exhibit distinct levels of adherence and effectiveness regarding preventive measures. Did the authors account for these potential differences in their analysis?  
**Notes**: We address differential intervention effects with the age interaction. Mike: track down information on adherence in DPP, any difference by age. If no information, then limitations section includes “we do not have information on whether...by important factors by age”

(2) Line 182: Please provide a reference for the data imputation method used in the study to ensure transparency and reproducibility.

**Notes**: Byron will add reference (nearest neighbor imputation using Gower’s distance.)

Citation: Gower, C. (1971) "A general coefficient of similarity and some of its properties," Biometrics, 857-871.

(3) Line 236: The choice of a three-year follow-up period seems relatively short, particularly considering that the risk of progression to type 2 diabetes (T2D) may differ by age. Older individuals, for example, might have a higher likelihood of developing T2D within a shorter timeframe. Did the authors consider stratifying their analysis by age groups to better capture potential differences in disease progression?  
**Notes**: The DPP sample was enrolled based on having high risk for type 2 diabetes based on weight and glucose status. This is evident from the risk among the placebo arm, almost 30% at 3 years. The original trial was stopped due to the success of the lifestyle intervention. Clinically speaking, it is significant to know whether starting an intervention (lifestyle or metformin) can have beneficial effects almost immediately without having to wait a long time. An age by treatment interaction is included in the prediction model which should account for differential risk and treatment effects by age.

Byron will add cumulative plots for DPP by age and some text.

Thank you for raising this question. We did not initially consider age-stratified analyses, but the reviewer’s comment spurred us to explore whether such an analysis would be feasible. We first explored the feasibility of age-stratified analyses by examining the cumulative incidence of diabetes by age groups (age < 55 years, age 55 to 65 years, and age > 65 years) in the DPP. We did not detect a difference in the cumulative incidence curves (Figure R1; p = 0.18), although we did see that the oldest age group tended to have higher estimated incidence than the younger age groups. We also did not detect an association between continuous age and incident diagnosis of diabetes in a Cox regression model with adjustment for sex, education, triglycerides, BMI, Hba1c, fasting plasma glucose, and randomized treatment (p = 0.67). Given the lack of an identified association between age and incident diabetes, we decided not to pursue age-stratified analyses.

Table R1. Comparison of the individualized model with standard model.

| **Evaluation statistic1** | **Internal Validation2** | | | **External Validation3** | | |
| --- | --- | --- | --- | --- | --- | --- |
| **FPG + Hba1c** | **Standard** | **Individualized** | **FPG + Hba1c** | **Standard** | **Individualized** |
| Overall | | | | | | |
| AUC | 66.8 (63.6 70.0) | 69.8 (66.7 73.0) | 70.7 (67.6 73.9) | 85.4 (82.8 88.0) | 85.8 (83.3 88.3) | 85.6 (83.0 88.1) |
| IPA | 6.1 | 9.9 | 10.4 | 18.0 | 18.8 | 18.6 |
| Women | | | | | | |
| AUC | 66.0 (62.1 70.0) | 68.4 (64.6 72.3) | 69.6 (65.8 73.4) | 85.5 (81.7 89.4) | 85.9 (82.1 89.7) | 85.7 (81.8 89.5) |
| IPA | 5.6 | 8.2 | 9.1 | 18.8 | 19.2 | 18.9 |
| Men | | | | | | |
| AUC | 69.2 (63.6 74.8) | 72.8 (67.3 78.2) | 73.2 (67.7 78.7) | 85.2 (81.8 88.7) | 85.9 (82.5 89.2) | 85.6 (82.3 89.0) |
| IPA | 7.2 | 13.3 | 13.1 | 17.2 | 18.4 | 18.4 |
| NH-Black Race | | | | | | |
| AUC | 64.9 (55.9 74.0) | 69.1 (60.8 77.5) | 71.6 (63.6 79.7) | 83.9 (78.8 88.9) | 85.4 (80.7 90.1) | 85.2 (80.5 90.0) |
| IPA | 4.0 | 6.7 | 7.1 | 16.8 | 17.5 | 17.4 |
| NH-White Race | | | | | | |
| AUC | 68.9 (64.9 72.8) | 70.9 (67.0 74.8) | 71.2 (67.3 75.2) | 87.4 (83.3 91.5) | 87.8 (83.9 91.6) | 87.5 (83.6 91.4) |
| IPA | 7.4 | 11.0 | 11.5 | 19.2 | 19.9 | 19.7 |
| Hispanic Race | | | | | | |
| AUC | 63.7 (56.0 71.4) | 67.4 (59.8 75.1) | 69.4 (62.0 76.7) | 84.5 (79.4 89.6) | 84.1 (78.6 89.5) | 83.8 (78.3 89.4) |
| IPA | 4.2 | 9.1 | 9.6 | 16.0 | 17.8 | 17.5 |
| Other Race | | | | | | |
| AUC | 61.5 (47.2 75.8) | 65.1 (50.1 80.1) | 69.0 (54.5 83.6) | 85.5 (77.0 94.0) | 85.7 (77.4 94.0) | 85.5 (77.1 93.9) |
| IPA | 1.3 | 7.5 | 8.7 | 21.5 | 21.2 | 20.6 |
| Abbreviations: AUC = Area underneath the receiver-operator characteristic curve; and IPA = Index of prediction accuracy | | | | | | |
| 1Table values are scaled by a factor of 100 for ease of interpretation. | | | | | | |
| 2Internal validation results are based on 10-fold cross-validation in the Diabetes Prevention Program data | | | | | | |
| 3External validation results are based on application of models fitted to the Diabetes Prevention Program data to the Multi-Ethnic Study of Atherosclerosis data. | | | | | | |

Figure R1: Cumulative incidence of diabetes diagnosis in the DPP by age group.

A graph showing a line graph

AI-generated content may be incorrect.