Supplementary Material to "A Probabilistic Approach to Extract Qualitative Knowledge for Early Prediction of Gestational Diabetes"

Athresh Karanam*[1], Alexander L. Hayes*[2], Harsha Kokel[1], David M. Haas[2], Predrag Radivojac[3], and Sriraam Natarajan[1]

[1] The University of Texas at Dallas, USA
[2] Indiana University Bloomington, USA
[3] Northeastern University, USA

1 Data Description

The nuMoM2b database contained more than $7,000$ variables across the study participants. It's unlikely that specific knowledge exists for every factor's influence on Gestational Diabetes Mellitus (GDM), so we first performed feature selection. We took the intersection of features selected for discriminatively predicting $P(Y = GDM|X)$ with recursive feature elimination and those found by Lasso—then added *Gravidity* and *Education* to this set. *Gravidity* was an influential variable in a previous study that mined electronic health records for GDM risk factors,[4] *Education* can be a weak indicator of socioeconomic status and we believed there could be background knowledge for how it influenced other factors. This resulted in the set of features in Table 1.

Attribute	Type	Categories
GDM Diagnosed	Boolean	True or False
Gravidity	Ordinal	1, 2, 3+
Ever used tobacco	Boolean	True or False
Smoked in the last three months	Boolean	True or False
Highest education level completed	Ordinal	Six levels from High School to post-graduate
Race Category	Category	Eight categories
Age	Ordinal	< 21, 21–25, 25–29, 29–32, ≥ 32
BMI	Ordinal	low, medium, high

Table 1. Summary of the variables. The "eight Race categories" were: Non-Hispanic White, Non-Hispanic Black, Hispanic, American Indian, Asian, Native Hawaiian, Other, Multiracial.

[*] Equal contribution

[4] Gravidity did not appear to be an informative factor during our feature selection. The study mentioned focused on GDM risk factors for women with parity ≥ 0, whereas the nuMoM2b population was nulliparous (parity $= 0$). Combining these two pieces of information may suggest that gravidity is only informative in a parity > 0 population, but we cannot verify with the data we have available.

We interpreted all variables as ordinal since monotonicity and synergy deal with increasing values (e.g. "BMI increasing implies GDM increasing"). For ordinal variables, this was implicit. For the boolean variables, increasing meant $False \rightarrow True$. We assume *Race* is categorical, but there are known cases where people identifying with a certain category tend to have higher risk of GDM—as mentioned in the caption this order was: Non-Hispanic White, Non-Hispanic Black, Hispanic, American Indian, Asian, Native Hawaiian, Other, and Multiracial.

2 Extracting Qualitative Rules

Algorithm 1 had four hyperparameters. We chose the following settings: $\epsilon_m = 0.003$, $\epsilon_s = 0.02$, $T_M = 0.005$, $T_S = 0.001$

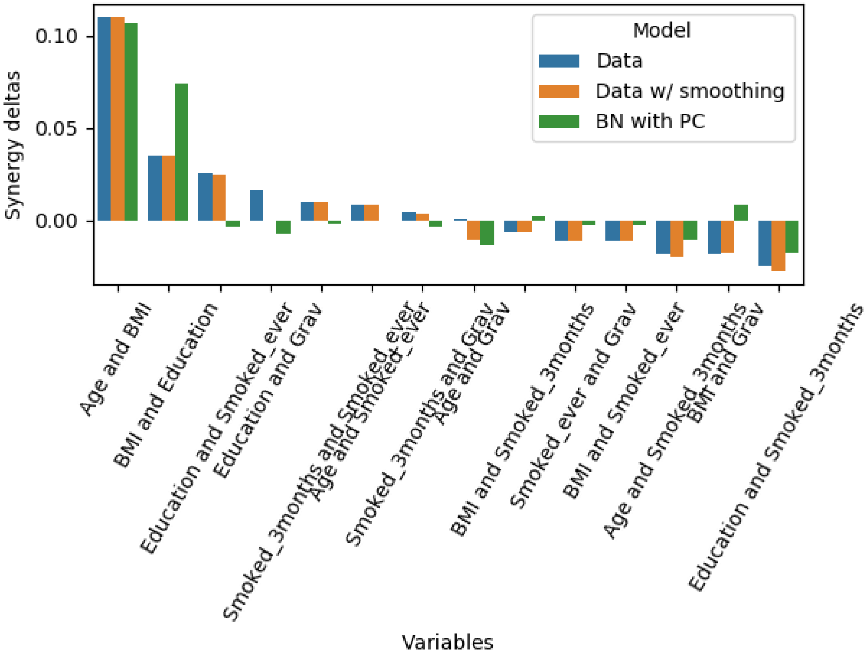

Fig. 1. A positive delta corresponds to a *synergic influence* while a negative delta corresponds to a *sub-synergic influence*. The synergic "Age/BMI" and sub-synergic "Age/Education" were strongly expected ahead of time, and are the two most likely cases extracted by our method, regardless of the technique used to model the joint distribution.