Exploiting Domain Knowledge as Causal Independencies in Modeling Gestational Diabetes

Saurabh Mathur[1], Athresh Karanam[1], Predrag Radivojac[2],David M. Haas[3], Kristian Kersting[4] and Sriraam Natarajan[1]

[1]*Department of Computer Science, University of Texas at Dallas, Richardson, TX 70580, USA*

[2]*Northeastern University, Boston, MA 02115, USA*

[3]*Indiana University School of Medicine Indianapolis, IN 46202, USA*

[4]*Department of Computer Science, TU Darmstadt, and Hessen Center for AI (hessen.AI), Darmstadt, Germany*

We consider the problem of modeling gestational diabetes in a clinical study and develop a domain expert-guided probabilistic model that is both interpretable and explainable. Specifically, we construct a probabilistic model based on causal independence (Noisy-Or) from a carefully chosen set of features. We validate the efficacy of the model on the clinical study and demonstrate the importance of the features and the causal independence model.

Keywords: Probabilistic Models, Bayesian networks

1. Introduction

We consider the problem of predicting the onset of gestational diabetes mellitus (GDM) from a combination of risk factors and a polygenic risk score. To this effect, we consider data from the **Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be** (nuMoM2b[1]) study and develop a probabilistic model for modeling GDM. While the success of deep learning methods[2] in medical tasks[3] has significantly increased the interest in machine learning based methods, these models suffer from the twin problems of being data-hungry and uninterpretable. While quite powerful in their classification abilities, these models are not easy to be employed in decision-making systems that require human interaction.

Consequently, we propose a probabilistic learning method that can effectively and efficiently incorporate domain knowledge. Inspired by previous work in probabilistic learning with expert knowledge,[4,5] we develop a framework for modeling GDM from a few risk factors including Age, BMI, metabolism, family history, blood pressure, etc, and combine the results with a polygenic risk score.

Specifically, our work considers two types of knowledge - causal independencies and quali-

tative influences. Causal independencies[6–9] specify sets of risk factors (called random variables in probabilistic learning terminology) that are independent of each other when affecting the target. The idea here is that each of these variables has an independent effect on the target – for instance, BMI and age affect GDM independently – and their effects can be combined by a probabilistic combination function. One such example is Noisy-Or. The advantage of such independencies lies in the fact that they lead to a drastic reduction in the number of parameters needed to learn the model.

While powerful, specifying only causal independencies could be insufficient. As an example, consider age and BMI as risk factors for GDM. While both these risk factors could be independent, when they both are higher, the risk of GDM could be increased. This information is not captured by simple causal independencies. To model such knowledge, earlier methods employ the use of qualitative constraints.[4,10,11] A qualitative constraint could be a monotonic statement of the form *as X increases Y increases*. For instance, in our task, it is easy to specify that as age increases the risk of GDM increases.

Inspired by our prior work,[5] we combine these two types of domain knowledge to learn a probabilistic model for predicting GDM from the nuMoM2b data and employ the use of polygenic risk score to provide a prior over the incidence of GDM. Specifically, we take the view of a temporal model due to Heckerman and Breese[6] and combine the influence due to the different risk factors using Noisy-Or. For each of these risk factors, we also employ monotonicity constraints whenever applicable. Our empirical evaluations demonstrate that the proposed method with the knowledge from domain experts outperforms probabilistic learning only from data and is comparable with the best machine learning methods that are not interpretable or interactive.

To summarize, we make the following key contributions: (1) We view the problem of modeling GDM using a probabilistic lens and in the presence of domain expert knowledge in the form of qualitative constraints and causal independencies. (2) We take the temporal view and derive the gradients for learning the probabilistic model. (3) We evaluate the algorithm on a real GDM study and establish its effectiveness.

The rest of the paper is organized as follows: after briefly reviewing the nuMoM2b dataset, causal independencies and qualitative influences, we present the derivation of our algorithm for learning a probabilistic model. We then present our empirical evaluation before outlining the areas of future research.

2. Data description

The **Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be** (nuMoM2b[1]) study was established to study individuals without previous pregnancy lasting 20 weeks or more (nulliparous) and to elucidate factors associated with adverse pregnancy outcomes. The study enrolled a racially/ethnically/geographically diverse population of $10,038$ nulliparous women with singleton gestations. The enrolled participants were followed for the duration of their pregnancy and visits were scheduled four times during the pregnancy: 6 weeks 0 days through 13 weeks 6 days estimated gestational age (EGA), 16 weeks 0 days through 21 weeks 6 days EGA, 22 weeks 0 days through 29 weeks 6 days EGA, and at the time of delivery.

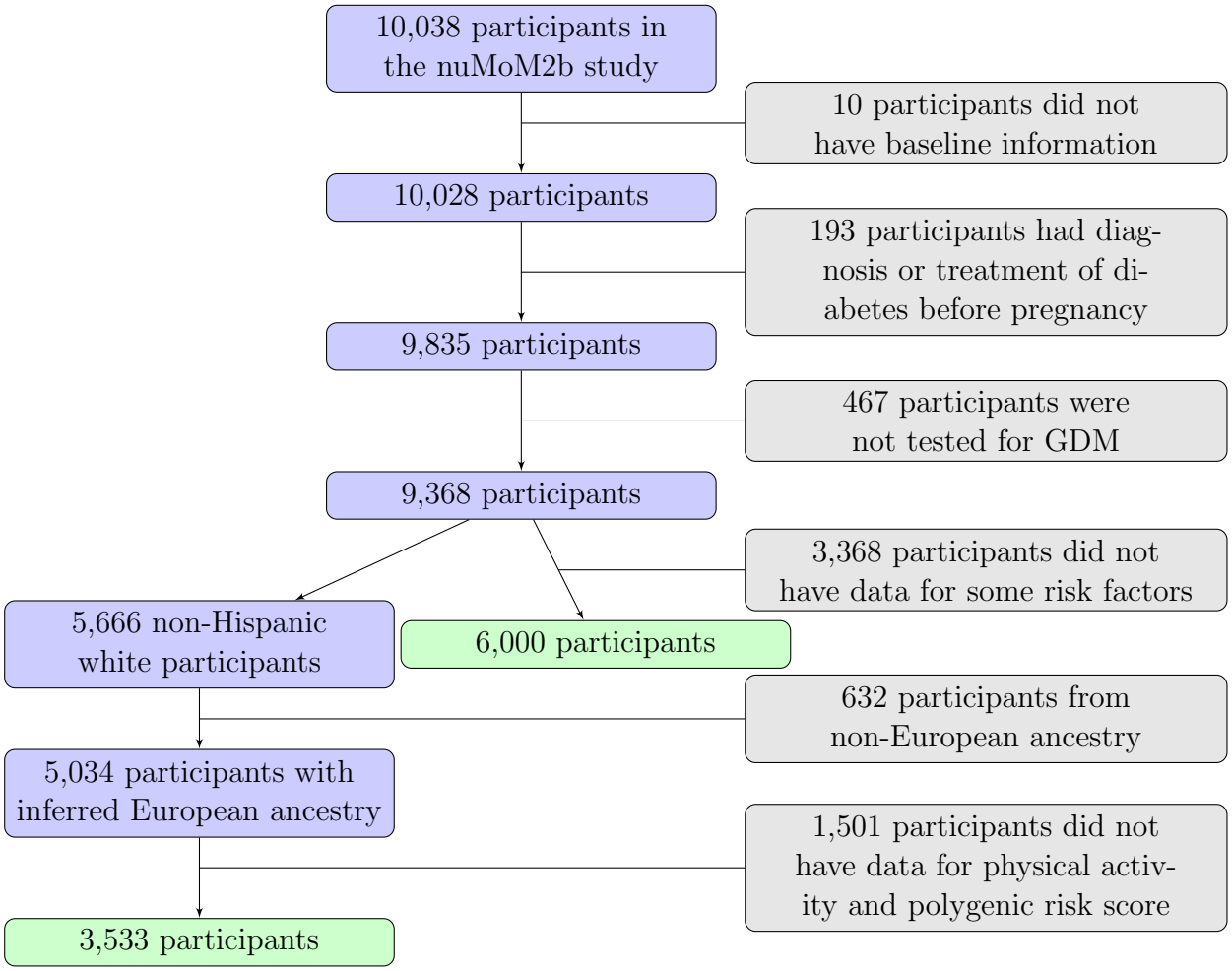

Fig. 1. Flowchart illustrating the process of selecting the cohorts for our experiments. The two sub-cohorts used in our experiments are indicated in green.

Our subset has 7 variables - *BMI, PRS, METs, Age, Hist, PCOS, HiBP*.

For our work, we excluded 193 cases where women were diagnosed with pregestational diabetes. Additionally, $3,368$ cases with missing features in the dataset were excluded. In our experiments, we use two cohorts. Figure 1 illustrates the mechanism for choosing these cohorts. A sub-cohort of $3,533$ non-Hispanic white participants with European ancestry was used for experiments involving *PRS* and a cohort of $6,000$ participants was used for experiments not involving *PRS*. Of the 7 variables, *Hist, PCOS, HiBP* are binary, *Age* is discrete while *BMI, PRS and METs* are continuous. *Age* was categorized into 4 values based on quantiles to limit the number of possible values. The continuous variables *BMI, PRS, and METs* were discretized into 5 categories based on quantiles.

3. Background: Knowledge-guided learning

We now present the necessary background on the two types of expert knowledge that we consider in this work – qualitative influences and causal independencies.

3.1. *Qualitative influence*

A qualitative influence (QI) statement[10] indicates the effect of change in one or more factor(s) on a target.[5] We focus on one particular type of QI: *monotonicty. Monotonicity* represents a direct relationship between two variables: "As BMI increases, neck circumference increases" indicates that the probability of greater neck circumference increases with an increase in BMI. Note that while the QI statements do not directly specify the quantitative relationships (i.e., the precise probabilities), they specify how the conditional distribution ($P(circumference \mid BMI)$ changes as the value of BMI changes. Such statements are quite natural to be specified in many domains, and more so, in medicine. Formally, a *monotonic influence* (MI) of variable X on variable Y, denoted $X \underset{\prec}{}^{M+} Y$ (or its inverse $X \underset{\prec}{}^{M-} Y$), indicates that higher values of X stochastically result in higher (or lower) values of Y.

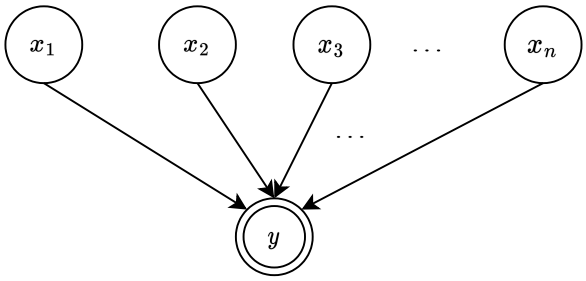

Fig. 2. A belief network for multiple causes and a single effect

3.2. *Causal Independence*

Causal independence, in simple terms, states that (1) the effect is independent of the order in which causes are introduced, and (2) the impact of a single cause on the effect does not depend on what other causes have previously been applied. This definition facilitates a (probabilistic) belief network representation that is consistent with a set of causal independence statements.[6,7] The Noisy-Or model, illustrated in Figure 3, belongs to a class of causal interactions which are characterized by the independence of causal inputs. The belief network in Figure 2 represents a general multiple-cause interaction wherein n causes influence a single effect (target variable) y. While this representation provides an intuitive way to capture the causal interaction between the risk factors x_i and the target variable, it requires 2^n parameter assessments for binary variables - one parameter for each instantiation of the causes. This leads to an exponentially large number of examples required to learn a robust conditional distribution. Akin to conditional independence assumptions in Bayesian networks, causal independence assumptions allow efficient parameter learning by causing an exponential reduction in the total number of model parameters as compared to the case of general multiple-cause interaction. Concretely, the number of parameter assessments required in the Noisy-Or model in Figure 3 is linear in the number of causes, n, while it is exponential in the original model.

Causal independence statements, in conjunction with qualitative influence statements, allow the injection of rich domain knowledge into an interpretable model while ensuring feasible

parameter learning from data. We build upon prior work[5] in employing this knowledge in the context of GDM modeling.

4. Causal independencies with qualitative constraints for modeling GDM

Given: A set of causally independent risk factors $\mathbf{X}$ for the target GDM Y and a set of qualitative influences C
To Do: Learn an interpretable model $\mathfrak{m}$ that models the conditional probability of a target variable given the risk factors.

As mentioned earlier, $\mathbf{X}$ is the set of risk factors $\langle BMI, PRS, METs, Age, Hist, PCOS, HiBP\rangle$ while Y denotes GDM. So the goal of our work is to learn $P(GDM \mid \mathbf{X})$ given the constraints C. In the rest of this section, we use $\mathbf{X}$ and Y instead of specific risk factors and GDM to demonstrate the generality of the approach.

In the Noisy-Or model, the target variable is activated if any of the causes is active, unless the active causes are inhibited. Formally, the probability of a cause being active is called the *link probability* and we parameterize it using the sigmoid function σ, i.e., $P(Y_i = 1 \mid X_i = x_i) = \sigma(w_i x_i + b_i),\ \forall i \in \{1, \ldots, n\}$. The key assumption of the Noisy-Or model is that the inhibitory effect for each cause is independent. Consequently, we parameterize these *inhibition probabilities* as $P(Y = 0 \mid Y_i = 1) = \sigma(q_i),\ \forall i \in \{1, \ldots, n\}$. Finally, the target variable may still be activated even if none of the causes are active. This is called *leakage* and represents all other possible causes that are not included as risk factors. We parameterize the leak probability as $P(Y = 1 \mid Y_1 = 0, \ldots, Y_n = 0) = \sigma(q_l)$. Thus, the target probability under the Noisy-Or model is given as:

$$P(Y = 1 \mid \mathbf{X} = x) = 1 - (1 - q_l) \prod_{i=1}^{n} (P(Y_i = 1 \mid X_i = x_i) q_i + P(Y_i = 0 \mid X_i = x_i)) \tag{1}$$

Following previous work,[4,5] we define positive (or negative) monotonic influence $X_i \overset{M+}{\prec} Y$ (or $X_i \overset{M-}{\prec} Y$) as $P(Y_i = 0 \mid X_i = a) \leq P(Y_i = 0 \mid X_i = b)\ \ \forall a, b \in domain(X_i), a > b$ (or a < b). The Noisy-Or model with monotonic influences is shown in Figure 3.

4.1. *Learning the parameters of the Noisy-Or model using Monotonic Influences*

The log-likelihood under the Noisy-Or model can be written as:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; \mathcal{D}) &= \sum_{j=1}^{N} \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)}) \\ &= \sum_{j=1}^{N} y^{(j)} \log(1 - P(Y = 0 \mid \mathbf{X} = x^{(j)})) + (1 - y^{(j)}) \log P(Y = 0 \mid \mathbf{X} = x^{(j)}) \end{aligned} \tag{2}$$

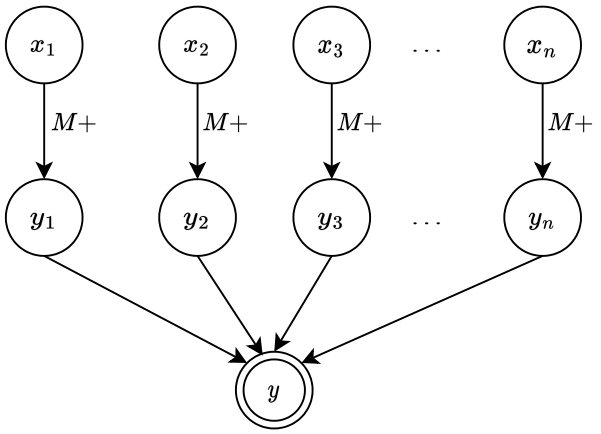

Fig. 3. The Noisy-Or model

We encode the monotonic influences as the margin constraints $\delta_i^{a,b} \leq 0$ where:

$$\delta_i^{a,b} = \begin{cases} P(Y_i = 0 \mid X_i = a) - P(Y_i = 0 \mid X_i = b) + \epsilon & X_i \prec^{M+} Y \in C \\ -P(Y_i = 0 \mid X_i = a) + P(Y_i = 0 \mid X_i = b) + \epsilon & X_i \prec^{M-} Y \in C \\ 0 & otherwise \end{cases}$$

Intuitively, if the monotonicity constraint is satisfied, $\delta \leq 0$ while if the constraint is violated, $\delta > 0$. ϵ is a small margin. Now using these constraints, we define the following penalty function

$$\zeta_i^{a,b} = I_{\delta_i^{a,b} > 0} \delta_i^{a,b\,2}$$

Intuitively, the above penalty is applied if the constraint is violated and is equal to the square of the magnitude of the violation. Essentially, the model will not penalize the cases where the constraints are satisfied (for instance, if the constraint on BMI is satisfied when the parameters are learned, the penalty for that parameter = 0).

Including the penalty function, the final objective that is to be maximized is

$$J(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; D) = \mathcal{L}(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; D) - \lambda \sum_{i=1}^{n} \sum_{a>b} \zeta_i^{a,b}$$

where, λ is the penalty weight. The first term is the classic loglikelihood that is computed using the different conditional distributions and the second term is simply the sum of the non-zero penalties weighted by a constant λ. Recall that $\mathbf{w}$ and $\mathbf{b}$ are the link probability parameters, and $\mathbf{q}$ and q_l are inhibition probability and the leak probability parameters respectively.

Intuitively, the penalty function serves as a regularizer that forces the model to satisfy the constraints as much as possible given the data.

Note that the advantage of this formalism is that since it is a weighted combination,**the data could be noisy or the constraints could be incorrect**. The model can simply trade-off between the data and constraints accordingly. Exploring the case when both data and domain expert are noisy is outside the scope of this work. Thus, the model isr robust to both data noise and expert advice noise. λ could be chosen by cross-validation, but, in our

experiments and in prior work,[5] the model is robust to the choice of λ as long as it is not close to 0 or 1.

4.2. *Derivation of the gradients of the log-likelihood term*

In order to compute the gradients of the log-likelihood term, we define the following intermediate gradient terms:

$$
\begin{aligned}
U_j &= \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})} = \frac{-y^{(j)}}{P(Y = 1 \mid \mathbf{X} = x^{(j)})} + \frac{1 - y^{(j)}}{P(Y = 0 \mid \mathbf{X} = x^{(j)})} \\
Q_{lj} &= \frac{\partial P(Y = 0 \mid X = x^{(j)})}{\partial q_l} = -\frac{P(Y = 0 \mid X = x^{(j)})\sigma'(q_l)}{1 - q_l} \\
Q_{ij} &= \frac{\partial P(Y = 0 \mid X = x^{(j)})}{\partial q_i} = \frac{P(Y = 0 \mid \mathbf{X} = x^{(j)}))P(Y_i = 1 \mid X_i = x_i^{(j)})\sigma'(q_i)}{P(Y_i = 1 \mid X_i = x_i^{(j)})q_i + P(Y_i = 0 \mid X_i = x_i^{(j)})} \\
V_{ij} &= \frac{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})}{\partial P(Y_i = 1 \mid X_i = x_i^{(j)})} = \frac{P(Y = 0 \mid \mathbf{X} = x^{(j)})(q_j - 1)}{P(Y_i = 1 \mid X_i = x_i^{(j)})q_i + P(Y_i = 0 \mid X_i = x_i^{(j)})} \\
W_{ij} &= \frac{\partial P(Y_i = 1 \mid X_i = x_i^{(j)})}{\partial w_i} = \sigma'(w_i x_i + b_i)x_i \\
B_{ij} &= \frac{\partial P(Y_i = 1 \mid X_i = x_i^{(j)})}{\partial b_i} = \sigma'(w_i x_i + b_i)
\end{aligned}
$$

Here, U_j is the gradient of the log likelihood of the jth data example with respect to the probability that the target Y is 0 (i.e., the case where $GDM = false$). $Q_l j$ and $Q_i j$, $V_i j$ are the gradients of the probability that the target is 0 ($GDM = false$) for the jth data example with respect to the leak parameter q_l, the inhibition parameter q_i, and the link probability $P(Y_i = 1 \mid X_i = x_i^{(j)})$ respectively. $W_i j$ and $B_i j$ are the gradients of the link probability $P(Y_i = 1 \mid X_i = x_i^{(j)})$ with respect to its parameters w_i and b_i respectively. Finally σ' is the gradient of the sigmoid function $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

The gradients of the log-likelihood function with respect to the link parameters w_i and b_i can be computed in terms of U_j, V_{ij}, W_{ij} and B_{ij} as

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; \mathcal{D})}{\partial w_i} &= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial w_i} \\
&= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})} \frac{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})}{\partial P(Y = 1 \mid X_i = x_i^{(j)})} \frac{\partial P(Y_i = 1 \mid X_i = x_i^{(j)})}{\partial w_i} \\
&= \sum_{j=1}^{N} U_j V_{ij} W_{ij}
\end{aligned}
\tag{3}
$$

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; \mathcal{D})}{\partial b_i} &= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial b_i} \\ &= \sum_{j=1}^{N} \frac{\log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial P(Y = 0 \mid X = x^{(j)})} \frac{\partial P(Y = 0 \mid X_i = x^{(j)})}{\partial P(Y_i = 1 \mid X_i = x_i^{(j)})} \frac{\partial P(Y_i = 1 \mid X_i = x_i^{(j)})}{\partial b_i} \\ &= \sum_{j=1}^{N} U_j V_{ij} B_{ij}\end{aligned} \tag{4}$$

The gradients of the log-likelihood function with respect to the inhibition and leak parameters q_i and q_l can be computed in terms of U_j, Q_{ij} and Q_{lj} as

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; \mathcal{D})}{\partial q_i} &= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial q_i} \\ &= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})} \frac{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})}{\partial q_i} \\ &= \sum_{j=1}^{N} U_j Q_{ij}\end{aligned} \tag{5}$$

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{w}, \mathbf{b}, \mathbf{q}, q_l; \mathcal{D})}{\partial q_l} &= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial q_l} \\ &= \sum_{j=1}^{N} \frac{\partial \log P(Y = y^{(j)} \mid \mathbf{X} = x^{(j)})}{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})} \frac{\partial P(Y = 0 \mid \mathbf{X} = x^{(j)})}{\partial q_l} \\ &= \sum_{j=1}^{N} U_j Q_{lj}\end{aligned} \tag{6}$$

4.3. *Derivation of the gradients of the penalty term*

The gradients of the penalty function are given by

$$\begin{aligned}
\frac{\partial \zeta_i^{a,b}}{\partial w_i} &= \frac{\partial \zeta_i^{a,b}}{\delta_i^{a,b}} \frac{\delta_i^{a,b}}{\partial w_i} \\
&= I_{\delta_i^{a,b}>0}\ 2\delta_i^{a,b} \frac{\delta_i^{a,b}}{\partial w_i} \\
&= I_{\delta_i^{a,b}>0} \begin{cases} \frac{P(Y_i=0|X_i=a)}{\partial w_i} - \frac{P(Y_i=0|X_i=b)}{\partial w_i} + \epsilon & X_i \prec^{M+} Y \in C \\ -\frac{P(Y_i=0|X_i=a)}{\partial w_i} + \frac{P(Y_i=0|X_i=b)}{\partial w_i} + \epsilon & X_i \prec^{M-} Y \in C \\ 0 & otherwise \end{cases} \\
&= I_{\delta_i^{a,b}>0} \begin{cases} \sigma'(w_i a + b_i)a - \sigma'(w_i b + b_i)b + \epsilon & X_i \prec^{M+} Y \in C \\ -\sigma'(w_i a + b_i)a + \sigma'(w_i b + b_i)b + \epsilon & X_i \prec^{M-} Y \in C \\ 0 & otherwise \end{cases}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\frac{\partial \zeta_i^{a,b}}{\partial b_i} &= \frac{\partial \zeta_i^{a,b}}{\delta_i^{a,b}} \frac{\delta_i^{a,b}}{\partial w_i} \\
&= I_{\delta_i^{a,b}>0}\ 2\delta_i^{a,b} \frac{\delta_i^{a,b}}{\partial b_i} \\
&= I_{\delta_i^{a,b}>0} \begin{cases} \frac{P(Y_i=0|X_i=a)}{\partial b_i} - \frac{P(Y_i=0|X_i=b)}{\partial b_i} + \epsilon & X_i \prec^{M+} Y \in C \\ -\frac{P(Y_i=0|X_i=a)}{\partial b_i} + \frac{P(Y_i=0|X_i=b)}{\partial b_i} + \epsilon & X_i \prec^{M-} Y \in C \\ 0 & otherwise \end{cases} \\
&= I_{\delta_i^{a,b}>0} \begin{cases} \sigma'(w_i a + b_i) - \sigma'(w_i b + b_i) + \epsilon & X_i \prec^{M+} Y \in C \\ -\sigma'(w_i a + b_i) + \sigma'(w_i b + b_i) + \epsilon & X_i \prec^{M-} Y \in C \\ 0 & otherwise \end{cases}
\end{aligned} \tag{8}$$

Using these gradients, we solve the maximization problem using the L-BFGS-B algorithm, increasing the value of λ until the solution satisfies all the constraints. The high-level flowchart of our model construction is presented in Figure 4. Given the entire GDM data set, after preprocessing and obtaining the causal independencies, we construct the smaller data set where we learn the model such that the qualitiative constraints are satisfied. The final model is then evaluated on the test set and the results are presented in the next section.

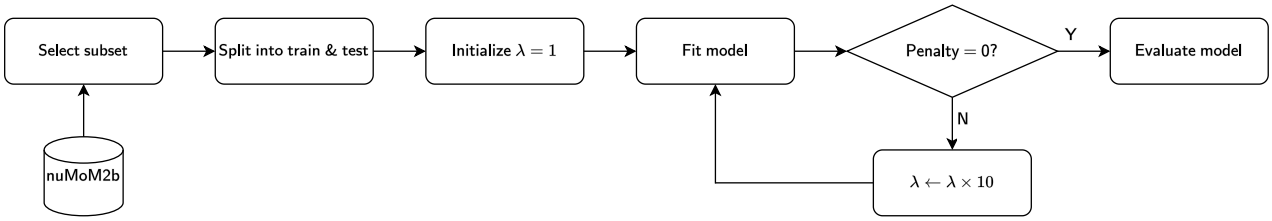

Fig. 4. Flowchart for the Noisy-Or model construction process

5. Experimental evaluation

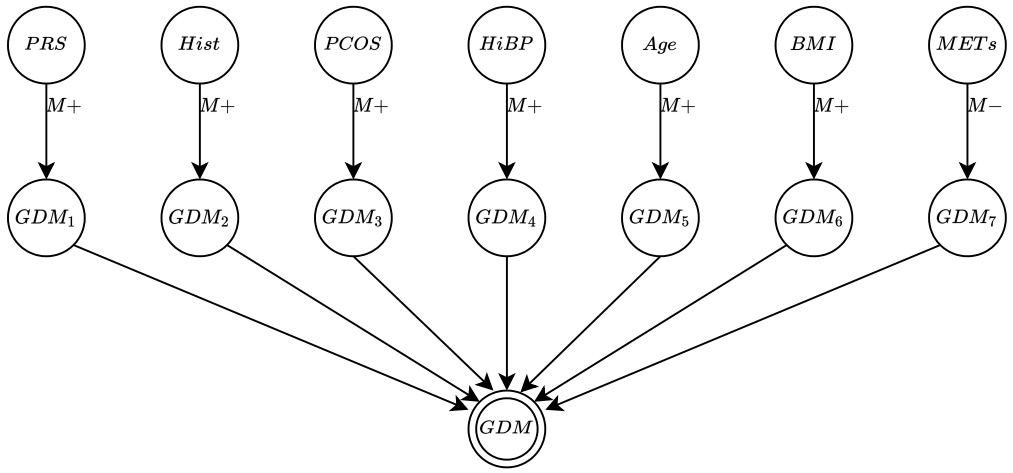

Fig. 5. Noisy-OR model used for the GDM dataset. Both QIs and causal independence knowledge are incorporated in this model. This representation shows that *PRS, Hist, PCOS, HiBP, Age and BMI* have a positive monotonic influence on GDM whereas *METs* have a negative monotonic influence. Additionally, all the risk factors are causally independent in this model.

Our experiments explicitly aim at answering the following questions,

Q1: Does inclusion of QIs improve model performance over a base model that does not have background knowledge in the form of QIs?
Q2: Can our proposed model incorporate causal independencies to efficiently estimate model parameters without significantly losing performance?

We evaluate our proposed approach on two sub-cohorts in the nuMoM2b study - one sub-cohort with *PRS* as a risk factor and one without it - as described in section 2. The domain knowledge in the form of causal independencies and QIs were provided by our domain expert Dr. Haas. Figure 5 presents our proposed noisy-OR model that incorporates this domain knowledge for the task of GDM prediction given the 7 risk factors.

To answer the first question, we train noisy-OR models for the two cohorts with and without the inclusion of QIs. Figure 6 presents the AUC-ROC[12] for our model trained on each of the sub-cohorts. In the case of the sub-cohort using the *PRS* (bottom in Figure 6), it can be clearly noted that incorporating QIs improves AUC-ROC from 0.6543 to 0.7376. In the sub-cohort not using the *PRS*, incorporating QIs improves the AUC-ROC from 0.6577 to 0.7018. It is evident from these charts that the inclusion of QIs as domain knowledge improves model performance. This analysis helps us answer Q1. Our proposed approach can effectively incorporate QIs to improve model performance.

To answer the second question, we compare our proposed approach to a strong discriminative baseline: gradient boosted trees (GBT). Figure 6 presents a comparison of our model with the baseline for the two sub-cohorts (top-left and top-right). GBT achieves AUC-ROC scores of 0.76 and 0.6982 for the sub-cohort with and without *PRS*, respectively. This is comparable

to the performance of our proposed approach when QIs are incorporated. However, unlike the noisy-OR model, GBT does not make any causal independence assumptions and hence has no causal meaning and is much more difficult to interpret. This analysis helps us answer Q2. Our proposed model can incorporate causal independencies to allow feasible parameter learning without losing model performance as compared to models that do not make causal independence assumptions.

To summarize, our experiments on two sub-cohorts of the GDM dataset suggest that our proposed approach can leverage domain knowledge in the form of QIs and causal independencies to effectively and efficiently learn an interpretable model without losing model performance as compared to a strong discriminative baseline that is uninterpretable.

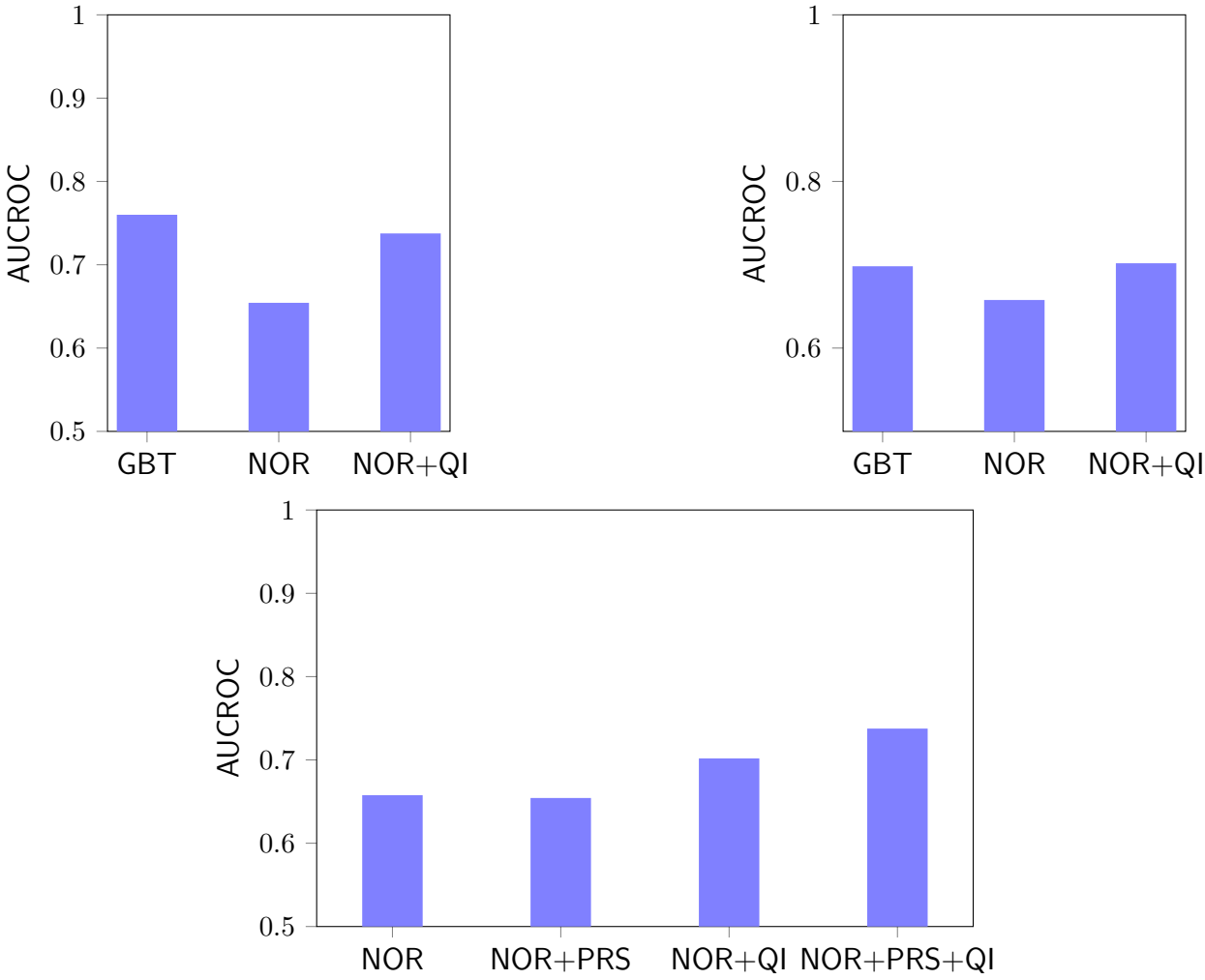

Fig. 6. The AUC-ROC scores for the Noisy OR model (NOR) as compared to the Gradient Boosted Trees model (GBT) with PRS (top left) and without PRS (top right). The AUC-ROC scores for the Noisy OR model (NOR) in the presence of PRS and Qualitative Influences (bottom)

6. Conclusion

We adapted the use of qualitative constraints and causal independencies to build an interpretable and explainable probabilistic model for modeling GDM given a **small number of risk factors**. We presented the learning method that learned the parameters of the model. Our empirical evaluations on nuMoM2b dataset clearly demonstrated that the use of the two types of constraints yielded better results than learning only from data and most importantly, exhibit similar performance as the state-of-the-art machine learning algorithm. Extending the model to include more risk factors is an immediate research direction. Learning a fully generative model such as Bayesian network would provide valuable insights in the interactions between risk factors. Finally, evaluating the learned models on larger and diverse data such as EHRs remains an interesting future direction.

Acknowledgements

The authors acknowledge the support by the NIH grant R01HD101246. KK acknowledges the support of the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) in Germany, project "The Third Wave of AI".

References

1. D. M. Haas, C. B. Parker *et al.*, A description of the methods of the nulliparous pregnancy outcomes study: monitoring mothers-to-be (numom2b), *American journal of obstetrics and gynecology* **212** (2015).
2. I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning* (MIT Press, 2016). `http://www.deeplearningbook.org`.
3. S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert and R. M. Summers, A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises (2020).
4. E. E. Altendorf, A. C. Restificar and T. G. Dietterich, Learning from sparse data by exploiting monotonicity constraints, in *UAI*, (AUAI Press, 2005).
5. S. Yang and S. Natarajan, Knowledge intensive learning: Combining qualitative constraints with causal independence for parameter learning in probabilistic models, in *ECML-PKDD*, (Springer, 2013).
6. D. Heckerman and J. S. Breese, A new look at causal independence, *CoRR* **abs/1302.6814** (2013).
7. S. Srinivas, A generalization of the noisy-or model, in *UAI*, (Morgan Kaufmann, 1993).
8. J. Vomlel, Exploiting functional dependence in bayesian network inference, *CoRR* **abs/1301.0609** (2013).
9. J. Pearl, *Probabilistic reasoning in intelligent systems - networks of plausible inference*Morgan Kaufmann series in representation and reasoning, Morgan Kaufmann series in representation and reasoning (Morgan Kaufmann, 1989).
10. M. P. Wellman, Fundamental concepts of qualitative probabilistic networks, *Artif. Intell.* **44**, 257 (1990).
11. A. J. Feelders and L. C. van der Gaag, Learning bayesian network parameters with prior knowledge about context-specific qualitative influences, in *UAI*, (AUAI Press, 2005).
12. J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve., *Radiology* **143**, 29 (1982).