# The Causal Health Classification Data set

The causal health classification data set is an extension to the causal health data set by Zečević et al. [2021], where three new binary "Diagnosis" variables, in a one-hot configuration, are introduced. In this case one of the diagnose variables is set to true and the remaining ones are set to false for every data point.

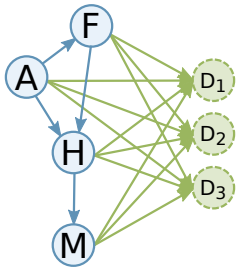


Figure 1: The DAG of the structural causal model of the Causal Health Classification Data set. Variables of the already existing Causal Health data set (blue) with the new diagnosis variables (green/dashed) added. (Best viewed in color.)

For each sample three multivariate polynomial functions are evaluated to determine the activate diagnose. Each function which depends on a subset of the original variables $A, M, F$ and $H$:

$$f_1(A) := \begin{cases} 0.00108A^3 - 0.08862A^2 + 1.337A + \mathcal{N}(25, 10) & \text{if } A \leq 4.09837 \\ \mathcal{N}(5, 10), & \text{otherwise} \end{cases}$$

$$f_2(F, M) := 0.0175F + 0.525M + \mathcal{N}(0, 5)$$

$$f_3(A, H) := 0.00013857A^3 - 0.0135A^2 + 0.2025A + 0.2025H + \mathcal{N}(17.1714, 0.2A)$$

The state of each diagnose variable $D_i$ is determined by taking the argmax over all three functions.

$$f_{Di}(A, F, H, M) := \begin{cases} \mathit{true} & \text{if argmax}(f_1(A), f_2(M, F), f_3(A, H)) = i \\ \mathit{false} & \text{otherwise} \end{cases}$$

The resulting SCM, as shown in Figure 1, consists of the Causal Health SCM with three additional Diagnose variables. Connections from $A, F, H$ and $M$ to every $D_i$ are introduced, since $f_{Di}(A, F, H, M)$ depends on all four original variables.

All interventions are carried out to be perfectly atomic, while every intervention sets the affected variable to a uniform distribution.

## References

Matej Zečević, Devendra Singh Dhami, Athresh Karanam, Sriraam Natarajan, and Kristian Kersting. Interventional sum-product networks: Causal inference with tractable probabilistic models. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.