Assignment 1: Diabetes Interventions

NWI-IMC012: Bayesian Networks and Causal Inference

Authors:

Daan Kersten
S1053750

Andrew Schroeder
S1111686

Sven Meijboom
S1054374

Radboud University, Nijmegen

November 22, 2024

Contents

1 Introduction

1.1 Problem Domain

Diabetes is one of the most significant chronic health diseases in the United States, with impacts on millions of Americans and representing a massive financial burden on the economy. Diabetes is caused by an inability to regulate sugar levels in the blood. More specifically, when food is consumed, digestion breaks the food down into simpler sugars which are released into the bloodstream. The sugars in the bloodstream then trigger the pancreas to release insulin, which enables cells to use the sugars in the bloodstream for energy. Diabetes disrupts this process by either causing insufficient insulin to be produced by the pancreas (Type 1) or by preventing effective utilization of the insulin (Type 2).

Heart disease, vision loss, kidney disease, and a variety of other adverse health effects are associated with consistently high levels of blood sugar. Concerning Type 2 diabetes, there is no cure though its impacts can be reduced by making a variety of lifestyle changes including losing weight, eating healthy, and staying active. Early diagnosis is an effective tool against Type 2 diabetes as it can lead to lifestyle changes and more effective treatment. As a result both predictive and causal diabetes models are advantageous tools that can be used by public health officials to inform policy and interventions. In this research project, we will attempt to develop such a causal model.

1.2 Research Questions

We will use data related to lifestyle and diabetes in the United States which is collected by the Center for Disease Control (CDC) to answer the following questions:

1. What is the causal relationship between the education level of a person and their chance of developing Type II diabetes?

2. What variables are most significant in determining an individual's chance of developing diabetes?

2 Data Description

To answer the questions, this research makes use of the CDC Diabetes Health Indicators dataset hosted on the UCI Machine Learning Repository. This dataset was created by the Center for Disease Control (CDC) to better understand the relationship between lifestyle and Type II diabetes in the US. Each row of the dataset consists of a single person participating in the study and there are no missing values. The dataset contains 253680 samples each containing values in all 21 variables. These variables are the following:

Variable Name	Type	Levels/Range	Description
Diabetes	Nominal	0 = no, 1 = yes	Person has diabetes or not
HighBP	Nominal	0 = no, 1 = yes	Person has high blood pressure
HighChol	Nominal	0 = no , 1 = yes	Person has high blood cholesterol
CholCheck	Nominal	0 = no, 1 = yes	Description
BMI	Numeric	12.0 - 98.0	Person's Body Mass Index
Smoker	Nominal	0 = no, 1 = yes	Person smoked at least 100 cigarettes in their life
Stroke	Nominal	0 = no, 1 = yes	Person has ever had a stroke
HeartDiseaseorAttack	Nominal	0 = no, 1 = yes	Person has ever had heart disease or a heart attack
PhysActivity	Nominal	0 = no, 1 = yes	Person has had physical activity in the past 30 days
Fruits	Nominal	0 = no, 1 = yes	Person consumes fruit daily
Veggies	Nominal	0 = no, 1 = yes	Person consumes vegetables daily
HvyAlcoholConsump	Nominal	0 = no, 1 = yes	Person is a heavy drinker (14 consumptions for men, 7 for women
AnyHealthcare	Nominal	0 = no, 1 = yes	Person has some kind of health care coverage
NoDocbcCost	Nominal	0 = no, 1 = yes	In last 12 months, person could not visit doctor due to cost
GenHlth	Ordinal	1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor	How good the person's general health is
MentHlth	Numeric	1 - 30 days	For last 30 days, how many days their mental health was not good
PhysHlth	Numeric	1 - 30 days	For last 30 days, how many days their physical health was not good
DiffWalk	Nominal	0 = no, 1 = yes	Person has serious difficulty walking or climbing
Sex	Nominal	0 = female, 1 = male	Person is a man or woman
Age	Ordinal	1 = 18-24, 9 = 60-64, 13 = 80 or older	Person's age
Education	Ordinal	scale 1-6 (1 = Never attended school, 6 = College graduate)	Person's education level
Income	Ordinal	scale 1-8 (1 = less than $10,000, 8 = $75,000 or more)	Person's income

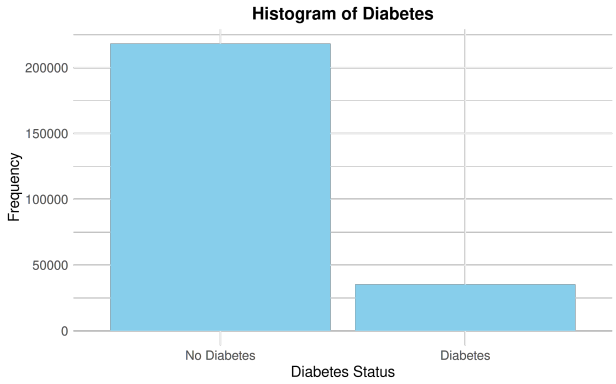
Table 1: Description of dataset variables

2.1 Data inspection

To better understand and manipulate the data, the distribution for some of the more interesting numeric and ordinal data is plotted. This should give some intuition for working with the data as well as provide early warnings if there are unexpected outliers or abnormalities.

2.1.1 Diabetes

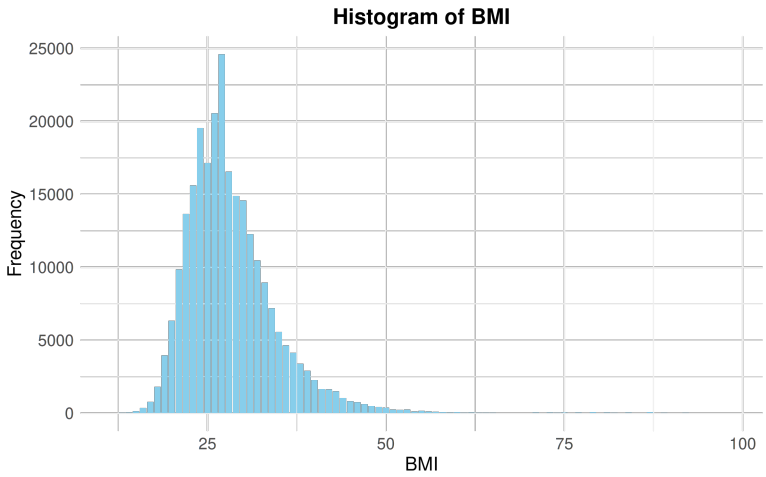
The first variable to be examined is the target variable, `Diabetes`.



As seen in the histogram above, all data points are within the expected binary range. Additionally, in the dataset about 16% of samples have diabetes which is above the US national percentage of 11.6% [CDC(2024)]. However we would expect the percentage to be higher than average due to selection bias in the population, that is, we expect those who already have diabetes to be more aware and willing to participate in a data collection by the CDC.

2.1.2 BMI

The second plot is that of `BMI`, which is an integer numeric value with a range from 12 to 98. In the histogram below it is clearly visible that there are almost no entries above a BMI of 60. At BMI levels of 98 there are even fewer instances. Given the low frequency of these extreme values, they do seem reasonable because such high BMI values are unlikely but not impossible.



When further inspecting these very high BMI values, there are 13737 people in our dataset with a BMI of 40 or higher. This corresponds to 5.42% of the dataset which aligns with research done on obesity in the US [Flegal et al.(2010)].

2.1.3 Age, MentHlth, and PhysHlth

Due to space constraints, in this section we combine the inspection for multiple, less interesting variables, which are still worth mentioning, specifically `Age`, `MenHlth` and `PhysHlth`. Histograms for these variables can be found in appendix section A.4. The Age variable shows a nice distribution with outliers. Each age group represents a 5 year span, starting from 18 and ending at 84. The majority of people in this dataset appear to be around group 8, which is approximately 60 years old. The second and third variables that were inspected included mental health and physical health over a thirty day span. People had to note how many days, out of the last thirty their mental and physical health was not good. It is visible in the plot that for most people this was zero days, or almost zero days. The second largest group was the group who said all of the 30 days were not good. This could be explained by people who are overall in bad state both mentally and physically. These things can often co-occur, this is also why these two distributions are so similar looking. Just like in the other plots, these don't show any abnormalities.

2.2 Data pre-processing

The data was pre-processed in such a way as to preserve both interpretability and compatability with the regression models we are using for our analysis, to the extent possible. As seen in Table 1, the majority of the dataset features are nominal data with two categories. When directly imported for use, R treats the data as numeric, which is appropriate for nominal data with two categories since guassian graphical models are applicable to numeric data and nominal dichotomous/binary data. Given that all the nominal data is treated as numeric, the only other features to consider are the ordinal and numeric variables.

The ordinal variables `GenHlth`, `Education`, and `Income` were encoded as ordered factors, each with it's respective integer levels ([1,5], [1,6], [1,8], respectively). Initially, `Age` was also encoded as an ordered factor, but lavaan provided a warning stating that the `Age` ordered categorical variables had more than 12 levels, which is not recommended as this can cause computational instability. There were three possible solutions to solve this problem:

1. **Keep `Age` as a 13-level ordered factor.** This probably would have been okay as it is only one over the limit.
2. **Combine some of the levels to reduce it below 13.** It was possible to combine the last two age categories, but we likely would have lost a bit of information doing this.
3. **Treat `Age` as a numeric variable.** This would make sense given that the variable is a binned version of the true underlying continuous distribution.

While the analysis probably would have still worked with a 13-level factor, it was decided that option 3 was optimal, given the underlying distribution of Age, thus it was encoded as numeric data.

The final step of preprocessing the data was to normalize our numeric variables to have a mean of zero and standard deviation of one. This was done to standardize the data such that comparisons between path coefficients can be made in a more interpretable manner later on. Thus `Age, BMI, MenthHlth,` and `PhysHlth` where all scaled.

3 Bayesian Network Model

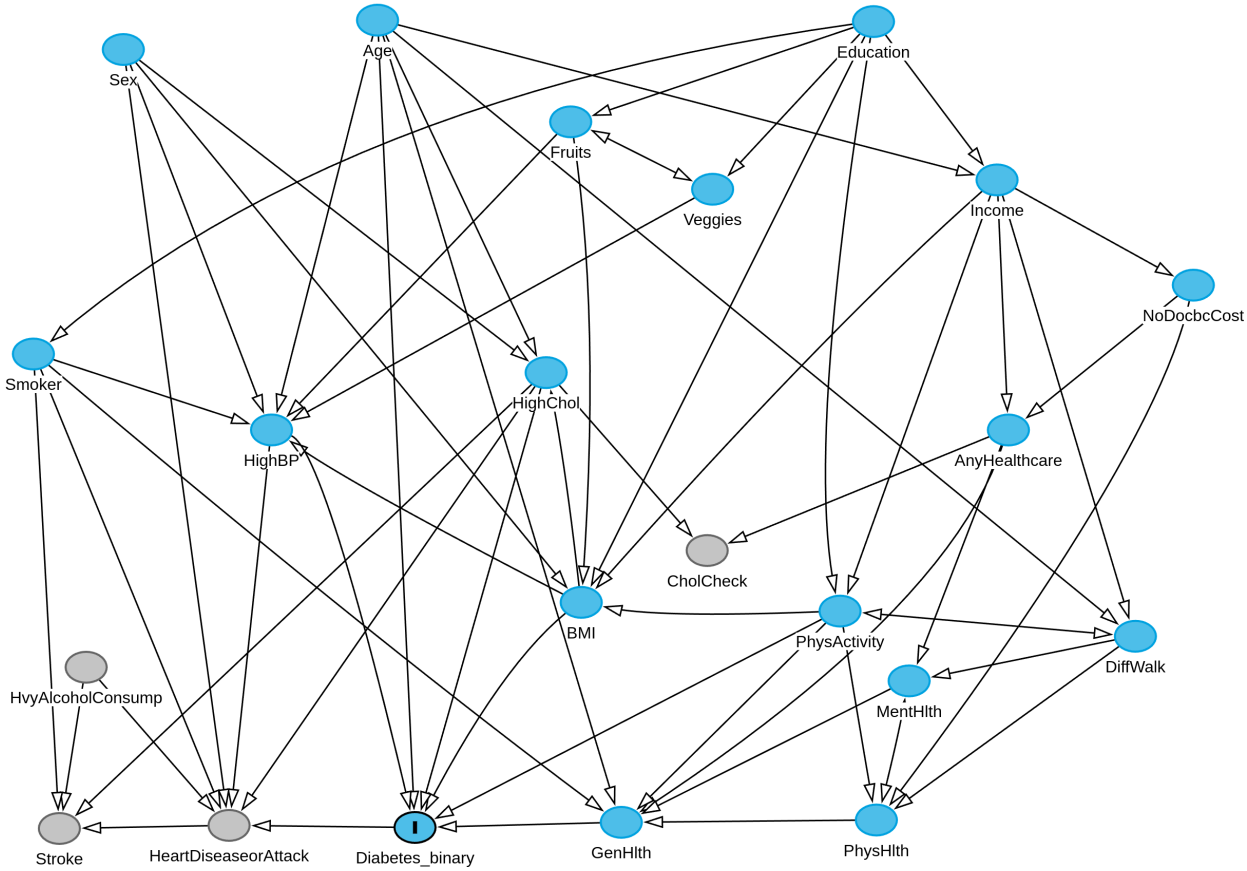


Figure 1: Bayesian Network Model of the Problem Domain

4 Methodology

4.1 Building the initial Bayesian Network

The initial version of the Bayesian network was built based on personal knowledge on diabetes and its health indicators. Additionally, we did internet research when there were doubts about the causes of certain variables. We initially established that Diabetes is directly influenced by a person's age, BMI, general health, whether they have had a heart disease or attack, whether they have high blood pressure and/or cholesterol, whether they partake in physical activities, if they smoke and if they have ever had a stroke before. For this reason, these are all edges going into the Diabetes variable. In a similar manner, the causal relations for all other variables

have been determined, where we discussed and theorized what causal variables could have an influence on the variable of interest.

4.2 Testing the Network Structure

4.2.1 Testing Network Implied Independence

To test the model for independence statements that were implied by the initial DAG but which are very unlikely considering the data, we followed a testing procedure similar to that performed in the R Companion document under the section **Model testing using polychoric correlations**. Specifically, we pre-processed the data as mentioned in section 2.2, such that polychoric correlations were appropriate (instead of the less interpretable canonical correlations). We then used the function `lavCor` to extract the polychoric correlation matrix, and then used the `localTests` function, limiting the maximum number of conditioning variables to zero or one. Given the size of our network, there were hundreds of independence tests when the number of conditioning variables was unrestricted, thus by limiting it to zero or one and extracting the worst 10 offenders for each set of tests, we prioritized the most egregious offenders first.

The results of the first set of local tests are presented in the Figure 2 and 3 below. There are several interesting observations, the first of which is that for the top 10 worst results, we see that the p-value for all of them is zero! If one were only looking at the p-value these results would suggest that none of these independence relations should hold. But the p-value is notoriously problematic, particularly for large datasets where it conflates information about dependence strength and sample size. For a large data set with over 250,000 samples, even when the effect size is small between two variables, the p-value would often still indicate a statistically significant result. Thus we specifically look at the effect size in the `estimate` column of Figure 3, which is also what is plotted.

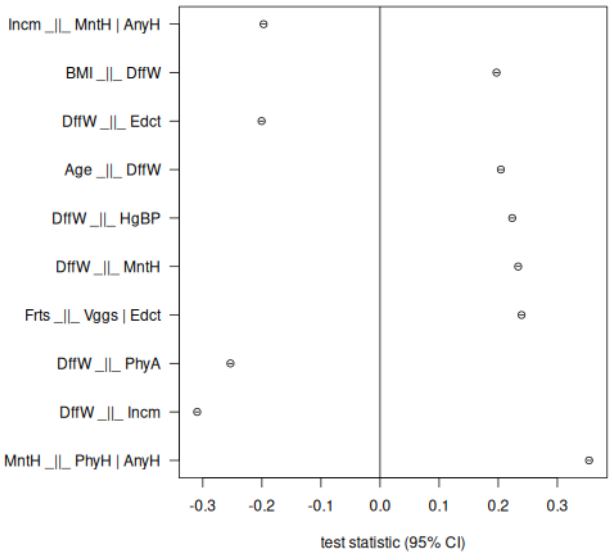


Figure 2: Test 1 Plot

Independence Test	estimate	p.value	2.5%	97.5%
MntH _||_ PhyH | AnyH	0.3537	0	0.3506	0.3575
DffW _||_ Incm	-0.3095	0	-0.3132	-0.3062
DffW _||_ PhyA	-0.2532	0	-0.2569	-0.2496
Frts _||_ Vggs | Edct	0.2394	0	0.2358	0.2432
DffW _||_ MntH	0.2337	0	0.2300	0.2374
DffW _||_ HgBP	0.2236	0	0.2200	0.2274
Age _||_ DffW	0.2045	0	0.2007	0.2082
DffW _||_ Edct	-0.2005	0	-0.2042	-0.1967
BMI _||_ DffW	0.1971	0	0.1934	0.2008
Incm _||_ MntH | AnyH	-0.1971	0	-0.2008	-0.1933

Figure 3: Test 1 Tabular

From the test results the worst offender is that MentHlth is independent of PhysHlth conditioned on AnyHealthcare. Reasoning about this, the first independence does not seem very likely - having health care coverage does not make mental health independent of physical health. In other words, knowing about someones physical health in addition to if they have health coverage is more informative about their mental health then if only their medical coverage is known. Thus it was deemed reasonable that there should be an edge connecting MentHlth and PhysHlth. But in which direction? It seems reasonable to suggest that better mental health would lead to better phsyical health, while better phsyical health could have a positive effect on mental health. Thus it was decided to include a bi-directional arrow between these two variables. Implicitly this bi-directed arrow represents some unobserved latent variable that is a common cause of both better physical health and mental health. One could argue that GenHlth is this common cause and indeed that may be the case, in which case it would be the parent of PhysHlth and MentHlth. But it seemed appropriate to treat GenHlth as a aggregate of other factors influencing health, such as smoking, age, physical health, and mental health. Thus GenHlth was not used as the common cause, and the latent variable was kept in the form of the bi-directed edge.

The second worst result is that DiffWalk is independent of Income. Again, it seems that this would not be true. It was reasoned that if someone had lower income they would have less access to medical coverage for possible surgery or health treatments, in addition to possibly leading a less active lifestyle with lower quality food, leading to long term health complications. As such an arrow from income to difficulty walking was added. The authors realized after writing the report and analyzing the results that actually the arrow probably should have been directed the other way, from difficulty walking to income, as that effect intuitively appears very strong. If someone has difficulty walking this would severely limit their income potential.

After making these adjustments, the same procedure was repeated four additional times, and a total of 10 arrows were added. For the sake of brevity not all the additional changes are discussed in detail, but each additional test identified some clear outliers, and when considering them it seemed obvious that an additional arrow was warranted. The authors realize this is more than the suggested 3 additional arrows, but felt the additional changes were justified, particularly considering that the initial model was fairly sparse. The test results after the 5th test are presented in Figure 4. From the first test to the last test the range on the test results changed from approximately [-0.3, 0.35] to [-0.15, 0.20].

4.2.2 Superfluous edges

After adding the 10 most significant missing edges, the network was then examined for superfluous edges. This was done by taking the DAG after the initial battery of tests, and fitting a structural equation model using the data from our dataset and the sem function provided by lavaan. Fitting the model provided the path coefficients, which give some indication of the direct effect size of one variable on another. All edges with path coefficients below the threshold of 0.01 were discarded. Specifically, these edges were the following:

1. AnyHealthcare → PhysActivity
2. AnyHealthcare → PhysHlth
3. Education → CholCheck

4. `Smoker` → `Diabetes_binary`

5. `Veggies` → `BMI`

The first three edges removed seem reasonable, however one might expect that being a smoker would significantly increase the probability of diabetes. Similarly it might be expected that eating more vegetables reduces the chance of developing diabetes. In these two cases the authors chose to follow the data rather than intuition, though it would make sense to have a medical expert check these edges. The final structural equation model after all network testing is presented in the appendix in figure 5.

4.3 Answering the Research Questions

Once the network structure testing was complete, the finalized DAG is used to answer the two research questions listed in the introduction: first, what is the causal effect of education level on diabetes and second and more broadly, which variables in the network have the largest causal impact on an individual's chance of developing diabetes?

To answer these questions, a structural equation model was once again fitted using the DAG, data, and the polychoric correlation matrix. The SEM model was in a lavaan data structure, thus it was cast to a dagitty structure with the path coefficients. This dagitty model was then passed through a function that iterates over each variable in the network and cuts all the incoming edges to that variable, saves the graph, and then proceeds to process the next variable. In this way, a new *interventional* dag is generated for each variable of the network. This function essentially applies the do-operator to each variable (except diabetes which is the target) in the network. By acquiring these new dags, the analysis moves from the observational regime to the interventional regime.

Once the interventional dags are acquired, another function is called to process each of these dags. For each interventional dag the implied covariance matrix is computed using `dagitty` and this matrix is used to extract out the causal effect of the intervention variable on diabetes (under the hood the trek rule would likely be applied). Note that when generating the implied covariance matrices, `standardized = True` was set, meaning each variable has variance of 1, which also means that the covariances are equivalent to correlations. The final result is a list that provides the implied covariance (causal effect) of each variable in the network on diabetes. This final list contains the answers to our causal inference questions. While the network is used for causal inference, it is also possible to utilize it for prediction, for example, if we have a set of observations for an individual the network could be used to predict the chances of developing diabetes.

5 Results and Discussion

The full results of the analysis from the previous section is presented in table 3 of the appendix. However, many of the implied covariances between the intervention variable and diabetes are vanishingly small and it is not useful to interpret such small effects. Thus only a subset of the more significant causal effects are presented and discussed in this section and can be found in table 2.

Variable	Cov(Variable, Diabetes)
Age	0.1864
GenHlth	0.1800
BMI	0.1667
HighBP	0.1200
HighChol	0.0860
PhysHlth	0.0756
DiffWalk	0.0524
MentHlth	0.0426
Sex	0.0168
Smoker	0.0159
Fruits	-0.0130
Income	-0.0346
Education	-0.0433
PhysActivity	-0.0701

Table 2: Implied covariance (causal effect) of network variables on diabetes.

Starting with the first research question: *What is the causal relationship between the education level of a person and their chance of developing Type II diabetes?* Examining the results, we can see that the implied covariance between education and diabetes is -0.0433. The negative sign of the effect is what is expected, since higher education is expected to lower the risk of developing diabetes as it may lead to healthier life choices in general. It is somewhat difficult to interpret what exactly this means in quantitative terms since education is an ordinal variable with 6 levels ranging from never attending school to college graduate. Additionally diabetes is binary yes/no variable. It is tempting to say that for each unit increase in education the probability of getting diabetes is 4% lower but this would be incorrect. All that can be said for certain is that education negatively correlates with diabetes, but this effect is actually quite weak considering that a Pearson's correlation of 0.1 to 0.3 is considered small in the social sciences [Brydges(2019)].

The second research questions asks: *What variables are most significant in determining an individual's chance of developing diabetes?*. Examining the results, again we can see that Age has the biggest effect on developing diabetes with a positive (thought still weak) effect of 0.1864. Again the sign of the effect is positive which is expected - as an individual ages, the probability of them developing type II diabetes increases, particularly after age 35. Looking at the results more broadly all causal effects seem to be fairly small, with none above 0.2. Not surprisingly, body mass index (BMI), high blood pressure, and high cholesteral all have positive correlations with diabetes. Those with negative correlations are also not surprising - it is intuitive that eating fruits, higher income, and physical activity would all decrease the probability of developing diabetes. Phyiscal health and mental health may appear counter-intuitive at first - there seems to be a positive correlation between good health and diabetes? Looking at table 1, the results make sense - physical health is measured by how many days in the past thirty days an individual's health was *not good*, meaning that a higher value in this variable is actually worse health and worse health positively correlates with diabetes. The same is true for mental health.

To conclude the discussion, it is interesting to note there is one potentially significant variable that is not included in the dataset - genetics. It is possible that genetics is the largest

variable impacting diabetes development. The reader may recall in the testing section of the report that an implicit latent variable was created with a bi-directed arrow between physical health and mental health - perhaps genetics is that latent variable?

6 Conclusion

This project highlighted both the potential and the challenges of using Bayesian networks to model diabetes risk factors. The large number of variables in our dataset made it difficult to distinguish necessary edges from superfluous ones. Additionally, independence tests were limited to conditioning on a single variable at a time to avoid an unmanageable number of tests, which may have restricted the depth of our analysis. Some other limitations of our analysis that could cast some uncertainty on the results are the following:

1. After checking for superfluous edges, we removed some edges due to a very low path coefficient, including the edge between `Smoker` and `Diabetes_binary`. The low path coefficient between these variables runs counter to intuition. It is situations like these where a domain expert would prove beneficial in the analysis.

2. Some mistakes were discovered after the analysis was complete. One such mistake was mentioned in the testing section of the report, where an edge from income to difficulty walking was added. In retrospect this should likely have been drawn from difficulty walking to income.

3. Given that the authors were not domain experts in diabetes, the original network was quite sparse. During testing 10 additional edges were added based on the independence tests. Despite exceeding the recommended three changes, the authors suspect that even more edges may be justified. Specifically more edges originating from `HvyAlcoholConsump`.

4. The analysis could have been improved by introducing the latent variable `Genetics` into the network, as this variable may have a large effect on diabetes.

5. Variables like Age, Education, Income, and GenHlth are binned and could have reduced the model's accuracy.

Given these limitations, the authors urge caution when interpreting and drawing conclusions from the results. Despite the potential issues, it was demonstrated in the results section of the report that the majority of variables had effects with the correct sign. Though the effect sizes were all fairly weak, the results were enough to answer the two research questions.

7 References

References

[Brydges(2019)] Christopher R Brydges. 2019. Effect Size Guidelines, Sample Size Calculations, and Statistical Power in Gerontology. *Innovation in Aging* 3, 4 (Aug. 2019), igz036. https://doi.org/10.1093/geroni/igz036

[CDC(2024)] CDC. 2024. National Diabetes Statistics Report. https://www.cdc.gov/diabetes/php/data-research/index.html

[Flegal et al.(2010)] Katherine M. Flegal, Margaret D. Carroll, and Cynthia L. Ogden. 2010. Trends in Obesity and Extreme Obesity Among US Adults—Reply. *JAMA* 303, 17 (05 2010), 1695–1696. https://doi.org/10.1001/jama.2010.518 arXiv:https://jamanetwork.com/journals/jama/articlepdf/185788/jlt0505_1695a_1696.pdf

A Appendix: Implementation Details

A.1 Programming Language and Packages Used

The analysis was conducted using the R programming language. The following packages were utilized to support model building, testing, and data analysis:

- dagitty: for causal graph specification and visualization
- lavaan: for structural equation modeling and polychoric correlation computation
- bayesianNetworks: to assist in network modeling and Bayesian inference
- pastecs: for descriptive statistics and data summaries

A.2 Source Code Description

Within the project Github repository the majority of the soruce code is within the directory `src`. The main inference and testing tasks were performed using a series of functions defined in the `utilities.R` file, which was imported for use elsewhere. The utilities file provides these functionalities:

1. Installs all necessary packages such as `ggplot, dagitty, lavaan`, etc
2. Function `get_interventional_dags`: this function receives the base network and generates all interventional dags for finding the causal effect of all variables on the target variable.
3. Function `get_causal_effects`: using the generated interventional dags, the implied covariance between each of the interventions and the target variables can be calculated.
4. Function `load_data`: loads the raw data from the repository and encodes the variables appropriately while also scaling the numeric data.

5. Function `run_independence_tests`: using the loaded data and a provided dag, this function generates the polychoric correlation matrix and then runs `localTests` from the Daggity package to test the implied independence relationships. Used to test for missing edges in a DAG.

6. Function `get_superfluous_edges`: accepts a fitted dag and extracts all those path coefficients smaller than a provided threshold value. Used for identifying and removing superfluous edges in a DAG.

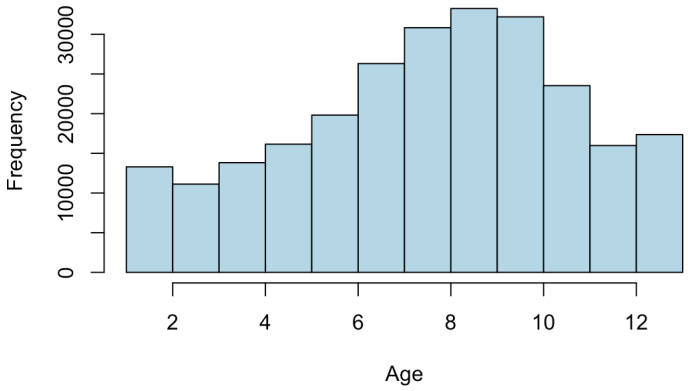
The testing of the DAG structure was done in `testing_causal_diagrams.R` which utilized several of the functions in the `utilities.R` file. After testing, the main analysis where the DAG was fitted and causal effects generated was done in `inference.R`, also utilizing the utility functions.

A.3 Link to Source Code

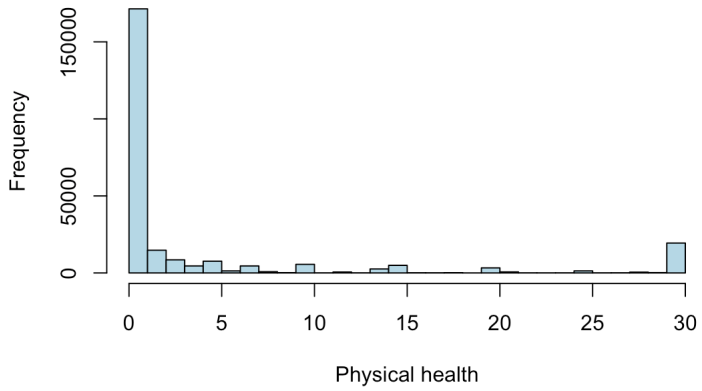
GitHub repository: `https://github.com/aschroede/BNCI`

A.4 Plots for data inspection

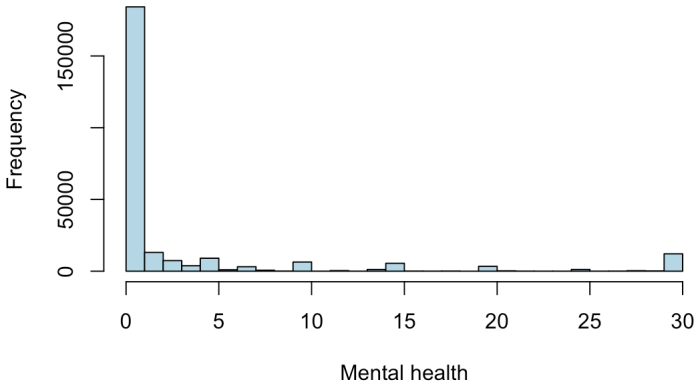
A.4.1 Age



A.4.2 Physical health



A.4.3 Mental heatlh



A.5 Path coefficients of all variables

Variable	Cov(Variable, Diabetes)
Age	0.1864
AnyHealthcare	-0.0083
BMI	0.1667
CholCheck	-0.0000
DiffWalk	0.0524
Education	-0.0433
Fruits	-0.0130
GenHlth	0.1800
HeartDiseaseorAttack	-0.0000
HighBP	0.1200
HighChol	0.0860
HvyAlcoholConsump	0.0001
Income	-0.0346
MentHlth	0.0426
NoDocbcCost	0.0059
PhysActivity	-0.0701
PhysHlth	0.0756
Sex	0.0168
Smoker	0.0159
Stroke	-0.0000
Veggies	-0.0051

Table 3: Implied covariance (causal effect) of network variables on diabetes.

A.6 Plotted results of model tests

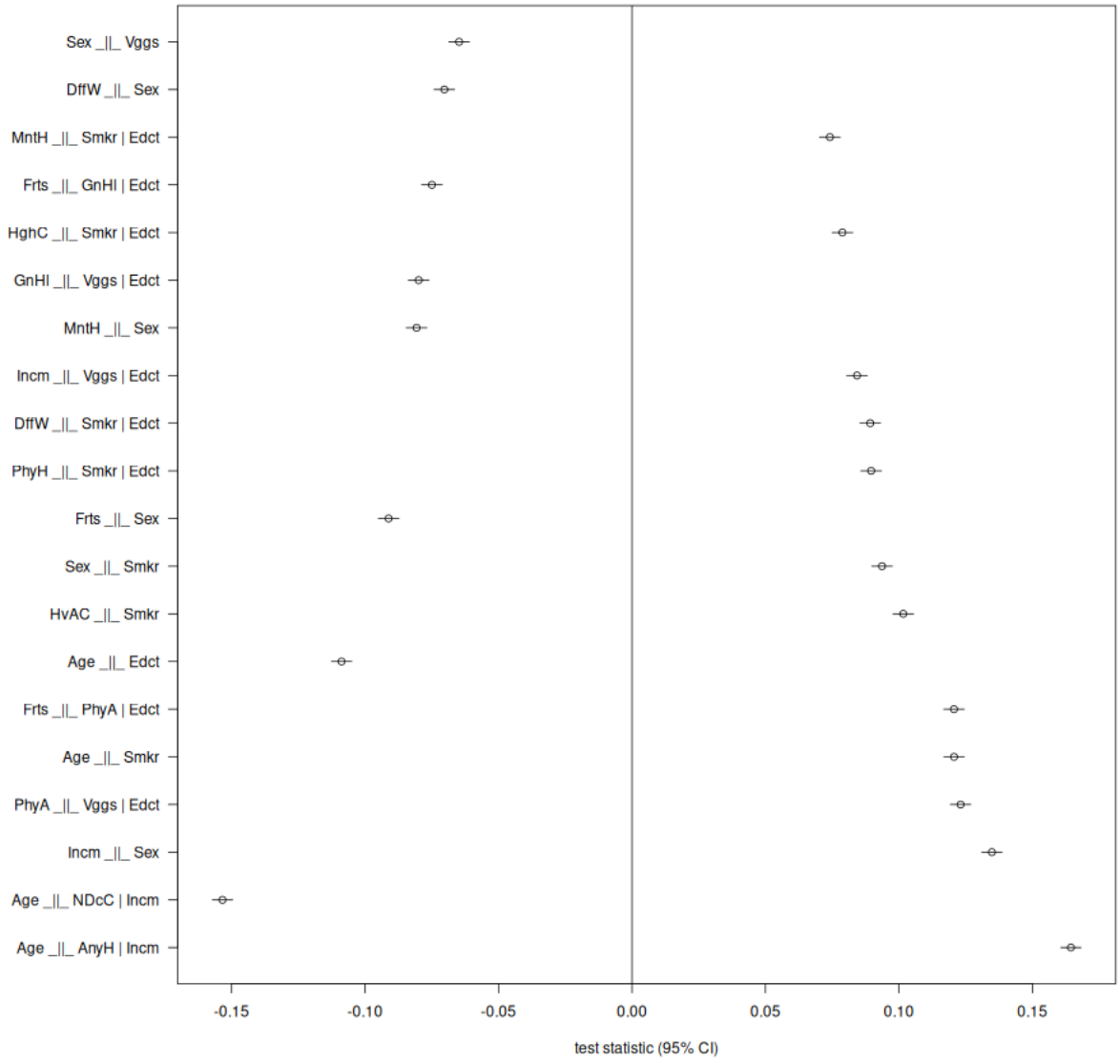


Figure 4: Results of the fifth and last test using polychoric correlations with maximum of one conditioning variable

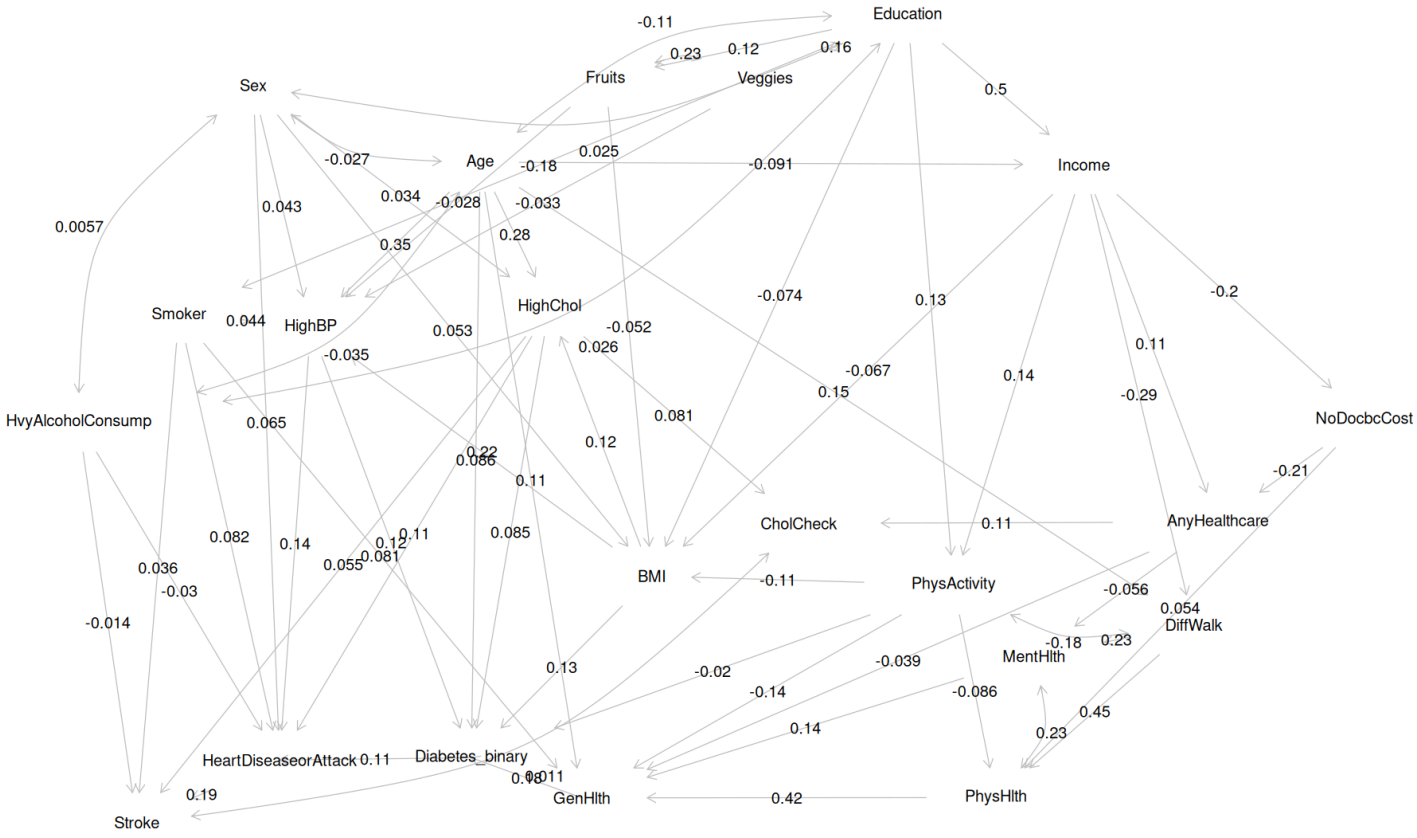


Figure 5: Finalised SEM after testing with path coefficients where edges with path coefficients less than 0.01 were removed.