The School of Mathematics

THE UNIVERSITY *of* EDINBURGH

Identifying Risk Factors for Dementia (Cognitive Impairment) using a Bayesian Network Approach

by

Jonathan Hoover

Dissertation Presented for the Degree of
MSc in Statistics with Data Science

July 2021

Supervised by
Dr. Sara Wade and Dr. Cecilia Balocchi

Executive Summary

1. Introduction

Dementia cases increase every year due to aging populations [1]. There is a critical need to prepare health systems for this increased burden. A potential solution is to identify preventable risk factors to target disease prior to onset. We use Bayesian networks, a class of graphical models that express dependence relationships between variables, to identify risk factors for cognitive impairment (our proxy for dementia) and suggest potential avenues for future study.

2. Data Source

The data in this analysis comes from the easySHARE release 8.0.0 [2]. The data contains information on age, sex, education, smoking, drinking, depression, social contact, BMI, and physical activity, which are known risk factors for dementia [3] [4]. The data does not include dementia status but includes recall and orientation measurements which serve as proxies for cognition.

3. Methods

We perform a cross-sectional study on baseline interviews from the easySHARE data. Following variable selection and feature engineering we produce a dataset with 14 risk factors. We then define a proxy for dementia from recall and orientation, called cognitive impairment, that splits cognition into severe, mild, and unimpaired groups.

To learn the structure of our network, we define a unique hybrid structure-learning approach that emphasizes learning connections between risk factors and cognitive impairment. We then design a two-phase, prediction-based validation strategy to validate both the learning phase and the generalizability of our results. Additionally, we create a classification score for validation that emphasizes learning relationships between risk factors and impairment. This score was used to select the final networks.

4. Results and Conclusions

We found two nearly equivalent Bayesian networks that identified age, sex, education, depression, and quality of life (or drinking in one of the networks) as direct risk factors (not necessarily causal) for cognitive impairment. Additionally, we identified interaction pathways between each risk factor in our data (except BMI) that led to cognitive impairment. The data emphasized the utility of Bayesian networks for identifying potential intervention points for cognitive impairment, but future studies are required for causality.

Lastly, external validation suggested that we cannot generalize to countries outside of those included in the training data. Future studies should focus on individual countries or regional analyses to identify risk factors specific to each region.

Acknowledgments

I am grateful to the supervisors of the project, Dr. Sara Wade and Dr. Cecilia Balocchi, for their guidance. I'd also like to thank the project experts, Dr. Graciela Muniz-Terrara and Dr. Anja Leist for their expertise and advice.

This paper uses data from SHARE Waves 1, 2, 3, 4, 5, 6, 7, 8 and 9 (DOIs: 10.6103/SHARE.w1.800, 10.6103/SHARE.w2.800, 10.6103/SHARE.w3.800, 10.6103/SHARE.w4.800, 10.6103/SHARE.w5.800, 10.6103/SHARE.w6.800, 10.6103/SHARE.w7.800, 10.6103/SHARE.w8.800, 10.6103/SHARE.w8ca.800, 10.6103/SHARE.w9ca800), see Börsch-Supan et al. (2013) for methodological details [5].

The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

This paper uses data from the generated easySHARE data set (DOI: 10.6103/SHARE.easy.800), see Gruber et al. [2] for methodological details. The easySHARE release 8.0.0 is based on SHARE Waves 1, 2, 3, 4, 5, 6, 7 and 8 (DOIs: 10.6103/SHARE.w1.800, 10.6103/SHARE.w2.800, 10.6103/SHARE.w3.800, 10.6103/SHARE.w4.800, 10.6103/SHARE.w5.800, 10.6103/SHARE.w6.800, 10.6103/SHARE.w7.800, 10.6103/SHARE.w8.800).

University of Edinburgh – Own Work Declaration

Name: Jonathan Hoover

Matriculation Number: s2113204

Title of work: Identifying Risk Factors for Dementia (Cognitive Impairment) using a Bayesian Network Approach

I confirm that all this work is my own except where indicated, and that I have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in quotes for all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

I understand that any false claim for this work will be penalised in accordance with the University regulations (https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct).

Signature:
Jonathan Hoover

Date: July 1, 2022

Contents

List of Tables

List of Figures

1 Introduction

Dementia is a condition defined by impaired memory, cognition, behavior, and emotional regulation that leads to a loss of autonomy [1]. As the world's population ages, the population living with dementia also rises; as of 2021, over 55 million people were living with dementia and by 2030 the estimated number is over 78 million [1]. With an aging population, there is a critical need to prepare health systems for the increased economic and health burden that comes with more dementia cases. As there are few treatments for dementia, one way to ease the burden is to identify preventable risk factors that can be targeted prior to disease onset.

Recent studies have identified 12 potentially modifiable risk factors: Education, hearing loss, traumatic brain injury, hypertension, alcohol consumption, obesity, smoking, depression, social isolation, physical inactivity, air pollution, and diabetes [4]. Each factor may have a direct association with dementia, but evidence suggests there are likely complex interactions at play between the variables [6]. Traditional analyses, like regression, can study interaction through interaction terms. However, these methods fail to elicit the probability dependence structures (or pathways) needed to understand the complex relationships between risk factors and dementia. These structures are critical for identifying potential intervention points. Fortunately, Bayesian networks, a graphical class of models with nodes and edges expressing direct dependence relationships, are ideally suited to identifying interactions in complex systems like dementia [7].

In this report, we use Bayesian networks and the open-source clinical dataset easySHARE [2] to learn the probability relationships between cognitive impairment (our proxy metric for dementia), unmodifiable risk factors, and modifiable risk factors. The final list of factors is age, sex, education, physical health, physical activity, current smoking status, daily smoking status, BMI, recent drinking history, quality of life, depression, household size, marital status, and income. These were chosen to cover 7 of the 12 modifiable risk factors presented in [4], along with known demographic risk factors (age and education) [3] and a potential confounder (income [1]).

The structure of the report is as follows. First, we define our data and the data manipulation methods used to create our final dataset. Next, we explain Bayesian networks and the methods used to identify the structure of our network. Then we explain our model validation strategy, including both an internal strategy to validate the learning phase and an external strategy to assess generalizability. Lastly, we present our final network, analyze the relationship between the risk factors and cognitive impairment, and suggest potential avenues for future study and intervention.

2 Data Source

The data in this analysis comes from the easySHARE release 8.0.0 [2], which is a simplified dataset of the more complex SHARE 8.0.0 [8] [13][17][14][18][15][5][16][12]. release. easySHARE includes the same number of observations as the SHARE dataset, but with a restricted set of variables.

SHARE is a cross-national, interdisciplinary database containing information from interviews in 29 countries in survey waves spanning from 2004 to 2020. The study population is non-incarcerated people of relatively good health (not hospitalized) who live in (and speak the language of) one of the 29 countries included.

easySHARE itself is intended primarily for training and teaching purposes. While this limitation must be noted, the data contains variables with information spanning demography, household composition, social networks and support, childhood conditions, health and health behavior, functional limitation indices, and work and money. These variables cover at least 7 of the 12 modifiable risk factors for dementia presented in [4], so the data is sufficient for risk factor analysis.

3 Methods

3.1 Cross-Sectional Study Design

We performed a cross-sectional analysis to identify dementia risk factors for individuals in the study population. We subset only baseline interviews to prevent any bias resulting from repeat interviews. Baseline interviews were found by selecting the first wave available for any given participant. There are 140,125 baseline interviews. All further analysis was done only on baseline interviews.

3.2 Variable Selection, Missing Value Processing, and Feature Engineering

3.2.1 Response Variable Creation

Neither SHARE nor easySHARE contain a dementia diagnosis. Fortunately, the dataset contains recall, orientation, and numeracy measurements, which capture memory and cognition. We use this data to create a cognitive score as a proxy for the cognitive component of dementia based on the work in [20].

The score takes the sum of immediate recall (0-10), delayed recall (0-10), and orientation (0-4), so the cognitive score ranges from 0-24. Higher scores signify better cognition. We did not include numeracy because the questions were deemed too difficult for assessing impairment [17]. Please note the code to generate cognitive score was provided by Sara Wade. We adapted it to remove numeracy.

We also note that this cognitive score purely measures recall and orientation, which is *not* a robust measure of dementia. Proper clinical diagnosis is the gold standard, but we proceed with this definition given it is the only metric available.

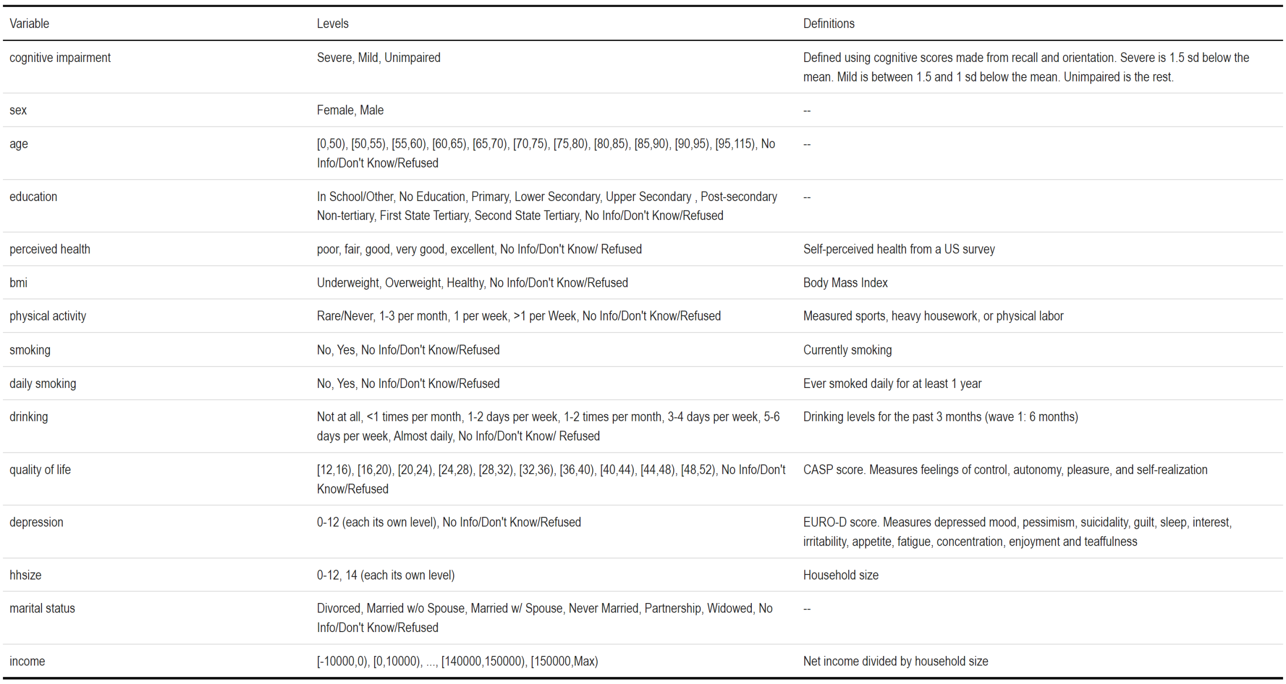
3.2.2 Risk Factor Selection

Not all variables in easySHARE are relevant dementia risk factors. We identified 21 potential variables covering known modifiable and unmodifiable risk factors. These contain information on education, smoking, drinking, depression, social contact, BMI, and physical activity, which were presented in [4].

We also selected age and sex, which are known unmodifiable risk factors [3], along with income, which is a potential confounder [1]. Lastly, we created a normalized per person income variable by dividing household income by household size. This left us with 22 potential risk factors.

Of these 22 variables, we identified 8 to drop. Education (numerical), doctor's visits, hospital stays, number of diseases, and net household income were dropped because they shared information with other variables in the data (categorical education, self-perceived health, and normalized per person income). Number of siblings, number of children, and partner in household were dropped because they only had a moderate relationship with cognitive score.

The final list of variables and their levels is shown in Table 1 (They have been discretized according to 3.2.3)



Variable	Levels	Definitions
cognitive impairment	Severe, Mild, Unimpaired	Defined using cognitive scores made from recall and orientation. Severe is 1.5 sd below the mean. Mild is between 1.5 and 1 sd below the mean. Unimpaired is the rest.
sex	Female, Male	--
age	[0,50), [50,55), [55,60), [60,65), [65,70), [70,75), [75,80), [80,85), [85,90), [90,95), [95,115), No Info/Don't Know/Refused	--
education	In School/Other, No Education, Primary, Lower Secondary, Upper Secondary , Post-secondary Non-tertiary, First State Tertiary, Second State Tertiary, No Info/Don't Know/Refused	--
perceived health	poor, fair, good, very good, excellent, No Info/Don't Know/ Refused	Self-perceived health from a US survey
bmi	Underweight, Overweight, Healthy, No Info/Don't Know/Refused	Body Mass Index
physical activity	Rare/Never, 1-3 per month, 1 per week, >1 per Week, No Info/Don't Know/Refused	Measured sports, heavy housework, or physical labor
smoking	No, Yes, No Info/Don't Know/Refused	Currently smoking
daily smoking	No, Yes, No Info/Don't Know/Refused	Ever smoked daily for at least 1 year
drinking	Not at all, <1 times per month, 1-2 days per week, 1-2 times per month, 3-4 days per week, 5-6 days per week, Almost daily, No Info/Don't Know/ Refused	Drinking levels for the past 3 months (wave 1: 6 months)
quality of life	[12,16), [16,20), [20,24), [24,28), [28,32), [32,36), [36,40), [40,44), [44,48), [48,52), No Info/Don't Know/Refused	CASP score. Measures feelings of control, autonomy, pleasure, and self-realization
depression	0-12 (each its own level), No Info/Don't Know/Refused	EURO-D score. Measures depressed mood, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment and teaffulness
hhsize	0-12, 14 (each its own level)	Household size
marital status	Divorced, Married w/o Spouse, Married w/ Spouse, Never Married, Partnership, Widowed, No Info/Don't Know/Refused	--
income	[-10000,0), [0,10000), ..., [140000,150000), [150000,Max)	Net income divided by household size

Table 1: Table listing final variables, levels, and definitions for those that are unclear.

3.2.3 Discretizing Variables

The software used to learn the Bayesian networks (bnlearn [23]) requires each variable to be either continuous or discrete for all associations to be considered. Since many of the risk factors are inherently discrete, we chose to discretize all data.

Discretization is not trivial: the grouping strategy can inform the relationships found between variables. Further analysis of binning is required for more robust results, but this is not considered here. The chosen bins are described below.

3.2.3.1 Discretizing the Outcome Variable

Cognitive score was split into three impairment categories: Severe, mild, and unimpaired. Severe was defined as those with cognitive scores 1.5 standard deviations below the mean. Mild was defined as those between 1.5 and 1 standard deviations below the mean. Unimpaired was anyone else. As such, severe was anything below 6.77, mild was anything between 6.77 and 8.72, and unimpaired was anything greater than or equal to 8.72.

The severe cutoff is reasonable given it requires cognitive score to be *far* below the mean; however, the mild cutoff is somewhat arbitrary, which emphasizes the need for further binning analysis.

3.2.3.2 Discretizing Risk Factors: Continuous Variables

Most age observations were between 50 and 95; we grouped these values into 5-year bins while values below 50 and above 95 were put into their own respective bins. Quality of life (casp) scores range from 12-48, so we split into bins of 4 to reduce categories. BMI was split into well-defined categories for underweight [0, 18.5), healthy [18.5, 25), overweight [25, 30), and obese [30, inf) [21].

Most normalized household income observations were between -10,000 and 150,000 (net incomes can be negative), so we split these observations into bins of size 10,000. Those above 150,000 were binned together due to low observations. Household size and depression (eurod) only have small ranges, so we categorized each value as its own level.

3.2.3.3 Discretizing Variables: Categorical Variables

For education, we grouped In School/Other together due to low observation counts; all other categories were kept in their original levels. The remaining categorical variables were kept in their original levels.

3.2.4 Missing Values

For the outcome variable (cognitive score), there can be no missingness, so we simply dropped all observations with missing cognitive scores. For the remaining variables, easySHARE codes missingness according to type. Most types are non-informative and can be dropped, but there are three potentially informative types that should not be dropped: 1) Not Applicable Filtered, 2) Don't Know/Refused 3) and No Info.

We identified 46,913 non-informative missing cases to drop. This left 93,212 observations. We then checked for informative missingness and found only "No Info" and "Don't Know/Refused." The proportion of the latter was low in the data, so we combined with "No Info" to create a single category.

3.2.5 Relationships between Risk Factors and Cognitive Impairment

To confirm the selected variables have a relationship with cognitive impairment, we analyzed the conditional probability of the risk factors given impairment status.

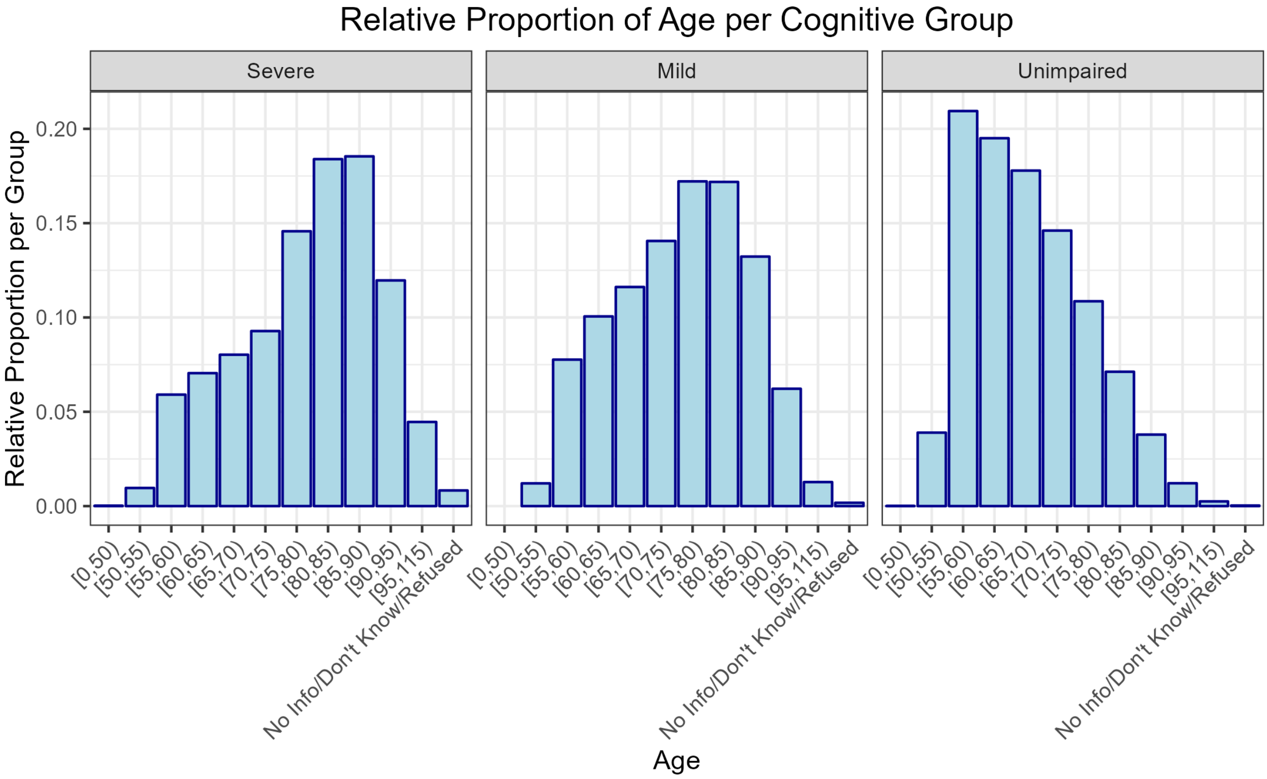


Figure 1: Probability of age given a specific impairment status. The proportions in each impairment status will sum to one.

Figure 1 clearly shows the distribution of age depends on impairment status: If a person is severely impaired, they are more likely to be older than if they were unimpaired. This suggests a strong relationship between age and cognitive impairment, which is expected from the literature [3]. The missingness variable does not appear to have much impact here.

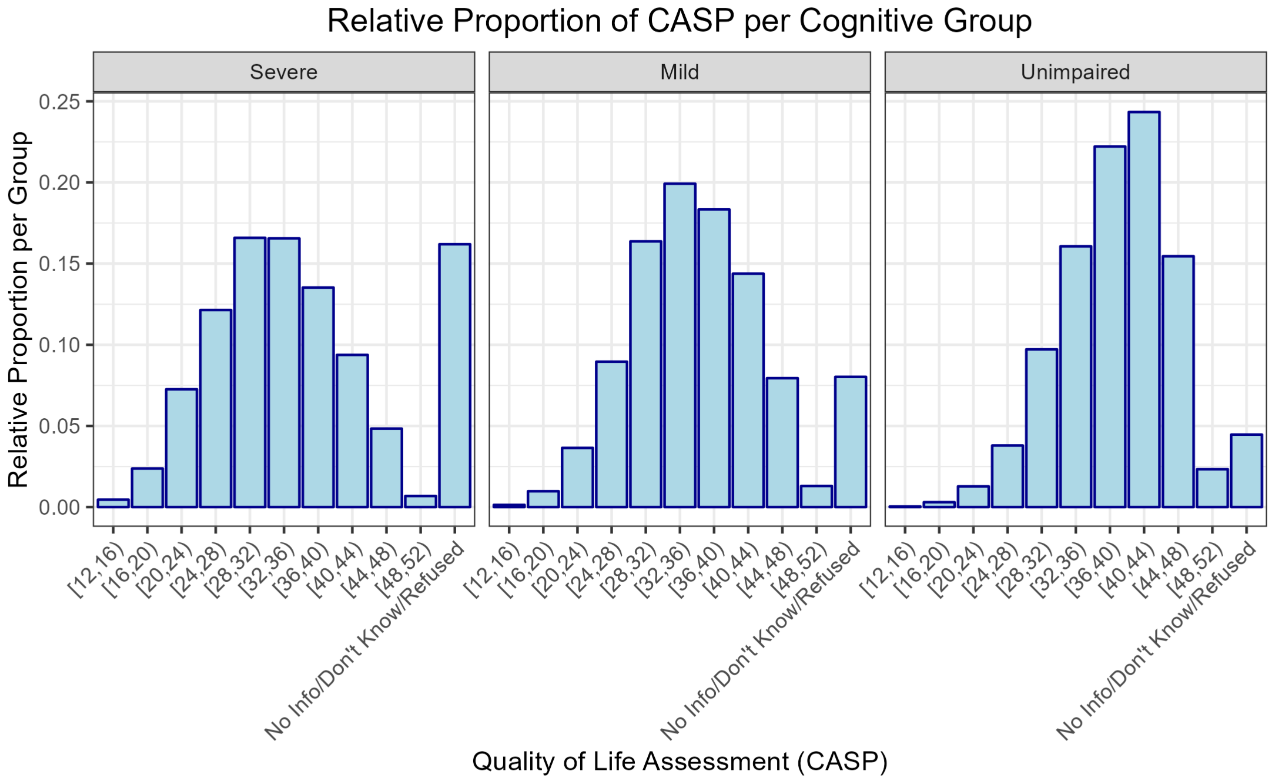


Figure 2: Probability of quality of life (CASP) given specific impairment status. The proportions in each impairment status will sum to one.

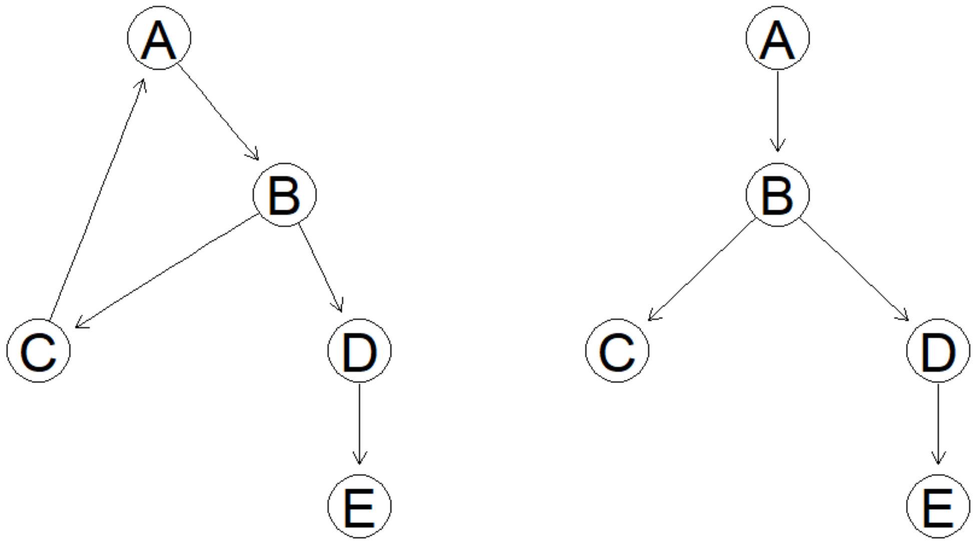
Figure 2 shows that the distribution of quality of life scores depends on impairment status. If a person is severely impaired, they are more likely to have low quality of life scores (particularly 12-24) than people in other impairment groups. Interestingly, the missingness variable appears to be indicative here. This information should improve modeling, which validates our choice to include missingness as a variable in the data.

Similar exploratory analysis was performed for all variables, and most appear to be relevant. However, we cannot tell if the relationships are direct or due to confounding. Bayesian networks will allow us to understand the direction and degree of these relationship. We do not include the other data for brevity.

3.3 Models

3.3.1 Bayesian Networks

Bayesian networks are graphical models that characterize the joint distribution of a set of random variables through a directed acyclic graph (DAG) [7]. Figure 3 shows a cyclic vs an acyclic graph (both directed). These DAGs visualize the underlying conditional independence structure in the data through nodes and arcs [7]. For instance, in Figure 3B, nodes A, B, C, D, and E are random variables and the arcs connecting them define the dependence relationships.



(A) **(B)**

Figure 3: Example of a directed cyclic graph (A) and a directed acyclic graph (B).

The structure of these DAGs induces a factorization [7] of the global joint distribution into local distributions such that:

$$P(\mathbf{X}|\mathcal{G},\Theta) = \prod_{i=1}^{N} P(\mathbf{X}_i|\Pi_{\mathbf{X}_i},\Theta_{\mathbf{X}_i}) \tag{3.1}$$

Where $\mathcal{G}$ is the graph, $\mathbf{X}$ is the set of random variables (with parameters Θ), and $\Pi_{\mathbf{X}}$ are the parents of each node in $\mathbf{X}$. As such, the global joint distribution of $\mathbf{X}$ decomposes into a local distribution for each $\mathbf{X}_i$ (with parameters $\Theta_{\mathbf{X}_i}$) conditional on its parents, $\Pi_{\mathbf{X}_i}$ [7]. This factorization implies that each random variable, conditioned on its parents, is independent of its non-direct descendants [22]. In other words, graphical separation implies independence. In our data, the nodes include the risk factors and outcome variable. For more details on Bayesian networks please see [7] and [22].

Note that our data prevents causal inference, particularly because it may not meet assumptions of acyclicity (physical health may affect depression which may affect physical health, etc.). As such, we are limited to inference on direct and indirect dependence.

3.3.2 Structure Learning

The first step to learning a Bayesian network is learning it's graphical structure. Many algorithms are available for structure learning. There are constraint-based algorithms that focus on conditional independence tests, score-based algorithms that explore relationships by minimizing a score, and many others.

Score-based algorithms are faster than constraint-based algorithms [19], but they find the optimal DAG by considering the relationship between *all* the variables. While important, we primarily care about relationships leading to cognitive impairment. As such, we designed a hybrid structure-learning approach. First, we use a score-based method (hill climb with Bayesian Information Criteria) to learn a base-DAG structure. Then we learn a set of potentially improved DAGs using conditional independence tests between each variable in the network and cognitive impairment. The conditional independence tests should identify relationships with impairment that might not be detected in score-based methods alone.

Structure learning algorithms are stochastic, so they are subject to random error. To reduce the chance of identifying a non-representative base-DAG, we bootstrapped the learning process 400 times and selected arcs present over 85% of the time. The direction of the arc was chosen as the direction present in over 50% of the bootstraps. Note that we employed a blacklist to prevent output arcs from cognitive score and input arcs to age and education. The first is because we only care about inputs to cognitive impairment and the second is because we assume demographics are not influenced by other variables.

The conditional independence testing procedure is best described with an example. Say age and education are identified as parent nodes of cognitive impairment in the base-DAG. We would then condition on age and education and test conditional independence between cognitive impairment and the remaining variables. For those found significant, we randomly select one to add as an arc, condition on this variable as a parent (along with the previous parents), and repeat this procedure until no other variables are significant. The resulting DAG is then included as a potential network. This procedure is stochastic due to the random arc selection.

To explore the space of potential arcs, we repeated this stochastic procedure 400 times. Any DAGs that introduced cyclicity were discarded. From all DAGs identified, we tested for equivalence and reported only the unique DAGs.

3.3.3 Model Fitting

The second step to learning a Bayesian network is to learn the local distributions (conditional probabilities for a node). This is done using the bn.fit function in the bnlearn R package [23]. We used the classical Bayesian posterior method with a uniform prior to estimate the conditional probability of an event.

We fit each DAG to the data to get a set of Bayesian networks and selected the best ones with the validation strategy described below.

3.3.4 Model Validation Strategy

We use a multi-class prediction strategy to validate our results. However, prediction is not the goal of this study; it is simply a method for comparison.

The validation strategy is two-fold. First, we train the model in a dataset containing multiple countries (randomly selected) and then predict on a test set that contains the same countries. This will allow us to find the most representative network for our selected countries without overfitting. We use an 80-20 split to create the train and test sets.

To determine whether the learned models are generalizable, we predict on an external validation set with countries different from those used for training and assess the results. This geographic external validation strategy tests for generalization since it evaluates on data generated under different cultural and systemic structures [11]. We randomly selected 65% of the countries to be used in the train-test validation described above, and the remaining 35% were left for external validation. The countries and sample sizes for each data set are shown in Supplementary Table 4.

3.4 Handling Data Imbalances

Our dataset is highly unbalanced for cognitive impairment; there are significantly more unimpaired cases than either mild or severe impairment.

The imbalance in our data is problematic for two reasons. First, imbalanced data can lead to poor classification results since the classifier often predicts the majority class to increase accuracy. Since prediction is our validation strategy, imbalances could bias the results. Second, Bayesian networks rely on conditional probabilities to learn their networks, so low case counts for a category will yield conditional probabilities close to zero. This makes identifying relationships difficult. As such, imbalanced data could prevent detection of important relationships with severe and mild impairment.

We take two approaches to counteract this problem. First, we oversample the minority classes (mild and severe) to generate a balanced dataset. Second, we define a classification score combining sensitivity for the mild and severe classes with specificity for the unimpaired class.

Oversampling is a simple strategy known to improve classification in unbalanced classification problems [10]. Oversampling amplifies signals in the minority class, which enables detection of relationships that may not be detected otherwise. However, oversampling may also amplify random signals. Fortunately, our internal validation strategy should let us select networks that minimize random signal detection. The oversampled training set contains 134,520 observations.

Note that we learned DAG structures using both the original and oversampled data. The oversampled data consistently outperformed the original data (Supplementary Tables 5 and 6), so all remaining analyses consider only the oversampled data.

We defined a classification score by combining sensitivity from the minority classes (severe and mild) with specificity of the majority class (unimpaired). The best performing model has the highest score. This scoring system emphasizes reduction of false negatives in the minority classes and false positives in the majority class. It also emphasizes learning relationships related to impairment rather than unimpairment, which may allow the structure learning algorithms to identify more potential risk factors than otherwise.

4 Results

4.1 Model Selection

We performed the hybrid structure-learning approach defined earlier on the oversampled training dataset and identified 47 potential DAGs including the base-DAG. We fit each of these models to the oversampled training set and tested prediction on the internal validation set (n = 10568). We extracted sensitivity, specificity, and balanced recall for each class and calculated the combined classification score. We then ranked each model according to highest combined score.

Table 2 shows the internal validation results for the base-DAG, the two best internally validated models, and the best externally validated model.

	Severe			Mild			Unimpaired			Combined	
Oversampled DAG	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Combined Score	Overall Rank
dag_39	0.750	0.882	0.816	0.669	0.803	0.736	0.689	0.885	0.787	2.305	1
dag_32	0.720	0.873	0.796	0.630	0.793	0.712	0.671	0.865	0.768	2.214	2
dag_31	0.685	0.843	0.764	0.505	0.791	0.648	0.647	0.847	0.747	2.038	20
base_dag	0.506	0.897	0.702	0.397	0.764	0.580	0.686	0.753	0.719	1.656	47

Table 2: Internal validation results for selected models trained on oversampled data. The combined score is calculated by adding sensitivity for severe and mild to specificity for unimpaired.

Table 2 shows that models trained on oversampled data detect severe and mild classes quite well under internal validation. Sensitivity for DAG 39 is 0.750 and 0.669 for severe and mild cases, respectively. Specificity is 0.885 for unimpaired cases. Additionally, balanced accuracy is above 0.730 for each class. DAG 32 is only slightly worse, so the networks are nearly equivalent. The base-DAG, however, performed the worst of the 47 models in the internal validation set, which confirms that the use of conditional independence testing improved our results.

Next we tested prediction on the external validation set (n = 27379) to assess the generalizability of the models we learned. Table 3 shows the external validation results for selected models. For full internal and external validation results see Supplementary Tables 6 and 7, respectively.

	Severe			**Mild**			**Unimpaired**			**Combined**	
Oversampled DAG	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Combined Score	Overall Rank
dag_31	0.507	0.830	0.669	0.385	0.756	0.571	0.608	0.777	0.692	1.669	1
base_dag	0.443	0.919	0.681	0.430	0.757	0.593	0.698	0.731	0.715	1.603	14
dag_32	0.422	0.856	0.639	0.381	0.765	0.573	0.643	0.725	0.684	1.528	42
dag_39	0.381	0.863	0.622	0.360	0.768	0.564	0.653	0.680	0.666	1.421	47

Table 3: External validation results for selected models trained on oversampled data. External dataset contained 7 countries.

Table 3 shows that DAGs 39 and 32 were ranked 47 and 42 in external validation, respectively, suggesting that these models are not generalizable outside the countries present in the training data. However, given the best external score (DAG 31: 1.669) is approximately the same as the worst internal score (base-DAG: 1.656), it is fair to say none of the models generalize particularly well. This is especially true since the range in external validation scores is only 0.248 while the range in the internal validation scores is 0.649. This makes sense considering countries and regions have unique cultural systems that may influence the interactions of our variables.

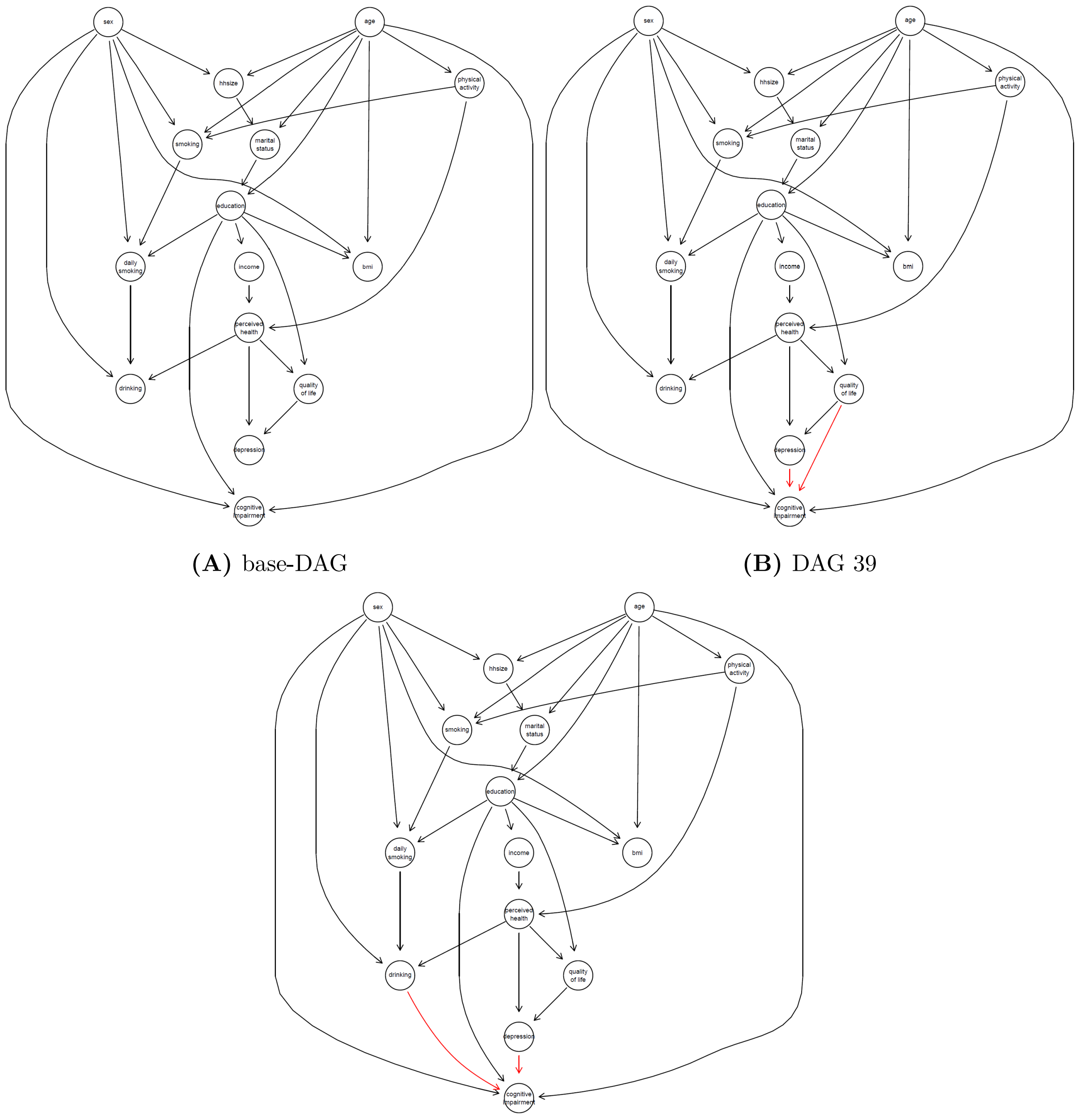
Consequently, our conclusions only apply to the countries in the internal data set. As such, we will only perform and in-depth analysis on DAG 39 and DAG 32. These are our selected models.

4.2 Direct Risk Factors for Cognitive Impairment

Our original goal was to use a Bayesian network to learn the dependence structure of our data, which would enable us to understand the direct and indirect relationships between potential risk factors and cognitive impairment. Figure 4 shows the most representative models (DAGs 39 and 32) we could find (See supplementary Figure 9 for DAG 31). The base-DAG is included as a reference.

All of the DAGs in Figure 4 are similar. Each of them identifies age, education, and sex as direct links to cognitive impairment because they were present in the base-DAG. The only differences are: 1) DAG 39 and 32 both add a direct relationship between depression and cognitive impairment, 2) DAG 39 adds a direct relationship between quality of life and cognitive impairment, and 3) DAG 32 adds a direct relationship between drinking and cognitive impairment.

Depression is clearly an important risk factor since it appears in both DAG 39 and 32. However, quality of life and drinking seem to play a similar role in their respective models. In fact, the conditional independence test results indicate that if quality of life is used as a risk factor, drinking loses its significance and vice versa. As such, they provide the same information, but quality of life is slightly better since its validation score is slightly higher. This makes sense since drinking levels and perceived quality of life are likely correlated.



(A) base-DAG

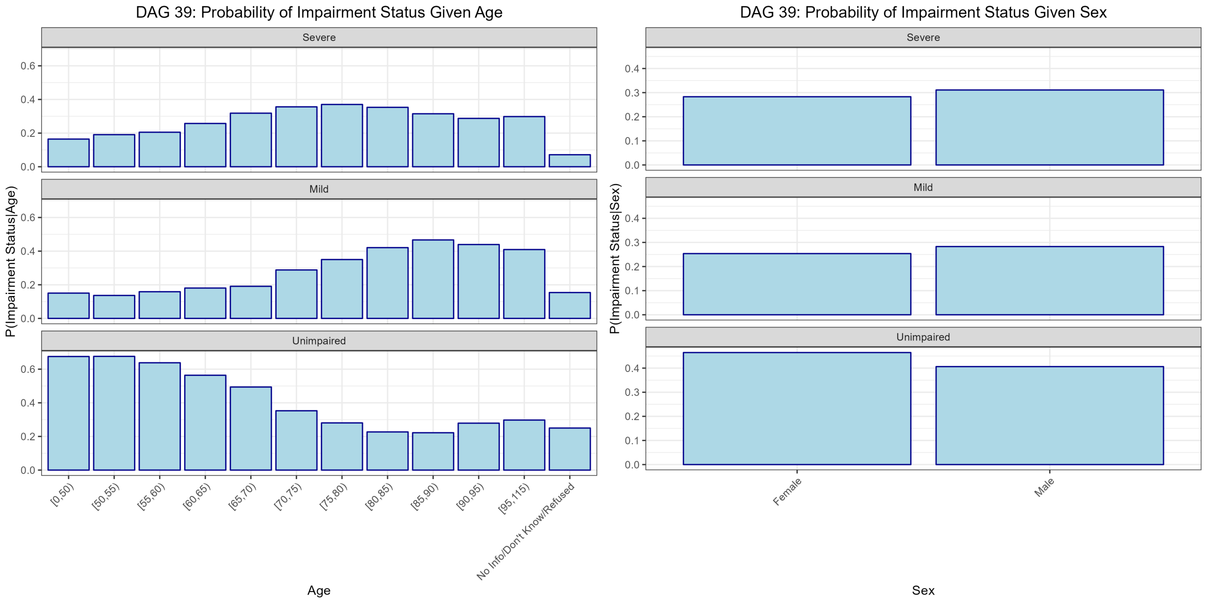
(B) DAG 39

(C) DAG 32

Figure 4: Baseline and selected Bayesian network models used to describe the probability dependence relationship between risk factors and cognitive impairment. A) base-DAG, B) DAG 39, and C) DAG 32. Red lines emphasize differences from the base-DAG.

Overall the results strongly indicate that the probability of cognitive impairment is directly dependent on age, education, sex, depression, and quality of life/drinking. To understand the direction of these relationships, we queried DAGs 39 and 32 to find the conditional probability of cognitive impairment given direct risk factors. These are not necessarily causal relations; we can only claim probabilistic dependence. Nonetheless, the relationships can guide future causal studies to test if these risk factors (particularly the modifiable ones) are targets for intervention.

Note that we only show the results for DAG 39 because the trends held for both DAGs with only minor changes in probability (except drinking; we show DAG 32 for drinking because the relationship was not in DAG 39).



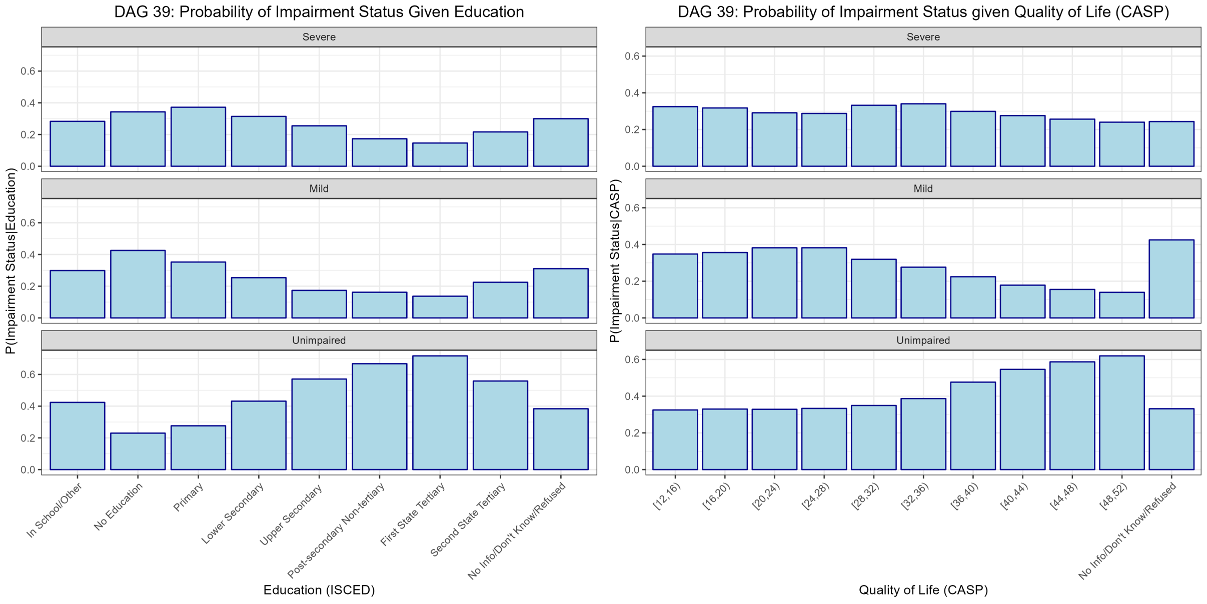
(A) **(B)**

Figure 5: Conditional probability of impairment status given A) age and B) sex. Results queried from DAG 39.

Figure 5A emphasizes that the probability of impairment increases with age as expected from literature [3]. Figure 5B shows that males are more likely to be impaired than females (although it's potentially not significant), which agrees with literature on vascular dementias but not Alzheimer's disease [3].

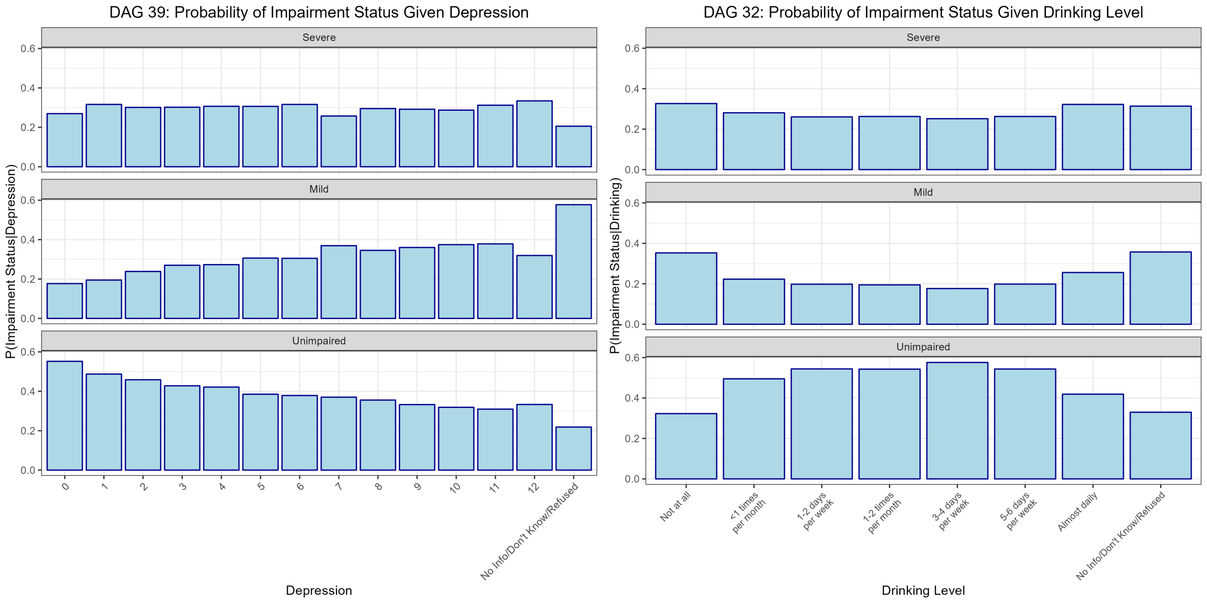
Age and sex are unmodifiable risk factors (sex is modifiable but not in a risk factor context), so there are no potential intervention targets here. However, it is worth noting that biological confounders (not in the data) may be the source of these relationships; if this were the case, the confounders themselves may be modifiable.

The remaining risk factors are modifiable, so they are potential targets for intervention.



(A) **(B)**

Figure 6: Conditional probability of impairment status given A) education and B) quality of life. Both education and casp increase from left to right. Results queried from DAG 39.



(A) Depression

(B) Drinking

Figure 7: Conditional probability of impairment status given A) depression and B) drinking. Both depression and drinking levels increase from left to right. Depression results queried from DAG 39 and drinking results from DAG 32.

Figure 6 emphasizes that the probability of impairment generally decreases with high education (Figure 6A) and high quality of life scores (Figure 6B). Figure 7A, shows that the probability of mild impairment increases with high depression while severe impairment does not appear affected. Probability of impairment also appears to increase with high levels of drinking and no drinking (Figure 7B). However, the effect of no drinking is likely confounded by either a history of drinking or health reasons.

The data shows that improving education, reducing drinking, combatting depression, and improving overall quality of life can reduce the probability of impairment. However, more rigorous studies are required for causal inference. Nonetheless, these networks offer useful targets for future study.

4.3 Indirect Risk Factors for Cognitive Impairment

Referring again to Figure 4, each of the variables in our study (except BMI) eventually flows into cognitive impairment. While they are not directly related, they affect the probability of direct links to cognitive impairment, so they still have an impact. As such, they are also potential targets for future studies and intervention.

For instance, Figure 7A showed that the probability of (mild) impairment decreases with lower depression scores. Since self-perceived health (sphus) is a direct link to depression, levels of perceived health associated with lower depression should be indirectly associated with lower impairment. Figure 8 shows the conditional probability of low, mild, and high depression scores given health.

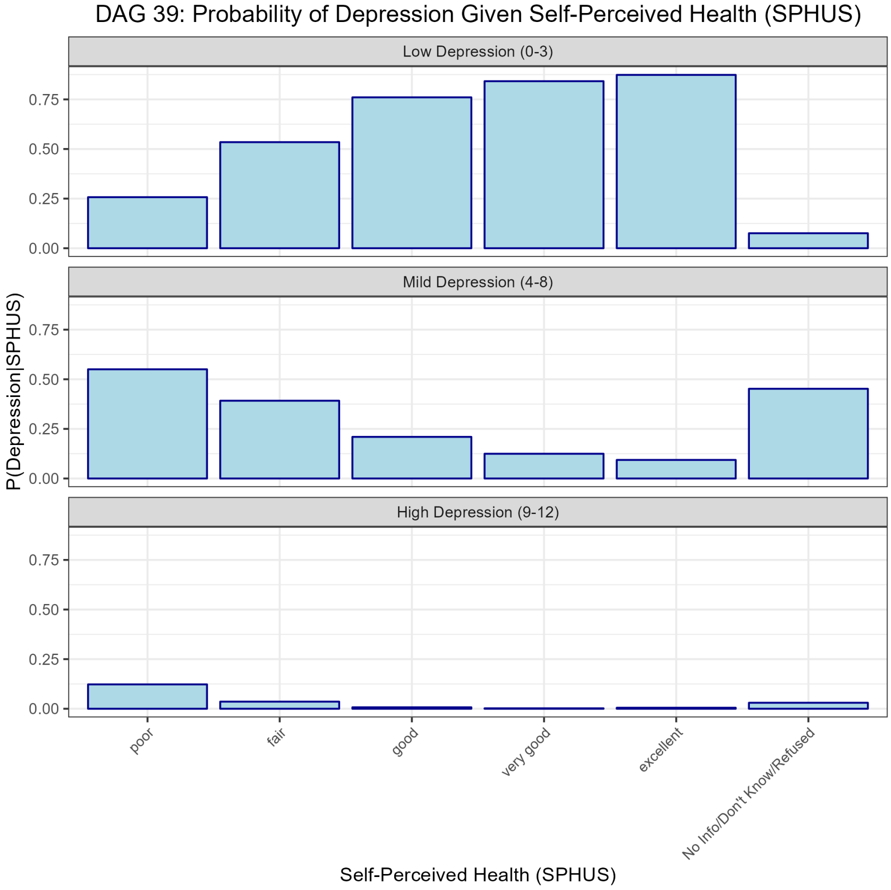


Figure 8: Conditional probability of depression given self-perceived health (SPHUS). Results queried from DAG 39 on each individual level in depression and then marginalized over levels in each group.

The results suggest that the probability of low depression increases as perceived health increases. As such, intervention strategies to increase perceived health could potentially (indirectly) reduce the probability of impairment. The same type of analysis can be performed with each of the remaining pathways in the Bayesian networks, but these analyses are left as future directions.

5 Conclusion and Discussion

Overall, we designed a hybrid structure-learning approach that performed better than score-based methods alone when testing on a population similar to the training population. We learned two very similar Bayesian networks (DAGs 39 and 32; Figure 4) with this approach that identified age, sex, education, depression, and quality of life/drinking as direct risk factors (not necessarily causal) for cognitive impairment. Additionally, we identified interaction pathways for each risk factor in our data (except BMI) that lead to cognitive impairment. Lastly, through an analysis of perceived health (sphus) and depression, we showed that indirect risk factors are potential intervention targets to reduce the probability of impairment.

This type of pathway analysis is critical for future causal and intervention studies. However, for the scope of this report, we did not assess every interaction in the network. For now, the Bayesian networks we provide offer especially useful targets for future study.

5.1 Limitations and Future Directions

While this analysis provides novel information for further analysis, there are important limitations. First, and most critical, we use cognitive impairment (based on recall and orientation alone) as a proxy for dementia. This does not fully capture the symptoms of dementia, so our conclusions only apply to recall and orientation-based cognitive impairment. Future studies should emphasize data that contains a dementia label.

Similarly, the bins used to define both cognitive impairment and our risk factors play a role in the interactions found, so the conclusions only apply to our specific definitions of cognitive impairment and risk factors. Future analysis is required to understand the impact of grouping strategy.

Lastly, the study population is non-incarcerated people of relatively good health (not hospitalized) who live in the 13 countries included in the training data. As the external validation results suggest, we cannot generalize to countries outside of those in our data. Future studies should focus on individual countries or regions to identify region specific risk factors.

There are potentially more limitations due to missing data, data entry errors, etc. However, the original data covers 16 years and 29 countries, so there are bound to be issues. All considering, our analysis provides incredibly valuable targets for future studies. We simply note the limitations to emphasize the need for continued analysis.

References

[1] S. Gauthier, P. Rosa-Neto, J. Morais, and C. Webster, "World alzheimer report 2021: Journey through the diagnosis of dementia," 2021.

[2] A. Börsch-Supan and S. Gruber, "easyshare. release version: 8.0.0," 2022.

[3] J. L. Podcasy and C. N. Epperson, "Considering sex and gender in alzheimer disease and other dementias," *Dialogues in Clinical Neuroscience*, vol. 18, pp. 437–446, 12 2016.

[4] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S. G. Costafreda, A. Dias, N. Fox, L. N. Gitlin, R. Howard, H. C. Kales, M. Kivimäki, E. B. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E. L. Sampson, Q. Samus, L. S. Schneider, G. Selbæk, L. Teri, and N. Mukadam, "Dementia prevention, intervention, and care: 2020 report of the lancet commission," *The Lancet*, vol. 396, pp. 413–446, 8 2020.

[5] F. Malter and A. Börsch-Supan, "Share wave 4: Innovations & methodology," 2013.

[6] K. J. Anstey, R. Peters, M. E. Mortby, K. M. Kiely, R. Eramudugolla, N. Cherbuin, M. H. Huque, and R. A. Dixon, "Association of sex differences in dementia risk factors with sex differences in memory decline in a population-based cohort spanning 20–76 years," *Scientific Reports*, vol. 11, p. 7710, 12 2021.

[7] M. Scutari, "Bayesian network models for incomplete and dynamic data," *Statistica Neerlandica*, vol. 74, pp. 397–419, 8 2020.

[8] A. Börsch-Supan, M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber, "Data resource profile: The survey of health, ageing and retirement in europe (share)," *International Journal of Epidemiology*, vol. 42, pp. 992–1001, 8 2013.

[9] N. Coley, M. P. Hoevenaar-Blom, J. Dalen, E. P. M. van Charante, M. Kivipelto, H. Soininen, S. Andrieu, and E. Richard, "Dementia risk scores as surrogate outcomes for lifestyle-based multidomain prevention trials—rationale, preliminary evidence and challenges," *Alzheimer's & Dementia*, vol. 16, pp. 1674–1685, 12 2020.

[10] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," pp. 243–248, IEEE, 4 2020.

[11] K. G. M. Moons, A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, and M. Woodward, "Risk prediction models: Ii. external validation, model updating, and impact assessment," *Heart*, vol. 98, pp. 691–698, 5 2012.

[12] F. Malter and A. Börsch-Supan, "Share wave 5: Innovations & methodology," 2015.

[13] A. Börsch-Supan, "Survey of health, ageing and retirement in europe (share) wave 1. release version: 8.0.0.," 2022.

[14] A. Börsch-Supan, "Survey of health, ageing and retirement in europe (share) wave 2. release version: 8.0.0.," 2022.

[15] A. Börsch-Supan, "Survey of health, ageing and retirement in europe (share) wave 4. release version: 8.0.0.," 2022.

[16] A. Börsch-Supan, "Survey of health, ageing and retirement in europe (share) wave 5. release version: 8.0.0.," 2022.

[17] A. Börsch-Supan and H. Jürges, "Wave 1: The survey of health, ageing and retirement in europe – methodology," 2005.

[18] A. Börsch-Supan, A. Brugiavini, H. Jürges, A. Kapteyn, J. Mackenbach, J. Siegrist, and G. Weber, "Wave 2: First results from the survey of health, ageing and retirement in europe (2004-2007). starting the longitudinal dimension," 2008.

[19] M. Scutari, C. E. Graafland, and J. M. Gutierrez, "Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms?," *Proceedings of Machine Learning Research*, vol. 72, pp. 416–427, 2018.

[20] E. M. Crimmins, J. K. Kim, K. M. Langa, and D. R. Weir, "Assessment of cognition using surveys and neuropsychological assessment: The health and retirement study and the aging, demographics, and memory study," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 66B, pp. i162–i171, 7 2011.

[21] D. Gallagher, S. B. Heymsfield, M. Heo, S. A. Jebb, P. R. Murgatroyd, and Y. Sakamoto, "Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index," *The American Journal of Clinical Nutrition*, vol. 72, pp. 694–701, 9 2000.

[22] N. Ruozzi, "Directed graphical models." https://personal.utdallas.edu/ nrr150130/gmbook/bayes.html.

[23] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, pp. 22–1, 2010.

Supplementary Tables and Figures

Dataset	Countries	Proportion of Countries	N
Internal Training	Poland, Germany, Austria, Greece, Switzerland, Belgium, Slovenia, Portugal, Spain, Estonia, Netherlands, France, Sweden	0.65	52668
Internal Validation	--	--	10568
External Validation	Czech Republic, Denmark, Hungary, Ireland, Italy, Israel, Luxembourg	0.35	27379

Table 4: Internal and external dataset composition. The internal sets were randomly subsetted, so individual country composition may randomly differ from what is presented here.

	Severe			Mild			Unimpaired			Combined	
DAG	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Combined Score	Overall Rank
dag_4	0.218	0.986	0.602	0.038	0.998	0.518	0.990	0.162	0.576	0.418	1
dag_2	0.205	0.985	0.595	0.025	0.999	0.512	0.989	0.151	0.570	0.382	2
dag_1	0.209	0.985	0.597	0.004	0.999	0.502	0.990	0.141	0.565	0.353	3
dag_5	0.186	0.985	0.586	0.006	0.999	0.503	0.991	0.134	0.562	0.326	4
dag_3	0.152	0.986	0.569	0.004	1.000	0.502	0.991	0.107	0.549	0.262	5
base_dag	0.150	0.986	0.568	0.000	1.000	0.500	0.991	0.108	0.549	0.258	6

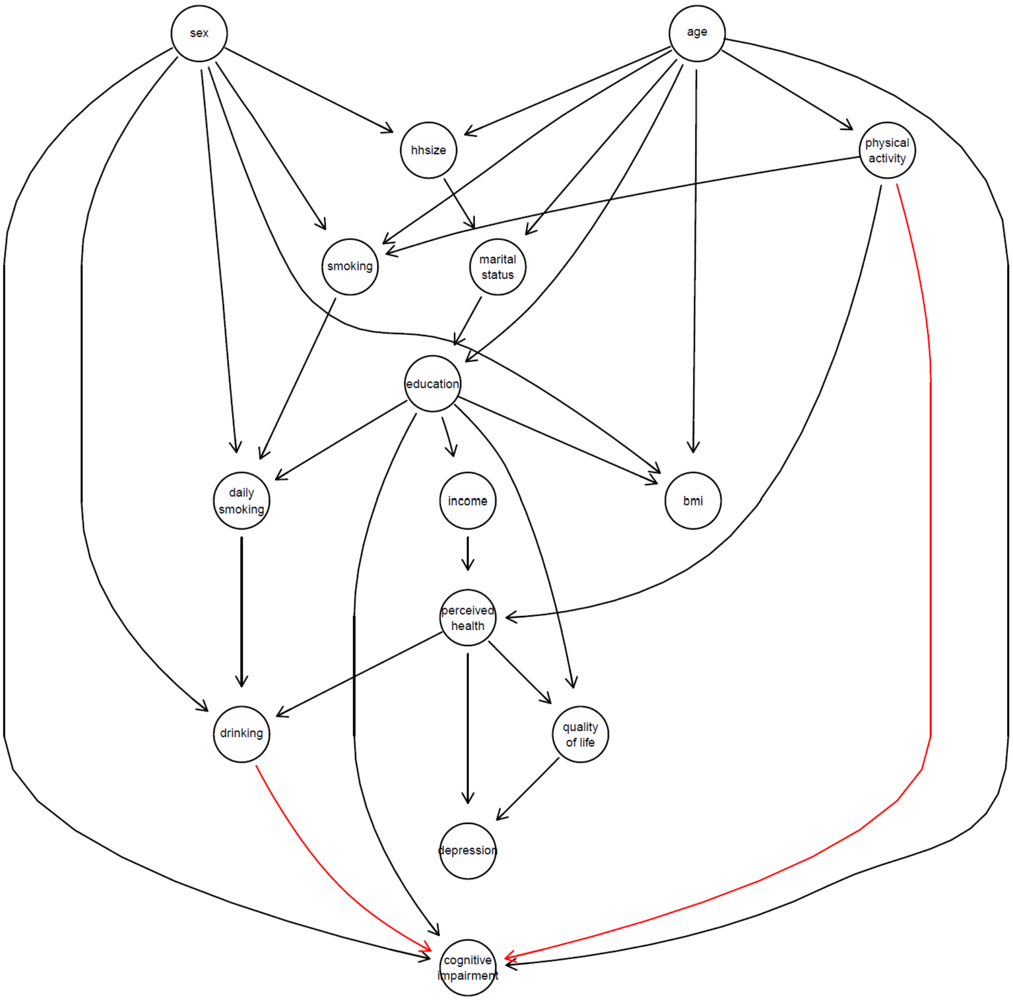
Table 5: Full internal validation results for models trained on original (not-oversampled) data.

	Severe			Mild			Unimpaired			Combined	
Oversampled DAG	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Combined Score	Overall Rank
dag_39	0.750	0.882	0.816	0.669	0.803	0.736	0.689	0.885	0.787	2.305	1
dag_32	0.720	0.873	0.796	0.630	0.793	0.712	0.671	0.865	0.768	2.214	2
dag_44	0.692	0.871	0.782	0.621	0.794	0.707	0.673	0.859	0.766	2.172	3
dag_24	0.698	0.858	0.778	0.592	0.797	0.695	0.666	0.873	0.770	2.163	4
dag_7	0.703	0.842	0.773	0.568	0.789	0.679	0.644	0.882	0.763	2.154	5
dag_1	0.668	0.856	0.762	0.605	0.780	0.693	0.645	0.862	0.754	2.135	6
dag_9	0.647	0.867	0.757	0.621	0.783	0.702	0.660	0.859	0.759	2.127	7
dag_34	0.665	0.881	0.773	0.592	0.790	0.691	0.681	0.847	0.764	2.104	8
dag_17	0.670	0.868	0.769	0.589	0.794	0.691	0.671	0.845	0.758	2.104	9
dag_4	0.677	0.880	0.778	0.591	0.794	0.692	0.681	0.834	0.757	2.102	10
dag_13	0.688	0.871	0.780	0.568	0.795	0.682	0.676	0.845	0.761	2.102	11
dag_10	0.648	0.866	0.757	0.602	0.788	0.695	0.665	0.852	0.758	2.102	12
dag_25	0.683	0.848	0.765	0.552	0.794	0.673	0.655	0.866	0.761	2.101	13
dag_40	0.683	0.864	0.773	0.549	0.805	0.677	0.681	0.846	0.764	2.078	14
dag_22	0.666	0.864	0.765	0.553	0.787	0.670	0.665	0.853	0.759	2.072	15
dag_20	0.665	0.856	0.761	0.548	0.790	0.669	0.661	0.858	0.759	2.071	16
dag_33	0.661	0.851	0.756	0.541	0.784	0.663	0.650	0.866	0.758	2.068	17
dag_45	0.658	0.876	0.767	0.564	0.784	0.674	0.672	0.839	0.755	2.061	18
dag_16	0.639	0.867	0.753	0.579	0.787	0.683	0.665	0.842	0.753	2.060	19
dag_31	0.685	0.843	0.764	0.505	0.791	0.648	0.647	0.847	0.747	2.038	20
dag_36	0.644	0.866	0.755	0.545	0.777	0.661	0.657	0.849	0.753	2.038	21
dag_28	0.629	0.877	0.753	0.555	0.788	0.672	0.679	0.833	0.756	2.018	22
dag_3	0.679	0.856	0.767	0.493	0.792	0.643	0.663	0.837	0.750	2.008	23
dag_43	0.657	0.853	0.755	0.515	0.793	0.654	0.659	0.829	0.744	2.000	24
dag_38	0.633	0.863	0.748	0.527	0.786	0.656	0.662	0.829	0.745	1.989	25
dag_29	0.631	0.865	0.748	0.517	0.780	0.649	0.659	0.828	0.744	1.976	26
dag_14	0.616	0.868	0.742	0.521	0.782	0.651	0.666	0.828	0.747	1.965	27
dag_15	0.607	0.873	0.740	0.532	0.784	0.658	0.671	0.817	0.744	1.956	28
dag_21	0.628	0.880	0.754	0.517	0.785	0.651	0.679	0.809	0.744	1.954	29
dag_2	0.610	0.863	0.737	0.516	0.797	0.656	0.677	0.825	0.751	1.951	30
dag_27	0.616	0.857	0.736	0.499	0.785	0.642	0.661	0.835	0.748	1.950	31
dag_8	0.613	0.859	0.736	0.511	0.779	0.645	0.653	0.819	0.736	1.943	32
dag_5	0.602	0.873	0.737	0.510	0.788	0.649	0.679	0.824	0.752	1.936	33
dag_35	0.632	0.862	0.747	0.442	0.800	0.621	0.681	0.805	0.743	1.880	34
dag_41	0.583	0.861	0.722	0.488	0.778	0.633	0.657	0.809	0.733	1.879	35
dag_19	0.609	0.866	0.737	0.465	0.782	0.623	0.664	0.800	0.732	1.873	36
dag_12	0.587	0.872	0.729	0.480	0.787	0.634	0.677	0.796	0.737	1.864	37
dag_18	0.603	0.872	0.738	0.454	0.790	0.622	0.683	0.803	0.743	1.861	38
dag_46	0.587	0.875	0.731	0.464	0.772	0.618	0.667	0.803	0.735	1.854	39
dag_30	0.557	0.872	0.714	0.476	0.782	0.629	0.673	0.795	0.734	1.827	40
dag_23	0.558	0.877	0.717	0.472	0.784	0.628	0.680	0.788	0.734	1.818	41
dag_42	0.560	0.879	0.719	0.467	0.777	0.622	0.676	0.791	0.733	1.818	42
dag_26	0.555	0.872	0.714	0.454	0.786	0.620	0.679	0.786	0.732	1.796	43
dag_6	0.540	0.887	0.714	0.460	0.770	0.615	0.677	0.781	0.729	1.782	44
dag_11	0.561	0.878	0.720	0.428	0.790	0.609	0.689	0.773	0.731	1.762	45
dag_37	0.538	0.880	0.709	0.445	0.783	0.614	0.684	0.775	0.730	1.758	46
base_dag	0.506	0.897	0.702	0.397	0.764	0.580	0.686	0.753	0.719	1.656	47

Table 6: Full internal validation results for models trained on oversampled data.

	Severe			Mild			Unimpaired			Combined	
Oversampled DAG	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy	Combined Score	Overall Rank
dag_31	0.507	0.830	0.669	0.385	0.756	0.571	0.608	0.777	0.692	1.669	1
dag_14	0.480	0.862	0.671	0.402	0.754	0.578	0.638	0.760	0.699	1.642	2
dag_6	0.418	0.887	0.653	0.447	0.726	0.586	0.637	0.766	0.701	1.631	3
dag_29	0.482	0.856	0.669	0.394	0.757	0.575	0.634	0.754	0.694	1.629	4
dag_27	0.482	0.852	0.667	0.375	0.772	0.573	0.650	0.771	0.711	1.628	5
dag_41	0.437	0.868	0.652	0.427	0.748	0.588	0.639	0.762	0.701	1.626	6
dag_30	0.448	0.879	0.664	0.407	0.768	0.588	0.675	0.768	0.722	1.624	7
dag_25	0.476	0.837	0.656	0.379	0.773	0.576	0.633	0.759	0.696	1.614	8
dag_5	0.449	0.868	0.659	0.406	0.778	0.592	0.672	0.759	0.715	1.614	9
dag_33	0.464	0.844	0.654	0.384	0.766	0.575	0.635	0.760	0.697	1.608	10
dag_22	0.468	0.849	0.658	0.387	0.754	0.570	0.625	0.753	0.689	1.608	11
dag_46	0.420	0.861	0.640	0.422	0.737	0.579	0.622	0.765	0.693	1.607	12
dag_36	0.470	0.856	0.663	0.395	0.756	0.575	0.632	0.740	0.686	1.605	13
base_dag	0.443	0.919	0.681	0.430	0.757	0.593	0.698	0.731	0.715	1.603	14
dag_7	0.457	0.827	0.642	0.387	0.745	0.566	0.592	0.754	0.673	1.598	15
dag_8	0.440	0.859	0.650	0.408	0.753	0.581	0.634	0.747	0.691	1.595	16
dag_9	0.453	0.853	0.653	0.397	0.756	0.577	0.631	0.745	0.688	1.595	17
dag_18	0.463	0.875	0.669	0.390	0.773	0.581	0.671	0.740	0.706	1.594	18
dag_20	0.460	0.845	0.652	0.379	0.760	0.570	0.628	0.753	0.691	1.593	19
dag_43	0.447	0.853	0.650	0.398	0.769	0.583	0.643	0.741	0.692	1.586	20
dag_2	0.427	0.861	0.644	0.406	0.777	0.592	0.663	0.750	0.707	1.583	21
dag_24	0.449	0.840	0.644	0.389	0.748	0.568	0.608	0.743	0.675	1.581	22
dag_19	0.422	0.873	0.648	0.416	0.748	0.582	0.644	0.741	0.692	1.579	23
dag_15	0.423	0.867	0.645	0.401	0.776	0.589	0.670	0.752	0.711	1.576	24
dag_16	0.414	0.858	0.636	0.406	0.766	0.586	0.649	0.751	0.700	1.571	25
dag_28	0.431	0.881	0.656	0.410	0.779	0.595	0.683	0.726	0.704	1.566	26
dag_35	0.453	0.866	0.660	0.380	0.779	0.579	0.669	0.731	0.700	1.565	27
dag_38	0.420	0.861	0.640	0.400	0.747	0.573	0.631	0.745	0.688	1.564	28
dag_17	0.429	0.852	0.640	0.392	0.763	0.577	0.638	0.740	0.689	1.561	29
dag_21	0.422	0.882	0.652	0.408	0.779	0.593	0.685	0.727	0.706	1.557	30
dag_37	0.443	0.888	0.666	0.388	0.769	0.579	0.680	0.724	0.702	1.556	31
dag_11	0.432	0.884	0.658	0.393	0.776	0.584	0.684	0.723	0.704	1.548	32
dag_3	0.426	0.850	0.638	0.384	0.769	0.576	0.642	0.735	0.688	1.545	33
dag_42	0.398	0.890	0.644	0.417	0.760	0.589	0.674	0.728	0.701	1.544	34
dag_13	0.437	0.866	0.652	0.380	0.770	0.575	0.658	0.723	0.691	1.541	35
dag_40	0.460	0.856	0.658	0.361	0.776	0.568	0.653	0.720	0.687	1.540	36
dag_26	0.412	0.885	0.648	0.403	0.776	0.590	0.686	0.723	0.704	1.538	37
dag_45	0.416	0.878	0.647	0.407	0.775	0.591	0.675	0.714	0.694	1.537	38
dag_23	0.428	0.888	0.658	0.387	0.773	0.580	0.686	0.722	0.704	1.536	39
dag_1	0.403	0.844	0.624	0.394	0.736	0.565	0.602	0.737	0.670	1.534	40
dag_44	0.410	0.860	0.635	0.399	0.776	0.587	0.658	0.722	0.690	1.532	41
dag_32	0.422	0.856	0.639	0.381	0.765	0.573	0.643	0.725	0.684	1.528	42
dag_10	0.386	0.856	0.621	0.402	0.769	0.586	0.650	0.735	0.693	1.523	43
dag_34	0.400	0.880	0.640	0.408	0.774	0.591	0.676	0.706	0.691	1.514	44
dag_12	0.415	0.884	0.649	0.382	0.773	0.578	0.680	0.703	0.691	1.501	45
dag_4	0.374	0.879	0.626	0.396	0.776	0.586	0.676	0.689	0.683	1.460	46
dag_39	0.381	0.863	0.622	0.360	0.768	0.564	0.653	0.680	0.666	1.421	47

Table 7: Full external validation results for models trained on oversampled data.



(A) DAG 31

Figure 9: Most generalizable DAG from all models learned. DAG 31