Opening the Black Box: Discovering and Explaining Hidden Variables in Patient Modelling

IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018)

Leila Yousefi, Stephen Swift, Mahir Arzoky, Allan Tucker

Brunel University London

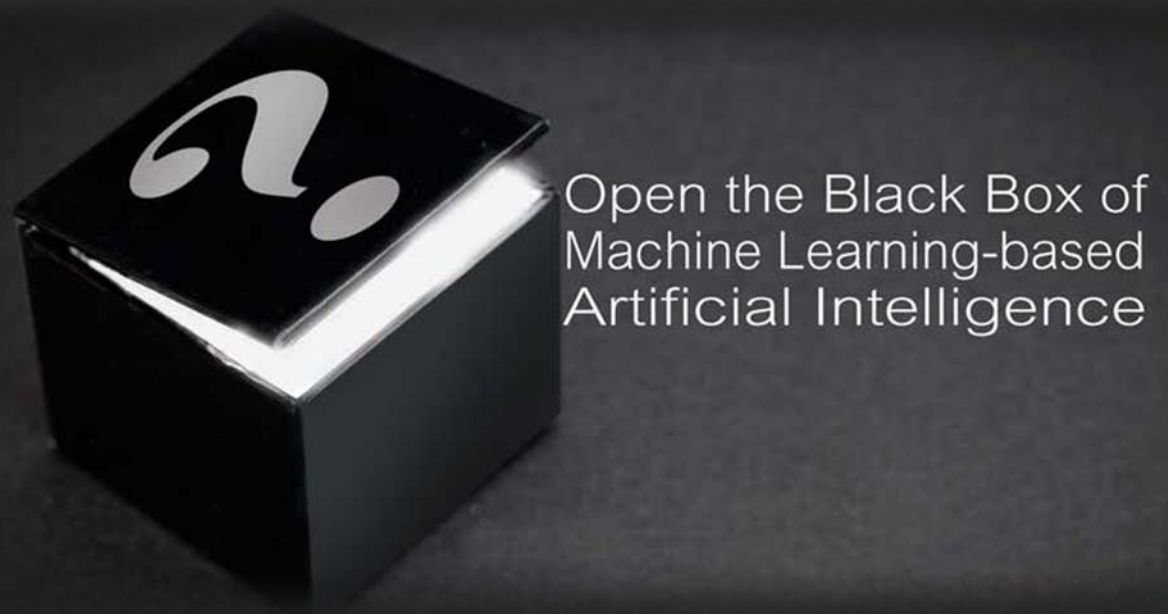
Lucia Saachi, Luca Chiovato

University of Pavia, Instituti Maugeri, Italy

Type 2 Diabetes Mellitus (T2DM)

Mortality due to diabetes age 20-79 in 2017 (in millions)

Outline



- Motivation
- Data
- Problem
- Solution
- Hidden variable discovery approach
  - Over-sampling and Enhanced Stepwise approach
  - Stratifying patients based on their hidden variable
- Results
- Conclusions and future works

Leila.Yousefi@brunel.ac.uk

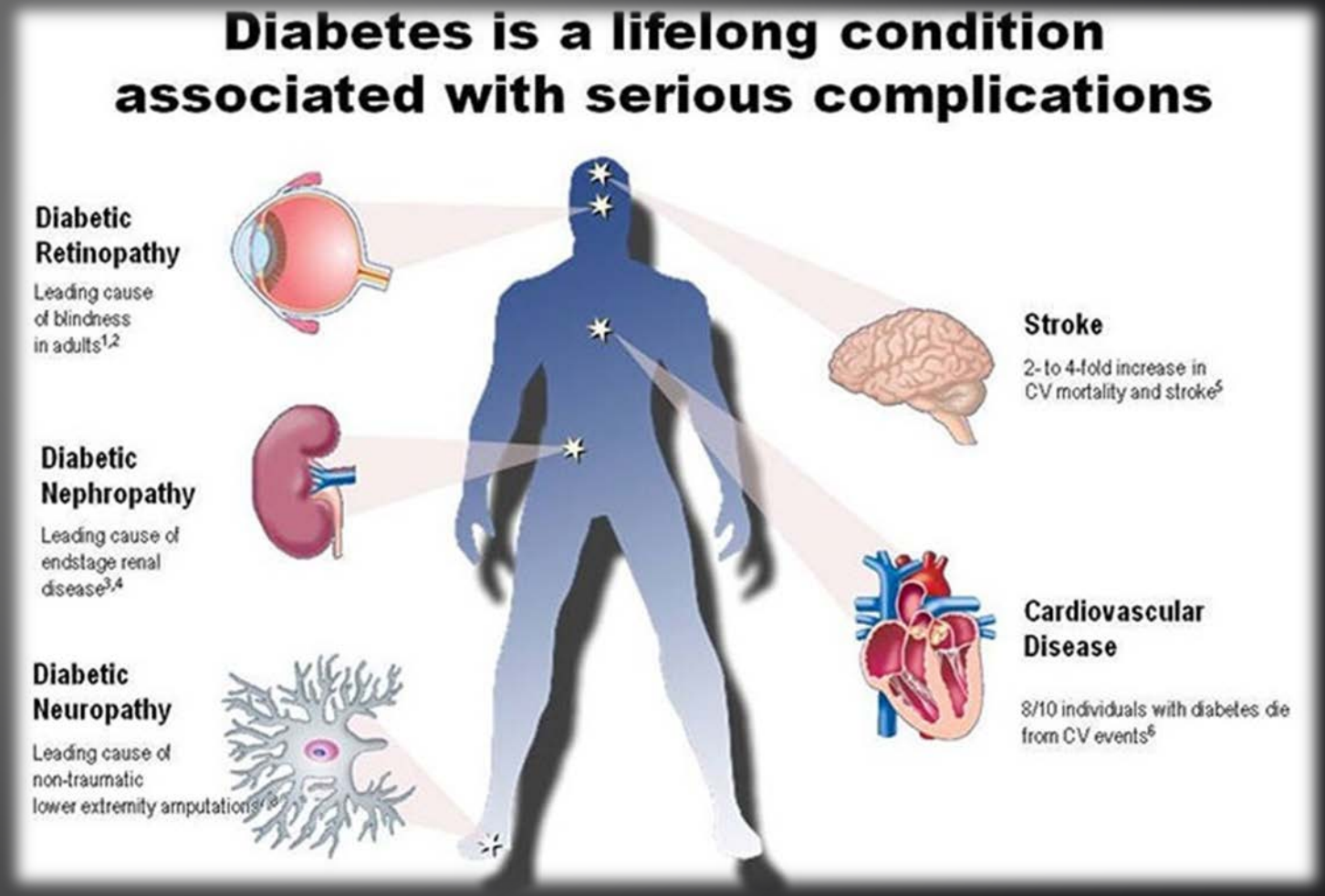
Type 2 Diabetes

The Data – Clinical Features

Clinical feature		Risk factors	compilation
HbA1c (\%)	6.6 \pm 1.2	YES	NO
Retinopathy	{0,1}	NO	NO
Neuropathy	{0,1}	NO	NO
Nephropathy	{0,1}	NO	NO
Liver Disease	{0,1}	NO	NO
Hypertension	{0,1}	NO	NO
BMI (kg/m2)	26.4 \pm 2.4	YES	NO
Creatinine (mg/dL)	0.9 \pm 0.2	YES	NO
HDL cholesterol (mmol/l)	1.1 \pm 0.3	YES	NO
Systolic blood pressure(SBP) (mmHg)	148 \pm 19	YES	NO
Smoking Habit	{0,1,2}	YES	NO
Hidden variable	[0,1]	YES	YES

Main clinical risk factors of T2DM and control (Mean/pm SD)

- Type 2 Diabetes Mellitus (T2DM)
- Patients aged 25 to 65 years
- 2009 and 2013
- IRCCS Istituti Clinici Scientifici Maugeri of Pavia, Italy
- MOSAIC project funded by the European Commission



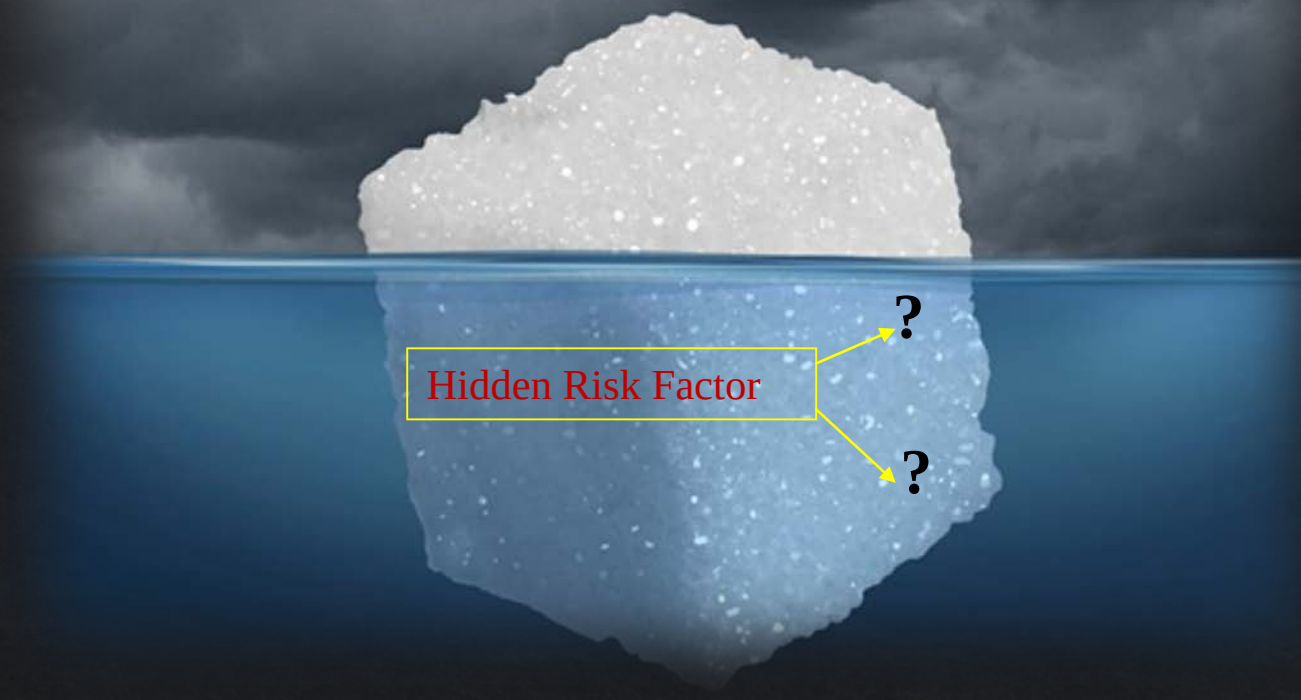
T2DM Data

Visit NO	Patient ID	HbA1c	Retinopathy	Neuropathy	Nephropathy	Liver disease	Hypertension	BMI	Creatinine	Cholestrol	HDL	DBP	SBP	SMK
1	885	0.769	0	0	0	0	1	0.286	-0.391	2.082	0.020	1.705	0.286	1.335
2	885	0.769	0	0	0	1	1	0.286	-0.391	2.082	0.020	1.705	0.286	1.335
3	885	0.769	1	0	0	1	1	0.286	-0.391	2.082	0.020	1.705	0.286	1.335
4	885	0.769	1	0	1	1	1	0.286	-0.391	2.082	0.020	1.705	0.286	1.335
1	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**2.782**	**-0.511**	**-0.149**	**-0.053**	**0.297**	**0.286**	**1.335**
2	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**2.782**	**-0.511**	**-0.149**	**-0.053**	**0.297**	**0.286**	**1.335**
4	**894**	**-0.056**	**0**	**0**	**1**	**1**	**1**	**2.937**	**-0.511**	**-0.017**	**-0.343**	**0.297**	**0.794**	**1.335**
5	**894**	**-0.056**	**0**	**0**	**1**	**1**	**1**	**2.937**	**-0.511**	**-0.017**	**-0.343**	**0.297**	**0.794**	**1.335**
6	**894**	**-0.056**	**0**	**0**	**1**	**1**	**1**	**2.937**	**-0.511**	**-0.017**	**-0.343**	**0.297**	**0.794**	**1.335**
7	**894**	**-0.262**	**0**	**0**	**1**	**1**	**1**	**2.782**	**-0.511**	**0.534**	**-0.488**	**0.297**	**0.540**	**1.335**
8	**894**	**-0.262**	**0**	**0**	**1**	**1**	**1**	**2.782**	**-0.511**	**0.534**	**-0.488**	**0.297**	**0.540**	**1.335**
9	**894**	**-0.262**	**0**	**0**	**1**	**1**	**1**	**2.782**	**-0.511**	**0.534**	**-0.488**	**0.297**	**0.540**	**1.335**
10	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**2.906**	**-0.511**	**0.744**	**-0.488**	**-0.642**	**-0.223**	**1.335**
11	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**2.906**	**-0.511**	**0.744**	**-0.488**	**-0.642**	**-0.223**	**1.335**
12	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**2.906**	**-0.511**	**0.744**	**-0.488**	**-0.642**	**-0.223**	**1.335**
13	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**3.557**	**-0.391**	**0.376**	**0.455**	**-0.642**	**-0.223**	**1.335**
14	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**3.557**	**-0.391**	**0.376**	**0.455**	**-0.642**	**-0.223**	**1.335**
15	**894**	**0.151**	**0**	**0**	**1**	**1**	**1**	**3.557**	**-0.391**	**0.376**	**0.455**	**-0.642**	**-0.223**	**1.335**
16	**894**	**0.013**	**0**	**0**	**1**	**1**	**1**	**3.324**	**-0.235**	**0.744**	**-0.125**	**-0.642**	**-0.223**	**1.335**
1	1010	1.388	0	0	1	0	0	0.162	-0.630	2.450	-0.779	2.175	2.827	1.335
2	1010	1.388	0	0	1	0	1	0.162	-0.630	2.450	-0.779	2.175	2.827	1.335
3	1010	1.388	0	0	1	0	1	0.162	-0.630	2.450	-0.779	2.175	2.827	1.335
4	1010	1.388	0	0	1	0	1	0.162	-0.630	2.450	-0.779	2.175	2.827	1.335
5	1010	2.350	0	0	1	0	1	0.206	-0.511	0.875	0.818	0.297	-0.223	1.335
6	1010	2.350	0	0	1	0	1	0.206	-0.511	0.875	0.818	0.297	-0.223	1.335
7	1010	2.350	0	0	1	0	1	0.206	-0.511	0.875	0.818	0.297	-0.223	1.335
8	1010	2.350	0	0	1	0	1	0.206	-0.511	0.875	0.818	0.297	-0.223	1.335
9	1010	2.350	0	0	1	0	1	0.078	-0.750	1.636	-0.053	-0.642	0.286	1.335

The Problem

- Predicting comorbidities at the earliest from time-series data is challenging
- Each patient has a dynamic and unique profile
- Complications interact
- Many unmeasured effects, being ‘Black Box in nature’
- Mining time-series data in the prognosis of disease with rare positive results
- Unexpected development of complications and varying responses of patients to disease over time.
- Need to find different groups of patients

sharing similar profile of risk factors



Leila.Yousefi@brunel.ac.uk

The solution - Personalising Medicine

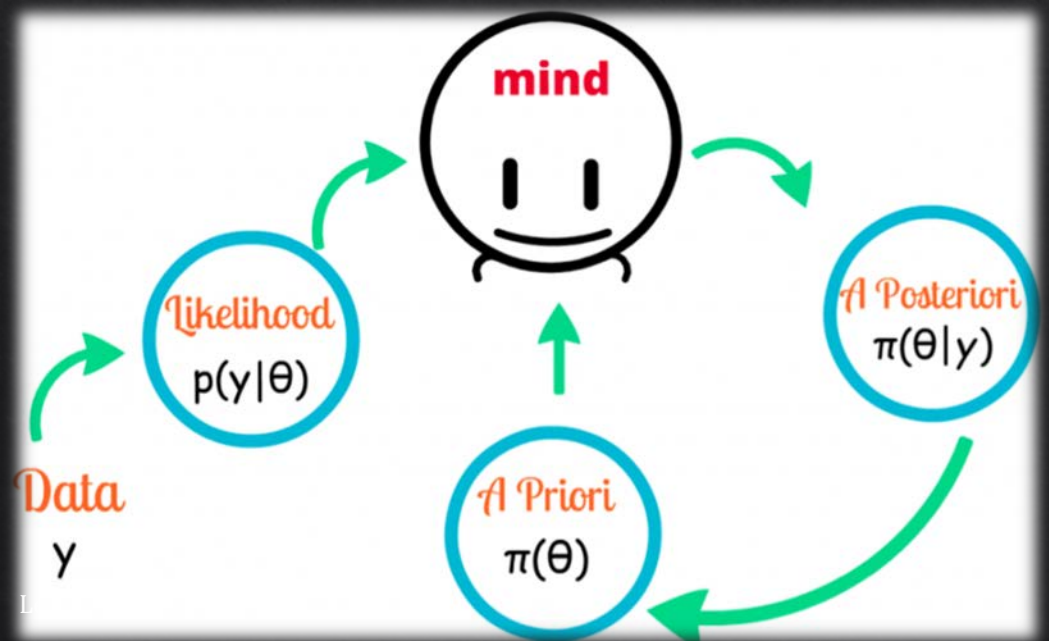
Hidden Variable Discovery Approach

- Finding methods to assess the influences of these latent variables
- Discover the dependencies between the latent variable and the observed variables
- Discover Diabetic trigger and eliminate diabetes forever!
- Determining the precise position of the latent variable
- Identifying and understanding groups of patients' with similar disease profiles (based on discovered hidden variables)

Leila.Yousefi@brunel.ac.uk

Dynamic Bayesian Networks

- Ideal for clinical data:
  - Flexibility in continuous and discrete variable
  - Handling uncertainty through the modelling of probability distributions
  - Enables prediction through inference
  - No limit for minimum sample size
  - Transparent (querying the model, graphical structure, etc.)
  - It can naturally facilitate latent variables …

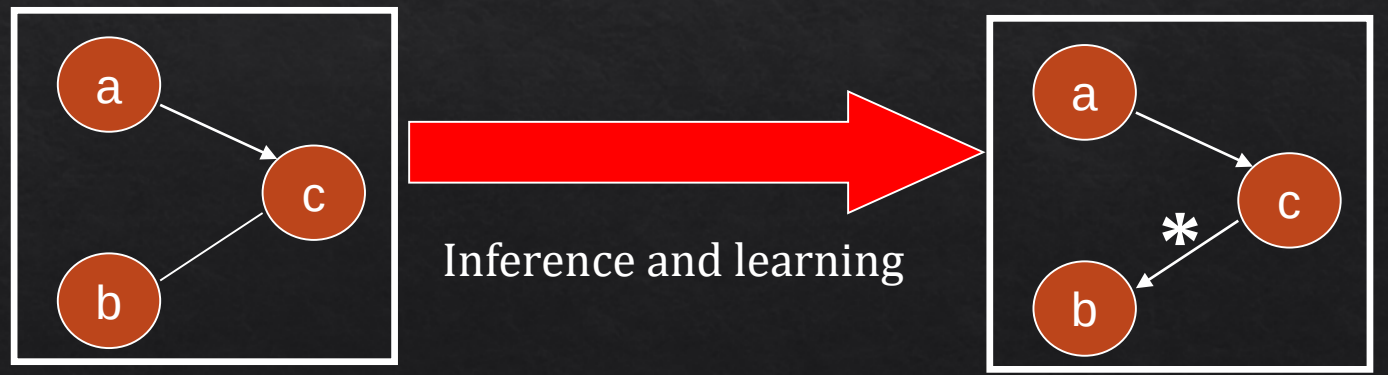


IC (Inductive Causation) algorithm

- Conditional independence analyses to infer causal structures
- IC* algorithm (an extension of IC) learns a partially oriented Directed Acyclic graph (pattern) with latent variables.

Whenever a then b
but not vice versa

Possibly
a => b



Causal structure- DAG

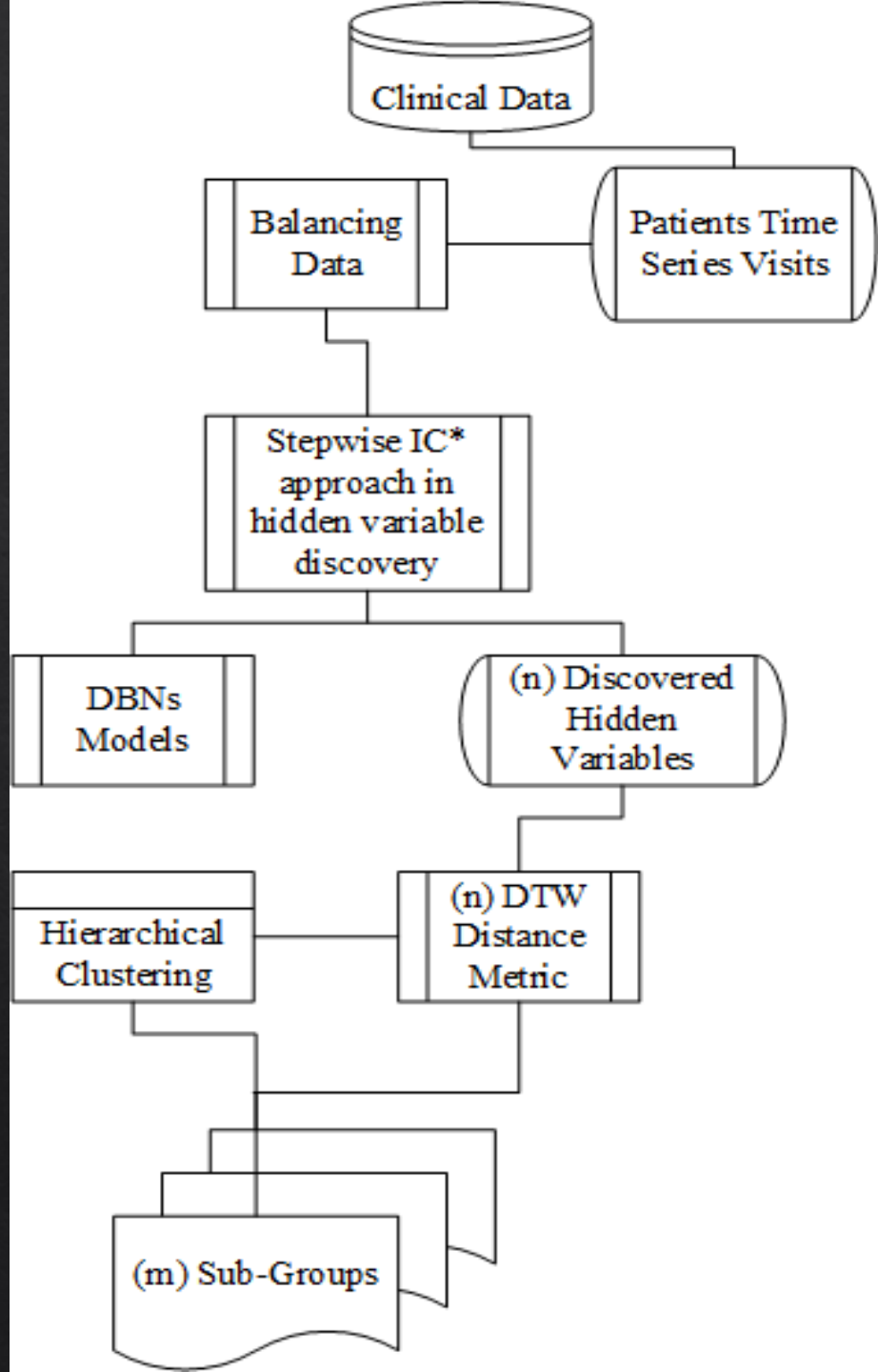
Leila.Yousefi@brunel.ac.uk

See Pearl, "Causality: Models, Reasoning, and Inference", 2000, p52 for more details.

Enhanced Stepwise method

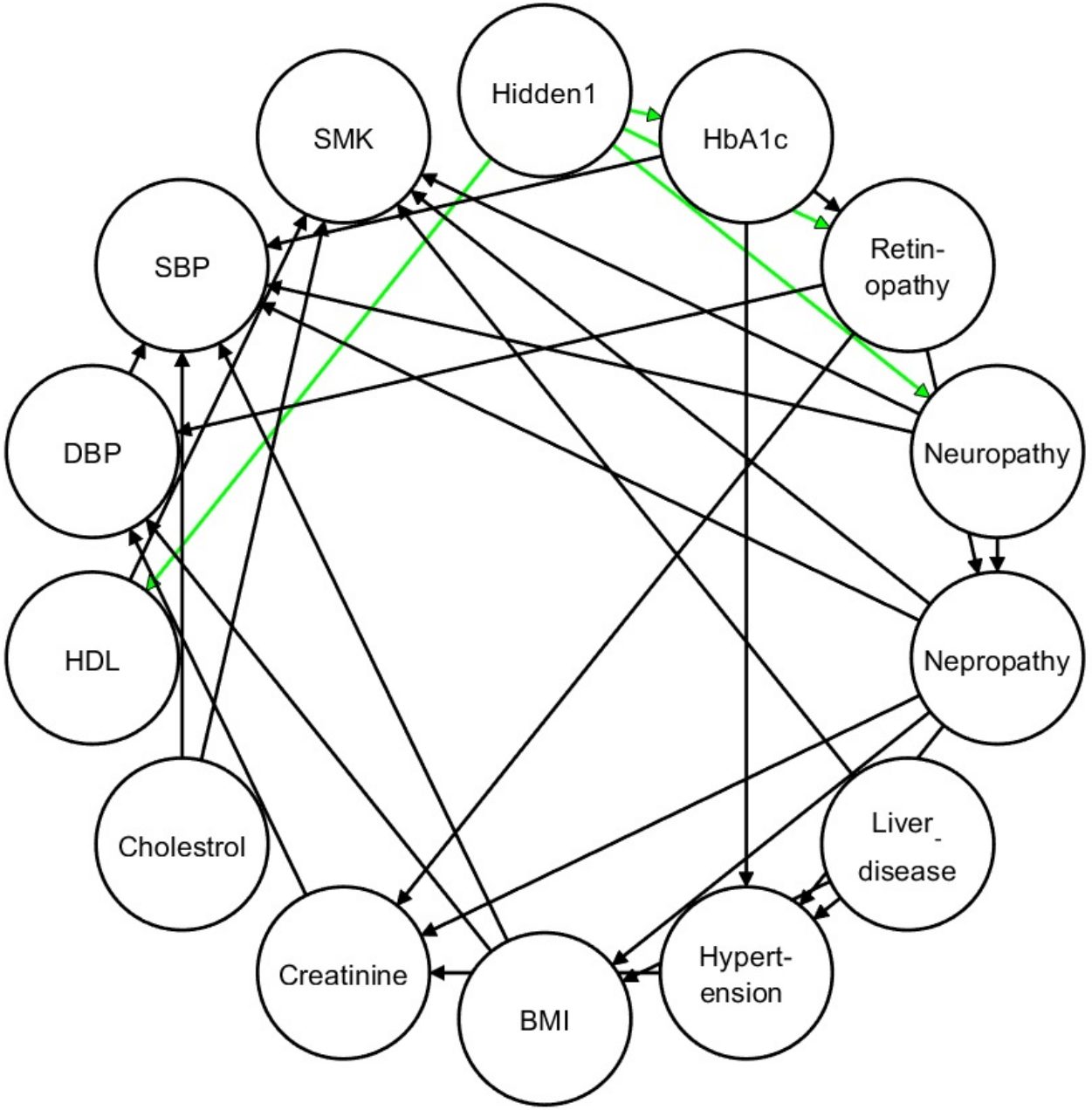
Incrementally identifying hidden variables

1. **B**alance the data based upon a specific complication using oversampling method on the random patients in a minority class (Positive cases)
2. Apply IC* algorithm
3. Provide parameter by applying inference rules on all discovered hidden variables.
4. Treat the discovered hidden variable as an observed variable.
5. Re-apply the IC* and repeat step 2, 3 and 4 until no new hidden variable is discovered.
6. If no hidden variable was found, or chain connections between hidden variables are destroyed then stop

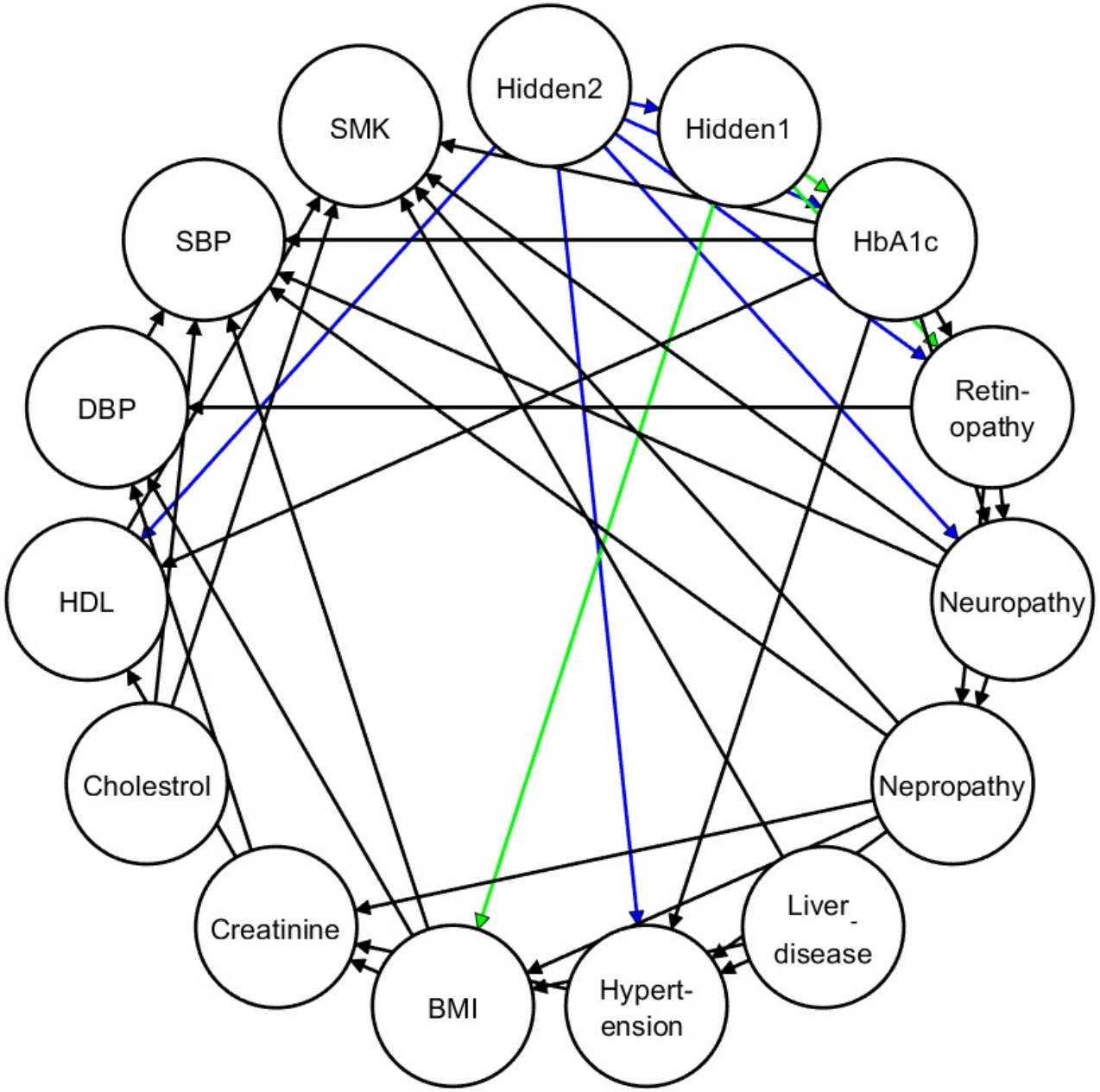


Leila.Yousefi@brunel.ac.uk

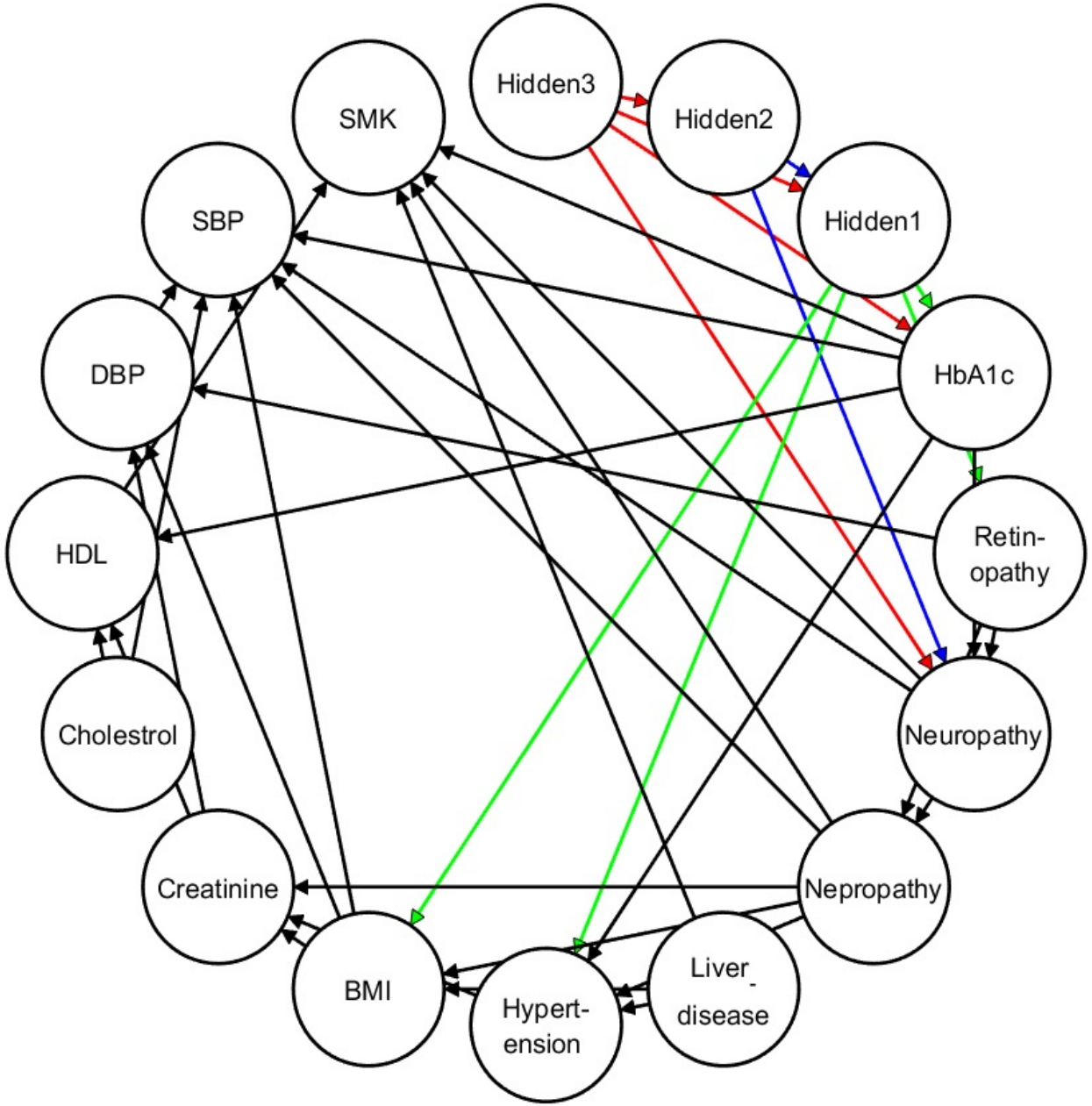
Step 1:


Hidden1
SMK
HbA1c
SBP
Retin-
opathy
DBP
Neuropathy
HDL
Nepropathy
Cholestrol
Liver_
disease
Creatinine
BMI
Hypert-
ension

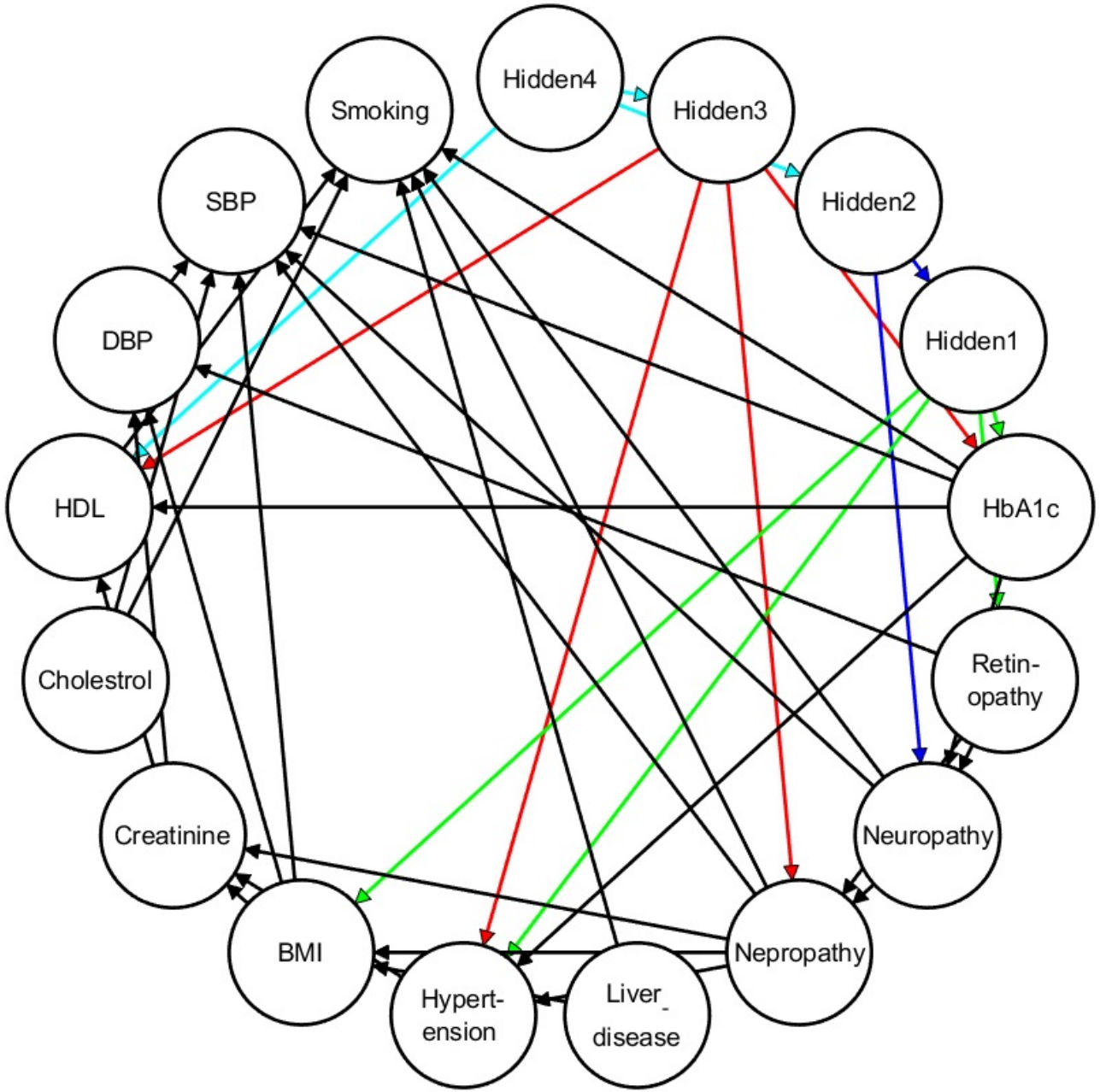
Step 2


Hidden2
SMK
Hidden1
SBP
HbA1c
DBP
Retin-
opathy
HDL
Neuropathy
Cholestrol
Nepropathy
Creatinine
Liver_
disease
BMI
Hypert-
ension

Step 3


SMK
Hidden3
Hidden2
SBP
Hidden1
DBP
HbA1c
HDL
Retin-
opathy
Cholestrol
Neuropathy
Creatinine
Nepropathy
BMI
Hypert-
ension
Liver_
disease

Step 4


Smoking
Hidden4
Hidden3
SBP
Hidden2
DBP
Hidden1
HDL
HbA1c
Retin-
opathy
Cholestrol
Creatinine
Neuropathy
BMI
Nepropathy
Hypert-
ension
Liver_
disease

Hierarchical Clustering and discovering Phenotypes

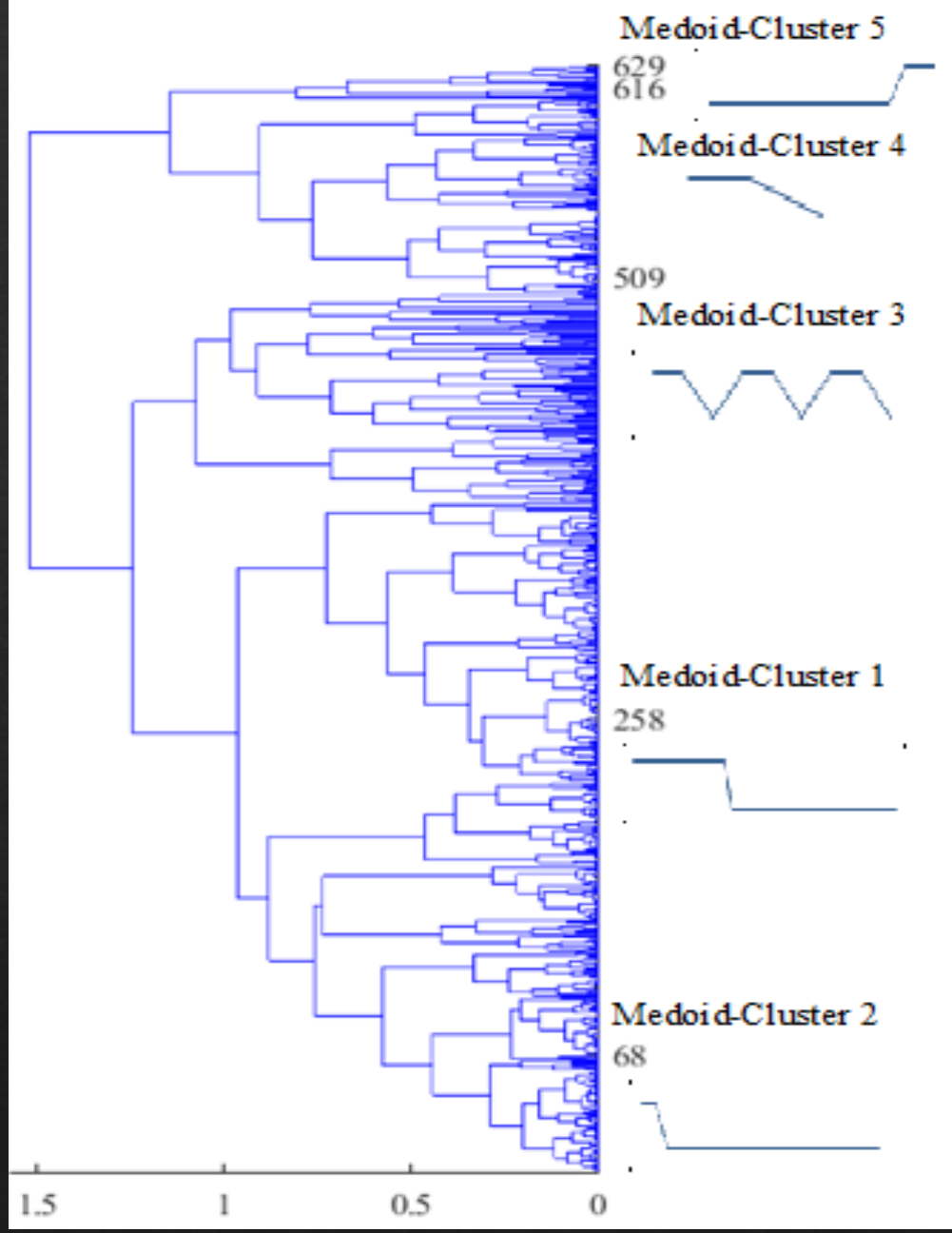
C1	Temporal Phenotype For Hidden variable pattern 1
Cluster1	
Cluster2	
Cluster3	
Cluster4	
Cluster5	

C2	Temporal Phenotype For Hidden variable pattern 2
Cluster1	
Cluster2	
Cluster3	
Cluster4	
Cluster5	

C3	Temporal Phenotype For Hidden variable pattern 3
Cluster1	
Cluster2	
Cluster3	
Cluster4	
Cluster5	

C4	Temporal Phenotype For Hidden variable pattern 4
Cluster1	
Cluster2	
Cluster3	
Cluster4	
Cluster5	

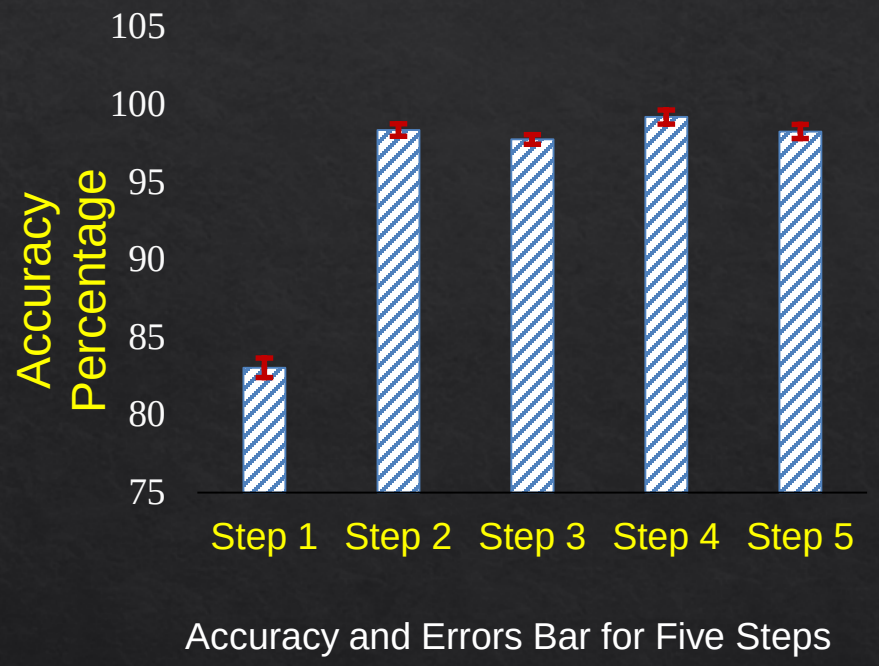
Using the Medoid hidden variable cluster profile for "deep temporal phenotype"



Dendogram of complete linkage cluster analysis and Temporal phenotypes (The First Hidden Clusters "Profiles"-C1)

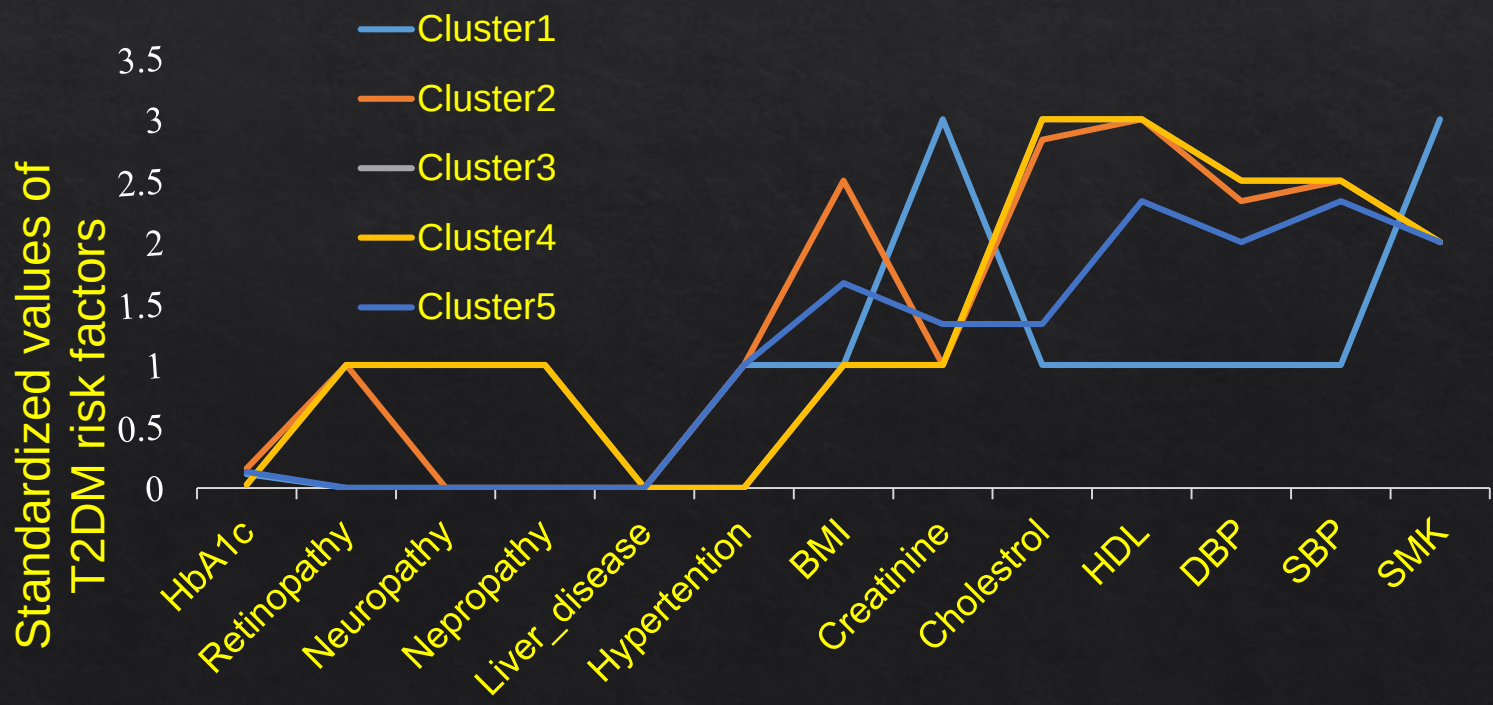
Leila.Yousefi@brunel.ac.uk

Results



Accuracy and Errors Bar for Five Steps

Node	Accuracy	Sensitivity	Specificity	Precision
No Hidden variable	0.40	0.50	0.40	0.40
Stepwise (Step1)	0.60	0.40	0.80	0.70
Enhanced Stepwise(Step1)	0.82	0.30	1.00	0.99
Stepwise (Step2)	0.80	1.00	0.60	0.70
Enhanced Stepwise (Step2)	0.97	0.82	0.98	0.88
Stepwise (Step3)	0.80	1.00	0.60	0.70
Enhanced Stepwise (Step3)	0.97	0.83	0.98	0.84
Enhanced Stepwise (Step4)	0.97	0.83	0.99	0.94
Enhanced Stepwise (Step5)	0.97	0.84	0.99	0.87



Mean values of T2DM risk factors and complications clusters based on the Fourth Hidden variable (C4).

Leila.Yousefi@brunel.ac.uk

Conclusion and Future Works

- Effectively integrates Bayesian methods with latent variables

by adapting the prior probability of the event occurrence for future time points

- The proposed method is more accurate than using one of hidden variable step
- Avoiding overfitting in the structure learning, using a stronger stopping rule in the step-wise approach
- Interpreting the impact of hidden variables in finding temporal phenotypes in the presence of unmeasured diabetic disorders
- Exploiting mutual information metrics (Ebert, 2007) to filter some of the hidden variable relationships
- Discovering interesting dependencies between the latent variable and the observed variables
- Applying the methodology on a different data
- Exploring Temporal Abstraction and multi-label classification

Leila.Yousefi@brunel.ac.uk

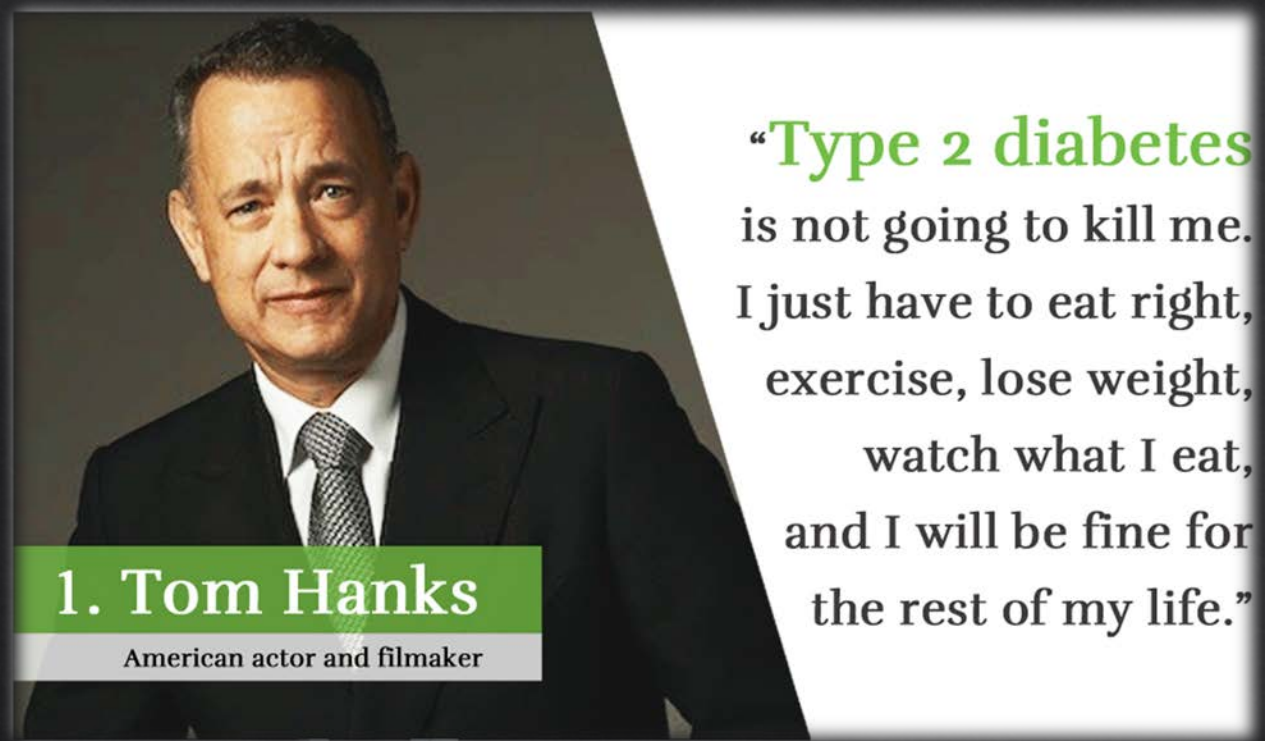
References

- P. Spirtes, C. N. Glymour, and R. Scheines, Causation, prediction, and search. MIT press, 2000.
- I. Ebert-Uphoff, “Measuring connection strengths and link strengths in discrete bayesian networks,” Georgia Institute of Technology, Tech. Rep., 2007.
- J. Pearl, “Probabilistic reasoning in intelligent systems. 1988,”San Mateo, CA: Kaufmann, vol. 23, pp. 33–34.
- L. Yousefi, L. Saachi, R. Bellazzi, L. Chiovato, and A. Tucker, “Predicting comorbidities using resampling and dynamic Bayesian networks with latent variables,” in 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), June 2017, pp. 205–206.
- L. Yousefi, L. Saachi, R. Bellazzi, L. Chiovato and A. Tucker “Predicting Disease Complications Using a Step-Wise Hidden Variable Approach for Learning Dynamic Bayesian Networks” in 2018 IEEE 31th International Symposium on Computer-Based Medical Systems (CBMS), June 2018

Leila.Yousefi@brunel.ac.uk

Thank you for listening!

Any Question?



Leila.Yousefi@brunel.ac.uk