IISE ANNUAL CONFERENCE & EXPO 2022

Type 2 Diabetes Risk Scoring via Bayesian Neural Networks

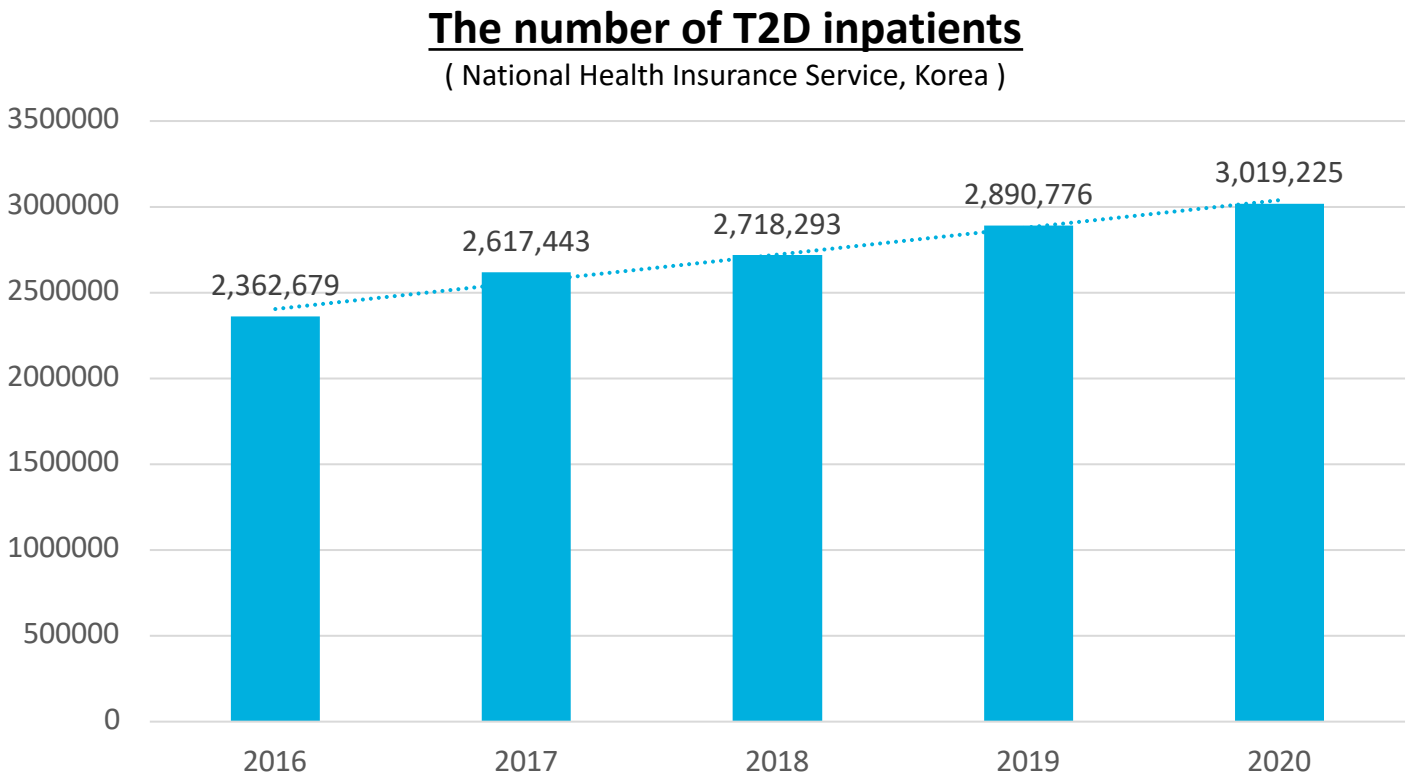
Hyewon Cho, Sujin Jeon, Sunghoon Lim
Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST)
hyewon@unist.ac.kr

WWW.IISE.ORG/ANNUAL

#IISEANNUAL2022

INDEX

Type-2 diabetes (T2D) mellitus in Korea : One of the main chronic diseases, with fast increasing prevalence



- Within last 5 years, the number of inpatients, diagnosed with ***Type 2 Diabetes has increased approximately 28%***
- ***The risk of cardiovasculr disease increases*** with rising number of T2D patients
- T2D ranked ***the 6th highest*** cause of death in Korea, followed by cancer[1)], cardiovascular[2)], pneumonia[3)], brain-cerebovascular disaese[4)], and suicidal death[5)].

Source : National Health Insurace Sharing Service (2021), Korean Statistical Information of Service (2021)

Usually, T2D tends to bring out other complications or can be developed from underlying diseases

Type 2 Diabetes Mellitus

Macro-vascular diseases

- **Cardiovascular diseases** : Heart attack,
- **Cerebrovascular diseases** : stroke, …

Micro-vascular diseases

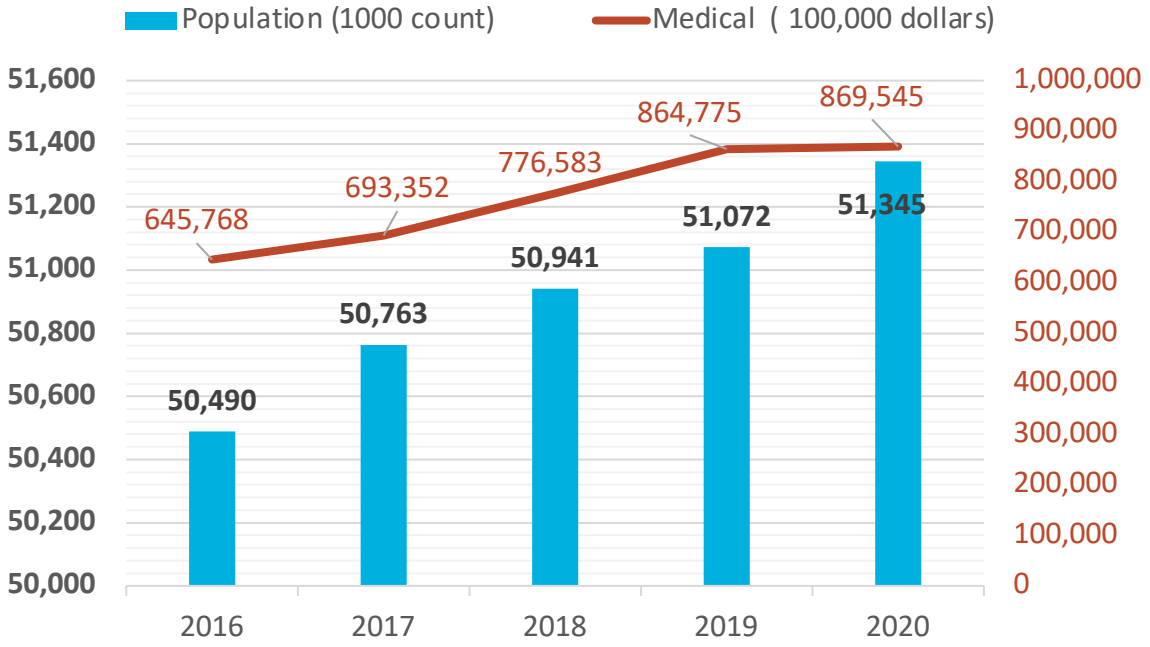
- **Opthamological diseases** : Glaucoma, cataracts, …
- **Neuropathy** : Diabetic foot, peripheral, …
- **Nephropathy**

Hence, understanding T2D disease landscape along with
landscape of its complications is essential for analyzing and studying T2D risk factors

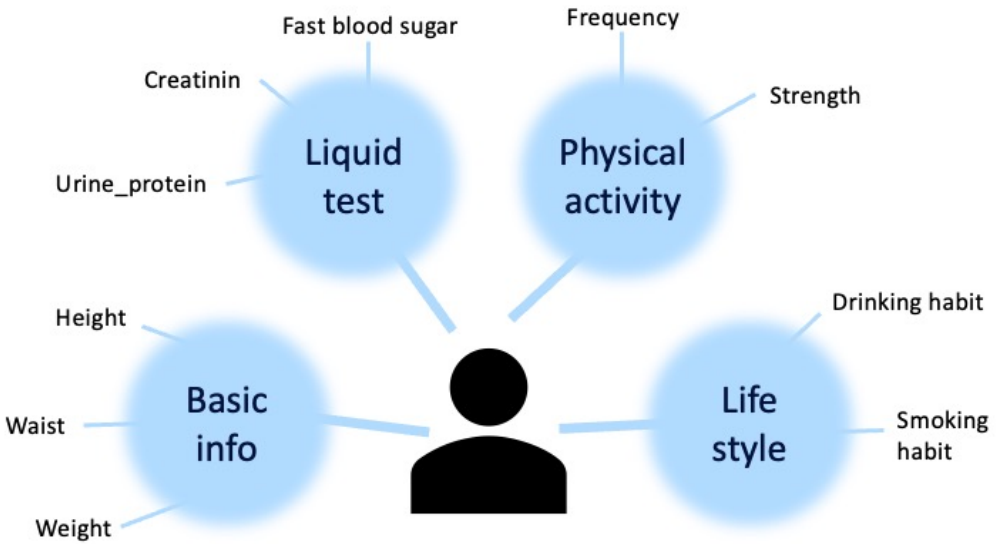
Source : Samsung hospital (2022)

Over 95% of Korean people have national health insurance, and go through health examination annually

NHISS* statistics



NHISS supported health examination



Source : National Health Insurace Sharing Service (2021), Korean Statistical Information of Service (2021)

Electric Health Records (EHR) database approach: Diagnosis information and health examination information

< Data Overview>

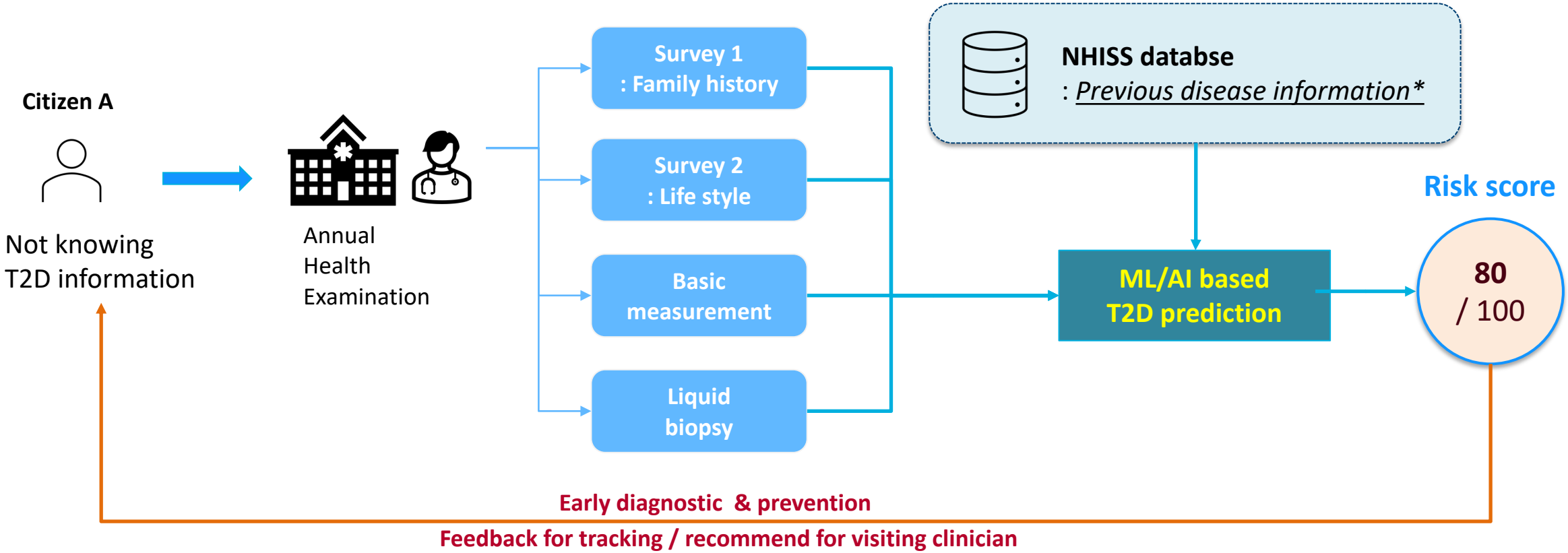
1. Diagnosis information

KEY	Date	Main Disease	Sub Disease
A	2009	E10	E10.1
A	2010	D10	D20
A	2014	…	…
A	2015		
B	2011		
B	2014		
B	2015		
C	2009	…	…
C	2015	C18	C18.3

2. Health checkup information

KEY	Date	Basic	History	Family histoty	Life style	result
A	2011	Age,..	Prev, …	Mom/dad, …		measured
A	2012					
A	2013					
A	2015					
B	2010					
B	2014					
B	2015					
C	2009					
C	2015					

Therefore, exploring T2D landscape via health examination along with complications can bring out social health improvement



Conventional studies of risk scoring with EHR

Chronic diseases

Disease	Title	Published	Data	Note
T2DM	Early detection of type 2 diabetes mellitus using machine learning-based prediction models	2020	EHR(Electronic health record) collected from 10 institutions	Stacked Deoising Autoencoder, Boosting (Adaboost, RF)
Hypertension	Predicting hypertension using machine learning: Findings from Qatar Biobank Study	2020	Qatar Biobank study data	Randon Forest, 5 fold cross validation

Cancer

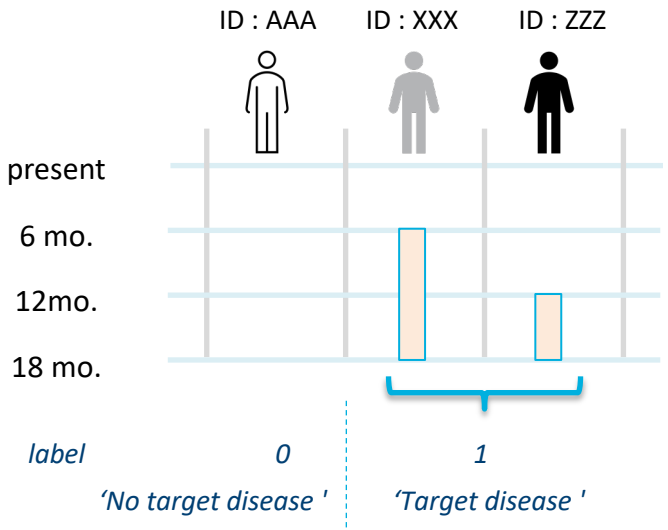
Disease	Title	Published	Data	Note
Breast Cancer	Predicting factors for survival of breast cancer patients u sing machine learning techniques	2019	Hospital based dataset from Malay Med Cent (8066)	-Decision tree (CART, Random Forset) - MLP based ANN - SVM
Gastric Cancer	Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning	2020	Chinese PLA hospital data w/ stained cell imaging	- CNN

Pandemic

Disease	Title	Published	Data	Note
COVID-19	A Bayesian machine learning approach for spatio-temporal prediction of COVID-19 cases	2022	Aggregation of 245 healthzones in community of Spain	Bayesian with Graphical, LSTM model

T2D risk scoring modeling summary

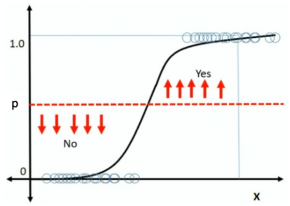
(1) Patient data preprocess - diagnostic data & checkup data



"Once disease diagnosed, after-all data are consisdered as disease"

(2) T2D risk logistic reg. model without complications

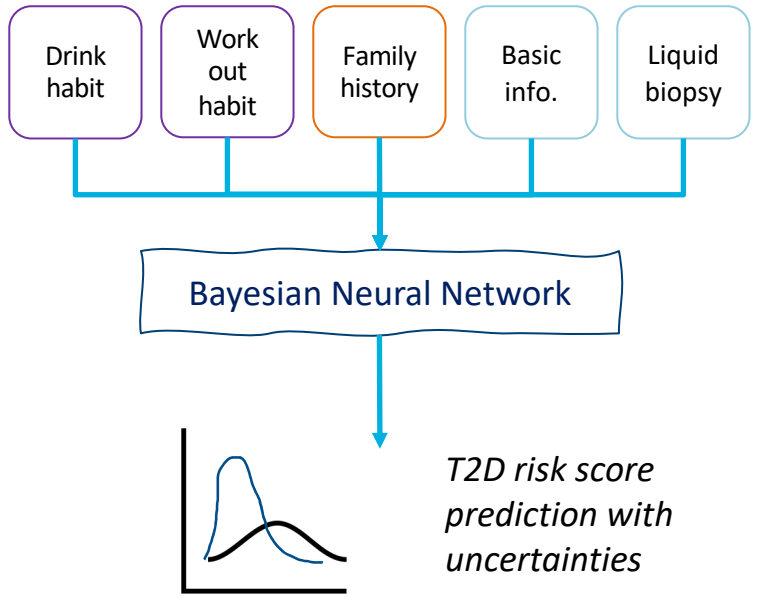
Pt ID	label	FBS	CRTN	BMI
AAA	0	100	0.3	19
XXX	0	90	0.7	21
XXX	1	125	0.8	24
ZZZ	1	120	0.9	27



T2D risk score prediction

"Apply logistic regression model for basic risk score analysis upon tabular data"

(3) BNN based T2D risk scoring with complications



T2D risk score prediction with uncertainties

(1) Prepare dataset: label including 12-month ahead health examination

A. Merge dataset from NHISS : there are 2 database which need to be merged _ diagnosis and health checkup information

1. Diagnosis information

KEY	Date	Main Disease	Sub Disease
A	2009	E10	E10.1
A	2010	D10	D20
A	2014	…	…
A	2015		
B	2011		
B	2014		
B	2015		
C	2009	…	…
C	2015	C18	C18.3

2. Health checkup information

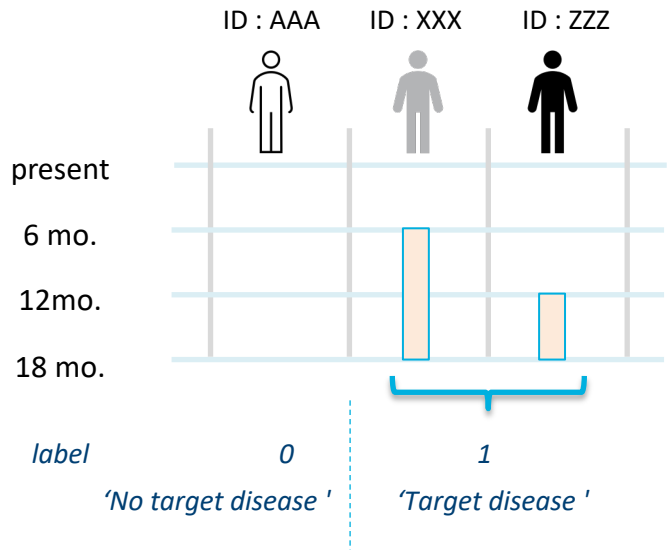
KEY	Date	Basic	History	Family histoty	Life style	result
A	2011	Age,..	Prev, …	Mom/dad, …		measured
A	2012					
A	2013					
A	2015					
B	2010					
B	2014					
B	2015					
C	2009					
C	2015					

➢ **Merge Data**

Pt ID	label	FBS	CRTN	BMI
AAA	0	100	0.3	19
XXX	0	90	0.7	21
XXX	1	125	0.8	24
ZZZ	1	120	0.9	27

(1) Prepare dataset: label including 12-month ahead health examination

A. Merge dataset from NHISS : there are 2 database which need to be merged _ diagnosis and health checkup information



- National health examination is conducted ***annually***
- Once the person gets ***diagnosed T2D 12 months after the test***, the person is ***considered to have diabetes.***
- If T2D occurred to one person, we consider that ***sample has the disease afterwards,*** for all the time.

➢ **Merge Data**

Pt ID	label	FBS	CRTN	BMI
AAA	0	100	0.3	19
XXX	0	90	0.7	21
XXX	1	125	0.8	24
ZZZ	1	120	0.9	27

(1) Prepare dataset: label including 12-month ahead health examination

B. Missing data imputation : there are missing values; the more missing values, the less accurate the model learns

Features*	Missing ratio
WSTC (waist)	0.02 %
BMI	0.03 %
HGB	0.01 %
FBS	0.01 %
TG	0.01 %
Total Cholesterol	0.01 %
GFR	28.1 %
LDL	0.69 %
HDL	0.01 %

* Candidate of features (filtered later)

- **Impuation methods for missing value**

Method	MICE	Linear regression	Ridge	Lasso	Elastic Net
RMSE	0.0265	0.0197	0.0197	0.0200	0.0199

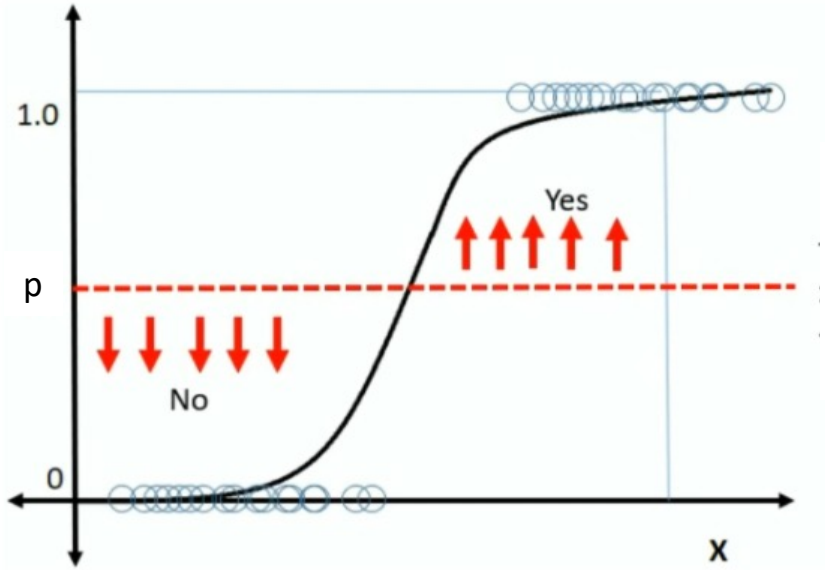
Method	GBM	XGB	Random Forest	Denosing AE
RMSE	0.0261	0.0195	0.0199	0.011

** Autoencoder

(2) Disease-risk score : logistic regression-based approach

✓ What is **Logistic regression** (multi-variate)

$$\ln\frac{p}{1-p} = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \ldots + \beta_n * x_n$$



Threshold value of 'p'%' probability

- p% success over threshold t
- (100-p)% failure under threshold t

(2) Disease-risk score : logistic regression-based approach

✓ What is **Logistic regression** (multi-variate)

$$\ln \frac{p}{1-p} = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \ldots + \beta_n * x_n$$

✓ **Odd Ratio** : Compared to reference group, how experimental features are different

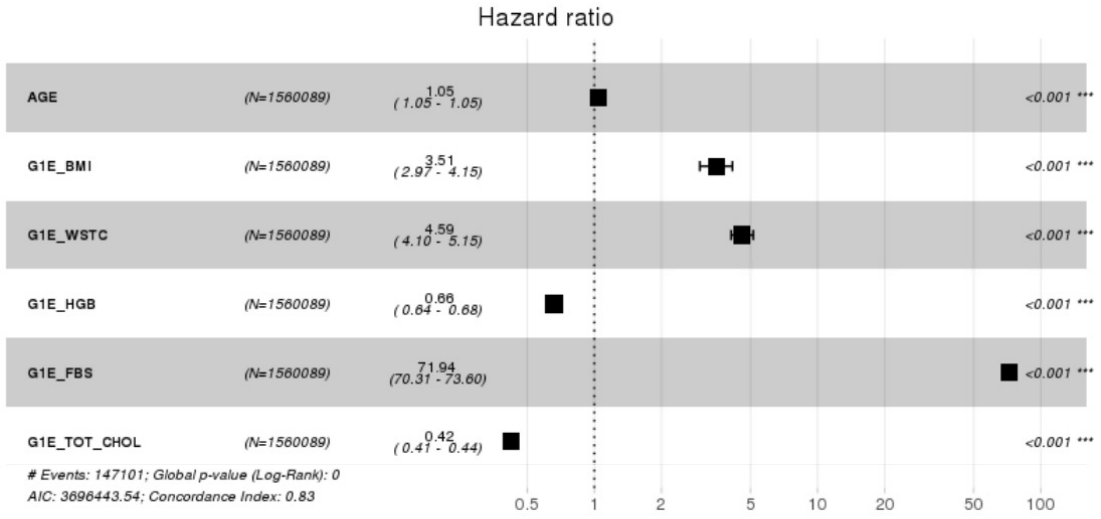
$$Odd\ ratio = \frac{experimental}{reference} = \frac{\frac{p'}{1-p'}}{\frac{p}{1-p}} = \exp(\beta_1)$$

- This allows for analyzing each risk factors.
- Relative risk scoring per features / control groups
- Useful for healthcare data analysis

(2) Disease-risk score : logistic regression-based approach

✓ Result of logistic regression (1) : Reliability on features

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -8.6183743  0.0420381 -205.013  < 2e-16 ***
SEX         -0.0815289  0.0104250   -7.820 5.26e-15 ***
G1E_WSTC     2.8005984  0.1063096   26.344  < 2e-16 ***
G1E_BMI      1.3978214  0.1495281    9.348  < 2e-16 ***
G1E_HGB     -0.5525061  0.0326036  -16.946  < 2e-16 ***
G1E_FBS      8.6436891  0.0321555  268.809  < 2e-16 ***
G1E_TOT_CHOL -2.1096979  0.0255285  -82.641  < 2e-16 ***
G1E_TG       0.7152311  0.0420040   17.028  < 2e-16 ***
AGE          0.0616911  0.0003573  172.675  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Features selected showed p-value under 0.001 which implies that the features are related to the disease prediction with high confidence.

Also, FBS (Fast blood sugar) feature is the most relative factor, followed by WSTC (waist length), BMI and CRTN

(2) Disease-risk score : logistic regression-based approach

✓ Result of logistic regression (2) : Mean and standard deviation distribution of risk score in age groups

Ref. age	20's	30's	40's	50's	60's	70's
Mean	0.008	0.02	0.05	0.11	0.20	0.30
std	0.02	0.05	0.10	0.16	0.21	0.24

small → BIG

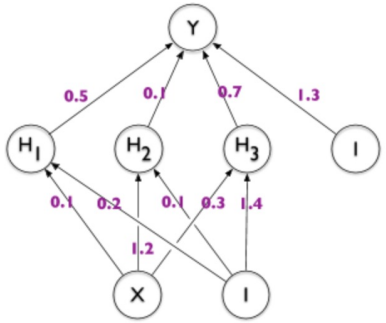
- ***The older reference group is, the higher risk score*** distributes according to increasing mean value
- ***The older reference group is, the wider risk score*** distributes according to increasing standard deviation value

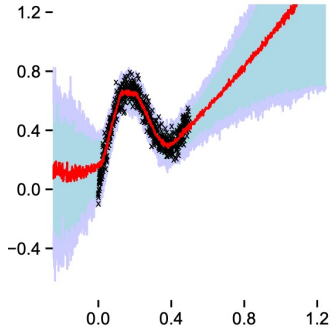
For the elders, overall, their health functions lower, which may accompany complications.
It may lead to the high variance of risk score's distribution, as shown in the elderly groups.

[Development plan] complications and distribution wise risk analysis

✓ What is a **Bayesian Neural Network (BNN)** ?

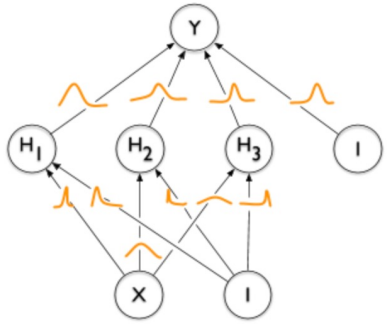
Deterministic Neural Network

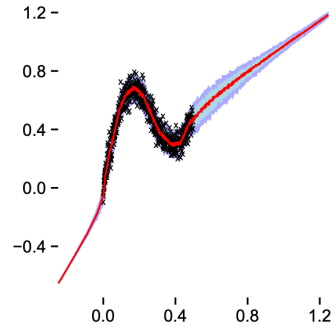




- ***A deterministic neural network*** works by ***maximizing the likelihood of the seen data*** using backpropagation (point-estimation)
- The model can be overfitted for 'observed data only', and ***easy to fail on 'generalization'***

Bayesian Neural Network





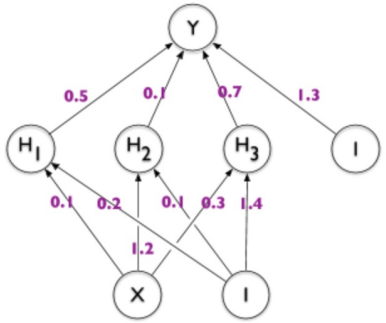
- ***A Bayesian neural network*** uses ***'Bayes rule'*** with seen data to estimate ***a full posterior distribution*** of the parameters.
- ***NN learns 'the distribution' of parameters,*** unlike determinisited NN (distribution-estimation)
- ***Ratio based on each events can be estimated***, hence, it can provide some insight for understanding data such as healthcare industries.

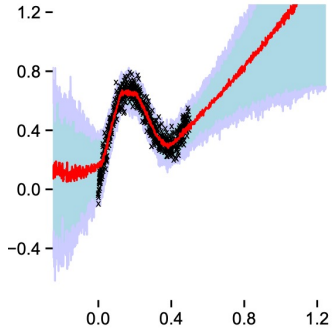
Source : Weight Uncertainty in Neural Networks(2015)

[Development plan] complications and distribution wise risk analysis

✓ What is **Bayesian Neural Network(BNN)** ?

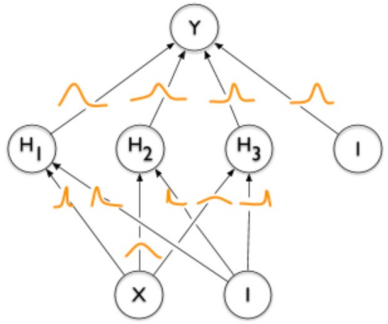
Deterministic Neural Network

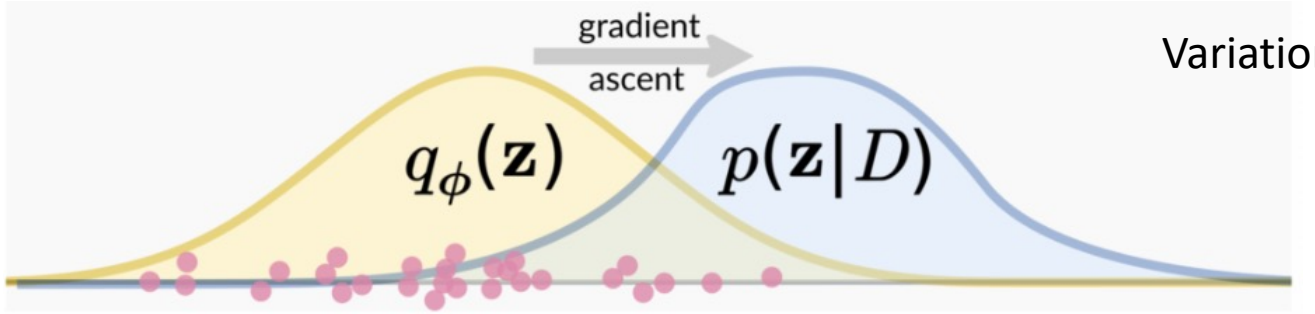




- ***A deterministic neural network*** works by ***maximizing the likelihood of the seen data*** using backpropagation (point-estimation)
- The model can be overfitted for 'observed data only', and ***easy to fail on 'generalization'***

Bayesian Neural Network



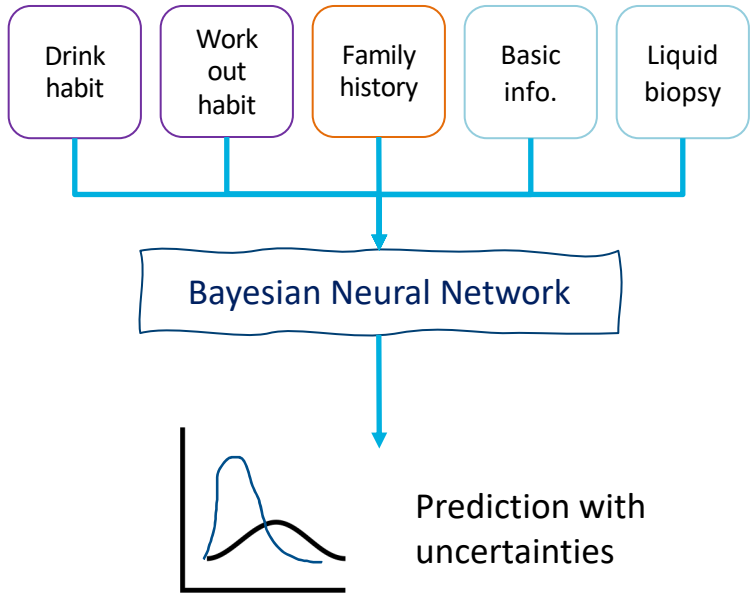


Variational Inference

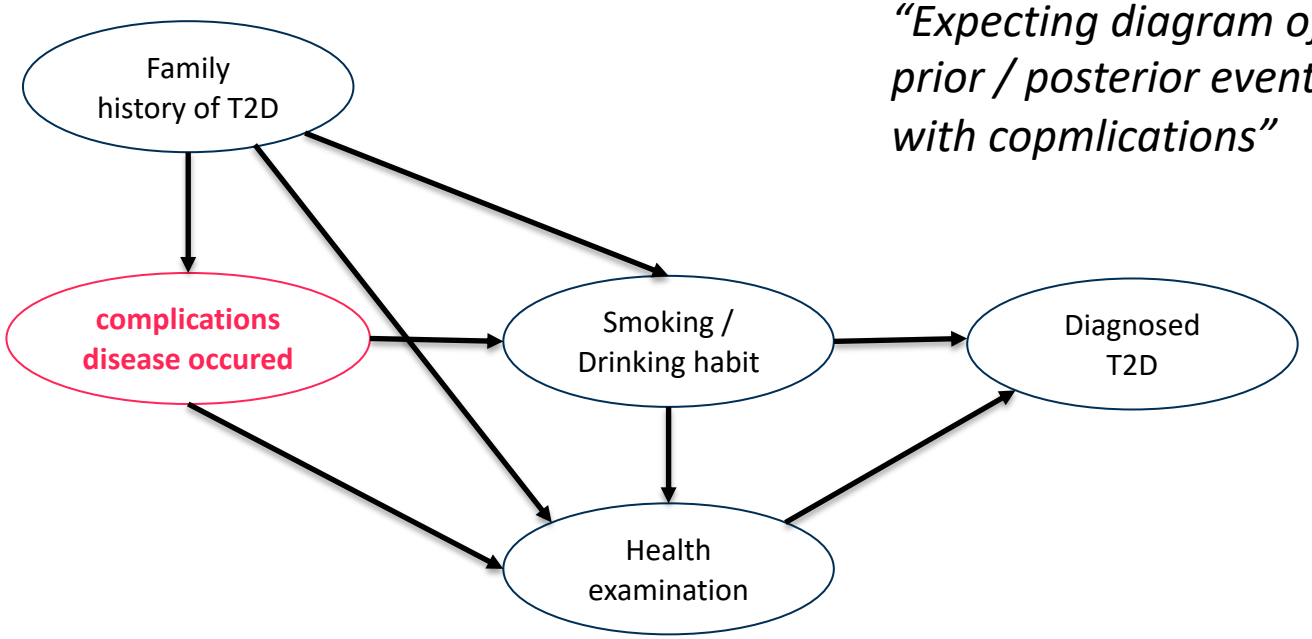
Source : Weight Uncertainty in Neural Networks(2015)

[Development plan] complications and distribution wise risk analysis

✓ Furure plan with BNN



"Expecting diagram of prior / posterior events with copmlications"



hyp **" complications considered T2D risk scoring can provide better estimation"**

Source : Weight Uncertainty in Neural Networks(2015)

IISE ANNUAL CONFERENCE & EXPO 2022

Q&A

hyewon@unist.ac.kr

Industrial Engineering || Ulsan National Institute of Science and Technology (UNIST)

Remember to complete your evaluation for this session within the app!

WWW.IISE.ORG/ANNUAL | #IISEANNUAL2022