

의학연구자를 위한 Machine Learning

분당차병원,
연구역량 강화를 위한 통계 교육

2021년 9월 8일(수)

17:30-18:30

서울성심병원 김지형

- 동기 부여 영상- 미리 보기
- Prediction of Neurological Outcomes in Out-of-hospital Cardiac Arrest Survivors Immediately after Return of Spontaneous Circulation: Ensemble Technique with Four Machine Learning Models
- <https://jkms.org/src/jkms-summary-AVF/jkms-36-e187-AVF.html> (10:55)

다양한 머신 러닝

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

한국형 당뇨 진단 모형

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

한국형 당뇨 예측 모형

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

Cardiovascular Events 예측 모형

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

유방암 치료 후 생존율 모형

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

위암 치료 후 생존 모델

위암 치료 후 생존 모델

다양한 예측 모형

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

예측모형들의 비교

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

모형의 활용

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

NRI의 단점

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

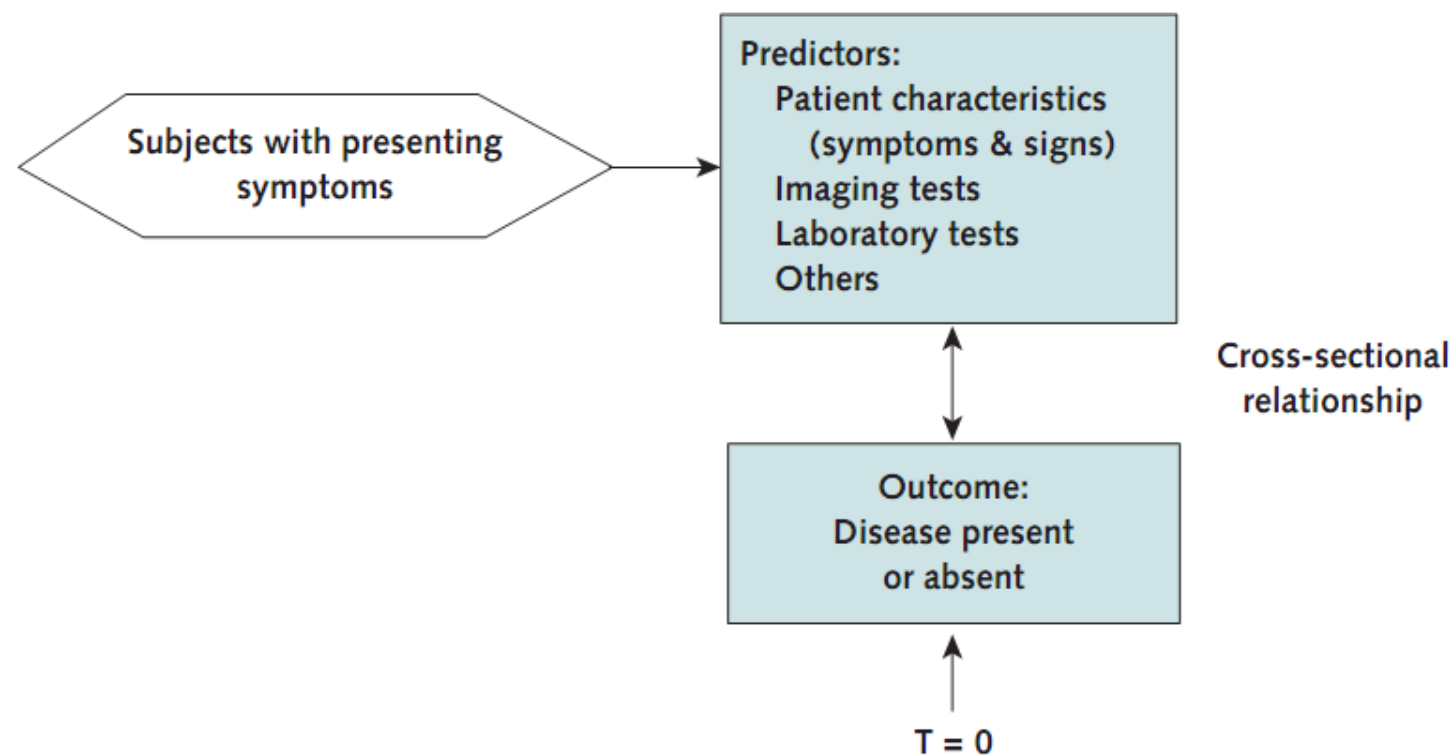
결론

In God we trust; all others must bring data (protocol and SAS output). — W. Deming

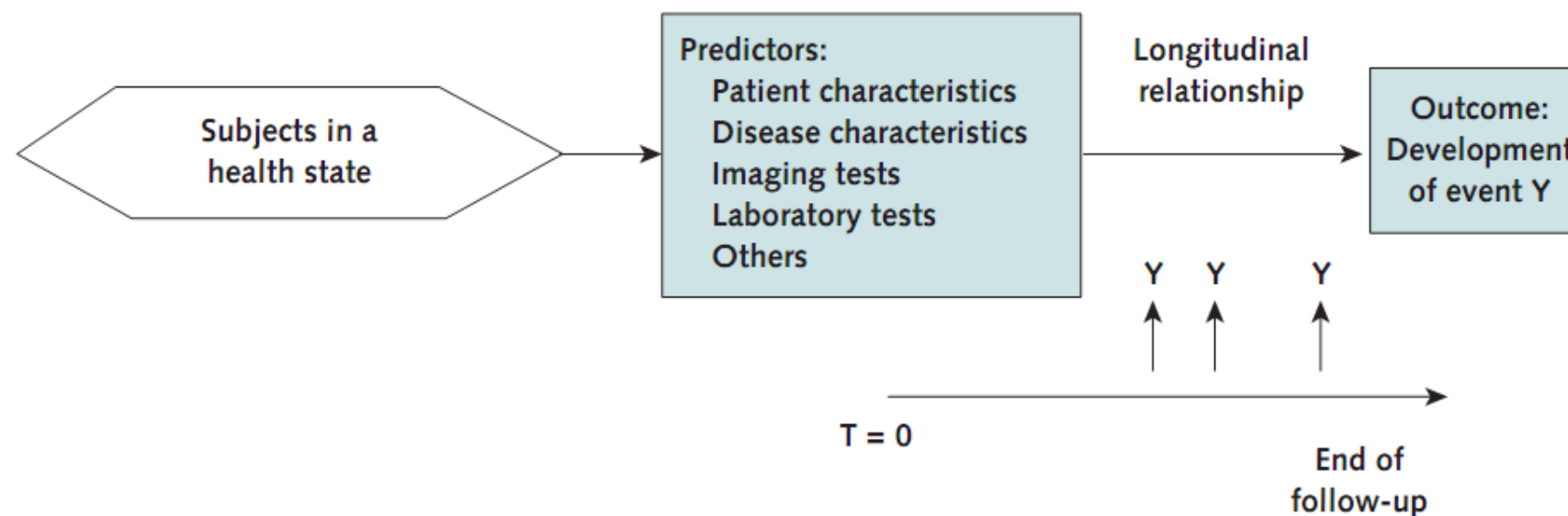
다양한 머신 러닝

• 진단 모형 : 예측 모형

Diagnostic multivariable modeling study

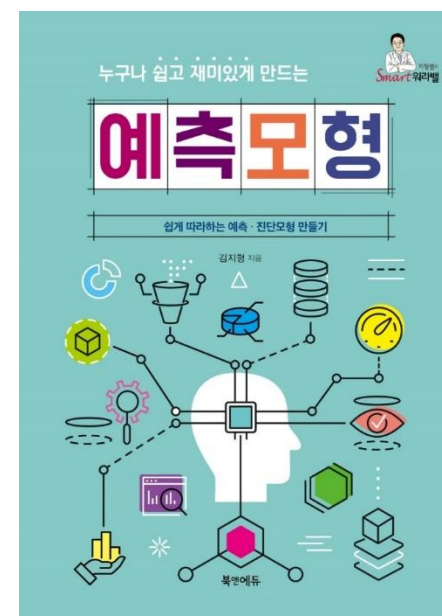


Prognostic multivariable modeling study



- 예측 모형 : 예측 도구(tools)
 - AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010
 - Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the United Kingdom. Pediatrics. 2006
 - Conflicts at the heart of the FRAX tool. CMAJ. 2014

- 예측(진단) 모형 : 머신 러닝
 - 의학에서 : 공학에서
 - 진단도구 예측도구(2014),
예측 모형('19) : 머신 러닝('20)
 - Logistic regression : 기타



- Predictors → outcome
 - test results
 - Risk indicators
 - prognostic factors
 - Covariates
 - independent variables
(more statistically)

- Predictors?

- 머리카락이 길이를 보니 여자
임에 틀림없어
- 사람들이 우산을 들고 다니는
걸 보니 비가 올 건가 봐
- 우산을 쓰고 다니는 걸 보니 비
가 오나 봐

- Predictors?
 - 감염을 진단하기 위한 CRP
(감염에 의한 단백질의 증가)
 - 간염 측정하는 OT/PT
(간세포가 죽어서 세포내 효소)

- 예측 모형/머신 러닝
 - 일상에서 늘 접하는 문제
 - 남자/여자, 어른/아이
 - 스팸메일 or not
 - 구매할까 말까
 - 질병 진단 등

- Train data
= Learning data
= Development data
- Test data $\doteq$ Validation data
- Internal Validation data
 $\longleftrightarrow$ External Validation data



Iris Setosa



Iris Versicolor



Iris Virginica

<https://tinyurl.com/Machine-Learning-EZ>

“다양한 머신러닝의 방법들”

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	outcome
5.10	3.50	1.40	0.20	setosa
4.90	3.00	1.40	0.20	setosa
4.70	3.20	1.30	0.20	setosa
4.60	3.10	1.50	0.20	setosa
5.00	3.60	1.40	0.20	setosa
5.40	3.90	1.70	0.40	setosa
4.60	3.40	1.40	0.30	setosa
5.00	3.40	1.50	0.20	setosa
4.40	2.90	1.40	0.20	setosa
4.90	3.10	1.50	0.10	setosa
5.40	3.70	1.50	0.20	setosa
4.80	3.40	1.60	0.20	setosa
4.80	3.00	1.40	0.10	setosa
4.30	3.00	1.10	0.10	setosa
5.80	4.00	1.20	0.20	setosa
5.70	4.40	1.50	0.40	setosa

- 예측 모형/머신러닝은 다양하다
- 예측 모형/머신러닝은 일상적이다
- 예측 모형/머신러닝은 쉽다
- 그러나.....,

한국형 당뇨 진단 모형

- 한국형 당뇨 진단 모형
 - A simple screening score for diabetes for the Korean population: development, validation, and comparison with other scores.
 - 2001- 2005년
 - 당뇨로 알려지지 않은 20세 이상 9602명

• 한국형 당뇨 진단 모형

The screenshot shows a web browser window with the URL www.cdc.gov/nchs/tutorials/nhanes/Preparing/Download/intro_iii.htm. The page header includes the CDC logo and the text "Centers for Disease Control and Prevention" and "CDC 24/7: Saving Lives. Protecting People.™". A navigation bar contains an "A-Z Index" and letters from A to Z. The main heading is "Download Data Files". On the left, a sidebar menu lists the tutorial steps: "NHANES III Web Tutorial", "Using the Tutorial", "Survey Orientation", "Preparing an Analytic Dataset", "4. Locate Variables", "5. Download Data" (highlighted with a green arrow), "6. Merge & Append Datasets", "7. Review Data & Create New Variables", and "Survey Design Factors". The main content area shows a breadcrumb trail: "NCHS Home > Tutorials > NHANES III Web Tutorial > Preparing an Analytic Dataset". Below this are social media buttons for "Recommend", "Tweet", and "Share". A row of tabs indicates the current selection: "Continuous", "NHANES III" (highlighted in yellow), "NHANES II", and "NHANES I". The "Purpose" section states: "NHANES III data files are available for download from the website as data (.DAT) files along with the SAS code which is provided for your use to create a SAS dataset, and the documentation explaining how the data are arranged in the data files. To use these files you need to [create a directory](#) to save them, [download the data files, SAS code, and documentation](#), and then [run the SAS code to create the SAS datasets](#)."

Task 1: Create a directory for NHANES III files

- 한국형 당뇨 진단 모형
 - 9602명 중
 - 218명 (자신도 몰랐던)
당뇨 진단

- 키, 체중, 성별, 가족력, 흡연, 음주, 고혈압, 허리 둘레 등등

➔ 당뇨

id	키	체중	성별	가족력	흡연	음주	고혈압	허리 둘레		당뇨
홍길동										0
손오공										1
이성계										1
김유신										0

Table 2—Logistic regression analyses for related factors for undiagnosed diabetes

Variables	β -Coefficient	Odds ratio (95% CI)	P	Score assigned
Intercept	−5.608			
Age (years)				
<35		Reference		0
35–44	1.068	2.91 (1.74–4.88)	<0.0001	2
≥45	1.305	3.69 (2.23–6.11)	<0.0001	3
Family history of diabetes				
No		Reference		0
Yes	0.621	1.86 (1.29–2.68)	0.0008	1
Hypertension				
No		Reference		0
Yes	0.417	1.52 (1.17–1.97)	0.0018	1
Waist circumference (male/female) (cm)				
<84/77		Reference		0
84–89.9/77–83.9	0.779	2.18 (1.47–3.24)	0.0001	2
≥90/84	1.161	3.19 (2.20–4.64)	<0.0001	3
Smoking status				
Never or ex-smoker		Reference		0
Current smoker	0.386	1.47 (1.08–2.01)	0.0155	1
Alcohol intake (drinks/day)				
<1		Reference		0
1–4.9	0.493	1.64 (1.16–2.32)	0.0055	1
≥5	0.795	2.21 (1.42–3.45)	0.0004	2

AUC = 0.730; maximal score is 11. Model was fitted in the development dataset, KNHANES 2001 and 2005.

- 한국형 당뇨 진단 모형

$$k = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \cdots + \beta_n \times X_n$$

$$p = \frac{e^k}{1 + e^k}$$

<http://me2.do/Fi6rzUKu>

• 한국형 당뇨 진단 모형

k

D12	f_x	=EXP(E12)/(1+EXP(E12))	2	
A	B	C	D	E
2				
3		Intercept	1	-5.608
4		age	44	1.068
5		Family history of diabetes	no	0
6		Hypertension	yes	0.417
7		Waist circumference (cm): male	89	0.779
8		Waist circumference (cm): female	0	0
9		smoking	no	0
10		Alcohol(daily)	0	0
11				
12			0.0340922	1 -3.344
13		probability of DM	3	3.41%

X_n

Variables	β -Coefficient	Odds ratio (95% CI)	P	Score assigned
Intercept	-5.608			
Age (years)				
<35		Reference		0
35-44	1.068	2.91 (1.74-4.88)	<0.0001	2
≥ 45	1.305	3.69 (2.23-6.11)	<0.0001	3
Family history of diabetes				
No		Reference		0
Yes	0.621	1.86 (1.29-2.68)	0.0008	1
Hypertension				
No		Reference		0
Yes	0.417	1.52 (1.17-1.97)	0.0018	1
Waist circumference (male/female) (cm)				
<84/77		Reference		0
84-89.9/77-83.9	0.779	2.18 (1.47-3.24)	0.0001	2
$\geq 90/84$	1.161	3.19 (2.20-4.64)	<0.0001	3
Smoking status				
Never or ex-smoker		Reference		0
Current smoker	0.386	1.47 (1.08-2.01)	0.0155	1
Alcohol intake (drinks/day)				
<1		Reference		0
1-4.9	0.493	1.64 (1.16-2.32)	0.0055	1
≥ 5	0.795	2.21 (1.42-3.45)	0.0004	2

- 한국형 당뇨 진단 모형
 - We created a weighted scoring system by **rounding down** odds ratios (ORs) in the final model to the nearest integer.
For example, OR 1.52 was rounded to 1 and OR 3.19 was rounded to 3.

Question	Answer (Score)	Enter your Score (Enter 0 if you don't know)								
1. Your age group?	< 35 y (0 point) 35–44 y (2 points) ≥ 45 y (3 points)									
2. Have either of your parents or siblings been diagnosed with diabetes?	No (0 point) Yes (1 point)									
3. Are you currently taking medication for hypertension or do you have hypertension (i.e.g, blood pressures greater than 140/90 mmHg)?	No (0 point) Yes (1 point)									
4. What is your waist circumference (taken below the ribs, usually at the level of the navel) ?	<table border="0"> <tr> <td>Men</td> <td>Women</td> </tr> <tr> <td>< 84 cm (33 inch) (0 point)</td> <td>< 77 cm (30 inch) (0 point)</td> </tr> <tr> <td>84–89.9 cm (33–34.9 inch) (2 points)</td> <td>77–83.9 cm (30–32.9 inch) (2 points)</td> </tr> <tr> <td>≥ 90 cm (35 inch) (3 points)</td> <td>≥ 84 cm (33 inch) (3 points)</td> </tr> </table>	Men	Women	< 84 cm (33 inch) (0 point)	< 77 cm (30 inch) (0 point)	84–89.9 cm (33–34.9 inch) (2 points)	77–83.9 cm (30–32.9 inch) (2 points)	≥ 90 cm (35 inch) (3 points)	≥ 84 cm (33 inch) (3 points)	
Men	Women									
< 84 cm (33 inch) (0 point)	< 77 cm (30 inch) (0 point)									
84–89.9 cm (33–34.9 inch) (2 points)	77–83.9 cm (30–32.9 inch) (2 points)									
≥ 90 cm (35 inch) (3 points)	≥ 84 cm (33 inch) (3 points)									
5. Do you currently smoke cigarettes on a daily basis?	Never or Ex-smoker (0 point) Current smoker (1 point)									
6. How much alcohol do you drink on a daily basis? (regardless of types of alcohols)	Never or less than 1 drink / day (0 point) 1–4.9 drinks /day (1 point) ≥ 5 drinks /day (2 points)									
TOTAL SCORE (add points from questions 1–6)										
If the TOTAL SCORE is ≥5, you are at high risk for diabetes, so see your doctor for a blood test.										

Your total score from questionnaire	Probability of having diabetes now
≤ 4	2%
5~7	6%
8~9	12%
≥ 10	19%

Figure 1—Self-assessment screening questionnaire for undiagnosed participants, recommended for use by health care providers and laypersons. y, years.

• 설문지 4점 2%

age	44
Family history of diabetes	no
Hypertension	no
Waist circumference (cm): male	89
Waist circumference (cm): female	0
smoking	no
Alcohol(daily)	0

probability of DM	2.27%
-------------------	-------

• 설문지-7점 → 6%

age	46
Family history of diabetes	no
Hypertension	no
Waist circumference (cm): male	89
Waist circumference (cm): female	0
smoking	yes
Alcohol(daily)	4

probability of DM	6.63%
-------------------	-------

- 설문지-4점 → 2%

age	43
Family history of diabetes	no
Hypertension	no
Waist circumference (cm): male	0
Waist circumference (cm): female	66
smoking	no
Alcohol(daily)	0

probability of DM	1.06%
-------------------	-------

- 한국형 당뇨 진단 모형 (9602명)

	민감도	특이도
4점이상 기준	89%	41%
5점이상 기준	81%	54%
6점이상 기준	65%	67%

- 한국형 당뇨 진단 모형
 - Development, validation, and comparison with other scores

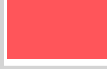
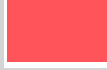
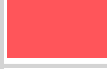
- 한국형 당뇨 진단 모형
 - Consistent results were observed when we applied this score to the KNHANES 2007-2008 validation dataset.
 - 2001- 2005 development dataset

Table 3—Performance of the new and existing diabetes screening methods in the development and validation datasets

Method, by dataset	High risk (%)	Sensitivity	Specificity	PPV	NPV	Positive likelihood ratio	Negative likelihood ratio	Youden index	AUC
Development dataset (KNHANES 2001 and 2005)									
≥4	60	89	41	5	99	1.52	0.27	30	0.730
≥5*	47	81	54	6	99	1.75	0.36	35	0.730
≥6	34	65	67	7	98	2.00	0.51	33	0.730
ADA questionnaire II†	21	41	79	7	97	2.01	0.74	21	0.604
U.S. screening score‡	14	33	86	8	97	2.44	0.77	20	0.685
Rotterdam model§	29	53	72	7	98	1.89	0.65	25	0.661
Qingdao diabetes risk score	36	62	65	6	98	1.77	0.59	27	0.693
Thai risk score¶	46	74	55	6	98	1.64	0.48	29	0.689
Validation dataset (KNHANES 2007–2008)									
After imputation ≥5**	48	80	53	4	99	1.68	0.39	32	0.729
Without imputation ≥5	46	79	55	4	99	1.73	0.39	33	0.729
Without imputation ≥4	59	89	41	4	99	1.52	0.27	30	0.729
ADA questionnaire II†	28	51	73	5	98	1.87	0.67	24	0.618
U.S. screening score‡	17	34	83	5	98	2.01	0.79	17	0.686
Rotterdam model§	33	55	68	4	98	1.71	0.67	23	0.651
Qingdao diabetes risk score	41	63	59	4	98	1.54	0.63	22	0.691
Thai risk score¶	48	79	53	4	99	1.66	0.40	31	0.722

*Best cut point; AUC of the receiver-operating curve: 0.730 (95% CI 0.720–0.739), $P = 0.0001$. †ADA diabetes questionnaire II: A function of age, waist, weight, height, gestational diabetes, parental and sibling diabetes history, high blood pressure, and exercise. See ref. 18 for a graphical presentation of this algorithm. Data on history of gestational diabetes were not collected, so this condition was omitted when assessing the ADA questionnaire II. ‡U.S. screening score: score_U.S. ≥5, in which score_U.S. = (40 years ≤ age < 50 years) × 1 + (50 years ≤ age < 60 years) × 2 + (60 years ≤ age) × 3 + (male) × 1 + (family history of diabetes) × 1 + (history of hypertension) × 1 + (overweight) × 1 + (obese) × 2 + (extremely obese) × 3 – (exercise) × 1 (from ref. 5). §Rotterdam model: score_Rotterdam >6, in which score_Rotterdam = 2 per 5-year increment from 55 year + (male) × 5 + (use of antihypertensive medications) × 4 + (BMI ≥30 kg/m²) × 5 (from ref. 19). ||Qingdao diabetes risk score: score_Qingdao ≥14, in which score_Qingdao = [male: (waist ≤ 2.3 chi) × 1 + (2.4 chi ≤ waist ≤ 2.6 chi) × 4 + (2.7 chi ≤ waist ≤ 2.9 chi) × 8 + (3.0 chi ≤ waist) × 12] + [female: (waist ≤ 2.0 chi) × 1 + (2.1 chi ≤ waist ≤ 2.3 chi) × 3 + (2.4 chi ≤ waist ≤ 2.6 chi) × 6 + (2.7 chi ≤ waist) × 9] + (age ≤ 35) × 1 + (36 years ≤ age ≤ 45 years) × 3 + (46 years ≤ age ≤ 55 years) × 6 + (56 years ≤ age < 65 years) × 9 + (65 years ≤ age) × 12 + (diabetes in parents or siblings) × 8. 1 chi ≈ 33 cm (from ref. 11). ¶Thai risk score: score_Thai ≥6, in which score_Thai = (45 years ≤ age ≤ 49 years) × 1 + (50 years ≤ age) × 2 + (male) × 2 + (23 kg/m² ≤ BMI < 27.5 kg/m²) × 3 + (27.5 kg/m² ≤ BMI) × 5 + (90 cm ≤ waist in male) × 2 + (80 cm ≤ waist in female) × 2 + (history of hypertension) × 2 + (family history of diabetes) × 4 (from ref. 9). **After imputing the missing data of family history of diabetes.

한국형 당뇨 진단 모형

							Positive	Negative		
	High risk						likelihood	likelihood	Youden	
	(%)	Sensitivity	Specificity	PPV	NPV		ratio	ratio	index	AUC
Development dataset										
(KNHANES 2001 and 2005)										
≥4	 60	 89	 41	5	99		 1.52	 0.27	 30	 0.73
≥5*	 47	 81	 54	6	99		 1.75	 0.36	 35	 0.73
≥6	 34	 65	 67	7	98		 2	 0.51	 33	 0.73
ADA questionnaire II	 21	 41	 79	7	97		 2.01	 0.74	 21	 0.604
U.S. screening score	 14	 33	 86	8	97		 2.44	 0.77	 20	 0.685
Rotterdam modelx	 29	 53	 72	7	98		 1.89	 0.65	 25	 0.661
Qingdao diabetes risk score	 36	 62	 65	6	98		 1.77	 0.59	 27	 0.693
Thai risk score	 46	 74	 55	6	98		 1.64	 0.48	 29	 0.689
Validation dataset (KNHANES 2007–2008)										
After imputation ≥5**	 48	 80	 53	4	99		 1.68	 0.39	 32	 0.729
Without imputation ≥5	 46	 79	 55	4	99		 1.73	 0.39	 33	 0.729
Without imputation ≥4	 59	 89	 41	4	99		 1.52	 0.27	 30	 0.729
ADA questionnaire II	 28	 51	 73	5	98		 1.87	 0.67	 24	 0.618
U.S. screening score	 17	 34	 83	5	98		 2.01	 0.79	 17	 0.686
Rotterdam modelx	 33	 55	 68	4	98		 1.71	 0.67	 23	 0.651
Qingdao diabetes risk score	 41	 63	 59	4	98		 1.54	 0.63	 22	 0.691
Thai risk score	 48	 79	 53	4	99		 1.66	 0.4	 31	 0.722

- 로지스틱 회귀분석을 이용해서 진단 모형을 만들 수 있다.
- 회귀식 또는 계산표를 이용해서 확률 계산
- Validation and Comparison