RESEARCH

Using optimal subset regression to identify factors associated with insulin resistance and construct predictive models in the US adult population

Rongpeng Gong[1,*], **Yuanyuan Liu**[1,*], **Gang Luo**[1], **Jiahui Yin**[2], **Zuomiao Xiao**[3] and **Tianyang Hu**[4]

[1]Medical College of Qinghai University, Xining, People's Republic of China
[2]College of Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Jinan, China
[3]Department of Clinical Laboratory, The Affiliated Ganzhou Hospital of Nanchang University, Ganzhou, China
[4]Precision Medicine Center, The Second Affiliated Hospital, Chongqing Medical University, Chongqing, China

Correspondence should be addressed to T Hu: hutianyang@stu.cqmu.edu.cn

*(R Gong and Y Liu contributed equally to this work)

Abstract

Background: In recent decades, with the development of the global economy and the improvement of living standards, insulin resistance (IR) has become a common phenomenon. Current studies have shown that IR varies between races. Therefore, it is necessary to develop individual prediction models for each country. The purpose of this study was to develop a predictive model of IR applicable to the US population.
Method: In total, 11 cycles of data from the NHANES database were selected for this study. Of these, participants from 1999 to 2010 (n = 14931) were used to establish the model, and participants from 2011 to 2020 (n = 13,646) were used to validate the model. Univariate and multivariable logistic regression was used to analyze the factors associated with IR. Optimal subset regression was used to filter the best modeling variables. ROC curves, calibration curves, and decision curve analysis were used to determine the strengths and weaknesses of the model.
Results: After screening the variables by optimal subset regression, variables with covariance were excluded, and a total of seven factors (including HDL, LDL, ALB, GLB, GLU, BMI, and waist) were finally included to establish the prediction model. The AUCs were 0.851 and 0.857 in the training and validation sets, respectively, and the Brier value of the calibration curve was 0.153.
Conclusion: The optimal subset predictive model proposed in this study has a great performance in predicting IR, and the decision curve analysis shows that it has a high net clinical benefit, which can help clinicians and epidemiologists easily detect IR and take appropriate interventions as early as possible.

Key Words
- insulin resistance
- optimal subset regression
- calibration curves
- NHANES
- HOMA-IR

Endocrine Connections (2022) **11**, **e220066**

Background

Insulin resistance (IR) is a systemic disorder of glucose metabolism that results in changes in multiple organs and insulin regulatory pathways (1, 2, 3). The clinical significance of IR has become clear over the past few decades, a disorder characterized by markedly elevated insulin levels (hyperinsulinemia) and decreased insulin

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-22-0066

function (4). Type 2 diabetes develops when an IR individual fails to secrete enough insulin to overcome the deficiency (5, 6). A large number of studies have shown that IR is the main pathophysiological factor in the development of the metabolic syndrome and cardiovascular diseases (7, 8, 9, 10). Meanwhile, people with IR have a higher risk of death. In an 18.9-year prospective cohort study in the United States, Pan *et al.* found a 26% increased risk of death in postmenopausal women with IR (11). A cohort study by Lee *et al.* involving 1687 US population found a 79% increase in all-cause mortality in patients with IR after a median follow-up of 4.5 years (12). Since IR could lead to a large number of adverse effects, early detection of IR and intervention will certainly reduce the number of people with IR in advance and reduce various risks including developing diabetes and death.

Currently, there are several methods for predicting IR directly or indirectly. Among them, the hyperinsulinemic-euglycemic clamp (HEC) test, originally developed by DeFronzo (13), is considered the gold standard method for diagnosing IR. However, this method is costly, invasive, and inconvenient and is often extremely time-consuming in clinical applications, making it unreasonable for large-scale use in the general population. Given this, the HOMA-IR index for evaluating IR models came into being (14). Unfortunately, the HOMA-IR index lacks a standardized method for measuring insulin, and currently, its clinical application is still limited. HOMA-IR is not routinely measured in clinical practice when used to diagnose IR. Furthermore, HOMA-IR is rarely measured in large-scale physical examinations and broad population surveys. Thus, from a clinical point of view, a simple, easy-to-follow, and inexpensive predictor to identify IR may effectively help clinicians and epidemiologists identify subjects with IR early.

At present, several predictive models for IR have been established. For example, in 2018, Boursier *et al.* used triglycerides and glycated hemoglobin to predict IR in an obese population (15). Yeh *et al.* proposed a predictive model for IR in elderly Taiwanese in 2019 using triglyceride/high-density lipoprotein (TG/HDL) that is readily available in the clinic (16). These models were clinically proven to be available. However, multiple studies have shown that the occurrence of IR is ethnic-specific (17, 18, 19, 20). Our research aims to collect clinically accessible indicators and screen out the most appropriate variables by optimal subset regression in the US adult population to establish an easy-to-use IR predictive model.

Method

Database

This study selected all cycle data collected since 1999 by the National Health and Nutrition Examination Surveys (NHANES) project of the US National Health Center, with a total of 11 cycles of 22 years (21). The NHANES project focuses on health examinations and healthy eating in the US, including comprehensive data on diet, nutritional status, and chronic diseases. Continuous data collection began in 1999, with new data released in 2-year cycles of approximately 10,000 participants per cycle, all selected by a complex multi-stage hierarchical probabilistic design with unique demographic weights applied, with a sample representative of the entire US population. Detailed data were divided into five categories: demographic informatics data, dietary data, body measurement data, laboratory data, and questionnaire data. All NHANES-based studies were approved by the National Health Statistics Research Ethics Review Board. Ethical approval, and more detailed information can be found on the website of the Ethics Review Board of the National Center for Health Statistics (https://www.cdc.gov/nchs/nhanes/irba98.htm) (22).

Study population and study design

This study initially enroled 116,876 participants who completed interviews and exams at Mobile Examination Centers (MECs). After rigorous screening, a total of 28,577 participants were included (Fig. 1). The exclusion criteria are as follows: (1) under the age of 18 (n = 47,979); (2) not participating in the detection of fasting blood glucose and insulin items (n = 39,465); and (3) taking anti-hyperglycemic agents (n = 855) (details of the names of drugs in Supplementary Table 1, see section on supplementary materials given at the end of this article).

The purpose of this study was to use the optimal subset regression to screen the best modeling factors to establish a predictive model of IR based on clinically readily available variables. The diagnostic criteria of IR refer to the HOMA-IR index recognized by international experts. The calculation formula of HOMA-IR is as follows: fasting plasma glucose level (FPG, mmol/L) × fasting insulin level (FINS, μU/mL)/22.5. In the US population, HOMA-IR ≥ 2.73 is positive for IR. FPG and FINS were measured by the University of Missouri-Columbia Diabetes Diagnostic Laboratory using the Primus CLC330. FPG was measured by the hexokinase method, and FINS was measured by the insulin RIA.

The population who participated in the NHANES project from 1999 to 2010 was used to establish the model,

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-22-0066

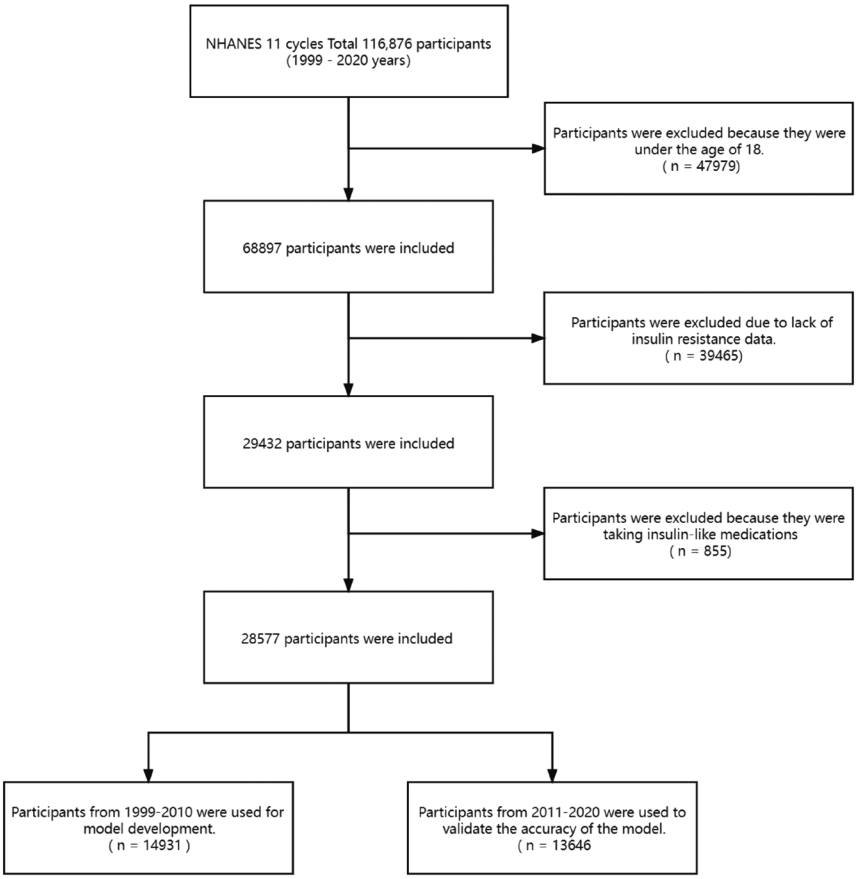

Figure 1
Flowchart of participant selection.

while the population who participated in the NHANES project from 2011 to 2020 was used to validate the model.

Data collection

Data were collected by professionally trained and qualified personnel in the MECs, including demographics (age, gender, ethnicity, education, etc.), anthropometric measurements (height, waist circumference, weight, BMI, etc.), health-related behaviors (smoking and alcohol consumption), laboratory tests (ALT, HDL, etc.), etc. Among them, the serum samples were sent to the US Centers for Disease Control and Prevention National Center for Environmental Health Laboratory Science Department and designated to authorized institutions for analysis under cold-chain conditions after scientific storage management (22).

Demographic data

Participants were measured for age, gender, ethnicity, education, smoking, and alcohol consumption in the mobile test vehicle, and other necessary interviews were conducted. In our study, educational level was divided into three categories: low education (no education or education up to grade 11), secondary education (education level of high school), and high education (education level of college and above).

Smoking status was divided into three categories: current smokers (smoked ≥ 100 cigarettes in the past in total and reported smoking on several days or days at the time of interview), ex-smokers (smoked < 100 cigarettes in the past but did not currently smoke), and non-smokers (smoked < 100 cigarettes in the past). Alcohol consumption was classified as drinking and non-drinking according to the recommendations of the US Department of Health and Human Services: >1 drink per day for women and >2 drinks per day for men were defined as drinkers (23).

Blood pressure of the participants was measured after a 5-minute rest, with a 5-minute interval between the next measurements, taking the average of two or more blood pressure measurements. SBP > 140 or DBP > 90 was diagnosed as hypertension. Diabetes was diagnosed by meeting any of the following criteria: FPG ≥ 7.0 mmol/L, glycated hemoglobin > 6.5%, the doctor told the participant to have diabetes, self-reported diabetes

for a long time, random blood glucose, or 2 h-OGTT test ≥ 11.0 mmol/L, was taking diabetes-related drugs. BMI was calculated based on height and weight, and its formula is BMI = weight/height-squared (kg/m^2).

Laboratory data

The collection of laboratory data was carried out in a mobile test vehicle, and the samples were scientifically stored at −20°C or −30°C after collection and transported to the laboratory for analysis at appropriate time. Samples prior to 2007 were analyzed by the Johns Hopkins University laboratory, and from 2007, by the University of Minnesota laboratory. Detailed processing steps can be found in the description of plasma sample components on the NHANES official website (https://www.cdc.gov/nchs/nhanes/about_nhanes.htm). Total cholesterol (TC) and triglyceride (TG) were measured enzymatically, and high-density lipoprotein (HDL) was measured by two methods: heparin-manganese precipitation or direct immunoassay. Other laboratory tests were measured by conventional biochemical spectroscopy, using a Hitachi Model 704 multichannel analyzer (Boehringer Mannheim Diagnostics, Indianapolis, IN, USA). Meanwhile, the NHANES project team employs several different approaches to test the quality of assays performed by the laboratory, including but not limited to conducting a second examination of previously examined participants.

Statistical analysis

All data in this study were analyzed using R software (version 4.1.2; packages: magrittr, dply, corrplot, leaps, rms, InformationValue, etc.). ROC curves were plotted using MedCalc software (version 15.6.1).

After checked for normality, continuous variables that obey the normal distribution were expressed as mean ± standard deviation (M ± S.D.) and compared by the independent sample *t*-test; if not, expressed as the median with interquartile range (IQR) and compared by the Mann–Whitney *U* test. Categorical variables were represented by counts and weighted percentages and were compared using the chi-square test. The multiple imputation method was used to impute missing variables in order to maximize statistical power and minimize bias. In addition, to determine whether the generated complete data differed significantly from the original data, a sensitivity analysis was performed. The results showed that the data after multiple imputation was not significantly different from the original data, and there was no statistical significance ($P > 0.05$). Univariate and multivariable logistic regression analyses were used to analyze which factors were closely associated with IR. In univariate logistic regression models, variables with effect value greater than 10% or $P < 0.1$ were included in multivariable logistic regression. The likelihood ratio test was used to select the relevant factors for constructing the predictive model in the training set by means of optimal subset regression. The discriminative power of different models was quantified and compared using the area under the receiver operating characteristic (ROC) curve (AUC). The calibration curve was evaluated by the unreliability *U* test. Use the 'rms' package of R software to draw calibration curves. We additionally introduce the Brier score (calculated by R software) to evaluate the accuracy of model predictions in classification tasks.

Decision curve analysis (DCA) was performed to determine the clinical net benefit of the model (24, 25). DCA is combining accuracy measures and clinical applicability by integrating clinical consequences associated with a test result. The net benefit is calculated by the difference between the proportion of relative harms of false positives and false negatives weighted by the odds of the selected threshold for high-risk designation, in other words, the difference between the expected benefit and the expected harm.

All statistical tests were two-sided and *P* values < 0.05 were considered significant.

Results

Baseline characteristics of the participating population

A total of 14,931 participants from 1999 to 2010 were included to establish the predictive model. The age of the participants was 46.9 ± 19.7, of whom 6444 participants were diagnosed with IR and were older than IR-negative participants (48.7 vs 45.5). In addition, IR-positive and IR-negative participants differed significantly in the following variables: gender, race, education, BMI, waist circumference, hypertension, smoking, alcohol consumption, HDL, LDL, TC, TG, alanine aminotransferase (ALT), aspartate aminotransferase (AST), glutamyl transpeptidase (GGT), total bilirubin (TBIL), blood urea nitrogen (BUN), lactate dehydrogenase (LDH), albumin (ALB), globulin (GLB), creatinine (Cre), uric acid (UA), Na, and glucose (GLU), with all *P* values <0.001 (Table 1).

A total of 13,646 participants from 2011 to 2020 were included to validate the accuracy of the model, and the age of the participants was 48.3 ± 18.3. 6114 participants

Table 1 Basic crowd information description of the training set.

Variables	Total (n = 14,931)	IR-negative (n = 8487)	IR-positive (n = 6444)	*P*-value
Age, mean ± S.D.	46.9 ± 19.7	45.5 ± 20.0	48.7 ± 19.2	<0.001
Gender, *n* (%)				<0.001
Male	7231 (48.4)	3969 (46.8)	3262 (50.6)	
Female	7700 (51.6)	4518 (53.2)	3182 (49.4)	
Race, *n* (%)				<0.001
Mexican American	3266 (21.9)	1610 (19)	1656 (25.7)	
Other Hispanic	2935 (19.7)	1588 (18.7)	1347 (20.9)	
Non-Hispanic White	7110 (47.6)	4402 (51.9)	2708 (42)	
Non-Hispanic Black	1011 (6.8)	515 (6.1)	496 (7.7)	
Other races	609 (4.1)	372 (4.4)	237 (3.7)	
Education, *n* (%)				<0.001
Poorly educated	4520 (30.3)	2306 (27.2)	2214 (34.4)	
Moderately educated	3536 (23.7)	1974 (23.3)	1562 (24.2)	
Highly educated	6875 (46.0)	4207 (49.6)	2668 (41.4)	
BMI, mean ± S.D.	28.3 ± 6.5	25.7 ± 4.7	31.7 ± 6.9	<0.001
Waist, mean ± S.D.	97.0 ± 15.8	90.4 ± 12.7	105.8 ± 15.2	<0.001
Hypertension, *n* (%)				<0.001
No	5354 (35.9)	2389 (28.1)	2965 (46)	
Yes	9577 (64.1)	6098 (71.9)	3479 (54)	
DM, *n* (%)				<0.001
No	12,823 (85.9)	7987 (94.1)	4836 (75)	
Yes	2108 (14.1)	500 (5.9)	1608 (25)	
Smoking status, *n* (%)				<0.001
Never smoking	7895 (52.9)	4508 (53.1)	3387 (52.6)	
Former smokers	3879 (26.0)	2055 (24.2)	1824 (28.3)	
Current smoker	3157 (21.1)	1924 (22.7)	1233 (19.1)	
Alcohol, *n* (%)				<0.001
No	10,565 (70.8)	5764 (67.9)	4801 (74.5)	
Yes	4366 (29.2)	2723 (32.1)	1643 (25.5)	
HDL, median (IQR)	1.3 (1.1, 1.6)	1.4 (1.2, 1.8)	1.2 (1.0, 1.4)	<0.001
LDL, median (IQR)	2.9 (2.4, 3.6)	2.9 (2.3, 3.5)	2.9 (2.4, 3.6)	0.01
TC, median (IQR)	5.0 (4.3, 5.7)	5.0 (4.3, 5.7)	5.0 (4.3, 5.8)	0.004
TG, median (IQR)	1.2 (0.8, 1.8)	1.0 (0.7, 1.5)	1.5 (1.0, 2.2)	<0.001
ALT, median (IQR)	21.0 (16.0, 28.0)	19.0 (15.0, 25.0)	23.0 (17.8, 33.0)	<0.001
AST, median (IQR)	23.0 (19.0, 27.0)	22.0 (19.0, 26.0)	23.0 (19.0, 28.0)	<0.001
GGT, median (IQR)	20.0 (14.0, 31.0)	17.0 (13.0, 25.0)	24.0 (17.0, 37.0)	<0.001
TBIL, median (IQR)	12.0 (10.3, 15.4)	12.0 (10.3, 15.4)	12.0 (8.6, 13.7)	<0.001
BUN, median (IQR)	4.3 (3.2, 5.4)	4.3 (3.2, 5.4)	4.3 (3.6, 5.7)	<0.001
LDH, median (IQR)	130.0 (114.0, 148.0)	128.0 (113.0, 147.0)	132.0 (116.0, 150.0)	<0.001
ALB, median (IQR)	42.0 (40.0, 45.0)	43.0 (40.0, 45.0)	42.0 (40.0, 44.0)	<0.001
GLB, median (IQR)	30.0 (27.0, 33.0)	29.0 (27.0, 32.0)	31.0 (28.0, 34.0)	<0.001
Cre, median (IQR)	70.7 (61.9, 88.4)	70.7 (61.9, 88.4)	71.6 (61.9, 88.4)	0.074
UA, median (IQR)	315.2 (261.7, 374.7)	297.4 (243.9, 356.9)	339.0 (285.5, 398.5)	<0.001
Na, median (IQR)	139.0 (138.0, 140.8)	139.0 (138.0, 141.0)	139.0 (138.0, 140.5)	<0.001
Cl, median (IQR)	104.0 (102.0, 105.2)	104.0 (102.0, 105.0)	104.0 (102.0, 105.6)	0.37
Ca, median (IQR)	2.4 (2.3, 2.4)	2.4 (2.3, 2.4)	2.4 (2.3, 2.4)	0.069
GLU, median (IQR)	5.1 (4.7, 5.6)	4.9 (4.6, 5.3)	5.5 (5.0, 6.2)	<0.001

ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; Ca, calcium; Cl, chlorine; Cre, creatinine; GGT, glutamyl transpeptidase; GLB, globulin; GLU, glucose; HDL, high-density lipoprotein; LDH, lactate dehydrogenase; LDL, low-density lipoprotein; Na, sodium; TBIL, total bilirubin; TC, total cholesterol; TG, triglyceride; UA, uric acid.

were diagnosed with IR and were older than those who were IR-negative (49.8 vs 47.1). There were still significant differences between IR-positive and IR-negative participants in the following variables: gender, race, education, BMI, waist circumference, hypertension, smoking, alcohol consumption, HDL, LDL, TC, TG, ALT, AST, GGT, TBIL, BUN, LDH, ALB, LB, Cre, UA, Na, and GLU, with all P values < 0.001 (Table 2).

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-22-0066

Table 2 Basic crowd information description of the test set.

Variables	Total (n = 13,646)	IR-negative (n = 7532)	IR-positive(n = 6114)	*P*-value
Age, mean ± S.D.	48.3 ± 18.3	47.1 ± 18.6	49.8 ± 17.8	<0.001
Gender, *n* (%)				0.012
Male	6628 (48.6)	3585 (47.6)	3043 (49.8)	
Female	7018 (51.4)	3947 (52.4)	3071 (50.2)	
Race, *n* (%)				<0.001
Mexican American	1857 (13.6)	826 (11)	1031 (16.9)	
Other Hispanic	3030 (22.2)	1667 (22.1)	1363 (22.3)	
Non-Hispanic White	4997 (36.6)	2914 (38.7)	2083 (34.1)	
Non-Hispanic Black	1433 (10.5)	721 (9.6)	712 (11.6)	
Other races	2329 (17.1)	1404 (18.6)	925 (15.1)	
Education, *n* (%)				<0.001
Poorly educated	2884 (21.1)	1479 (19.6)	1405 (23)	
Moderately educated	3061 (22.4)	1637 (21.7)	1424 (23.3)	
Highly educated	7701 (56.4)	4416 (58.6)	3285 (53.7)	
BMI, mean ± S.D.	29.1 ± 7.2	26.1 ± 5.3	32.9 ± 7.5	<0.001
Waist, mean ± S.D.	99.0 ± 16.9	91.4 ± 13.4	108.3 ± 16.2	<0.001
Hypertension, *n* (%)				<0.001
No	5259 (38.5)	2318 (30.8)	2941 (48.1)	
Yes	8387 (61.5)	5214 (69.2)	3173 (51.9)	
DM, *n* (%)				<0.001
No	11,161 (81.8)	6921 (91.9)	4240 (69.3)	
Yes	2485 (18.2)	611 (8.1)	1874 (30.7)	
Smoke, *n* (%)				<0.001
Never smoking	7946 (58.2)	4393 (58.3)	3553 (58.1)	
Former smokers	3110 (22.8)	1579 (21)	1531 (25)	
Current smoker	2590 (19.0)	1560 (20.7)	1030 (16.8)	
Alcohol, *n* (%)				<0.001
No	6193 (45.4)	3307 (43.9)	2886 (47.2)	
Yes	7453 (54.6)	4225 (56.1)	3228 (52.8)	
HDL, median (IQR)	1.3 (1.1, 1.6)	1.5 (1.2, 1.8)	1.2 (1.0, 1.4)	<0.001
LDL, median (IQR)	2.8 (2.2, 3.4)	2.7 (2.2, 3.4)	2.8 (2.2, 3.5)	<0.001
TC, median (IQR)	4.8 (4.1, 5.5)	4.8 (4.1, 5.5)	4.8 (4.1, 5.5)	0.167
TG, median (IQR)	1.1 (0.8, 1.7)	1.0 (0.7, 1.3)	1.4 (1.0, 2.0)	<0.001
ALT, median (IQR)	19.0 (14.0, 27.0)	17.0 (13.0, 23.0)	22.0 (16.0, 31.0)	<0.001
AST, median (IQR)	21.0 (17.0, 26.0)	21.0 (17.0, 25.0)	21.0 (17.0, 27.0)	<0.001
GGT, median (IQR)	20.0 (14.0, 30.0)	17.0 (12.0, 25.0)	23.0 (17.0, 35.0)	<0.001
TBIL, median (IQR)	8.6 (6.8, 12.0)	10.3 (6.8, 13.7)	8.6 (6.8, 12.0)	<0.001
BUN, median (IQR)	4.6 (3.6, 5.7)	4.6 (3.6, 5.7)	4.6 (3.9, 5.7)	<0.001
LDH, median (IQR)	136.0 (118.0, 158.0)	135.0 (117.0, 157.0)	138.0 (120.0, 159.0)	<0.001
ALB, median (IQR)	42.0 (39.0, 44.0)	42.0 (40.0, 44.0)	41.0 (39.0, 43.0)	<0.001
GLB, median (IQR)	29.0 (27.0, 33.0)	29.0 (26.0, 32.0)	30.0 (27.0, 33.0)	<0.001
Cre, median (IQR)	73.4 (61.9, 86.6)	73.4 (61.9, 86.6)	73.4 (61.0, 87.5)	0.944
UA, median (IQR)	315.2 (261.7, 374.7)	297.4 (249.8, 356.9)	339.0 (285.5, 398.5)	<0.001
Na, median (IQR)	140.0 (138.0, 141.0)	140.0 (138.0, 141.0)	140.0 (138.0, 141.0)	0.04
Cl, median (IQR)	103.0 (101.0, 105.0)	103.0 (101.0, 105.0)	103.0 (101.0, 105.0)	0.006
Ca, median (IQR)	2.3 (2.3, 2.4)	2.3 (2.3, 2.4)	2.3 (2.3, 2.4)	0.709
GLU, median (IQR)	5.3 (4.9, 5.8)	5.0 (4.7, 5.4)	5.6 (5.2, 6.4)	<0.001

ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; Ca, calcium; Cl, chlorine; Cre, creatinine; GGT, glutamyl transpeptidase; GLB, globulin; GLU, glucose; HDL, high-density lipoprotein; LDH, lactate dehydrogenase; LDL, low-density lipoprotein; Na, sodium; TBIL, total bilirubin; TC, total cholesterol; TG, triglyceride; UA, uric acid.

Logistic regression analysis

The optimal subset regression analysis identified eight candidate factors, and after exclusion of collinearity interference, a total of seven factors were finally included, as shown in the nomogram (Figs 2 and 3). The included factors were HDL, LDL, ALB, GLB, GLU, BMI, and waist circumference, and their correlation with IR is shown in Table 3. After adjusting for potential confounders, HDL (OR: 0.37, 95% CI 0.33–0.41) and LDL (OR: 0.88, 95% CI 0.84–0.92) are inversely correlated with IR, while ALB (1.04, 95% CI 1.03–1.05), GLB (OR: 1.07, 95% CI

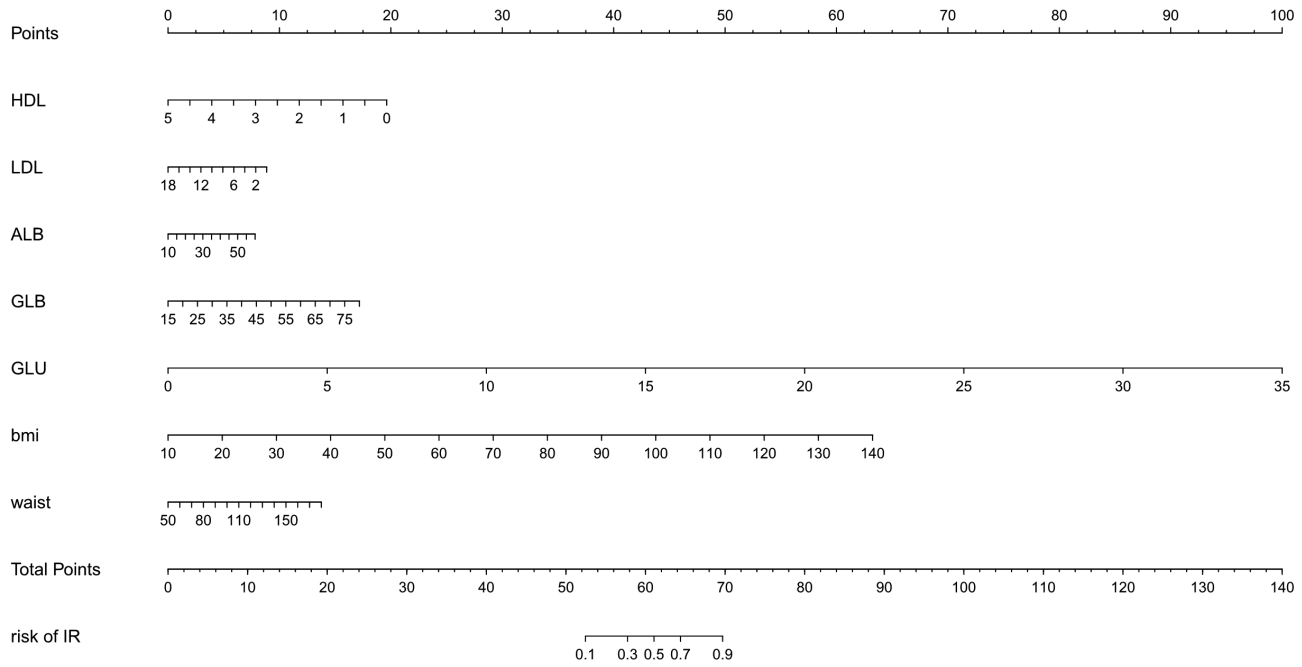

Figure 2
The nomogram of the optimal subset regression to predict the risk of incident insulin resistance.

1.06–1.08), GLU (OR: 2.07, 95% CI 1.96–2.19), BMI (OR: 1.13, 95% CI 1.12–1.15), and waist circumference (OR: 1.03, 95% CI 1.02–1.03) are positively correlated with IR. Among them, the correlations between HDL/GLU and IR are more significant.

Analysis of full factor modeling and optimal subset modeling

We included all variables (age, gender, race, education, BMI, waist, hypertension, smoking, alcohol consumption, HDL, LDL, TC, TG, ALT, AST, GGT, TBIL, BUN, LDH, ALB, LB, Cre, UA, Na, and GLU) to establish a full factor prediction model and plotted ROC curves (Fig. 4B1 and B2). In the training set, the AUC of the full factor modeling is 0.870 (95% CI: 0.864–0.875) and of the validation set is 0.874 (95% CI: 0.869–0.880) (Table 4). The calibration curve is shown in Fig. 5, the predicted occurrence of IR in the full factor modeling is well-calibrated, and there is no significant difference between the predicted probability and the observed probability. In the temporal validation population, the mean difference between predicted and calibrated probabilities is 0.017, the maximum difference is 0.036, the *P*-value for the U-index is 0.406, and Brier is 0.143, suggesting that the *P*-value obtained by the full factor model is as expected, while the mean difference provides an adequate estimate of the target error when assessed using temporal validation data.

A total of seven factors (HDL, LDL, ALB, GLB, GLU, BMI, and waist) were included in the optimal subset model, and the ROC curves are shown in Fig. 4. In the training set, the AUC for the optimal subset modeling is 0.851 (95% CI: 0.845–0.857), and in the validation set, the AUC is 0.857 (95% CI: 0.851–0.863) (Table 3). The calibration curve is shown in Fig. 5B2, the prediction of the occurrence of IR in the optimal subset modeling is well-calibrated, and in the external validation population, the mean difference between predicted and calibrated probabilities is 0.019, and the maximum difference is 0.0368, The *P*-value for the U-index is 0.238 and the Brier is 0.153. Likewise, this also shows that the *P*-values obtained by the optimal subset model are as expected, while the mean difference provides an adequate estimate of the target error when evaluated using external data.

The decision curves show that the red line representing the full factor modeling is almost always above the blue line representing the optimal subset modeling, indicating that the net benefit of the full factor model was slightly higher than that of the optimal subset model (Fig. 6A and B).

Discussion

In recent decades, with the development of the global economy and the advancement of a large number of clinical

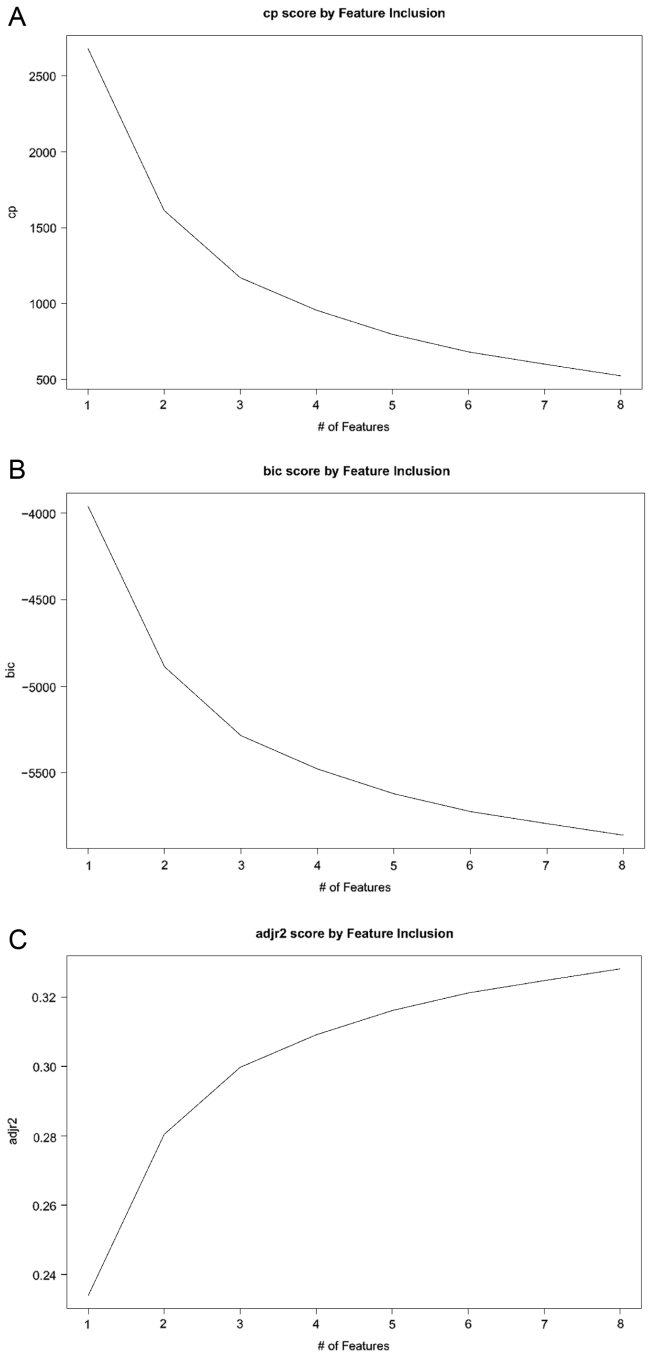

Figure 3
Number of included factor variables screened by the three criteria for the optimal subset. (A) Screening variables using the CP principle. (B) Screening variables using the BIC scoring principle. (C) Screening variables using the adjusted R-squared principle.

studies, most infectious diseases worldwide have been successfully conquered, and the main culprits affecting the health of people in the world have been gradually replaced by non-communicable diseases (26). Currently, insulin resistance has become a global health issue that cannot be ignored. The root causes for the widespread prevalence of IR can be roughly classified into two categories: one is the increase in the consumption of high-calorie, low-fiber diets; the other is the sedentary period caused by high-convenience tools, which in turn leads to a decrease in physical activity (27, 28, 29). The occurrence of IR will not only increase the incidence of diseases such as type 2 diabetes, hypertension, coronary artery disease, stroke, etc. but also increase all-cause mortality of patients with cancer, causing trillions of medical burdens worldwide (8, 30, 31, 32, 33, 34, 35). The incidence of IR, which remains on the rise, is largely attributable to the failure to recognize its occurrence early. In current clinical and epidemiological practice, there is also a lack of convenient tools for the early identification of IR. Thus, the development of a simple predictive model for the assessment of IR can effectively help clinicians and epidemiologists to identify subjects with IR early.

In this study, we established a full factor predictive model based on 25 clinical routine items. In the external validation set, the AUC value was 0.874, as AUC > 0.85 was generally considered to be excellent test performance. However, considering the large number of variables included, it is difficult to apply in clinical practice. Therefore, we screened the best modeling factors by means of the optimal subset regression and established the model with only seven factors. The AUC value in the external validation set was 0.857, which was also excellent, approaching the performance of the full factor modeling. Meanwhile, the subsequent calibration curves suggest that there was no difference between the prediction accuracy and the ideal accuracy of the optimal subset model, and the accuracy of the model was high. We performed the DCA and the results showed that both the full factor predictive model and the optimal subset regression predictive model performed excellently for the clinical net benefit. As expected, the clinical benefit of the optimal subset regression prediction model was not inferior to that of the full factor prediction model after a large number of variables were removed. Taken together, we believe that adopting a highly refined optimal subset model in the clinic is most beneficial for clinical practice.

A large number of IR prediction models have been established in previous studies; however, the performance of the models has been suboptimal and there is considerable ethnic heterogeneity. For example, a model's performance with an AUC of 0.841 proposed by Lechner *et al.* in 2021 with a German population ($n = 2231$) including age, gender, waste-to-height ratio (WHtR), FPG, and TG/HDL,

Table 3 Univariate and multivariate logistics regression analyses included indicators and insulin resistance.

Variable	Crude OR (95% CI)	Crude *P*-value	Adjusted OR (95% CI)	Adjusted *P*-value
(Intercept)	6.71 (5.9–7.63)	<0.001	0 (0–0)	<0.001
HDL	0.2 (0.18–0.22)	<0.001	0.37 (0.33–0.41)	<0.001
LDL	1.03 (1–1.07)	0.073	0.88 (0.84–0.92)	<0.001
ALB	0.95 (0.95–0.96)	<0.001	1.04 (1.03–1.05)	<0.001
GLB	1.08 (1.07–1.08)	<0.001	1.07 (1.06–1.08)	<0.001
GLU	2.8 (2.65–2.95)	<0.001	2.07 (1.96–2.19)	<0.001
bmi	1.22 (1.21–1.23)	<0.001	1.13 (1.12–1.15)	<0.001
waist	1.08 (1.08–1.09)	<0.001	1.03 (1.02–1.03)	<0.001

The variables included in multivariate logistic regression were those with $P < 0.1$ in univariate analysis or those included in previous studies.

is lower than our model (36). In a Chinese study, Liu *et al.* established a model by factors such as WHtR, waist circumference, WHR, BMI, etc. with an AUC value of 0.747 and poor performance (37). In addition, Yeh *et al.* (16) reported that a model established with TG/HDL to predict IR performed well in the elderly Taiwanese population, with an AUC of 0.729, but the AUC value was only 0.56 when applied to the overweight African-American population (38). Admittedly, the ethnic heterogeneity of IR prediction models is significant, and these results reinforce the motivation for our study. However, due to the complexity of racial makeup, developing a model for each race is clearly impractical. Even the same race, when living in different countries or regions, with different diets

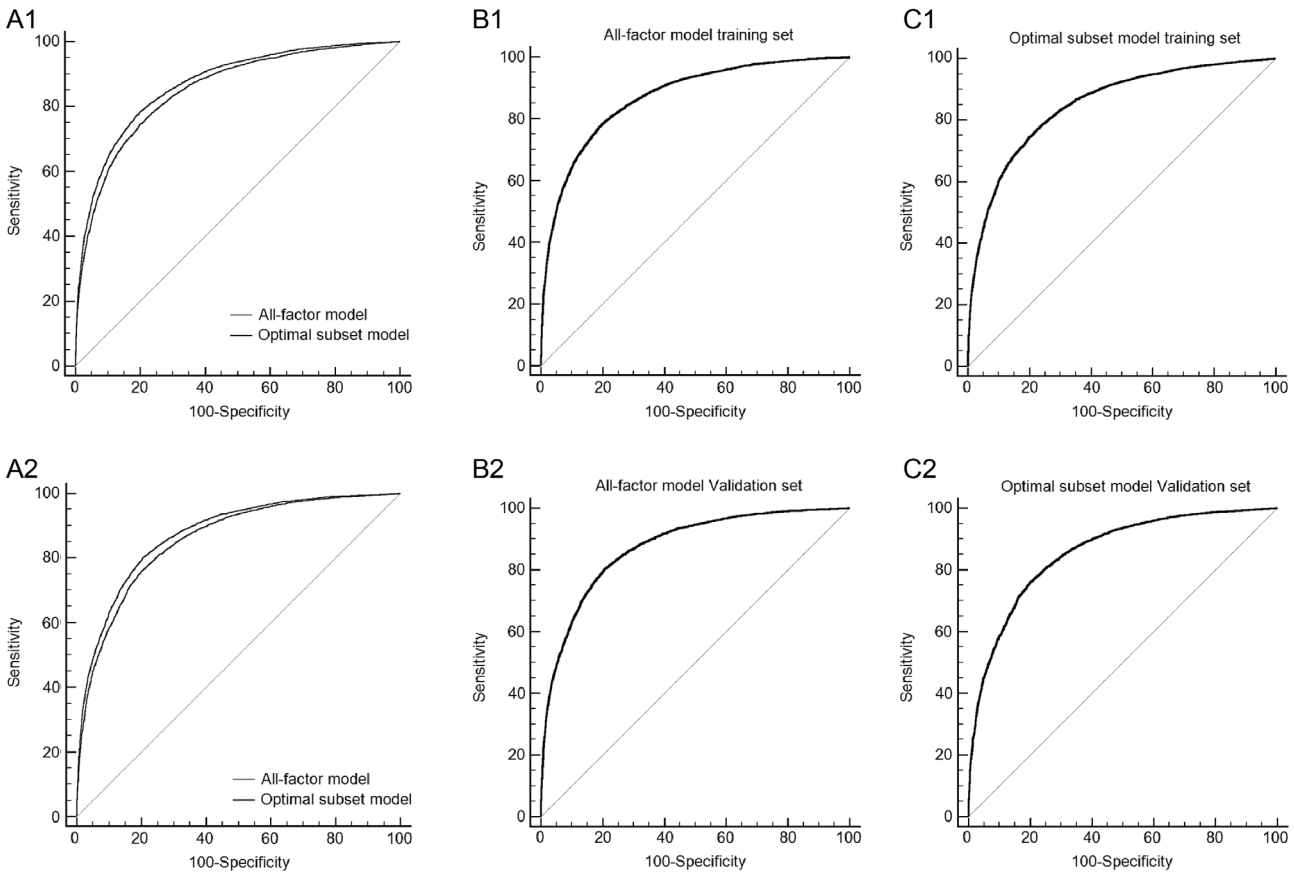

Figure 4
The ROC curves of the full model and optimal subset model in the training set and validation set. (A1) ROC curves of the full model and optimal subset model in the training set. (B1) ROC curve of the full model in the training set. (C1) ROC curve of the optimal subset model in the training set. (A2) ROC curves of the full model and optimal subset model in the validation set. (B2) ROC curve of the full model in the validation set. (C2) ROC curve of the optimal subset model in the validation set.

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-22-0066

Table 4 Evaluation of prediction model.

Scoring system	AUC	95% CI	Optimal cut-off	Sensitivity (%)	Specificity (%)	Youden's index
All-factor model						
Training set	0.870	0.864–0.875	0.422	78.3	80.1	0.584
Validation set	0.874	0.869–0.880	0.407	80.3	79.4	0.596
Optimal subset model						
Training set	0.851	0.845–0.857	0.450	74.4	80.2	0.446
Validation set	0.857	0.851–0.863	0.476	76.0	80.0	0.560

and living habits, has a completely different background for the eventual development of IR. Therefore, we believe that it may be more reasonable to divide the population by country or region. We first propose an insulin resistance model suitable for the American population and provide a theoretical basis for related research in other countries and regions in the future.

To the best of our knowledge, this study is the first to successfully establish a predictive model of IR by a large, cross-ethnic, nationally representative sample of the US population. The model benefits from the national population representation of the NHANES database, and our results apply to ethnically diverse US populations. However, it must be pointed out that our study has some limitations: First, our research is a cross-sectional study, which makes it difficult for us to explore the causal relationship between these factors and IR. Secondly, due to differences between races, as we have emphasized, the accuracy of the model still requires additional evaluation when applied to races in other countries or regions. In addition, the absence of lifestyle factors is also a limitation of this study. Therefore, the actual predictive value of this model still needs to be confirmed by prospective randomized controlled trials.

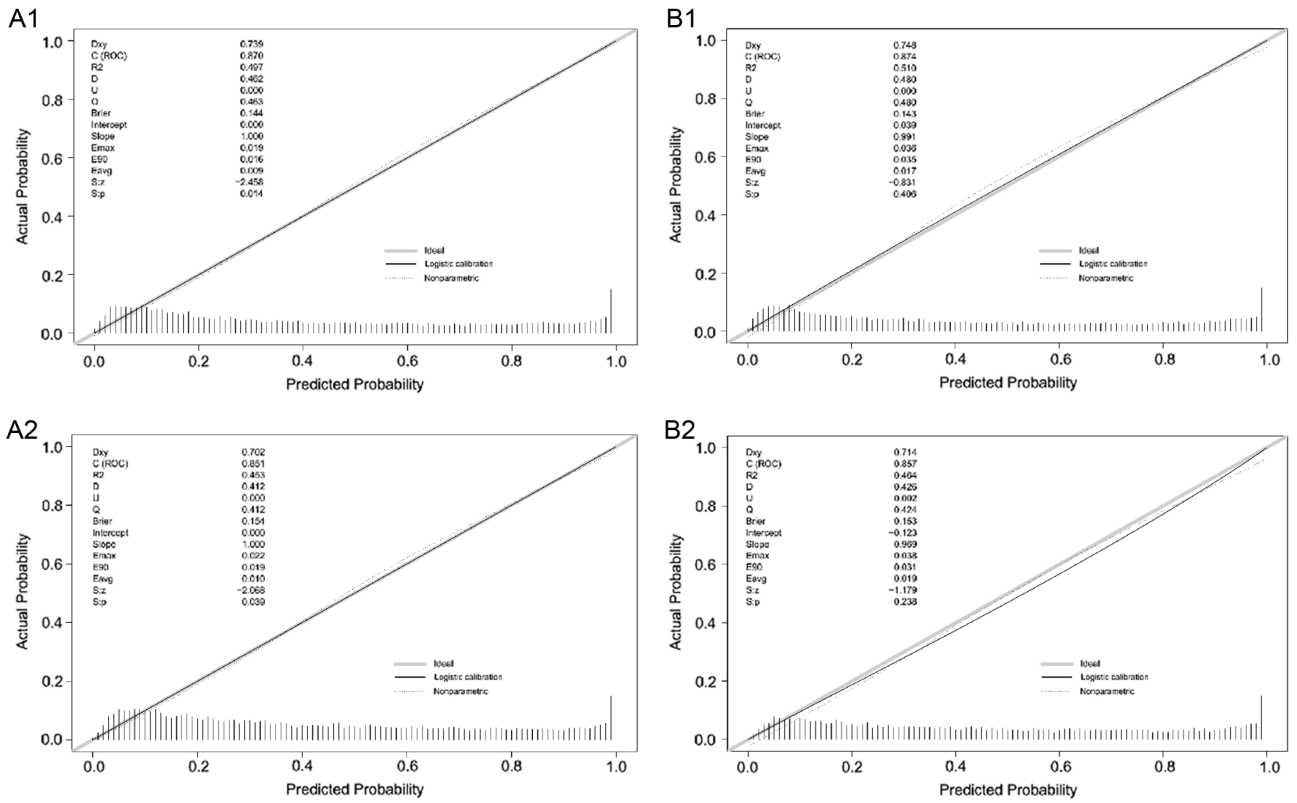

Figure 5
Calibration curves for the full factor model and the optimal subset model. (A1) Calibration curve for the full factor modeling cohort. (A2) Calibration curve for the full factor validation model cohort. (B1) Calibration curve for the optimal subset regression modeling cohort. (B2) Calibration curve for optimal subset regression validation cohort.

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-22-0066

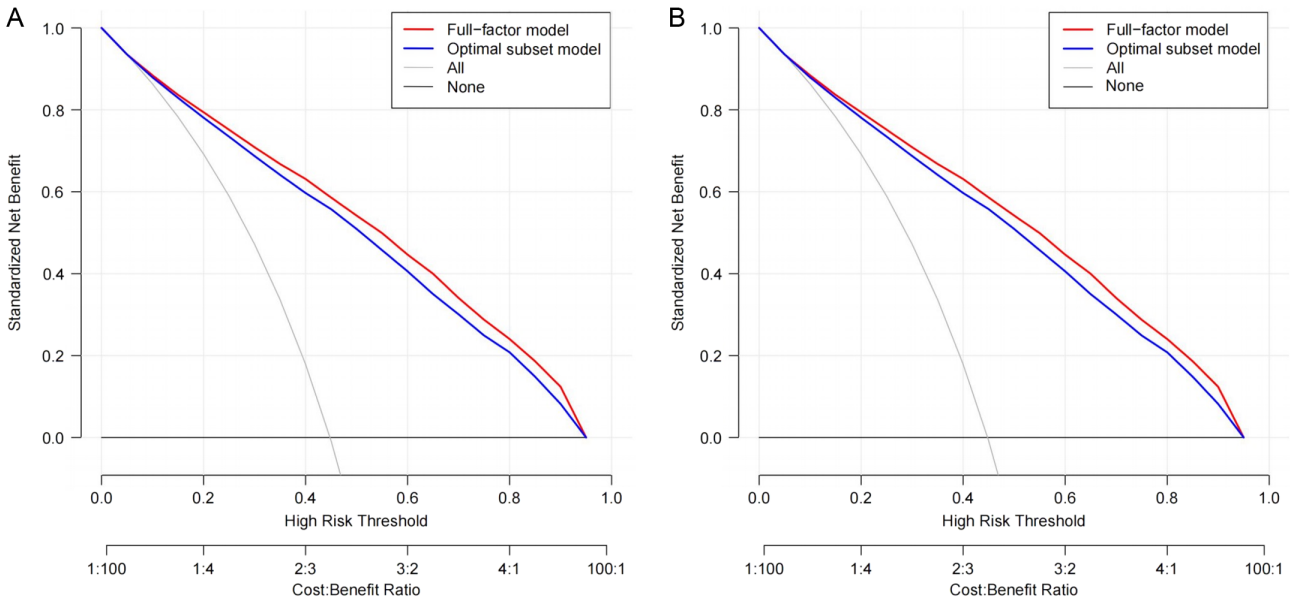

Figure 6
DCA curves of the full model (indicated by the red line) and optimal subset model (indicated by the blue line) in the training set (A) and in the validation set (B).

Conclusion

The optimal subset predictive model proposed in this study has great performance in predicting IR, and the decision curve analysis shows that it has a high clinical net benefit, which can help clinicians and epidemiologists to detect IR easily, and take appropriate interventions as early as possible. The model works across racially diverse US populations, but its predictive value in other countries or regions remains to be proven.

Supplementary materials
This is linked to the online version of the paper at https://doi.org/10.1530/EC-22-0066.

Declaration of interest
The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding
This work was supported by General Project of Natural Science Foundation of Qinghai Province (2020-ZJ-930).

Availability of data and materials
Data can be obtained from the NHANES database (https://www.cdc.gov/nchs/nhanes/).

Author contribution statement
Gong and Liu conceived the idea; Gong and Liu wrote the manuscript; Yin, Hu and Xiao collected and read the literature and revised the article; Hu and Gong read through and corrected the manuscript. All authors read and approved the final manuscript. Gong is the first author. Liu is the co-first author. Hu is the corresponding author of this paper.

References

1 Yaribeygi H, Farrokhi FR, Butler AE & Sahebkar A. Insulin resistance: review of the underlying molecular mechanisms. *Journal of Cellular Physiology* 2019 **234** 8152–8161. (https://doi.org/10.1002/jcp.27603)
2 Petersen MC & Shulman GI. Mechanisms of insulin action and insulin resistance. *Physiological Reviews* 2018 **98** 2133–2223. (https://doi.org/10.1152/physrev.00063.2017)
3 Carpentier AC. 100(th) anniversary of the discovery of insulin perspective: insulin and adipose tissue fatty acid metabolism. *American Journal of Physiology: Endocrinology and Metabolism* 2021 **320** E653–E670. (https://doi.org/10.1152/ajpendo.00620.2020)
4 Sbraccia P, D'Adamo M & Guglielmi V. Is type 2 diabetes an adiposity-based metabolic disease? From the origin of insulin resistance to the concept of dysfunctional adipose tissue. *Eating and Weight Disorders* 2021 **26** 2429–2441. (https://doi.org/10.1007/s40519-021-01109-4)
5 Brown AE & Walker M. Genetics of insulin resistance and the metabolic syndrome. *Current Cardiology Reports* 2016 **18** 75. (https://doi.org/10.1007/s11886-016-0755-4)
6 Krentz AJ. Insulin resistance. *BMJ* 1996 **313** 1385–1389. (https://doi.org/10.1136/bmj.313.7069.1385)
7 Huang PL. A comprehensive definition for metabolic syndrome. *Disease Models and Mechanisms* 2009 **2** 231–237. (https://doi.org/10.1242/dmm.001180)
8 Laakso M & Kuusisto J. Insulin resistance and hyperglycaemia in cardiovascular disease development. *Nature Reviews: Endocrinology* 2014 **10** 293–302. (https://doi.org/10.1038/nrendo.2014.29)

9 Eckel RH, Grundy SM & Zimmet PZ. The metabolic syndrome. *Lancet* 2005 **365** 1415–1428. (https://doi.org/10.1016/S0140-6736(05)66378-7)

10 Hill MA, Yang Y, Zhang L, Sun Z, Jia G, Parrish AR & Sowers JR. Insulin resistance, cardiovascular stiffening and cardiovascular disease. *Metabolism: Clinical and Experimental* 2021 **119** 154766. (https://doi.org/10.1016/j.metabol.2021.154766)

11 Pan K, Nelson RA, Wactawski-Wende J, Lee DJ, Manson JE, Aragaki AK, Mortimer JE, Phillips LS, Rohan T, Ho GYF, *et al.* Insulin resistance and cancer-specific and all-cause mortality in postmenopausal women: the Women's Health Initiative. *Journal of the National Cancer Institute* 2020 **112** 170–178. (https://doi.org/10.1093/jnci/djz069)

12 Lee CL, Liu WJ & Wang JS. Associations of low-carbohydrate and low-fat intakes with all-cause mortality in subjects with prediabetes with and without insulin resistance. *Clinical Nutrition* 2021 **40** 3601–3607. (https://doi.org/10.1016/j.clnu.2020.12.019)

13 DeFronzo RA, Tobin JD & Andres R. Glucose clamp technique: a method for quantifying insulin secretion and resistance. *American Journal of Physiology* 1979 **237** E214–E223. (https://doi.org/10.1152/ajpendo.1979.237.3.E214)

14 Bonora E, Targher G, Alberiche M, Bonadonna RC, Saggiani F, Zenere MB, Monauni T & Muggeo M. Homeostasis model assessment closely mirrors the glucose clamp technique in the assessment of insulin sensitivity: studies in subjects with various degrees of glucose tolerance and insulin sensitivity. *Diabetes Care* 2000 **23** 57–63. (https://doi.org/10.2337/diacare.23.1.57)

15 Boursier G, Sultan A, Molinari N, Maimoun L, Boegner C, Picandet M, Kuster N, Bargnoux AS, Badiou S, Dupuy AM, *et al.* Triglycerides and glycated hemoglobin for screening insulin resistance in obese patients. *Clinical Biochemistry* 2018 **53** 8–12. (https://doi.org/10.1016/j.clinbiochem.2017.12.002)

16 Yeh WC, Tsao YC, Li WC, Tzeng IS, Chen LS & Chen JY. Elevated triglyceride-to-HDL cholesterol ratio is an indicator for insulin resistance in middle-aged and elderly Taiwanese population: a cross-sectional study. *Lipids in Health and Disease* 2019 **18** 176. (https://doi.org/10.1186/s12944-019-1123-3)

17 do Vale Moreira NC, Ceriello A, Basit A, Balde N, Mohan V, Gupta R, Misra A, Bhowmik B, Lee MK, Zuo H, *et al.* Race/ethnicity and challenges for optimal insulin therapy. *Diabetes Research and Clinical Practice* 2021 **175** 108823. (https://doi.org/10.1016/j.diabres.2021.108823)

18 Tamayo T, Jacobs Jr DR, Strassburger K, Giani G, Seeman TE, Matthews K, Roseman JM & Rathmann W. Race- and sex-specific associations of parental education with insulin resistance in middle-aged participants: the CARDIA study. *European Journal of Epidemiology* 2012 **27** 349–355. (https://doi.org/10.1007/s10654-012-9691-9)

19 Paramsothy P, Knopp R, Bertoni AG, Tsai MY, Rue T & Heckbert SR. Combined hyperlipidemia in relation to race/ethnicity, obesity, and insulin resistance in the multi-ethnic study of atherosclerosis. *Metabolism: Clinical and Experimental* 2009 **58** 212–219. (https://doi.org/10.1016/j.metabol.2008.09.016)

20 Ford ES, Li C, Imperatore G & Cook S. Age, sex, and ethnic variations in serum insulin concentrations among U.S. youth: findings from the National Health and Nutrition Examination Survey 1999–2002. *Diabetes Care* 2006 **29** 2605–2611. (https://doi.org/10.2337/dc06-1083)

21 Sobus JR, DeWoskin RS, Tan YM, Pleil JD, Phillips MB, George BJ, Christensen K, Schreinemachers DM, Williams MA, Hubal EA, *et al.* Uses of NHANES biomarker data for chemical risk assessment: trends, challenges, and opportunities. *Environmental Health Perspectives* 2015 **123** 919–927. (https://doi.org/10.1289/ehp.1409177)

22 Ahluwalia N, Dwyer J, Terry A, Moshfegh A & Johnson C. Update on NHANES dietary data: focus on collection, release, analytical considerations, and uses to inform public policy. *Advances in Nutrition* 2016 **7** 121–134. (https://doi.org/10.3945/an.115.009258)

23 Breslow RA, Chen CM, Graubard BI, Jacobovits T & Kant AK. Diets of drinkers on drinking and nondrinking days: NHANES 2003–2008. *American Journal of Clinical Nutrition* 2013 **97** 1068–1075. (https://doi.org/10.3945/ajcn.112.050161)

24 Vickers AJ & Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 2006 **26** 565–574. (https://doi.org/10.1177/0272989X06295361)

25 Vickers AJ, Van Calster B & Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016 **352** i6. (https://doi.org/10.1136/bmj.i6)

26 The Lancet Global Health. Getting to the heart of non-communicable diseases. *Lancet: Global Health* 2018 **6** e933. (https://doi.org/10.1016/S2214-109X(18)30362-0)

27 Haileamlak A. Physical inactivity: the major risk factor for non-communicable diseases. *Ethiopian Journal of Health Sciences* 2019 **29** 810. (https://doi.org/10.4314/ejhs.v29i1.1)

28 Booth FW, Roberts CK & Laye MJ. Lack of exercise is a major cause of chronic diseases. *Comprehensive Physiology* 2012 **2** 1143–1211. (https://doi.org/10.1002/cphy.c110025)

29 Liu Y, Wang Y, Ni Y, Cheung CKY, Lam KSL, Wang Y, Xia Z, Ye D, Guo J, Tse MA, *et al.* Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. *Cell Metabolism* 2020 **31** 77.e5–91.e5. (https://doi.org/10.1016/j.cmet.2019.11.001)

30 Ndisang JF, Vannacci A & Rastogi S. Insulin resistance, type 1 and type 2 diabetes, and related complications 2017. *Journal of Diabetes Research* 2017 **2017** 1478294. (https://doi.org/10.1155/2017/1478294)

31 Nayyar M, Lastra G & Acevedo CM. Mineralocorticoids and cardiovascular disease in females with insulin resistance and obesity. *Current Hypertension Reports* 2018 **20** 88. (https://doi.org/10.1007/s11906-018-0887-6)

32 Saklayen MG. The global epidemic of the metabolic syndrome. *Current Hypertension Reports* 2018 **20** 12. (https://doi.org/10.1007/s11906-018-0812-z)

33 Kernan WN, Inzucchi SE, Viscoli CM, Brass LM, Bravata DM & Horwitz RI. Insulin resistance and risk for stroke. *Neurology* 2002 **59** 809–815. (https://doi.org/10.1212/wnl.59.6.809)

34 Wang F, Han L & Hu D. Fasting insulin, insulin resistance and risk of hypertension in the general population: a meta-analysis. *Clinica Chimica Acta: International Journal of Clinical Chemistry* 2017 **464** 57–63. (https://doi.org/10.1016/j.cca.2016.11.009)

35 Slater EE. Insulin resistance and hypertension. *Hypertension* 1991 **18** (3 Supplement) I108–I114. (https://doi.org/10.1161/01.hyp.18.3_suppl.i108)

36 Lechner K, Lechner B, Crispin A, Schwarz PEH & von Bibra H. Waist-to-height ratio and metabolic phenotype compared to the Matsuda index for the prediction of insulin resistance. *Scientific Reports* 2021 **11** 8224. (https://doi.org/10.1038/s41598-021-87266-z)

37 Liu T, Wang Q, Huang W, Tan J, Liu D, Pei T, Li X & Zhou G. Anthropometric indices to predict insulin resistance in women with polycystic ovary syndrome in China. *Reproductive Biomedicine Online* 2019 **38** 101–107. (https://doi.org/10.1016/j.rbmo.2018.10.001)

38 Sumner AE, Finley KB, Genovese DJ, Criqui MH & Boston RC. Fasting triglyceride and the triglyceride-HDL cholesterol ratio are not markers of insulin resistance in African Americans. *Archives of Internal Medicine* 2005 **165** 1395–1400. (https://doi.org/10.1001/archinte.165.12.1395)

Received in final form 19 May 2022
Accepted 9 June 2022
Accepted Manuscript published online 10 June 2022