J. Biomed. Eng. Res.
Vol. 22, No. 4, 369-375, 2001

제2형 당뇨병의 위험인자 분석을 위한 다층 퍼셉트론과 로지스틱 회귀 모델의 비교

서 혜 숙 · 최 진 욱 · 이 홍 규 *

서울대학교 의과대학 의공학교실, 내과학교실

(2001년 2월 20일 접수, 2001년 6월 28일 채택)

A comparison of Multilayer Perceptron with Logistic Regression for the Risk Factor Analysis of Type 2 Diabetes Mellitus

Hyesook Suh, Jinwook Choi, Hongkyu Lee *

Department of Biomedical Engineering, Department of Internal Medicine *,
College of Medicine, Seoul National University.

(Received February 20, 2001. Accepted June 28, 2001)

요 약 : 통계적 회귀 모델은 임상 자료 분석에서 가장 흔히 사용되는 방법 중의 하나이다. 회귀 모델은 자료의 선형성, 부가성, 그리고 정규 분포라는 기본 가정을 전제로 한다. 그러나, 의학 분야에서 대부분의 생물학적 자료는 비선형적이고 정규 분포를 따르지 않는다. 이러한 통계 모델의 기본 가정과 실제의 생물 자료 사이에 존재하는 모순을 극복하기 위해, 우리는 인공 신경망을 기반으로 한 새로운 분석 기법을 제안하고자 한다. 120명(정상 60명, 환자 60명)의 자료로 학습시킨 다층 퍼셉트론을 개발했다. 새로운 시험 자료를 적용하였을 때, 모델의 성능은 약 0.76 정도를 보였다. 다층 퍼셉트론의 신경망 모델과 통계적인 로지스틱 회귀 모델을 각각 사용하여 당뇨병의 위험 인자 분석도 또한 실시하였다. 다층 퍼셉트론에서 유의한 인자는 식후 2시간 혈당, 성별(남자) 공복시 혈당, 연령, 수축기혈압, 복위 그리고 허리엉덩이둘레비로 나왔다. 회귀 모델에서는 성별(남자), 식후 2시간 혈당, 연령 그리고 공복시 혈당이었다. 성별과 연령에 따른 분석을 다층 퍼셉트론에서 실시하였다. 다른 임상 연구의 결과와 유사하게 총 콜레스테롤과 체중이 위험 인자로 나왔다. 본 연구에서 우리는 상용되어온 통계 기법에 의해 분석되지 않았던 생물학적 자료의 위험 인자 분석에 다층 퍼셉트론이 적용될 수 있음을 알았다.

Abstract : The statistical regression model is one of the most frequently used clinical analysis methods. It has basic assumption of linearity, additivity and normal distribution of data. However, most of biological data in medical field are nonlinear and unevenly distributed. To overcome the discrepancy between the basic assumption of statistical model and actual biological data, we propose a new analytical method based on artificial neural network. The newly developed multilayer perceptron(MLP) is trained with 120 data set (60 normal, 60 patient). On applying test data, it shows the discrimination power of 0.76. The diabetic risk factors were also identified from the MLP neural network model and the logistic regression model. The significant risk factors identified by MLP model were post prandial glucose level(PP2), sex(male), fasting blood sugar(FBS) level, age, SBP, AC and WHR. Those from the regression model are sex(male), PP2, age and FBS. The combined risk factors can be identified using the MLP model. Those are total cholesterol and body weight, which is consistent with the result of other clinical studies.

From this experiment we have learned that MLP can be applied to the combined risk factor analysis of biological data which can not be provided by the conventional statistical method.

Key words : Neural network, Multilayer perceptron, Logistic regression, Risk factor analysis, Type 2 diabetes mellitus(NIDDM)

본 연구는 서울대학교병원 임상의학연구소 임상연구지원을 받아 (과제명: 신경망을 이용한 비선형모델의 질병요인분석) 수행된 연구결과임

통신저자 : 최진욱, (110-799) 서울특별시 종로구 연건동 28번지
서울의대 의공학교실
Tel. (02)760-3421, Fax. (02) 762-8464
E-mail. jinchoi@snu.ac.kr

서 론

1. 당뇨병의 임상적 의의

당뇨병은 고혈당으로 대표되는 대사 이상과 이에 동반되는

여러 가지 만성 혈관 합병증을 특징으로 하는 질환이다. 당뇨병은 흔히 제1형 당뇨병(Insulin-dependent diabetes mellitus; IDDM)과 제2형 당뇨병(Noninsulin-dependent diabetes mellitus; NIDDM)의 두 가지로 분류된다. 말초 조직에서 인슐린에 대한 저항과 인슐린 분비의 이상에 의해 야기되어지는 제2형 당뇨병은 성인에서 발견되는 당뇨병의 대부분을 차지한다[1].

제2형 당뇨병은 일반적으로 40세 이후에 발병하고 합병증의 예방은 가능하나 일단 발생하면 치료가 어려워 위험 인자를 분석하여 예방하는 것이 중요하다. 제2형 당뇨병의 위험 인자에 대한 회귀 분석 연구 결과는 그 대상에 따라 다양하게 보고되어 왔다. 남아프리카에서 진행된 연구 결과에 의하면 나이, 허리둘레, 총 에너지 소비량, 당뇨병 가족력이 위험 인자로 분석되었으며[2], 중국에서는 비만, 지방도, 가족력을 위험 인자로 보고한 연구 결과가 있으며[3], Abdella와 Farmer 등의 연구들에서는 비만과 가족력 외에 고혈압, 당뇨병 가족력이 있는 남자 형제, 나이, 체질량지수(BMI) 30이상을 제2형 당뇨병의 위험 인자로 보고하였다[4,5]. 또 Haffner의 연구에 의하면 미국에 있는 다양한 인종과 민족의 인슐린 민감도를 비교한 결과 아프리카계, 라틴계, 본국인이 각각 백인계보다 2, 2.5, 5배씩 더 큰 제2형 당뇨병의 발병 위험도를 보였다[6]. 국내에서 세계보건기구의 표준화된 방법에 의한 역학 조사가 1993년 경기도 연천군에서 처음 실시되었고, 그 결과 30세 이상 인구 집단의 당뇨병 유병률은 7.2%로 보고되었고, 위험 인자로는 허리/엉덩이 둘레비, 혈청 중성지방 농도, 혈청 GOT 농도, 연령, 수축기 혈압, 당뇨병의 가족력, 거주 지역의 특성(도시화의 정도) 등이 관찰되었다. 같은 지역에서 1995년 다시 역학 조사가 실시되었고 유병률은 10.1%(남자 12.3%, 여자 8.2%)로 증가하였다[7].

Table 1. The description of the 17 variables used in the study
표 1. 연구에 사용된 17 변수들의 설명

symbol	full description	measuring unit
AGE	age	years
SEX	sex(male=0, female=1)	-
FBS	Fasting blood sugar	mg/dL
PP2	Post prandial 2hr	mg/dL
HT	height	cm
WT	Weight	kg
AC	Abdominal circumference	cm
HC	Hip circumference	cm
SBP	Systolic blood pressure	mmHg
DBP	Diastolic blood pressure	mmHg
GOT	Glutamic oxaloacetic transaminase	U/L
GPT	Glutamate pyruvate transaminase	U/L
CHOL	Total cholesterol	mg/dL
TG	Triglyceride	mg/dL
HDL	High density lipoprotein	mg/dL
BMI	Body mass index	Kg/m^2
WHR	Waist/hip ratio	-

2. 신경망의 개요

1911년 Ramony Cajal이 인간의 뇌 구성요소로 뉴론(neurons)이라는 개념을 소개한 것은 인간의 뇌를 이해하려는 노력의 선구자적 업적이었다. 그의 이론에 의하면 뇌는 매우 복잡하고 비선형적이며 병렬 컴퓨터와 유사한 정보 전달 체계로 구성되어 있으며, 그 작용기전의 최소단위는 뉴론으로 구성되어 있다[8]. 신경망은 1950년대에 Minsky 등에 의하여 개발된 인공지능 이론의 하나로 인간의 지능처럼 패턴을 분류하는 문제를 해결하고자 하는 목적에서 연구가 시작되었으며, 다른 인공지능 연구와 함께 1980년대에는 신경망 이론에 관한 연구가 활기를 띄게 되었다[9].

신경망(neural network)이란 경험적 지식을 저장하고 사용 목적의 이용에 타고난 성향을 가진 집적 병렬 처리 장치이며, 두 가지 점에서 인간의 뇌와 유사하다. 첫째는 지식이 학습과정을 통해 신경망의 구조에 저장된다는 점이며, 둘째는 시냅스 가중치(synaptic weights)로 알려진 뉴론간 연결 강도가 지식을 저장하기 위해 사용된다는 점이다[10].

3. 신경망의 의료 분야 활용

최근 인공 신경망은 다양한 연구 주제와 다양한 분야에서 활용되고 있다. 인공 신경망 기법은 진단, 영상, 파형 분석, 성과 예측, 병리 표본 동정, 그리고 임상 약리학 분야에서 주로 사용되고 있다고 William은 보고하였다[11]. Adi Armoni의 연구에 의하면 인슐린 의존형 당뇨병(제1형 당뇨병)의 경우, 신경망이 선형적 회귀 모델보다 더 정확하게 질병을 예측하였다고 보고하였다[12]. 다른 연구들을 보면 유방암 진단 예측이나 각 질병의 수술 후 예후에 대한 예측 모델, 기상 예측 모델, 핵의학과 방사선 영상 그리고 병리 표본의 특징을 추출한 진단 알고리즘, 그리고 기초의학 분야에서 약리학이나 분자 생물학, 유전학에의 도입 사례에 대한 연구 보고가 있다.

4. 연구의 목적

대부분의 임상가나 의학 연구자들은 위험인자 분석 연구에 회귀분석법을 이용하는데 익숙해 있다. 대부분의 회귀분석법은 선형성, 부가성, 그리고 정규 분포라는 3가지 가정을 전제로 한다[13]. 그러나 보건의료자료의 대부분이 비선형적이며, 정규분포를 따르지 않는 특징을 가지고 있다. 따라서 이러한 가정에 근거한 통계방법을 위의 비선형적인 특성을 지니는 자료가 대부분인 임상자료에 적용할 때에는 한계가 있음을 고려하여야 한다.

본 연구에서는 공학에서 패턴인식 등에 흔히 사용되는 신경망 기법들 중에서 가장 기본적인 다층 퍼셉트론을 이용하여 당뇨병의 예측 모델을 개발하고 예측력과 위험 인자 분석 결

과를 기존의 통계 기법인 로지스틱 회귀 분석과 비교 평가함으로써 향후 보건의료자료 분석에 사용될 수 있는 새로운 분석 방법론을 개발하고 검증함을 목적으로 하였다.

연구 재료 및 방법

1. 연구 재료

한국 성인에서 발생되는 제2형 당뇨병에 대한 임상 자료로서 1993년과 1995년에 30세 이상 주민에게 실시한 경기도 연천군 역학 조사의 자료를 본 연구의 기본 자료로 사용하였다[7]. 1993년에 정상인 피검자 1193명 중에서 1995년에 당뇨병으로 진단된 67명 중 60명과 정상으로 남은 1126명 중 60명의 자료를 선택하였고, 학습을 위한 입력 변수는 기본 검사 항목 17개로 선정하였다(Table 1).

먼저 신경망을 적용하기 위한 첫 단계로 각 변수의 분포를 확인한 후 사용할 자료의 범위를 0과 1사이로 각각 정규화(normalization)할 수 있는 선형 사상을 정하였으며, 양쪽 경계를 넘는 최소값은 0, 최대값은 1이 되도록 하였다. 93년도에 조사된 정상 자료를 입력값(input data)으로 사용하고, 해당 환자에서 2년 후인 95년에 이르러 당뇨병의 발현 여부를 목표값(target data)으로 사용하였다. 신경망 학습을 위하여 모두 120명의 자료가 사용되었으며, 학습과정이 어느 한쪽 자료만으로 연속적인 영향을 받는 것을 피하기 위하여 환자와 정상 자료 60개씩을 짝수와 홀수로 교대로 섞어 120개의 전체 혼합자료를 만들었다[14].

통계적 기법에서는 결손 값이 없는 120명의 자료를 정규화하지 않고 그대로 사용하였으며, 연속형 자료는 모두 범주형으로 변환하여 로지스틱 회귀 분석 모델을 세웠다(Fig. 1). 93년 정상 자료를 원인 변수로 사용하고, 95년 당뇨병의 진단 유무

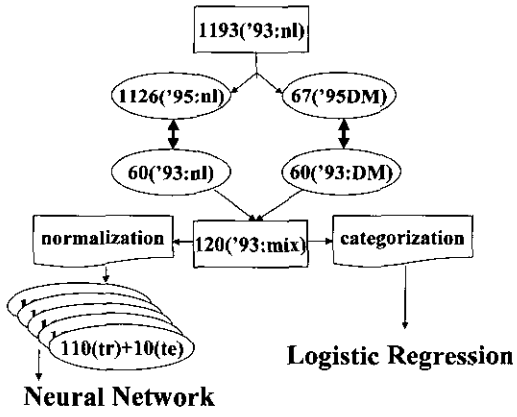

Fig. 1. The diagram of the data manipulation
nl : normal, DM : diabetes mellitus, tr : training data, te : test data
그림 1. 자료 사용 방법의 도식화
신경망 모델 : 자료의 표준화(0에서 1사이로), 로지스틱 회귀 모델 : 자료의 범주화

를 결과 변수로 사용하였다.

2. 연구 방법

1) 다층 퍼셉트론(multilayer perceptron, MLP)을 이용한 예측 모델 개발과 위험 인자 분석

본 연구에서 사용한 신경망의 구조는 입력층, 은닉층, 출력층의 3개 층으로 구성되었다. 은닉층의 전이 함수(transfer function)로는 -1부터 +1까지의 함수값을 갖는 tansigmoid 함수를 사용하였고, 출력층의 전이 함수로는 -∞부터 +∞까지의

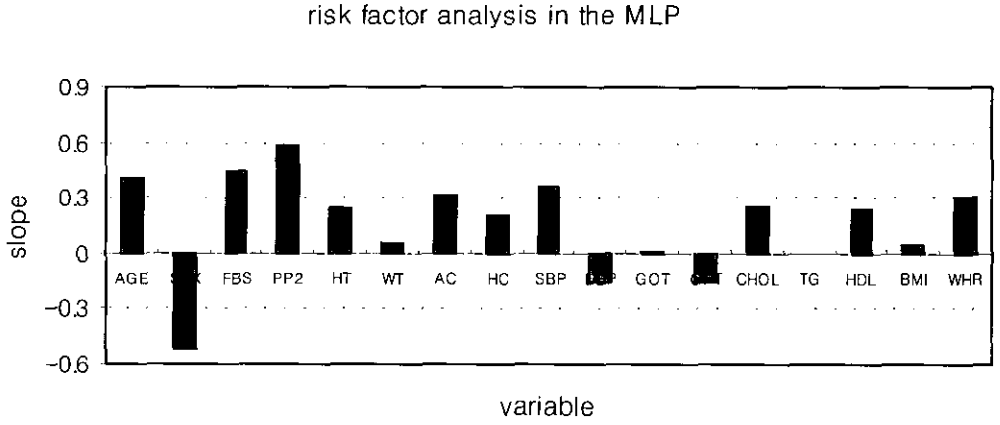

Fig. 2. The result of the risk factor analysis by the MLP model
그림 2. 다층 퍼셉트론을 이용한 위험 인자 분석의 결과
위험 인자 : 식후 2시간 혈당(PP2), 남자(SEX), 공복시 혈당(FBS), 연령(AGE), 수축기 혈압(SBP), 복위(AC), 허리/엉덩이 둘레비(WHR)

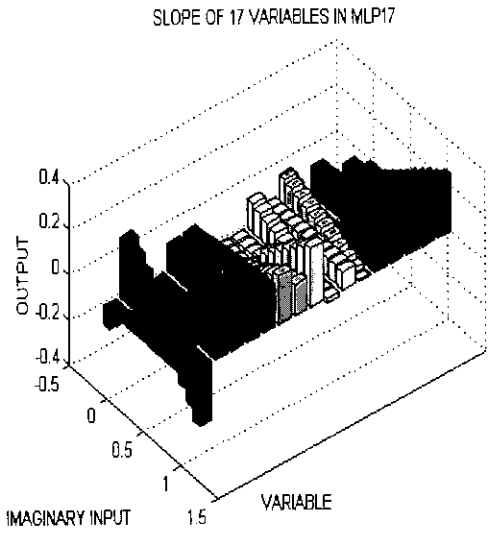

Fig. 3. 3-D graph of slopes predicted by the multilayer perceptron(MLP) model
그림 3. 다층 퍼셉트론 모델을 이용한 17 변수들의 3차원 기울기
기울기가 급경사일수록 위험 인자에 속한다.
양의 기울기는 입력 값이 클수록, 음의 기울기는 입력 값이 작을수록 위험 인자에 속한다.

함수값을 갖는 purelinear 함수를 사용하였다.

학습은 역전파 학습 알고리즘(back propagation learning algorithm)을 사용하여 연결 강도의 값을 계산하였다. 역전파 학습 알고리즘은 오차를 정정하는 방법에 의하여 학습이 이루어지는 모델로서, 입력에 대해 원하는 반응과 실제로 얻어진 결과값에 대한 차이를 신경망의 연결 강도에 반영하여 다음 번 계산 단계에서 오차를 줄여 나가는 학습 과정을 반복하게 된다[15]. 입력층에는 17개 변수를 입력받을 수 있도록 17개의 입력 노드를 구성하였으며, 은닉층에는 5개 노드를 구성하였다. 출력층은 정상과 환자의 이분성으로 분류하는 1개의 노드로 구성하였다. 이 때 과잉 학습(overtraining)으로 인한 성능 저하를 예방하기 위해, 5쌍에 대해 각 시험 자료(test data) 10명을 사용하여 최소의 에러(error=false positive + false negative)가 되는 최대 학습수(maximum epochs)를 찾아 신경망 모델의 학습 횟수로 선정하였다[14].

위험 인자 분석은 독립적 입력 변수 변환 방법을 사용하였다[16]. 이는 위와 같이 학습된 예측 모델(MLP)을 이용하여 분석하고자 하는 한 개의 변수만 선택하여 입력값이 최소값 0에서 최대값 1로 0.1씩 증가시켜 출력값의 변화를 살펴보는 것이다. 이 때 변화되는 1개의 입력 변수를 제외한 나머지 16개의 변수들은 모두 정규화된 자료의 중간값인 0.5로 고정시키고, 이와 같은 작업을 입력 변수 순서대로 진행시켜 입력 변수의 변화에 따라 출력 변수의 변화의 민감도를 측정하는 방법이다. 결과를 나타내는 지표로서는 입력값의 변화에 대한 출력값의 변화의 비를 나타내는 기울기를 사용하였고 기울기가 클수록 입력 변수의 출력에 대한 기여도가 큰 것으로 판단하였다. 각 17개 변수에 대한 기울기를 5쌍의 각기 다른 혼합자료에 대해 각각 구하고, 그 5쌍의 기울기들을 평균하여 최종 비교 지표로 사용하였다.

더 나아가 통계 기법의 층화 분석과 유사한 방법으로, 세밀한 분석을 위해 위의 기본 분석에서 높은 순위의 위험 인자이면서 임상적으로도 위험 인자로 보고된 변수들(성별, 연령)을 원하는 값에 고정하고 그 때의 다른 인자들의 기울기를 보는 방법도 시도하였다.

Table 2. The result using the logistic regression model.
표 2. 로지스틱 회귀 분석 모델의 결과
위험 인자 : 연령, 남자, 공복시 혈당(FBS)과 식후 2시간 혈당(PP2)

index / variable	95 % C.I	Probability	Estimate(OR)
AGE Ω 1	0.0997-3.1566	0.0420	1.5710(4.8115)
SEX B 1	0.4953-2.9079	0.0070	1.6409(5.1598)
FBS B 1	0.2746-2.2977	0.0142	1.2533(3.5019)
PP2 Ω 1	0.1805-3.0796	0.0316	1.5746(4.8288)
SBP Ω	-	-	-
BMI Ω	-	-	-
WHR Ω	-	-	-

AGE_Ω 1)66.5 yr., SEX_B 1 :male, FBS_B 1)103 ㎎/㎗, PP2 Ω 1)131.5 ㎎/㎗

Table 3. Summary of the risk factor analysis by the multilayer perceptron(MLP) and logistic regression(REG) model
표 3. 다층 퍼셉트론과 로지스틱 회귀 모델에 의한 위험 인자 분석의 요약
두 모델의 공통 위험 인자 : 연령, 남자, 공복시 혈당, 식후 2시간 혈당

model / order	MLP(slope≥0.30)	REG(OR)
1	PP2=0.5875	SEX=5.1598
2	SEX=-0.5214	PP2=4.8288
3	FBS=0.4473	AGE=4.8115
4	AGE=0.4116	FBS=3.5019
5	SBP=0.3585	-
6	AC=0.3156	-
7	WHR=0.2982	-

OR; odds ratio
Common risk factors of the two models were AGE, SEX(male), FBS and PP2.

2) 로지스틱 회귀 분석을 이용한 예측 모델 개발과 위험 인자 분석

의학 연구 자료 분석의 한 수학적 모형으로 결과 변수(질병 발생)에 관련되는 원인 변수의 다양한 조합의 관련 정도(발병 위험도)를 함수 식으로 표현한 것이 로지스틱 회귀 분석 모델(logistic regression model) 이다[17]. 본 연구에서는 이 모델을 세우기 위해 SAS 6.12 통계 패키지를 사용하였다. 각 원인

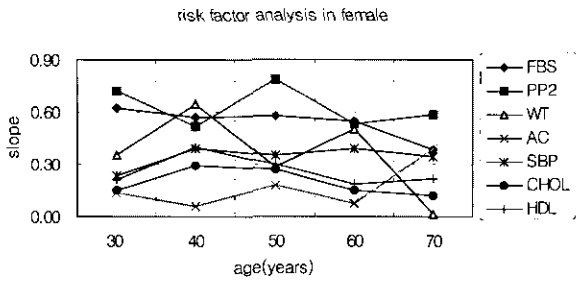

Fig. 4. The result of the risk factor analysis with the AGE in female by the MLP model
그림 4. 다층 퍼셉트론을 이용하여 여자에서 연령에 따른 위험 인자 분석의 결과
공복시 혈당(FBS)과 식후 2시간 혈당(PP2)은 나이에 상관없이 지속적으로 위험 인자에 속하는 반면, 복위(AC)는 60세 이후부터 위험 인자에 속하는 경향을 보여준다.

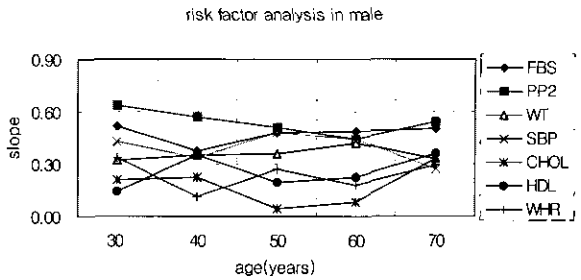

Fig. 5. The result of the risk factor analysis with the AGE in male by the MLP model
그림 5. 다층 퍼셉트론 모델을 이용한 남자에서 연령에 따른 위험 인자 분석의 결과
공복시 혈당(FBS), 식후 2시간 혈당(PP2), 수축기 혈압(SBP), 체중(WT)은 나이에 상관없이 지속적으로 위험 인자에 속하는 경향을 보여준다.

변수들 사이의 상호 작용 효과(interaction effect)를 알기 위해 범주형 자료에 대한 독립성 검정으로 카이 제곱 검정(Chi-Square test)을 실시하였다[18]. 17개의 예측 변수들을 선택하여 모델을 세우는 데 우도비 검정(likelihood ratio test)을 사용하였다[19]. 위험 인자 분석의 결과를 요약하는 지표로서는 구축된 모델의 추정치(estimate; β)를 사용하여 대응위험도(odds ratio, OR)를 다음과 같은 식에 의하여 계산하였다[20].

$$OR = e^{\beta}$$

결 과

1. 예측 모델들을 이용한 위험 인자 분석의 결과

1) 다층 퍼셉트론을 이용한 예측 모델과 위험 인자 분석의 결과

93년 정상인 자료를 사용하여 2년 후(95년) 당뇨병 발병을 예측하는 다층 퍼셉트론 모델에서 임의의 자료를 주었을 때 그 대상이 2년 후 제2형 당뇨병 환자로 발병할 가능성을 예측하는 모델의 성능은 민감도와 특이도를 이용한 ROC(Receiver Operating Curve)와 AUC(Area Under the Curve)를 사용하였고, AUC 약 76%로 나왔다.

제2형 당뇨병 발현에 영향을 미치는 위험 인자 분석을 위하여 본 연구에서는 기울기의 절대값을 소수 셋째 자리에서 반올림하여 약 0.30 이상인 변수들을 위험 인자로 보았다. 다층 퍼셉트론의 17변수 모델(MLP)에서 기울기의 절대값을 큰 것부터 순위에 따라 나열하면, 식후 2시간 혈당(PP2; post prandial 2 hours), 성별(SEX-> male), 공복시 혈당(FBS; fasting blood sugar), 나이(AGE), 수축기 혈압(SBP; systolic blood pressure), 복위(AC; abdomen circumference), 허리/엉덩이 둘레비(WHR; waist/hip ratio)의 순서이고, 나머지 10개 변수는 약 0.30 미만으로 작은 값을 보였다(Fig. 2).

위험 인자 분석에서 보여주는 음의 기울기는 입력 값이 증가함에 따라서 당뇨병의 발병 위험이 줄어듦을 의미한다. 성별의 경우 기본 자료에서 남자를 0으로, 여자를 1로 자료를 변환하였으므로 음의 기울기 값으로 절대값이 크게 나온 것은 남자의 경우가 더 당뇨 발병의 위험성이 큰 것을 의미한다. 17변수들에 대한 위험 인자 분석 결과 얻은 기울기를 3차원으로 나타낸 다층 퍼셉트론의 경우가 Fig. 3에 예시되고 있다.

2) 로지스틱 회귀 분석을 이용한 예측 모델과 위험 인자 분석의 결과

통계적 기법으로 세운 로지스틱 회귀 분석 모델(REG)의 예측 성능은 AUC 약 83%로 나왔고, 위험 인자 분석의 결과는 다음과 같았다.

95% 신뢰 구간(C.I.; confidence interval)에서 유의한 인자는 나이(AGE), 성별(SEX->male), 공복시 혈당(FBS), 식후 2시간 혈당(PP2)으로 나왔다(Table 2).

본 연구의 결과, 통계 모델(REG)에서 대응위험도가 높은 순위와 신경망 모델(MLP)에서 기울기의 절대값이 큰 순위는 동일하지 않았다. 그러나 순위는 달라도 통계에서 5% 유의수준으로 유의한 4개 인자(AGE, SEX, FBS, PP2)가 공통적으로 신경망 모델 및 통계 모델에서 동일하게 높은 위험 순위에 있음을 보여주었다(Table 3).

3) 신경망을 이용한 복합 조건에서의 위험 인자 분석

신경망 모델을 이용하여 임상적으로 궁금한 여러 인자가 복합적으로 작용하는 상황에서 다른 인자들의 위험 인자 분석을 시행하였다. 복합된 상황의 예로 성별 조건과 40세 이후의 연령층에서 당뇨병 발병에 영향을 미치는 위험 인자에 대하여 신경망을 통하여 분석하였다.

그 결과 다층 퍼셉트론 모델에 의하면 성별, 연령별의 모든 경우에서 공복시 혈당(FBS)과 식후 2시간 혈당(PP2)은 공통

적으로 당뇨병 발병에 영향을 미치는 위험 인자로 나왔다. 여자의 경우에는 공복시 혈당(FBS)과 식후 2시간 혈당(PP2) 외에 수축기 혈압(SBP)과 체중(WT)도 위험 인자로 나타났다(Fig. 4). 수축기 혈압은 30세에는 기울기가 0.30미만이지만 40세 이후에는 계속 0.30이상의 높은 기울기를 보이고 있어 중년 이후 고혈압이 제2형 당뇨병의 위험 인자임을 보여주고 있다. 체중 또한 위험 인자임을 보여주고 있으며, 특히 40세 여자에서 최대 위험 인자임을 보여주고 있다. 복위(AC)는 중년기에는 위험 인자로 인식되지 않다가 60세 이후에서 위험 인자로 나타났다.

남자의 경우 공복시 혈당(FBS)과 식후 2시간 혈당(PP2) 외에 수축기 혈압(SBP)과 체중(WT)도 위험 인자로 나타난 것은 여자의 경우와 같지만 양상은 약간 다르다(Fig. 5). 여자의 경우 연령에 따라 기울기가 크게 변하지만 남자의 경우에는 수축기 혈압과 체중 모두 중년 이후에서 일관된 형태의 위험 인자 특성을 보여주고 있다.

고 찰

제2형 당뇨병은 현대 성인에게 흔한 질병으로 국내에서 연간 발생률은 2.5%(남자 3.2%, 여자 1.5%)로, 유병률은 10.1%(남자 12.3%, 여자 8.2%)로 각각 1996년에 보고되었다[1]. 그러나 성인에서 흔히 발견되는 제2형 당뇨병의 원인으로는 스트레스를 비롯하여 체중, 혈압 등 환경적인 요인과 함께 가족력과 같은 유전적인 요인이 복합적으로 작용하는 것으로 보고되고 있다. 본 연구는 이러한 복잡 원인 관계에 있는 당뇨병의 원인과 그 발병 예측에 관한 상관 관계를 찾아보기 위하여 일반적인 통계 방법에 의하지 않고 비통계적인 방법으로 신경망을 이용하는 방법들을 제시하였다.

신경망이 의료분야에 활용되는 사례는 William, Armoni 등과 같은 사례가 있었으나 아직 본격적으로 활용되지는 않고 있으며, 최근 Frize의 연구 등과 같이 임상 의사 결정 지원을 위한 방법으로 활용하는 사례와 같이 현재에도 지속적으로 연구가 진행되고 있다[21].

본 연구의 대상은 93년과 95년도에 연천군에 거주한 사람들을 대상으로 하였기 때문에 한국인 전체를 일반화 할 수 있는 대표성을 지니지 못하는 점과 전체 자료 1193명 중에서 당뇨로 발현된 사람 60명과 발현되지 않은 사람 60명만을 사용하여 신경망의 학습이 이루어졌기 때문에 자료를 충분히 활용하지 못한 제약점을 지니고 있다.

신경망의 경우 예측력이 약 76%로 로지스틱 회귀 모델의 83%에 비하여 낮은 수치를 나타냈는데, 위에서 언급한대로 환자 60명과 정상인 60명만을 사용하여 학습 자료가 적은 점과 또한 회귀 모델에서는 120명 모두 학습 자료로 사용한데 비해, 신경망 모델에서는 학습 자료 110명과 시험 자료 10명을 분리하여 신경망의 학습이 이루어진 점등이 예측력이 낮게 나온 이유로 들 수 있다.

통계적 로지스틱 회귀 모델에서는 나이, 성별(남자), 공복시 혈당, 식후 2시간 혈당을 위험 인자로 설명하고 있다. 즉 66세 이상(p=0.0420), 남자(p=0.0070), 공복시 혈당이 103 mg/dL 이상(p=0.0142) 그리고 식후 2시간 혈당이 131.5 mg/dL 이상(p=0.0316)인 피검자는 현재 정상 범위에 속하지만, 향후 2년 뒤 제2형 당뇨병(인슐린 비의존성 당뇨병)으로 발병할 가능성이 더 높다고 할 수 있다. 대응위험도(odds ratio, OR)는 공복시 혈당이 103 mg/dL 이상인 경우 당뇨병으로 발현할 가능성이 발현되지 않을 가능성에 비하여 약 3.5배이고, 나머지 3개 위험 인자는 약 5배 정도로 높았다.

신경망을 이용한 위험 인자 분석 기법으로 기울기를 구한 결과, 통계 모델에서 대응위험도가 높은 순위와 신경망 모델(MLP)에서 기울기의 절대값이 큰 순위 사이에는 통계에서 유의한 4개 인자(AGE, SEX, FBS, PP2)가 모두 공통적으로 높은 위험 순위에 있음을 보여주었다.

통계에서는 유의하지 않던 인자들 중에도 신경망에서는 높은 위험 순위로 보여주고 있어 주목할 만하다. 즉 두 모델에서 공통적인 4개 인자를 제외한 위험 인자들은 수축기 혈압(SBP), 복위(AC) 그리고 허리/엉덩이 둘레비(WHR)의 변수였다. 그러나 신경망 모델의 특징적인 단점이 그러하듯 출력값을 내기 위해 그 안에서 일어나는 과정을 명확히 논리화하기 어렵다.

그러나 본 연구에서 예측한 결과들은 통계적인 방법으로는 유의성이 검증되지 않았으나, 다른 연구 결과에서 뒷받침하는 결과를 보여준다. 다른 연구에서 유의한 위험 인자로 보고된 바와 같이 나이가 많을 수록, 남자의 경우, 수축기 혈압이 높을수록, 복위가 클수록, 허리/엉덩이 둘레비가 클수록 제2형 당뇨병으로 발병할 가능성이 높은 것으로 나왔으며 본 연구에서도 위험 인자로 분석 결과가 나왔다.

한편, 신경망을 이용한 위험 인자 분석의 의의는 전체 환자의 자료가 60건 밖에 없는 경우에, 통계적 모델에 의하면 변인이 17개나 되기 때문에 각 변인의 결과에 대한 설명 정도가 강건하다고 판단하기 어렵다. 그러나 신경망 모델에서는 자료를 활용하여 이미 만들어진 모델을 이용하여 임의의 주어진 값에 대한 출력값을 관찰하기 때문에 전체 자료 크기가 작은 경우에도 영향을 적게 받는다고 할 수 있다. 위의 방법으로 다층 퍼셉트론에서 임의의 복합적인 상황을 선정하여 연령별, 성별 분석을 다시 한 결과 공복시 혈당, 식후 2시간 혈당, 수축기 혈압, 복위, 허리/엉덩이 둘레비 외에도 체중(WT)이나 총 콜레스테롤도 위험 인자로 나타났다. 이는 임상적으로 한 가지 인자만 작용한다기보다 여러 인자가 함께 상호 작용으로 성인병 발병에 기여한다고 밝힌 다른 임상 연구와도 유사한 결과를 보여주고 있다[22].

결 론

본 연구에서는 다층 퍼셉트론을 이용한 예측 모델의 개발과 그를 이용한 제2형 당뇨병 발현에 영향을 미치는 위험 인자를 분석하였으며, 기존의 위험 인자 분석 기법으로 널리 사용되는

회귀 분석 기법과 비교해 보았다. 예측 모델의 성능은 자료 이용의 특성상 예상한대로 성능 개선을 위한 노력에도 불구하고 다층 퍼셉트론이 통계적인 회귀 모델에 비하여 예측력이 더 떨어졌으나, 신경망을 이용한 위험 인자 분석에서는 통계 결과로 알 수 없었던 다른 위험 인자들을 확인할 수 있었다. 또한 구축된 신경망 모델을 이용하여 임상적으로 상호 작용할 것으로 보이는 여러 인자의 특정 복합된 조건하에서 당뇨병이 발현에 영향을 미치게 되는 상황을 재현할 수 있었으며 이를 통하여 얻은 결과들이 기타 임상 연구와 일치되는 분석 결과를 보여 주었다.

본 연구를 통하여 복잡한 생물학적 인과 관계를 지니는 의학 연구에서 신경망 모델을 이용하여 통계적으로 찾아 내지 못하였던 위험 인자를 찾아낼 수 있는 변별력을 확인할 수 있었으며, 추후 자료의 규모 등이 보강되면 좀 더 좋은 예측 모델을 개발할 수 있을 것으로 기대된다.

참 고 문 헌

1. 김응진, 민헌기, 최영길, 이태희, 허갑범, 신순현. 당뇨병학. 제 2판. 고려의학 : pp.22-5, 275-8, 1998
2. Levitt NS, Steyn K. Lambert EV, Reagon G, Lombard CJ, Fourie JM, Rossouw K, Hoffman M. "*Modifiable risk factors for Type 2 diabetes mellitus in a peri-urban community in South Africa*". Diabet Med ;16(11):pp.887-8, Nov. 1999
3. Zhou Y, Fan N, Yan J. "*A study on the risk factors of none-insulin-dependent type diabetes mellitus*". Chung Hua Liu Hsing Ping Hsueh Tsa Chih ;19(3): pp.152-3, Jun. 1998
4. Abdella N, Al Arouj M, Al Nakhi A, Al Assoussi A, Moussa M. "*Non-insulin-dependent diabetes in Kuwait: prevalence rates and associated risk factors*". Diabetes Res Clin Pract ;42(3):pp.187-96, Dec. 1998
5. Farmer AJ, Levy JC, Turner RC. "*Knowledge of risk of developing diabetes mellitus among siblings of Type 2 diabetic patients*". Diabet Med ;16(3):pp.233-7, Mar. 1999
6. Haffner SM. "*Epidemiology of type 2 diabetes: risk factors*". Diabetes Care ;21 Suppl 3:pp.C3-6, Dec. 1998
7. Shin CS, Lee HK, Koh CS, Kim YI, Shin YS, Yoo KY, Paik HY, Park YS, Yang BG. "*Risk factors for the development of NIDDM in Yonchon County, Korea*". Diabetes Care; vol 20, no 12:pp.1842-6, Dec. 1997
8. S. Haykin. *Neural Networks*:Macmillan, :pp.1-2, 363-365, 1994
9. Mattew Z. "*Neural Networks in Artificial Intelligence*". Ellis Horwood Ltd:pp.15-16, 1990
10. 김화수, 조용범, 최종욱. 전문가 시스템. 집문당:pp.253, 1995
11. William G. Baxt. "*Application of artificial neural networks to clinical medicine*". Lancet;346:pp.1135-8, 1995
12. Adi Armoni. "*Use of Neural Networks in Medical Diagnosis*". M.D. Computing; Vol 15, no 2,:pp.100-4, 1998
13. Frank E. Harrell, Jr, Kerry L. Lee, David B. Matchar, and Thomas A. Reichert. "*Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions*". Cancer Treatment Peprots ;Vol 69, no 10:pp.1071-7, 1985
14. N. K. Bose, P. Liang. "*Neural Network Fundamentals with Graphs*", Algorithms, and Applications.:pp.240-5, 1996
15. 김대수. 신경망 이론과 응용(I). 하이테크정보.:pp.92-3, 1996
16. Kyong-Sik Om, Hee-Chan Kim, Byoung-Goo Min, Chan-Soo Shin, and Hong-Kyu Lee. "*Statistical RBF network with applications to an expert system for characterizing diabetes mellitus*". EEIS transactions,; pp.355-65, 1997
17. 안윤옥, 유근영, 박병주. 실용 의학통계론:pp.169-81, 1997
18. 김우철, 김재주, 박병욱, 박성현, 송문섭, 이영조, 전종우, 조신섭. 일반통계학. 영지문화사. :pp.293-300, 1997
19. Neter, Kutner, Nachtsheim, "*Wasserman. Applied Linear Regression Models*". 3rd ed.:pp.585-90, 1990
20. 신영수, 안윤옥. 의학연구방법론. 서울대학교출판부.:pp.123-47, 159-68, 1997
21. M. Frize, C.M. Ennett, MASc. "*Improving the Potential Clinical Significance of Decision-Support Systems Using Artificial Neural Networks. Proc*". AMIA Annual Fall Symphosium.:pp.1011, 2000
22. Joe A. Florence, Bryan F. Yeager, Pharm D. "*Treatment of Type 2 Diabetes Mellitus*". AFP.:pp. 2835-42 , May. 1999
23. Lucila Ohno-Machado, Donald Bialek. "*Diagnosing Breast Cancer from FNAs:Variable Relevance in Neural Network and Logistic Regression Models*". MEDINFO98.:pp.537-40, 1998
24. 최진욱. 신경망을 이용한 유방암예측 모델의 개발. 대한의료정보학회지 제4권 제 1호:pp.83-7, 1998