مجله‌ی غدد درون‌ریز و متابولیسم ایران
دانشگاه علوم پزشکی و خدمات بهداشتی ـ درمانی شهید بهشتی
دوره‌ی دوازدهم، شماره‌ی ۱، صفحه‌های ۲۴ ـ ۱۶ (اردیبهشت ۱۳۸۹)

مقاله‌ی پژوهشی

پیش‌بینی ابتلا به دیابت با استفاده از رگرسیون منطقی

دکتر یداله محرابی[۱]، پروین سربخش[۲]، دکتر فرزاد حدائق[۳]، دکتر علی‌اکبر خادم معبودی[۲]

۱) گروه اپیدمیولوژی، دانشکده‌ی بهداشت، ۲) دانشکده‌ی پیراپزشکی، ۳) مرکز تحقیقات پیشگیری از بیماری‌های متابولیک، پژوهشکده‌ی علوم غدد درون‌ریز و متابولیسم، دانشگاه علوم پزشکی و خدمات بهداشتی ـ درمانی شهید بهشتی، **نشانی مکاتبه‌ی نویسنده‌ی مسئول**: تهران، اوین، دانشگاه علوم پزشکی شهید بهشتی، دانشکده بهداشت، گروه اپیدمیولوژی، دکتر یداله محرابی؛ e-mail:ymehrabi@gmail.com

چکیده

مقدمه: دیابت نوع ۲ یکی از بیماری‌های چندعاملی است که با توجه به اهمیت و بار فردی و اجتماعی، لزوم شناسایی افراد پرخطر برای ابتلا به آن مشهود است. تاکنون مطالعه‌های متعددی برای پیش‌بینی بروز دیابت با استفاده از مدل‌های آماری موجود انجام شده است ولی با وجود اهمیت بالینی اثر متقابل عوامل خطرساز بر بروز دیابت، امکان لحاظ کردن همه‌ی اثرهای متقابل ممکن در مدل‌های آماری فعلی وجود ندارد. در این مطالعه، به منظور یافتن ترکیبات منطقی مناسب از عوامل خطرساز مرتبط با دیابت نوع ۲ از روش رگرسیون لجستیک منطقی استفاده شد. مواد و روش‌ها: جمعیت مورد بررسی، از افراد بخش کوهورت مطالعه‌ی قند و لیپید تهران (TLGS) انتخاب شدند. ۳۵۲۳ نفر (۵۷/۸٪ زن و ۴۲/۲٪ مرد) وارد مطالعه شدند. تحلیل‌های مربوط با استفاده از روش رگرسیون لجستیک منطقی (Logic Logistic Regression) انجام شد. پارامترهای مدل با به کارگیری الگوریتم Annealing برآورد شد. به منظور اجتناب از بیش برآورد شدن، تعداد بهینه‌ی ترکیبات منطقی و متغیرهای مدل به روش اعتبار متقاطع تعیین شد. برای ارزیابی و مقایسه‌ی مدل منطقی به دست آمده با رگرسیون لجستیک حاصل از اثر اصلی، آماره‌ی انحراف، میزان حساسیت و ویژگی دو مدل محاسبه شد. هم‌چنین، مقایسه‌ی میزان پیش‌بینی بروز دیابت توسط مدل‌ها با استفاده از سطح زیر منحنی مشخصه‌ی عملکرد (ROC) انجام شد. نرم‌افزار R نسخه ۲/۸/۱ برای انجام تحلیل‌ها مورد استفاده قرار گرفت. یافته‌ها: با استفاده از الگوریتم Anealing مدل رگرسیون لجستیک منطقی با ۴ ترکیب بولی شامل ۵ متغیر برازش داده شد. آماره‌ی انحراف این مدل ۱۲۰۳/۳ به‌دست آمد که نسبت به مدل‌های دیگر و نیز مدل لجستیک پیشرو (آماره‌ی انحراف = ۱۲۰۶/۸۸) برازش بهتری داشت. عبارات بولی یافت شده در مدل با ۴ ترکیب شامل اختلال تحمل قند ناشتا (نسبت بخت: ۵/۵۳ و فاصله‌ی اطمینان ۹۵٪: ۷/۵۹ و ۴/۰۳)، اختلال تحمل قند دوساعته (نسبت بخت: ۵/۴۵ و فاصله‌ی اطمینان ۹۵٪: ۷/۴۹ و ۳/۹۶)، داشتن سابقه‌ی فامیلی دیابت با (نسبت بخت: ۱/۸۹و فاصله‌ی اطمینان ۹۵٪: ۲/۶۳ و ۱/۳۸)، تری‌گلیسرید بالا یا دور کمر بالا (نسبت بخت: ۲/۴ و فاصله‌ی اطمینان ۹۵٪: ۳/۳۲ و ۱/۷۳) بود (برای همه متغیرها P<۰/۰۰۱). سطح زیر منحنی راک مربوط به مدل ۰/۸۴۳ با فاصله‌ی اطمینان ۹۵٪ مجانبی (۰/۸۷۴ و ۰/۸۱۳) به دست آمد که با سطح زیر منحنی راک مدل لجستیک پیشرو (۰/۸۳۹) تفاوت معنی‌داری نداشت. نتیجه‌گیری: به نظر می‌رسد رگرسیون منطقی به عنوان یک روش جدید، قابلیت خوبی برای غربالگری بیماری‌های چند عاملی از جمله دیابت دارد زیرا در این رگرسیون، امکان شناسایی و لحاظ کردن اثر متقابل بین عوامل خطرساز وجود دارد.

واژگان کلیدی: رگرسیون منطقی، دیابت، اثرهای متقابل، پیش‌بینی بروز، منطق بولی، الگوریتم Annealing

دریافت مقاله: ۸۸/۴/۱ ـ دریافت اصلاحیه: ۸۸/۸/۲۰ ـ پذیرش مقاله: ۸۸/۹/۸

مقدمه

یکی از بیماری‌های چندعاملی که غربالگری آن در جامعه از اهمیت خاصی برخوردار است دیابت نوع ۲ است. عوامل خطرساز در ابتلا به دیابت نوع ۲ اضافه وزن و چاقی، بی‌تحرکی یا کم‌تحرکی، رژیم غذایی با چربی بالا و فبیر کم، نژاد، سابقه‌ی فامیلی، سن، وزن کم هنگام تولد و غیره است. هر چه تعداد عوامل خطرساز در فرد بیشتر باشد بیشتر در معرض خطر ابتلا به دیابت قرار می‌گیرد.[۱] افزایش شیوع دیابت نوع ۲ در همه‌ی دنیا به خصوص در کشورهای در حال توسعه از جمله ایران، نوعی اعلام خطر است.

دیابت نوع ۲ در بیشتر جوامع به یک اپیدمی تبدیل شده است و شواهد اپیدمیولوژی نشان می‌دهد که اگر اقدام‌های پیشگیرانه‌ی مؤثری انجام نشود، شیوع دیابت به طور جهانی افزایش خواهد یافت.[۱] براساس برآوردها، انتظار می‌رود در سال ۲۰۵۰ تعداد افراد دیابتی در دنیا به بیش از ۳۳۰ میلیون نفر برسد که این تعداد دوبرابر تعداد دیابتی‌ها در سال ۲۰۰۰ خواهد بود. هم‌چنین، بسیاری از موارد جدید ابتلا به دیابت مربوط به کشورهای در حال توسعه است که به نظر می‌رسد خاورمیانه بیشترین افزایش را در شیوع دیابت در سال ۲۰۳۰ خواهد داشت.[۱] تغییر عمده و سریع در سبک زندگی مردم این کشورها باعث افزایش شیوع چاقی و سایر عوامل خطرساز بیماری‌های غیر واگیر مانند فشارخون بالا و اختلال در چربی شده است که در سراسر دنیا به عنوان عمده‌ترین عوامل سبب‌شناختی مربوط به بروز دیابت نوع ۲ شناخته شده‌اند.[۲]

شناخت عوامل خطرساز مؤثر در بروز دیابت یک اقدام اساسی برای برنامه‌های پیشگیری از دیابت در هر جامعه‌ای است چرا که کاهش دادن این عوامل خطرساز باعث کاهش نرخ بروز دیابت خواهد شد. در این میان، یافتن معادله‌ای برای تعیین اثر عوامل خطرساز و شدت ارتباط آنها با ابتلا به دیابت، دارای اهمیت فراوان است.

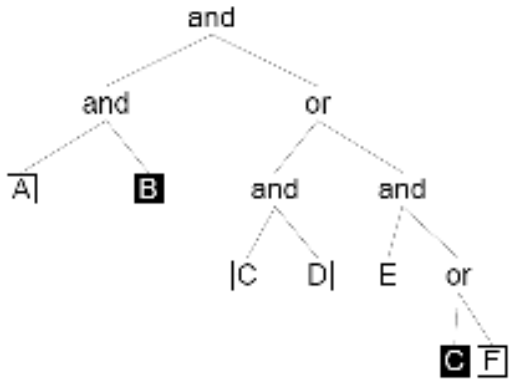
با توجه به اهمیت و بار فردی و اجتماعی این بیماری، لزوم شناسایی افراد در معرض خطر برای ابتلا به دیابت مشهود است. تاکنون، مطالعه‌های متعددی برای پیش‌بینی بروز دیابت با استفاده از مدل‌های آماری موجود، انجام شده است ولی در این مطالعه‌ها به دلیل تعدد عوامل خطرساز مؤثر در بروز دیابت، در مدل‌های چندگانه فقط اثرهای اصلی این عوامل بررسی شده است و باوجود اهمیت بالینی اثرهای متقابل این عوامل در بروز دیابت، امکان لحاظ کردن همه‌ی اثرهای متقابل ممکن، به‌ویژه در حالتی که تعداد قابل توجهی عامل خطرساز مورد بررسی قرار می‌گیرد، در مدل‌های معمول و رایج فعلی وجود ندارد.

در این مطالعه به منظور سادگی استفاده‌ی بالینی، هر کدام از عوامل خطرساز مرتبط با دیابت نوع ۲ به شکل دو حالتی تعریف شد و برای یافتن ترکیبات منطقی مناسب از متغیرها، از رگرسیون لجستیک منطقی[i] استفاده شد تا با در نظر گرفتن اثرهای برهم‌کنشی متغیرها، قدرت پیش‌بینی بهتری برای بروز دیابت به دست آید. هم‌چنین، روش مرسوم رگرسیون لجستیک با استفاده از خود متغیرها به عنوان متغیر پیش‌بینی کننده برای پیش‌بینی ابتلا به دیابت به کار گرفته شد و قدرت پیش بینی این مدل با روش رگرسیون لجستیک منطقی مقایسه شد.

رگرسیون منطقی[ii] یک روش رگرسیونی تعمیم یافته و جدید است که در آن متغیرهای پیش‌گو به صورت ترکیبات بولی از متغیرهای دو حالتی ساخته می‌شود. در رگرسیون منطقی، درصدد یافتن متغیرهای دو حالتی هستیم که حاصل ترکیب منطقی بولی مطلوب از متغیرهای دو حالتی اولیه هستند به طوری که استفاده از این ترکیبات جدید به عنوان متغیرهای پیش‌بینی‌کننده در مدل رگرسیونی، بهترین برازش را برای متغیر پاسخ داشته باشد. معیار مطلوب بودن در این جستجو، کمتر بودن تابع امتیاز[iii] متناسب با مدل رگرسیونی مورد نظر است. از روش رگرسیون منطقی برای برازش انواع مدل‌های رگرسیونی از جمله خطی، لجستیک و مدل مخاطرات کاکس می‌توان استفاده کرد.[۳]

هر ترکیب بولی، به صورت یک درخت منطقی[iv] متشکل از برگ‌هایی که متغیرهای مطلوب هستند، نشان داده می‌شود. برای مثال ترکیب $\{(A \wedge B^c) \wedge [(C \wedge D) \vee (E \wedge (C^c \vee F))]\}$ را که در آن ^ علامت "و"، $\vee$ علامت "یا" و c علامت "نه" (متضاد) هستند، می‌توان با یک درخت منطقی به‌صورت شکل ۱ نمایش داد.

i- Logic Logistic Regression
ii- Logic Regression
iii- Score function
iv- Logic Tree



شکل ۱- درخت منطقی عبارت
$\{(A \wedge B^c) \wedge [(C \wedge D) \vee (E \wedge (C^c \vee F))]\}$
مربع‌هایی که با رنگ سفید نشان داده شده‌اند به معنی حضور عامل و مربع‌های با رنگ تیره نشان‌دهنده‌ی عدم حضور عامل هستند. به عنوان مثال، شاخه‌ی اول بیان‌گر حضور عامل A و عدم حضور عامل B است.

درخت مذکور شامل عمل‌گرها (and, or, not) و برگ‌ها است. هر مربع نماینده‌ی یک برگ است و زمینه‌ی تیره برای برگ‌ها نشان‌دهنده‌ی متضاد بودن با حرف داخل آن است. رگرسیون منطقی تاکنون بیشتر برای داده‌های ژنی و SNP[i] به کار رفته است چرا که اثرهای متقابل SNPهای مختلف با یکدیگر، می‌تواند تأثیر قابل توجهی در بروز بیماری‌های مرتبط با ژنتیک داشته باشد. یافتن این نوع تقابل‌ها بسیار ارزشمند و حیاتی است. برای غربال‌گری بیماری‌های چند عاملی نیز به دلیل وجود اثر متقابل بین عوامل خطرساز متعدد، معمولاً رگرسیون منطقی یافته‌های قابل قبولی در پی دارد.[۳]

هدف از این مطالعه بررسی اثر متقابل بین عوامل خطرساز مختلف دیابت است به‌طوری که با شناسایی تقابل‌های موجود و گنجاندن آنها در مدل رگرسیون لجستیک منطقی، بتوانیم مدلی را ارایه دهیم که قابلیت پیش‌بینی بهتری برای ابتلا به دیابت داشته باشد.

مواد و روش‌ها

جمعیت مورد بررسی، افراد مطالعه‌ی قند و لیپید تهران (TLGS)[ii] بودند. این مطالعه در سه مرحله انجام شد. مرحله‌ی اول که مطالعه‌ی مقطعی بود از اسفند ۱۳۷۷ تا شهریور ۱۳۸۰ به طول انجامید و تعداد ۱۵۰۰۵ نفر از افراد بالای ۳ سال منطقه‌ی ۱۳ تهران در آن شرکت کردند (فاز ۱). سپس، افراد وارد فاز ۲ مطالعه شدند که مطالعه‌ای آینده‌نگر و شامل دو قسمت هم‌گروهی و مداخله‌ای بود و از مهر ۱۳۸۰ شروع شد و در شهریور ۱۳۸۴ به پایان رسید و بالاخره فاز سوم که از سال ۱۳۸۴ شروع و تا اسفند ۸۶ ادامه پیدا کرد.[۴]. افرادی که قند خون ناشتا بیشتر یا مساوی ۱۲۶ میلی‌گرم در دسی‌لیتر یا قند دوساعته‌ی بیشتر از ۲۰۰ میلی‌گرم در دسی‌لیتر داشتند یا داروی ضد دیابت مصرف می‌کردند به عنوان دیابتی در نظر گرفته شدند.[۵]

از بین ۶۴۳۷ فرد بالای ۲۰ سال گروه کوهورت، بعد از کنار گذاشتن افراد دیابتی فاز یک یا کسانی که داده‌های فاز یک آنها موجود نبود (۶۹۸ نفر) و ۶۲۵ نفر با داده‌ی گمشده قندخون ناشتا یا دوساعته، ۵۱۱۴ فرد غیردیابتی باقی ماند که حدود ۷۰٪ آنها در فازهای بعدی پیگیری شدند. موارد جدید ابتلا به دیابت، به افرادی اطلاق شد که در فاز یک غیر دیابتی بوده و تا انتهای فاز۲ یا ۳ به عنوان فرد دیابتی تشخیص داده شده‌اند. تعداد این افراد ۲۲۵ نفر بود. از افراد غیر دیابتی، ۳۳۱۰ نفر و از ۲۲۵ مورد جدید دیابت، ۲۱۳ نفر تمام متغیرهای مورد بررسی را دارا بودند. در مجموع، ۳۵۲۳ نفر وارد تحلیل آماری شدند.

متغیرهای مورد بررسی بر اساس مقادیر پایه‌ی افراد (مقادیر فاز یک مطالعه)، براساس تعریف عوامل خطرساز[۶] به متغیرهای دو حالتی (عامل خطرساز دارد/ ندارد) تبدیل شدند (جدول ۱).

برای مقایسه‌های مربوط به شاخص‌های زنان و مردان از آزمون‌های تی و مجذور خی استفاده شد. برای پیش‌بینی وضعیت ابتلا به دیابت، از رگرسیون منطقی با تابع پیوند لجستیک[iii] استفاده شد که در آن، آماره‌ی انحراف[iv] به عنوان تابع امتیاز، تعریف شد. متغیرهای مدل با به کارگیری الگوریتم‌های Annealing و Greedy برآورد شد. هم‌چنین، به منظور اجتناب از بیش برآورد شدن مدل، تعداد بهینه‌ی ترکیبات منطقی و متغیرهای مدل، به روش اعتبار متقاطع[v] تعیین شد.

i- Single Nucleotide Polymorphisms
ii- Tehran Lipid and Glucose Study

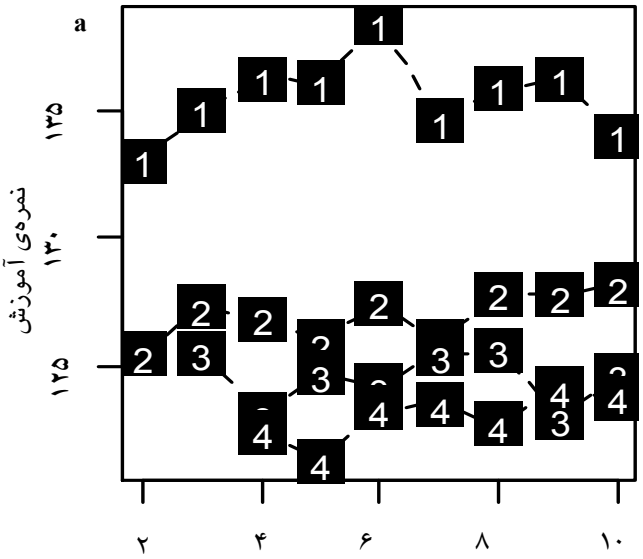
iii- Logistic Link function
iv- Deviance
v- Cross Validation Test

جدول ۱- شیوه‌ی دوحالتی کردن متغیرهای پژوهش

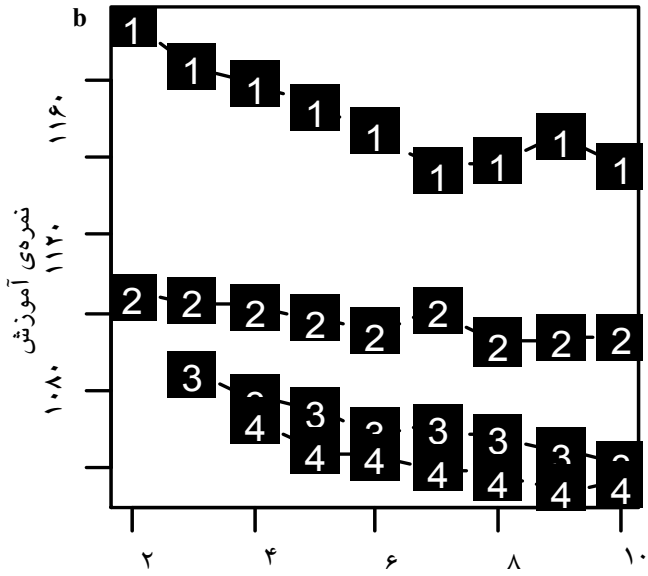
نماد	متغیر	حالت ۰	حالت ۱
DM	دیابت	سالم	دیابتی
SEX	جنس	زن	مرد
ACTIVITY	فعالیت فیزیکی	فعالیت معمولی	فعالیت کم
BMI	چاقی تنه‌ای	نمایه‌ی توده‌ی بدن زیر ۲۵ (کیلوگرم بر مترمربع)	نمایه‌ی توده‌ی بدن ۲۵ یا بالای آن (کیلوگرم بر مترمربع)
IFG	اختلال تحمل قندخون ناشتا	قند خون ناشتای کمتر از ۱۰۰	قندخون ناشتای بیشتر یا مساوی ۱۰۰ و کمتر از ۱۲۶ (میلی‌گرم بر دسی‌لیتر)
IGT	اختلال تحمل قندخون ۲ ساعته	قند خون دو ساعته‌ی کمتر از ۱۴۰	قندخون دو ساعته‌ی بیشتر یا مساوی ۱۴۰ و کمتر از ۲۰۰ (میلی‌گرم بر دسی‌لیتر)
AGE	سن	زیر ۴۰ سال	سن بیشتر یا مساوی ۴۰ سال
FH-DM	سابقه‌ی فامیلی	بدون سابقه‌ی فامیلی دیابت	دارای سابقه‌ی فامیلی دیابت
BP	فشارخون بالا	فشارخون سیستولی کمتر از ۱۴۰ میلی‌متر جیوه و دیاستولی کمتر از ۹۰ میلی‌متر جیوه و عدم مصرف داروی ضد فشارخون	فشارخون سیستولی بیشتر مساوی ۱۴۰ میلی‌متر جیوه یا دیاستولی بیشتر مساوی ۹۰ میلی‌متر جیوه یا مصرف داروی ضد فشارخون
WC	چاقی شکمی	دور کمر کمتر از ۱۰۲سانتی متر برای مردان و کمتر از ۸۸ سانتی متر برای زنان	دور کمر ۱۰۲ سانتی‌متر یا بزرگ‌تر از آن برای مردان و ۸۸ سانتی‌متر یا بزرگ‌تر از آن برای زنان
TG	تری‌گلیسرید	تری‌گلیسرید کمتریا مساوی ۲۵۰ (میلی‌گرم بر دسی‌لیتر)	تری‌گلیسرید بالای ۲۵۰ (میلی‌گرم بر دسی‌لیتر)
CHOL	کلسترول	کلسترول کمتر یا مساوی ۲۵۰ (میلی‌گرم بر دسی‌لیتر)	کلسترول بالای ۲۵۰ (میلی‌گرم بر دسی‌لیتر)
HDL	HDL	HDL بزرگ‌تر یا مساوی ۳۵ (میلی‌گرم بر دسی‌لیتر)	کلسترول HDL کمتر از ۳۵ (میلی‌گرم بر دسی‌لیتر)
Education	تحصیلات	تحصیلات دیپلم و زیر دیپلم	تحصیلات بالای دیپلم
Smoke	سیگار	غیر سیگاری	سیگاری

۱۴ عامل خطرساز دوحالتی در ارتباط با بروز دیابت وارد مدل رگرسیون لجستیک منطقی شدند. تأثیر تغییرات بعد مدل (تعداد متغیرهای مشمول در مدل) بر آماره‌ی انحراف مدل رگرسیون لجستیک منطقی برازش داده شده با ۳،۲،۱ و ۴ درخت و بُعدهای متفاوت از ۲ تا ۱۰ برگ، در شکل ۲ نشان داده شده‌اند. امتیازهای آزمون اعتبار متقاطع برای تعیین بعد مناسب مدل، انتخاب مدلی با ۴ درخت و ۵ برگ را می‌دهد.

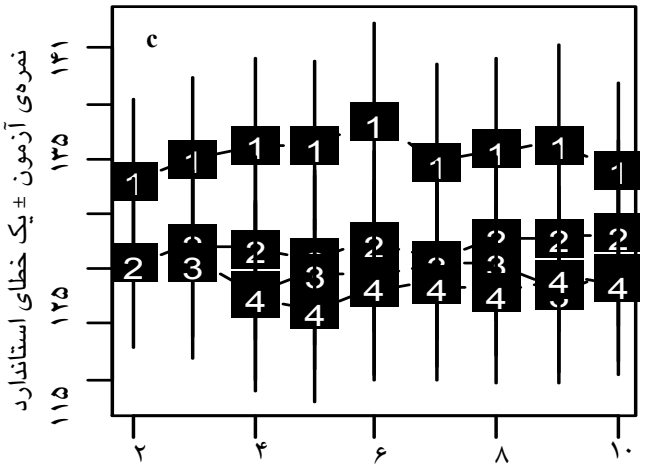
به این ترتیب، بعد از تعیین اندازه‌ی مناسب مدل با استفاده از آزمون اعتبار متقاطع، درصدد یافتن بهترین مدل با ۴ درخت و ۵ برگ بودیم. الگوریتم Annealing با جستجو در فضای حالت‌های چنین مدل‌هایی، ترکیبی از متغیرها را می‌یابد که کمترین آماره‌ی انحراف را دارند. از آن‌جا که این الگوریتم یک الگوریتم تصادفی و احتمالی است لازم نیست نتیجه منحصر به فرد باشد و یکتایی در جستجو حاصل نمی‌شود. بنابراین، مدلی که در ۱۰ بار اجرای الگوریتم با فراوانی نسبی بالایی مشاهده شد، به عنوان مدل منتخب الگوریتم معرفی گردید. شایان ذکر است که سایر مدل‌های مشاهده شده نسبت به مدل منتخب در این ۱۰ بار تغییرپذیری کمی داشتند. مدل حاصل در شکل ۲ ارایه شده است.



امتیازهای مربوط به مدل‌هایی با تعداد درخت‌ها و برگ‌های متفاوت در گروه آزمون



امتیازهای مربوط به مدل‌هایی با تعداد درخت‌ها و برگ‌های متفاوت در گروه آموزش



امتیازها ± خطای معیار مربوط به مدل‌هایی با تعداد درخت‌ها و برگ‌های متفاوت در گروه آزمون

شکل ۲- آزمون اعتبار متقاطع

برای ارزیابی و مقایسه‌ی مدل منطقی به دست آمده، آماره‌ی انحراف و میزان حساسیت و ویژگی مدل محاسبه شد و با مقادیر حاصل از رگرسیون لجستیک معمولی که در آن فقط اثرهای اصلی متغیرها وارد شدند، مقایسه شد. برای مقایسه‌ی درستی مدل‌ها در پیش‌بینی بروز دیابت منحنی مشخصه‌ی عملکرد ROC[i] برای هرکدام از آن‌ها رسم و سطح زیر منحنی محاسبه شد. برای ارزیابی قابلیت تعمیم‌دهی مدل‌ها نیز، فواصل اطمینانی برای سطح زیر منحنی‌های راک به دو روش معمول و خودگردان[ii] به دست آمد. هم‌چنین، نقطه‌ی برش بهینه برای یافتن حساسیت و ویژگی مدل‌ها طوری در نظر گرفته شد که عبارت زیر را مینیمم کند.[۷]

$$\sqrt{(1-\text{Sensitivity})^2+(1-\text{Specificity})^2}$$

از نرم‌افزارهای R نسخه‌ی ۱، ۸، ۲ برای رگرسیون لجستیک منطقی، SPSS نسخه‌ی ۱۵ برای رگرسیون لجستیک معمولی و STATA نسخه‌ی ۱۰ برای مقایسه‌ی سطح زیر منحنی‌های راک استفاده شد.

یافته‌ها

۳۵۲۳ نفر شامل ۲۰۳۸ زن (۵۷٪/۸) و ۱۴۸۵ مرد (۴۲٪/۲) مورد بررسی، ۸۰ نفر از مردان (۴/۵٪) و ۱۳۳ نفر از زنان (۵/۶٪) در طول مدت پیگیری به دیابت مبتلا شدند که تفاوت معنی‌داری در میزان ابتلا بین این دو گروه مشاهده نشد (P=۰/۱) ویژگی‌های پایه‌ی افراد مورد بررسی به تفکیک جنس در جدول ۲ آمده است. مقایسه‌ی عوامل خطرساز مرتبط با دیابت در دو گروه دیابتی و غیر دیابتی نشان داد که همه‌ی عوامل خطرساز غیر از فعالیت بدنی، جنس، سیگار کشیدن و کلسترولHDL بر بروز دیابت تأثیر معنی‌داری داشتند (جدول ۳).

i- Receiver Operating Charactristic
ii - Bootstrap

جدول ۲- ویژگی‌های (کمّی) پایه‌ی افراد مورد بررسی در ابتدای مطالعه به تفکیک جنس

مشخصات عمومی	زنان (تعداد = ۲۰۳۸)	مردان (تعداد = ۱۴۸۵)
سن (سال)	*۴۰/۴۱±۱۲/۶۲	۴۳/۴۱±۱۴/۱۰†
دور کمر (سانتی‌متر)	۸۷/۱۸±۱۲/۱۸	۸۸/۴۸±۴۸†
نمایه‌ی توده‌ی بدن (کیلوگرم بر مترمربع)	۲۷/۶۱±۴/۷۶	۲۵/۷۷±۳/۹۴†
فشارخون سیستولی (میلی متر جیوه)	۱۱۶/۷۹±۱۷/۳۸	۱۲۰/۰۷±۱۷/۴۷†
فشارخون دیاستولی (میلی متر جیوه)	۷۷/۴۵±۱۰/۱۳	۷۷/۹۴±۱۰/۶۵†
کلسترول (میلی‌گرم بر دسی‌لیتر)	۲۱۰/۴۸±۴۷/۳۱	۲۰۵/۰۸±۴۲/۶۵
تری‌گلیسرید (میلی‌گرم بر دسی‌لیتر)	۱۵۴/۴۷±۹۵/۰۱	۱۸۰/۷۹±۱۱۹/۵۸†
کلسترول HDL (میلی‌گرم بر دسی‌لیتر)	۴۵/۰۳±۱۰/۹۰	۳۸/۴۵±۹/۱۸†
قند خون ناشتا (میلی‌گرم بر دسی‌لیتر)	۸۹/۳۸±۹/۷۱	۹۱/۱۲±۹/۳۹†
قند خون دوساعته (میلی‌گرم بر دسی‌لیتر)	۱۰۹/۸۴±۲۸/۱۶	۱۰۲/۵۲±۵۲†

* اعداد به صورت میانگین ±انحراف معیار گزارش شده‌اند، † P کمتر از ۰/۰۰۱

جدول ۳- میزان شیوع پایه عوامل خطر در افراد مورد بررسی به تفکیک موارد جدید دیابتی و غیر دیابتی‌ها

عامل خطرساز	دیابتی (تعداد= ۲۱۳)	غیر دیابتی (تعداد=۳۳۱۰)
داشتن دور کمر بالا	*۱۱۹ (۵۵/۹)	۹۴۴ (۲۸/۵)†
نمایه‌ی توده‌ی بدن بالا	۱۷۶ (۸۲/۶)	۲۰۸۹ (۶۳/۱)†
قند خون ناشتا بالا	۱۲۴ (۵۸/۲)	۴۰۵ (۱۲/۲)†
قند خون دوساعته بالا	۱۲۱ (۵۶/۸)	۳۶۹ (۱۱/۱)†
سن بالای ۴۰ سال	۱۵۷ (۷۳/۷)	۱۶۲۵ (۴۹/۱)†
داشتن سابقه‌ی فامیلی دیابت	۹۳ (۴۳/۷)	۸۵۲ (۲۵/۷)†
داشتن فشارخون	۸۳ (۳۹/۰)	۶۰۲ (۱۸/۲)†
فعالیت فیزیکی کم	۹۳ (۴۳/۷)	۱۳۲۲ (۳۹/۹)
تری‌گلیسرید بالا	۶۲ (۲۹/۱)	۴۴۸ (۱۳/۵)†
کلسترول بالا	۱۴۳ (۶۷/۱)	۱۷۱۲ (۵۱/۷)†
کلسترول HDL پایین	۵۱ (۲۳/۹)	۶۵۱ (۱۹/۷)
تحصیلات زیر دیپلم	۱۹۷ (۹۲/۵)	۲۸۵۷(۸۶/۳)‡
کشیدن سیگار	۲۵ (۱۱/۷)	۴۳۳ (۱۳/۱)

* داده‌ها به صورت تعداد (درصد) گزارش شده‌اند، † P کمتر از ۰/۰۰۱، ‡ P کمتر از ۰/۰۵.

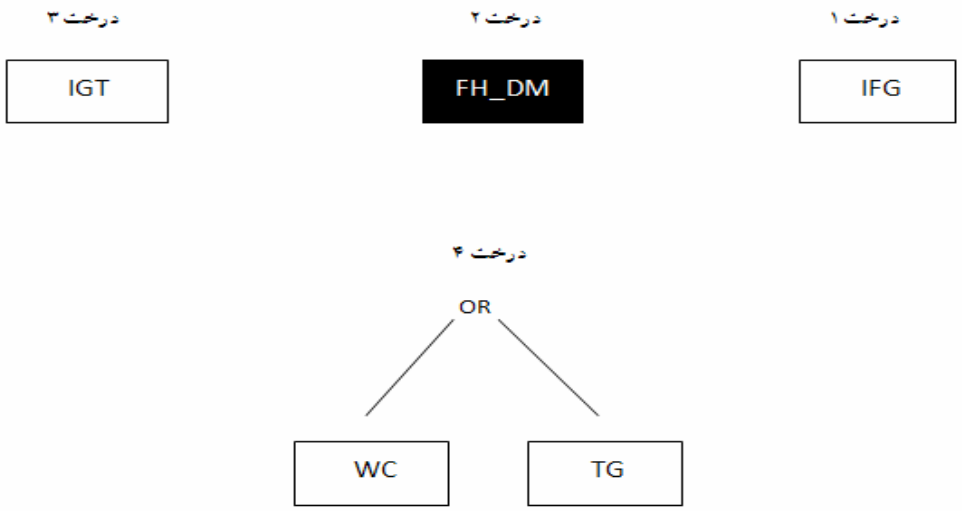
الگوریتم Anealing برای رگرسیون لجستیک منطقی با ۴ ترکیب بولی و ۵ متغیر مشمول در مدل، ترکیباتی به این صورت یافت: داشتن اختلال تحمل قندخون ناشتا با نسبت شانس۵/۵۳ و فاصله‌ی اطمینان (۷/۵۹ و ۴/۰۳)، داشتن اختلال تحمل قند دوساعته با نسبت شانس ۵/۴۵ و فاصله‌ی اطمینان (۳/۹۶ و ۷/۴۹)، نداشتن سابقه‌ی فامیلی دیابت با نسبت شانس ۰/۵۳ و فاصله اطمینان (۰/۷۲ و ۰/۳۸) (یا به طور معادل داشتن سابقه فامیلی دیابت با نسبت شانس ۱/۸۹ و فاصله اطمینان ۲/۶۳ و ۱/۳۹)، تری‌گلیسرید بالا یا دور کمر بزرگ با نسبت شانس ۲/۴ و فاصله‌ی اطمینان (۳/۳۲ و ۱/۷۳). همه‌ی این ترکیبات با مقدار P کمتر از ۰/۰۰۱ معنی‌دار بودند. سه تا از ترکیبات به صورت اثر اصلی و یکی از آنها به صورت اثر متقابل «تری‌گلیسرید بالا یا دور کمر بالا» ظاهر شد که در جدول ۴ نشان داده شده است. نمایش

درختی این مدل نیز در شکل ۳ آمده است. در این شکل، درخت اول قند خون ناشتای بالا داشتن، درخت دوم که به صورت ترکیب متضاد ظاهر شده و سابقه‌ی فامیلی دیابت نداشتن را نشان می‌دهد، درخت سوم قندخون دوساعته‌ی بالا و درخت چهارم دور کمر بالا یا تری‌گلیسرید بالا را (که شامل حالت هر دو غیر طبیعی باشد نیز می‌شود) به عنوان ترکیبات بولی مؤثر بر دیابت نشان می‌دهند. یافته‌های آخرین گام رگرسیون لجستیک مرسوم پیشرو نیز شامل همین متغیرها ولی فقط با اثر اصلی‌شان بود (یافته‌ها نشان داده نشده است).

جدول ۴- ترکیبات بولی یافت شده با الگوریتم Annealing و ضرایب مربوط به هر کدام در مدل لجستیک منطقی با ۴ درخت و ۵ برگ برای پیش‌بینی بروز دیابت

ترکیبات بولی	درخت	ضریب	خطای معیار	نسبت بخت	فاصله‌ی اطمینان ۹۵٪ برای OR	
					حد پایین	حد بالا
قندخون ناشتای مختل (IFG)	درخت ۱	۱/۷۱	۰/۱۶	۵/۵۳*	۴/۰۳	۷/۵۹
سابقه‌ی فامیلی دیابت (FH-DM)c	درخت ۲	-۰/۶۳	۰/۱۶	۰/۵۳*	۰/۳۸	۰/۷۲
اختلال تحمل گلوکز (IGT)	درخت ۳	۱/۶۹	۰/۱۶	۵/۴۵*	۳/۹۶	۷/۴۹
دور کمر (WC) یا تری‌گلیسرید (TG)	درخت ۴	۰/۸۷	۰/۱۶	۲/۴۰*	۱/۷۳	۳/۳۲
ثابت مدل	ثابت مدل	-۳/۸۵	۰/۱۸	۰/۰۲*		

* P کمتر از ۰/۰۰۱



شکل ۳- نمایش درختی ترکیبات بولی یافت شده با الگوریتم Annealing در مدل لجستیک منطقی با ۴ درخت و ۵ برگ برای پیش‌بینی بروز دیابت. متغیرهای مشاهده شده در درخت‌ها در جدول ۱ معرفی شده‌اند. متغیرهایی که با زمینه‌ی سیاه در درخت ظاهر شده‌اند در مدل به صورت نقیض آن متغیر تفسیر می‌شوند.

جدول مربوط به سطح زیر نمودار مدل‌ها و فواصل اطمینان متغیرهای پارامتری و غیرپارامتری نیز در جدول ۵ آمده است. در این فواصل به دلیل بالا بودن حجم نمونه، قضیه‌ی حد مرکزی به خوبی برقرار بود و فواصل اطمینان مشابهی در دو روش مجانبی و Bootstrap مشاهده شد.

آماره‌ی انحراف مدل لجستیک منطقی برابر ۱۲۰۳/۳ و برای مدل لجستیک پیشرو ۱۲۰۶/۸ محاسبه شد. برای مدل منطقی ۴ درختی، حساسیت مدل ۷۴٪ و ویژگی آن ۸۳٪ و برای مدل لجستیک پیشرو، حساسیت ۷۵٪ و ویژگی ۸۲٪ به دست آمد.

جدول۵- سطح زیر نمودار و فواصل اطمینان متغیرهای پارامتری و Bootstrap

مدل	سطح زیر نمودار	خطای معیار	مقدار P	فاصله‌ی اطمینان ۹۵٪ با فرض توزیع نرمال		فاصله‌ی اطمینان ۹۵٪ Bootstrap	
				حد پایین	حد بالا	حد پایین	حد بالا
رگرسیون لجستیک معمولی	۰/۸۳۹	۰/۰۱۶	<۰/۰۰۱	۰/۸۰۸	۰/۸۷۱	۰/۸۰۸	۰/۸۶۹
مدل منطقی با ۴ درخت	۰/۸۴۳	۰/۰۱۵	<۰/۰۰۱	۰/۸۱۳	۰/۸۷۴	۰/۸۱۲	۰/۸۷۲

بحث

براساس یافته‌های این مطالعه، متغیرهای وارد شده در آخرین گام لجستیک معمولی همان متغیرهای مشمول در مدل منطقی با ۴ درخت و ۵ برگ بودند با این تفاوت که در مدل لجستیک ۵ اثر اصلی وجود دارد در حالی که در مدل منطقی ۳ ترکیب به صورت اثر اصلی و یک ترکیب به صورت اثر متقابل (دور کمر یا تری‌گلیسرید) ظاهر شده است. با توجه به این تحلیل، شاید بتوان نتیجه گرفت که در مورد بروز دیابت با استفاده از عوامل خطرساز ذکر شده، وجود اثرهای متقابل و لحاظ نکردن آنها چندان نگران‌کننده نیست و آنچه مهم است اثرهای اصلی متغیرها است و به ظاهر متغیرها به‌صورت مستقل از هم در بروز دیابت نقش دارند. تنها اثر متقابل مشاهده شده در این مدل مربوط به دور کمر و تری‌گلیسرید است که در نظر گرفتن همین اثر متقابل نیز باعث کاهش آماره‌ی انحراف مدل از ۱۲۰۶/۸۸ برای لجستیک حاصل از اثرهای اصلی به ۱۲۰۳/۳۰ برای مدل منطقی و افزایش قدرت پیش‌بینی مدل شده است. سطح زیر نمودار مدل منطقی با ۵ برگ برابر ۰/۸۴۳ و سطح زیر نمودار لجستیک حاصل از اثرهای اصلی ۰/۸۳۹ است.

در مورد پیش‌بینی دیابت با رگرسیون لجستیک منطقی، مطالعه‌ی مشابهی یافت نشد ولی در مطالعه‌هایی که به بررسی قواعد تصمیم‌گیری در مورد دیابت پرداخته‌اند عوامل خطرساز مشابه با پژوهش حاضر به دست آمده است. از جمله ویلسون و همکاران[۸] در سال ۲۰۰۷ در مطالعه‌ای برای پیش‌بینی بروز دیابت در افراد بالای ۵۰ سال، عوامل خطرساز شامل سن بالا، دور کمر بالا، سابقه‌ی فامیلی دیابت، اختلال تحمل قندخون ناشتا، تری‌گلیسرید بالا و کلسترول HDL پایین را به عنوان متغیرهای پیش‌بینی‌کننده معرفی می‌کنند. در مطالعه‌ی دیگری که بارک و همکاران[۹] در سال ۱۹۹۸ انجام دادند نژاد، چاقی، بیماری‌های قلبی، تری‌گلیسرید بالا و اختلال تحمل قندخون ناشتا و دوساعته را به عنوان عوامل خطرساز مؤثر در بروز دیابت معرفی کرده‌اند.

دوحالتی کردن متغیرهای مطالعه به صورت وجود و یا عدم وجود عامل خطرساز، و ارایه مدلی با استفاده از این متغیرهای دوحالتی از نظر بالینی و سادگی استفاده از آن بدون نیاز به اطلاعات جزیی افراد، دارای اهمیت و ارزش فراوان است. اختلاف مدل منطقی مشاهده شده با لجستیک معمولی کم است که حاکی از ناچیز بودن اثرهای متقابل بین عوامل خطرساز دوحالتی دیابت است.

در جستجوی نگارندگان، مدلی که از رگرسیون منطقی برای پیش‌بینی بروز دیابت استفاده کند پیدا نشد مدل‌های پیش بینی فعلی از روش‌های مرسوم آماری از جمله

رگرسیون لجستیک استفاده کرده‌اند. هرچند یکی از محدودیت‌های رگرسیون منطقی ممکن است مشکل داده‌های گمشده باشد و در نهایت هم به شکل یکی از رگرسیون‌های موجود از جمله خطی، لجستیک یا کاکس، بسته به نوع مطالعه، تبدیل می‌شود ولی مزیتی که نسبت به روش‌هایی مانند مدل شبکه‌های عصبی مصنوعی دارد این است که کاملاً به شکل یک رگرسیون است و قابلیت تفسیر ضرایب، ارزیابی مدل با امتیازها و آماره‌های مربوط به نوع رگرسیون استفاده شده در آن وجود دارد. هم‌چنین، قابلیت لحاظ کردن اثرهای متقابل بین چند متغیر در قالب یک عبارت بولی و تلخیص متغیرها از مزایای این روش نسبت به روش‌های قبلی و فعلی است.[۳] دوحالتی کردن متغیرهای مطالعه به صورت وجود و یا عدم وجود عامل خطرساز، و ارایه مدلی با استفاده از این متغیرهای دوحالتی از نظر بالینی و سادگی استفاده از آن بدون نیاز به اطلاعات جزیی افراد، داری اهمیت و ارزش فراوان است. به عنوان مثال، با استفاده از مدل رگرسیون منطقی به دست آمده در این مطالعه و اطلاع از وضعیت متغیرهایی مانند طبیعی یا غیرطبیعی بودن قند خون ناشتا، قندخون دوساعته و وجود یا عدم وجود سابقه‌ی دیابت در خانواده‌ی فرد و در نهایت با اطلاع از وضعیت تری‌گلیسرید سرم و دور کمر فرد فقط در حد طبیعی یا غیر طبیعی بودن، می‌توان احتمال بروز دیابت را در آینده برآورد کرد.

در این مطالعه با توجه به تعداد کم موارد جدید ابتلا به دیابت نسبت به تعداد کل افراد وارد شده در مطالعه، امکان اعتبارسنجی مدل لجستیک معمولی نبود و این از محدودیت‌های مطالعه به شمار می‌رود. از طرفی به دلیل این‌که رگرسیون منطقی روش نوپایی است و تا به حال برای پیش‌بینی دیابت استفاده نشده است امکان مقایسه‌ی یافته‌های به دست آمده در مطالعه‌ی حاضر با مطالعه‌های مشابه وجود نداشت و تنها مقایسه‌ها با مطالعه‌هایی که از مدل‌های پیش‌بینی مرسوم مانند رگرسیون لجستیک معمولی استفاده کرده‌اند، انجام شد.

سپاسگزاری: در این مطالعه، از داده‌های طرح قند و لیپید تهران که توسط پژوهشکده‌ی علوم غدد درون‌ریز و متابولیسم دانشگاه علوم‌پزشکی و خدمات بهداشتی ـ درمانی شهید بهشتی اجرا شده است، استفاده شد. بر خود لازم می‌دانیم از همه‌ی کسانی که در طراحی و جمع‌آوری داده‌های طرح TLGS مشارکت داشتند، قدردانی نماییم. این پژوهش از پایان‌نامه‌ی کارشناسی ارشد آمار زیستی استخراج شده است.[۱۰]

References

1. Zimmet PZ. Diabetes epidemiology as a tool to trigger diabetes research and care. Diabetologia 1999; 42: 499-518.
2. Zimmet P. Globalization, coca-colonization and the chronic disease epidemic: can the Doomsday scenario be averted? J Intern Med 2000; 247: 301-10.
3. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. Journal of Computational and Graphical Statistics 2003; 12: 475 –511.
4. Azizi F, Rahmani M, Emami H, Mirmiran P, Hajipour R, Madjid M, et al. Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1). Soz Praventivmed 2002; 47: 408-26.
5. Genuth S, Alberti KG, Bennett P, Buse J, Defronzo R, Kahn R, et al. Follow-up report on the diagnosis of diabetes mellitus. Diabetes Care 2003; 26: 3160-7.
6. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III).Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation 2002, 106: 3143-421.
7. Perkins N , Schisterman E. The Inconsistency of "Optimal" Cut-points Using Two ROC Based Criteria. Am J Epidemiol 2006; 163: 670–5.
8. Wilson P, James B, Sullivan L, Fox C, Nathan D, DAgostino R. Prediction of Incident Diabetes Mellitus in Middle-aged Adult. Arch Intern Med 2007; 1068-74.
9. Burke JP, Haffner SM, Gaskill SP, Williams KL, Stern MP. Reversion from type 2 diabetes to nondiabetic status. Influence of the 1997 American Diabetes Association criteria. Diabetes Care 1998; 21: 1266-70.
10. Sarbakhsh P. "Logic regression and its application in predicting diabetes among 20 year old and over population in district 13 of Tehran," MSC thesis, Shahid Beheshti University of Medical Sciences, 2009.

Original Article

Prediction of Diabetes Using Logic Regression

Mehrabi Y[1], Sarbakhsh P[2], Hadaegh F[3], Khadem-Maboudi A[2]

[1]Department of Epidemiology, School of Public Health, [2]Department of Biostatistics, School of Paramedicine, [3]Prevention of Metabolic Disorders Research Center, Research Institute for Endocrine Science, Shahid Beheshti University of Medical Sciences, Tehran, I.R.Iran

e-mail:ymehrabi@gmail.com

Received: 22/06/2009 **Accepted**: 29/11/2009

Abstract

Introduction: Detection of population at risk of type II diabetes, as a multi-factorial disease, is an important issue because of its individual and social impacts. To date, several studies have been conducted to predict the incidence of diabetes, using different statistical methods. However, despite its clinical importance, it is highly difficult to consider all interactions among risk factors, in ordinary statistical models. This study aimed to extract appropriate logic combination of type 2 diabetes risk factors employing the recently introduced method, Logic regression. Materials and Methods: The study population was selected from a cohort of the Tehran Lipid and Glucose Study (TLGS). Data for 3523 participants, aged 20 years and over (57.8% female and 42.2% male) were entered into analysis, for which logistic logic regression method was used. The model parameters were estimated using the Annealing algorithm. To avoid overestimation, the optimal number of logic combinations was determined by the cross-validation method. Deviance, sensitivity and specificity measures were computed to evaluate the logic model and its comparison to ordinary logistic regression; the latter accommodated only the main effects. The prediction power of the two models was compared by Area under ROC curve. R software version 2.8.1 was employed for analyses. Results: Logistic logic regression with the 4 Boolean combination including 5 variables was fitted using the Annealing algorithm and resulted in in deviance of 1203.30. This model had better fit compared to other logic models and also ordinary logistic regression with forward procedure (deviance=1206.88). The Boolean combination of the above model included impaired fasting glucose (OR=5.53, 95%CI: 4.03-7.59), IGT (OR=5.54, 95%CI: 3.96-7.49), family history of diabetes (OR=1.89, 95%CI: 1.38-2.63), and interaction of high triglycerides or abnormal waist circumference (OR=2.4, 95%CI: 1.73-3.32); all p-values <0.001. The area under ROC curve for the model was 0.843 (95%CI: 0.813-0.874). Conclusion: This study showed that the logic regression as a newly introduced method has the ability of recognizing and modelling the interactions between different risk factors. Therefore, it is recommended as an appropriate tool for screening of the multi-factorial diseases such as diabetes.

Keywords: Logic regression, Diabetes, Interaction effects, Prediction of incidence, Boolean logic, Simulated annealing