مجله‌ی غدد درون‌ریز و متابولیسم ایران
دانشگاه علوم پزشکی و خدمات بهداشتی ـ درمانی شهید بهشتی
دوره‌ی چهاردهم، شماره‌ی ۴، صفحه‌های ۳۵۹ ـ ۳۵۲ (آبان ۱۳۹۱)

مقاله‌ی پژوهشی

استفاده از رگرسیون منطقی برای شناسایی اثرات متقابل برخی از پلی‌مورفیسم‌های ژنی و سایر عوامل خطر بر سطح پایین HDL : مطالعه‌ی قند و لیپید تهران

پروین سربخش[۱]، دکتر یدالله محرابی[۲]، دکتر مریم االسادات دانشپور[۳]، فرید زایری[۱]، مهشید نامداری[۱]، دکتر فریدون عزیزی[۴]

۱) گروه آمار زیستی، دانشکده‌ی پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، ۲) گروه اپیدمیولوژی، دانشکده‌ی بهداشت، دانشگاه علوم پزشکی شهید بهشتی، ۳) مرکز تحقیقات چاقی، پژوهشکده‌ی علوم غدد درون‌ریز و متابولیسم، دانشگاه علوم پزشکی شهید بهشتی، ۴) مرکز تحقیقات غدد درون‌ریز و متابولیسم، پژوهشکده‌ی علوم غدد درون‌ریز و متابولیسم، دانشگاه علوم پزشکی شهید بهشتی،
نشانی مکاتبه‌ی نویسنده‌ی مسئول: تهران، ولنجک، دانشگاه علوم پزشکی شهید بهشتی، دانشکده‌ی بهداشت، گروه اپیدمیولوژی،
e-mail: ymehrabi@gmail.com

چکیده

مقدمه: رگرسیون منطقی یک روش رگرسیونی تعمیم یافته است که می‌تواند اثرات متقابل پیچیده بین متغیرهای دوحالتی را تشخیص دهد. به دلیل اهمیت تقابل‌های ژنتیکی، این روش در بررسی‌های ژنتیکی با موفقیت استفاده شده است. هدف پژوهش حاضر، بررسی ارتباط بین کلسترول ـ HDL و برخی از پلی‌مورفیسم‌های مرتبط با آن، با استفاده از رگرسیون منطقی است. مواد و روش‌ها: داده‌های ۴۳۶ نفر (۱۷۲ مرد و ۲۶۴ زن) که به طور تصادفی از میان شرکت‌کنندگان فاز ۳ مطالعه‌ی قند و لیپید تهران با سن ۲۰ سال یا بیشتر انتخاب شده بودند، مورد تجزیه و تحلیل قرار گرفتند. برای تشخیص اثرات اصلی و متقابل پلی‌مورفیسم‌های ژنی مرتبط با کلسترول ـ HDL، از رگرسیون منطقی با تابع پیوند لجستیک استفاده گردید. برای جلوگیری از بیش‌برازش شدن مدل از آزمون اعتبار متقاطع، و برای یافتن ترکیب‌های منطقی مناسب و برآورد ضرایب آن‌ها از الگوریتم جستجوی Simulated Annealing استفاده شد. یافته‌ها: براساس یافته‌های آزمون اعتبار متقاطع، مدل منطقی با ۳ ترکیب بولی و ۴ پیش‌بینی‌کننده بهترین اندازه برای مدل منطقی بود. مدل منطقی برازش داده شده نشان داد افرادی با الل ε3 در پلی‌مورفیسم ژن ApoE یا تری‌گلیسرید بالا نسبت به افراد دیگر شانس ۲/۳۵ برابری با فاصله اطمینان ۹۵٪ (۴/۲۵و ۱/۳) برای داشتن سطح پایین کلسترول ـ HDL دارند. از سوی دیگر، داشتن تری‌گلیسرید بالا به تنهایی نیز قادر است شانس داشتن کلسترول ـ HDL پایین را ۲/۷۳ برابر افزایش دهد (فاصله اطمینان ۹۵٪: ۴/۵۳و ۱/۶۵). نتیجه‌گیری: یافته‌ها نشان داد برای داشتن کلسترول ـ HDL پایین، بین تری‌گلیسرید بالا و پلی‌مورفیسم ژن Apoe، اثر متقابل وجود دارد. رگرسیون منطقی به عنوان یک روش جدید قادر به تشخیص چنین اثرات متقابلی است.

واژگان کلیدی: اثرات متقابل، الگوریتم Annealing، پلی‌مورفیسم تک نوکلئوتیدی، رگرسیون منطقی، لیپوپروتئین با دانسیته‌ی بالا، مطالعه‌ی قند و لیپید تهران

دریافت مقاله: ۹۰/۱۱/۱۸ـ دریافت اصلاحیه: ۹۱/۲/۱۱ ـ پذیرش مقاله: ۹۱/۲/۱۳

مقدمه

بررسی عوامل ژنتیکی و محیطی که سبب ایجاد بیماری‌های چند عاملی می‌شوند و تعیین مقدار اثر این عوامل، از هدف‌های بررسی‌های ژنتیکی اخیر است. در بررسی‌ها شناسایی پلی‌مورفیسم تک نوکلئوتید (SNP[i]) و

i- Single-nucleotide polymorphism

برآورد اثرات متقابل دارای اهمیت فراوان می‌باشد، زیرا فرض بر این است که نه تنها اثرات اصلی SNP ها، بلکه اثرات متقابل بین آن‌ها و هم‌چنین بین SNP ها و عوامل محیطی در ایجاد بیماری‌های ژنتیکی انسان دخیل هستند.[۱] به همین منظور، به تازگی روش‌هایی برای تشخیص اثرات متقابل آماری بین SNP ها معرفی شده که سبب افزایش توان آماری مطالعه گردیده، هم‌چنین تفسیرهای بیولوژیکی مفیدی را ارایه می‌دهند.[۲]

یکی از روش‌های جدید آنالیز اثرات متقابل، رگرسیون منطقی[i] می‌باشد که به عنوان یک روش رگرسیونی تعمیم یافته در سال ۲۰۰۳ معرفی شده، و بیشتر برای بررسی اثرات متقابل مراتب بالاتر در بررسی‌های ژنتیکی استفاده می‌شود.[۳] هدف رگرسیون منطقی یافتن ترکیبات بولی[ii] از متغیرهای دوحالتی اولیه است، به‌طوری که این ترکیبات بتوانند پیامد مدنظر را به شکل بهتری پیش‌بینی نمایند. می‌توان هر ترکیب بولی یافت شده را در قالب یک درخت منطقی[iii] نمایش داد. متغیرهای پیش‌بین که در درخت یکسانی ظاهر می‌شوند به احتمال زیاد برای ایجاد بیماری، اثرات برهمکنشی با همدیگر دارند.

از آنجا که بررسی‌های پیشین نشان داده‌اند بین کاهش سطح لیپوپروتئین با دانسیته بالا (HDL) در خون و افزایش میزان بروز بیماری‌های قلبی ـ عروقی ارتباط معنی‌داری وجود دارد،[۴] شناسایی عوامل محیطی، بیوشیمیایی و ژنتیکی اثر گذار بر میزان کلسترول ـ HDL می‌تواند در پیشگیری از بیماری‌های قلبی ـ عروقی موثر باشد. عوامل متعددی مانند جنس، وزن، سن، رژیم غذایی، تحرک فیزیکی، عوامل محیطی و ژنتیکی بر سطح کلسترول ـ HDL تاثیر دارند[۵،۶] به علاوه، یکی از عوامل مهم تاثیرگذار تغییرات ژنتیکی فرد می‌باشد. ژن‌های متعددی کاندیدای این نوع بررسی‌ها هستند و پلی‌مورفیسم ژن‌هایی مانند آپولیپوپروتئین A1M1، آپولیپوپروتئین A1M2، آپو B، آپولیپوپروتئین AIV، آپولیپوپروتئین CIII، ABCA1SRB1، آپو E، با سطح پایین کلسترول ـ HDL خون ارتباط معنی‌داری را نشان داده‌اند.[۷-۹] تعیین نوع ارتباط این پلی‌مورفیسم‌ها و کلسترول ـ HDL می‌تواند در پیشگیری و درمان بیماری‌های قلبی ـ عروقی موثر باشد.

با توجه به شیوع بالای افزایش چربی خون در ایران[۱۰]؛ پژوهش حاضر سعی داشت سازوکار تاثیر SNP ها روی سطح کلسترول ـ HDL خون در جمعیت ایرانی را بررسی نماید، زیرا یافتن چنین ارتباطاتی می‌تواند در کنترل سطح کلسترول ـ HDL خون موثر باشد. در جمعیت ایرانی بررسی‌های متعددی در ارتباط بین پلی‌مورفیسم‌ها و لیپیدهای خونی صورت گرفته، ولی در بیشتر آن‌ها این ارتبط به صورت مجزا و تکی بررسی شده، و اثر دسته جمعی و تقابلی آن‌ها بر لیپید و اثرات برهمکنشی ممکن بررسی نگردیده است.[۱۱-۱۳] در این راستا، پژوهش حاضر تاثیر حضور همزمان پلی مورفیسم‌ها بر سطح کلسترول ـ HDL را بعد از تعدیل برای متغیرهای مداخله‌گر سن، جنس، چاقی شکمی، تری‌گلیسرید بالا، فشار خون بالا ، گلوکز ناشتای بالا و سیگار به وسیله‌ی مدل رگرسیون لجستیک منطقی بررسی نمود.

مواد و روش‌ها

آزمودنی‌ها در بررسی مقطعی حاضر از میان شرکت‌کنندگان فاز ۳ مطالعه‌ی قند و لیپید تهران[iv] انتخاب شدند.[۱۴] مطالعه‌ی قند و لیپید تهران یک مطالعه‌ی آینده‌نگر می‌باشد که روی یک نمونه از جمعیت منطقه‌ی ۱۳ تهران انجام شده، و هدف آن تعیین شیوع بیماریهای غیر واگیر و ترویج سبک زندگی سالم در این جمعیت می‌باشد. این مطالعه در ۳ مرحله انجام گردیده، مرحله‌ی اول که یک مطالعه‌ی مقطعی بود در طول سال‌های ۸۰-۱۳۷۸ انجام گرفت، و تعداد ۱۵۰۰۵ نفر از افراد بالای ۳ سال منطقه ۱۳ تهران که به روش نمونه‌گیری خوشه‌ای تصادفی انتخاب شده بودند در آن شرکت نمودند (فاز ۱). سپس افراد وارد فاز ۲ مطالعه شدند که مطالعه‌ای آینده‌نگر و شامل دو قسمت هم‌گروهی و مداخله‌ای بود، که از ۱۳۸۱ شروع شد و در ۸۴ به پایان رسید، سپس فاز سوم اجرا شد که از سال ۸۵ تا ۸۷ ادامه پیدا کرد.

تعداد ۴۳۶ نفر (۱۷۲ مرد و ۲۶۴ زن) از میان شرکت‌کنندگان فاز ۳ TLGS با سن ۲۰ سال یا بیشتر با داده‌های کامل از SNP ها و متغیرهای دیگر، به منظور بررسی SNP های مرتبط با کلسترول ـ HDL پایین، به صورت تصادفی انتخاب، و مورد بررسی قرار گرفتند.[۱۵]

i- Logic regression
ii-Boolean cobmination
iii- Logic tree

iv -Tehran lipid and glucose study

مقادیر کلسترول ـ HDL کمتر از ۴۰ میلی‌گرم در صد میلی‌لیتر برای مردان و کمتر از ۵۰ میلی‌گرم در صد میلی‌لیتر برای زنان به عنوان سطح پایین کلسترول ـ HDL تعریف شد.[۱۶] متغیرهای مداخله‌گر شامل چاقی شکمی (اندازه‌ی دور کمر ≤ ۹۵ سانتی متر برای مردان و زنان ایرانی)، تری‌گلیسرید ≤ ۱۵۰ میلی‌گرم در صد میلی‌لیتر، فشار خون ≤ ۱۳۰/۸۵ میلی‌متر جیوه یا درمان برای فشار خون بالا، گلوکز ناشتای پلاسما ≤ ۱۱۰ میلی‌گرم در صد میلی‌لیتر یا مصرف داروی پایین آورنده‌ی قند خون براساس تعریف مولفه‌های سندروم متابولیک بر اساس معیار ATPIII تعریف گردید.[۱۷] هم‌چنین، افرادی که در حال حاضر به صورت روزانه یا گه‌گاه سیگار مصرف می‌کنند ، به عنوان سیگاری در نظر گرفته شدند.

پلی‌مورفیسم‌های xbal در ژن آپو B، SstI در ژن آپولیپوپروتئین CIII، MspI در ژن آپولیپوپروتئین A1M1، Xagl در ژن ABCA1، پلی‌مورفیسم AluI در ژن SRB1، پلی‌مورفیسم ژن ApoE، پلی‌مورفیسم ژن آپولیپوپروتئین A1M2 و پلی‌مورفیسم ژن آپولیپوپروتئین AIV که به احتمال زیاد با اختلال در سطح کلسترول ـ HDL مرتبط هستند،[۷-۹] بررسی گردیدند.

در روش رگرسیون منطقی لازم است داده‌های اولیه به صورت دوحالتی بوده، و متغیرهای پیش‌بین جدید به صورت ترکیب‌های بولی از متغیرهای دو حالتی اولیه ساخته شوند.[۱۸] بنابراین، به منظور تجزیه و تحلیل داده‌ها، هر SNP به صورت متغیر تصادفی X با مقادیر ۰، ۱ و ۲ در نظر گرفته شد (به عنوان نمونه، نوکلئوتید AA، GA / AG و GG، به ترتیب با ۰ و ۱ و ۲ کدگذاری شدند). سپس، این متغیر به دو متغیر دوحالتی با عنوان‌های ژن غالب[i] (X_D) و ژن مغلوب[ii] (X_R) تبدیل شد. متغیر ژن غالب به این صورت تعریف می‌شود: $X_D = 1$ اگر $X \leq 1$ باشد و $X_D = 0$ اگر $X = 0$. متغیر ژن مغلوب نیز به این صورت تعریف می‌شود: $X_R = 1$ اگر $X = 2$ باشد و $X_R = 0$ اگر $X \leq 1$ باشد.[۳] به این ترتیب، تعداد 2p پیش‌بینی‌کننده‌ی دوحالتی از P تا SNP به‌دست می‌آید.

آمار توصیفی داده‌ها به صورت میانگین ±انحراف معیار برای متغیرهای کمی، و تعداد (درصد) برای متغیرهای کیفی گزارش شد. مقایسه‌ی ویژگی‌های جمعیت‌شناختی و بیوشیمیایی بین زنان و مردان برای متغیرهای کمی دارای توزیع نرمال با آزمون تی و برای متغیرهای غیرنرمال با آزمون من ـ ویتنی انجام شد. هم‌چنین، برای مقایسه‌ی متغیرهای کیفی در دو گروه از آزمون مجذور خی استفاده گردید. برای یافتن اثرات متقابل و مدل‌بندی داده‌ها نیز از روش رگرسیون منطقی استفاده شد که در ادامه توضیح داده می‌شود.

فرض کنید $X1, X2, \ldots, Xk$ متغیرهای پیشگوی دو حالتی و y متغیر پاسخ است. هدف برازش مدل رگرسیونی منطقی به این فرم است:

$$g(E(y)) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j$$

که در آن Lj یک عبارت بولی از متغیرهای پیشگوی Xi بوده و g[E(y)] یک تابع پیوند است. کارایی هر مدل منطقی به وسیله‌ی یک تابع امتیاز[iii] مرتبط با مدل رگرسیونی انتخاب شده ارزیابی می‌شود که نشان‌دهنده‌ی کیفیت مدل مفروض می‌باشد. برآورد βj به طور همزمان، با جستجو برای عبارت Lj با استفاده از الگوریتم جستجوی Simulated Annealing پیدا می‌شود. الگوریتم Simulated Annealing یک الگوریتم جستجوی تصادفی است و در فضای حالت‌های ممکن ترکیبات منطقی، بر مبنای تابع امتیاز تعیین شده دنبال بهترین ترکیب می‌گردد.[۳]

وجود ارتباط سیستماتیک و غیرتصادفی بین پاسخ و متغیرهای مستقل را می‌توان به‌وسیله‌ی آزمون تصادفی‌سازی مدل صفر[iv] بررسی نمود. این آزمون، با مقایسه امتیازهای به دست آمده از برازش تصادفی پاسخ و بهترین مدل منطقی یافت شده به وسیله‌ی الگوریتم Annealing ، وجود ارتباط بین متغیرهای مستقل و پاسخ را ارزیابی می‌نماید.

به منظور انتخاب مدل بهینه و جلوگیری از بیش برازش شدن مدل، از آزمون اعتبار متقاطع[v] استفاده گردید.[۳] تعداد کل متغیرهای موجود در مدل منطقی اندازه مدل نامیده شده، و به عنوان معیار پیچیدگی مدل در نظر گرفته شد. زمانی که دنبال بهترین مدل از لحاظ امتیاز می‌گردیم ممکن است به مدلی برسیم که تعداد متغیرهای بیشتری از آن‌چه مدل بهینه دارد. می‌توان با مقایسه‌ی عملکرد بهترین مدل‌ها در ابعاد

i- dominant
ii- recessive

iii- Score function
iv- Null model randomization test
v- Cross-validation test

مختلف، مدل با اندازه‌ی بهینه را انتخاب نمود. زمانی که داده به اندازه‌ی کافی موجود باشد می‌توان از روش مجموعه‌ی آموزش -آزمون[i] استفاده نمود. زمانی که داده‌ی کافی برای یک مجموعه‌ی مستقل آزمون وجود ندارد می‌توان با تقسیم کردن داده‌ها به m گروه مساوی، از روش اعتبار متقاطع برای پیدا کردن مدل با اندازه‌ی بهینه استفاده نمود.[۳]

برای تعیین ارتباط بین ترکیباتی شامل اثرات اصلی و متقابل SNPها با سطح پایین کلسترول ـ HDL در حضور متغیرهای مداخله‌گر، از رگرسیون لجستیک منطقی (Logic Logistic Regression) با تابع پیوند $g[E(y)]=\log\left[\frac{E(y)}{1-E(y)}\right]$ و تابع امتیاز "آماره‌ی انحراف" $D=\sum d(y_i,\hat{\pi}_i)^2$ استفاده شد. آزمون تصادفی‌سازی برای بررسی وجود ارتباط بین متغیرهای پیش بین و پاسخ انجام شد. برای جلوگیری از بیش برازش شدن مدل و انتخاب بهینه‌ترین مجموعه از SNPها برای پیش‌بینی بهتر متغیر پاسخ، از روش اعتبار متقاطع استفاده، و شاخص‌های مدل با به کارگیری الگوریتم Annealing برآورد گردید. هم‌چنین، به منظور برازش مدل رگرسیون منطقی، از نرم‌افزار R نسخه‌ی ۲،۱۳،۰ و برنامه‌ی Logic Reg استفاده شد.

یافته‌ها

در پژوهش حاضر، از میان ۴۳۶ نفر (۱۷۲ مرد و ۲۶۴ زن) مورد بررسی، ۲۹۳ نفر (۶۷/۲ ٪) دارای کلسترول ـ HDL پایین بودند. ویژگی‌های جمعیتی و شاخص‌های خونی این افراد به تفکیک جنس در جدول ۱ ارایه شده است. فراوانی آلل‌ها و ژنوتیپ‌ها در جدول ۲ نشان می‌دهد توزیع ژنوتیپ‌ها از تعادل هاردی ووایبرگ تبعیت نموده است.

آزمون تصادفی‌سازی مدل صفر فقط به منظور بررسی و تایید وجود ارتباط سیستماتیک بین متغیرهای پیش‌بین و پاسخ به کار می‌رود، تا در صورت تایید وجود چنین ارتباط سیستماتیک و غیر تصادفی بین داده‌ها، به دنبال یافتن مدل منطقی مناسب داده‌ها توسط الگوریتم جستجوی Annealing باشیم. این آزمون با برازش مدل‌هایی بر مبنای داده‌هایی که ترتیب آن‌ها به طور تصادفی بهم خورده، مدل‌های تصادفی را برازش می‌دهد و هیستوگرام مربوط به امتیازات این مدل‌های تصادفی را رسم می‌نماید، سپس بر مبنای داده‌های با ترتیب واقعی نیز بهترین مدل منطقی را برازش می‌دهد، و امتیاز آن را به عنوان امتیاز بهترین مدل رسم می‌کند.

جدول ۱- ویژگی‌های پایه‌ی افراد حاضر در پژوهش*

متغیر	زنان (تعداد= ۲۶۴)	مردان (تعداد= ۱۷۲)	کل (۴۳۶)	P
سن (سال)	۴۴/۸±۱۵/۲	۴۷/۰۹±۱۶/۳	۴۵/۷±۱۵/۷	۰/۱۵†
نمایه‌ی توده بدن[ii] (کیلوگرم بر مترمربع)	۲۸/۲±۵/۹	۲۶/۳۰±۳/۶۶	۲۷/۴±۴/۶	<۰/۰۰۱
سطح HDL	۴۴/۵۵±۹/۹۸	۳۷/۱±۷/۹	۴۱/۶±۹/۸	<۰/۰۰۱
فشار خون بالا	۶۹ (۲۶/۱)	۴۹ (۲۸/۵)	۱۱۸ (۲۷/۱)	۰/۵۸
دور کمر بالا	۹۷ (۳۶/۷)	۹۰ (۵۲/۳)	۱۸۷ (۴۲/۹)	۰/۰۰۱
تری‌گلیسرید بالا	۹۹(۳۷/۵)	۷۵(۴۳/۶)	۱۷۴ (۳۹/۹)	۰/۲
قند خون ناشتا بالا	۳۲(۱۲/۱)	۱۸(۱۰/۵)	۵۰(۱۱/۵)	۰/۵۹
سیگاری	۴(۱/۵)	۳۵(۲۰/۳)	۳۹(۸/۹)	<۰/۰۰۱

*داده‌ها به صورت تعداد (درصد) برای داده‌های کیفی و انحراف معیار ± میانگین برای داده‌های کمی گزارش شده‌اند. † مقدار P<۰/۰۵ از نظر آماری معنی‌دار است.

از سوی دیگر امتیاز مربوط به مدلی که در آن هیچ متغیر پیش‌بینی وجود ندارد را نیز تحت عنوان امتیاز مدل صفر رسم می‌نماید. با مقایسه‌ی امتیازات این سه گروه مدل پیرامون قدرت پیش‌بینی‌کنندگی متغیرهای پاسخ می‌توان اظهارنظر نمود، زیرا اگر امتیاز هیستوگرام مدل‌های تصادفی با بهترین مدل اختلاف داشته باشد می‌توان گفت امتیاز مدلی با ترتیب واقعی داده‌ها با متغیرهای پیش‌بین استفاده شده، از هیستوگرام امتیاز مدل‌های تصادفی بهتر است و متغیرهای موجود در مدل واقعی یا بهترین مدل دارای اثر پیش‌بینی‌کننده برای متغیر پاسخ هستند.

یافته‌های آزمون تصادفی سازی مدل صفر برای بررسی وجود ارتباط بین متغیرهای مستقل و متغیر پاسخ (کلسترول ـ HDL پایین) که در شکل ۱ نشان داده شده نشان می‌دهد بعد از برازش صد مدلی که در آن‌ها ترتیب داده ها به طور تصادفی باهم جابجا شده بودند هیستوگرام مربوط به امتیازات این مدل‌های تصادفی بیشتر (بدتر) از مدل برازش یافته با پاسخ‌های واقعی بود و بیان‌گر این است که بین متغیرهای پیش‌بین مورد بررسی و سطح پایین کلسترول ـ HDL ارتباط سیستماتیک و غیرتصادفی وجود دارد و از این متغیرها می‌توان برای مدل‌بندی سطح پایین کلسترول ـ HDL استفاده نمود.

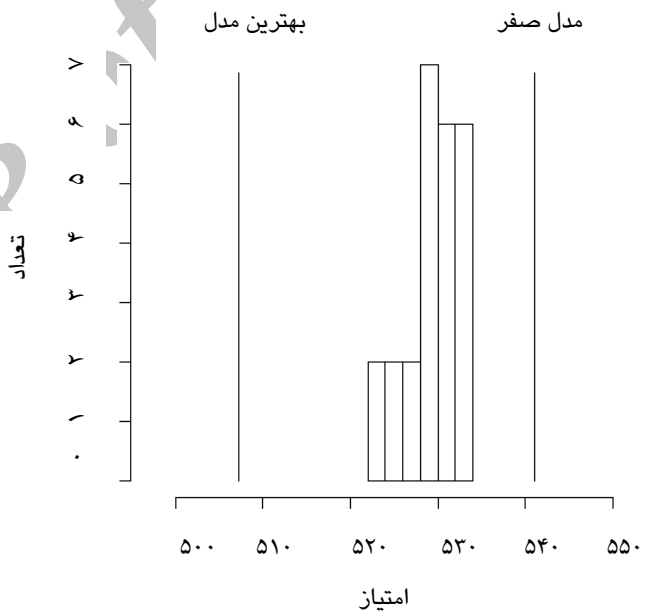
i- Training and test set

ii -Body mass index

جدول ۲- فراوانی اللها و ژنوتیپ‌های افراد مورد بررسی*

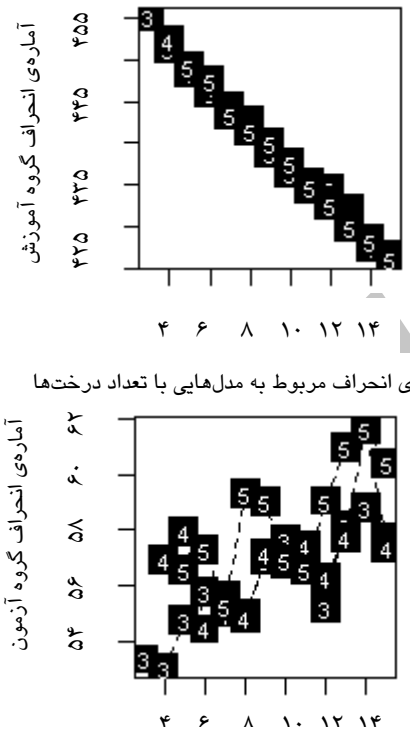
فراوانی اللها (٪) تعداد			
ApoE	e2	e3	e4
	۴۴(۱۰/۱)	۳۴۳(۷۸/۷)	۴۹(۱۱/۲)
فراوانی ژنوتیپ‌ها تعداد (٪)			
ApolipoproteinAIV_SNP_FnuHI	TT	GG	GT
	۱(۰/۲)	۳۶۵(۸۳/۷)	۷۰(۱۶/۱)
ApolipoproteinCIII_SNP_SstI	CC	GG	CG
	۳۰۳(۶۹/۵)	۱۴(۳/۲)	۱۱۹(۲۷/۳)
ABCA1_SNP_XagI	GG	AA	GA
	۱۵۵(۳۵/۶)	۵۹(۱۳/۵)	۲۲۲(۵۰/۹)
SRB1_SNP_AluI	GG	AA	GA
	۳۶۰(۸۲/۶)	۴(۰/۹)	۷۲(۱۶/۵)
ApolipoproteinA1M1_SNP	+/+	-/-	+/-
	۳۱۴(۷۲)	۷(۱/۶)	۱۱۵(۲۶/۴)
ApolipoproteinA1M2_SNP	+/+	-/-	+/-
	۳۹۶(۹۰/۸)	۶(۱/۴)	۳۴(۷/۸)
ApolipoproteinB_SNP_XbaI	+/+	-/-	+/-
	۴۱(۹/۴)	۲۲۳(۵۱/۱)	۱۷۲(۳۹/۴)

*داده‌ها به صورت تعداد (درصد) گزارش شده‌اند.



شکل ۱- آزمون تصادفی سازی مدل صفر به منظور بررسی وجود ارتباط بین پاسخ و متغیرهای مستقل

یافته‌های آزمون اعتبار متقاطع نیز در شکل ۲ نشان داده شده است که در این شکل اعداد داخل مربع‌های سیاه نشان‌گر تعداد درخت‌ها (ترکیبات منطقی) و اعداد روی محور افقی نشان‌گر تعداد برگ‌های (متغیرهای دو حالتی) مشمول در مدل است.



شکل ۲ - آزمون اعتبارمتقاطع. اعداد داخل مربع‌های سیاه نشان‌گر تعداد درخت‌ها (ترکیبات منطقی) و اعداد روی محور افقی نشان‌گرتعداد برگ‌های (متغیرهای دو حالتی) مشمول در مدل است.

اعداد روی محور عمودی نیز نشان‌گر مقدار امتیاز برای مدل‌های مختلف می‌باشد. شکل مربوط به گروه آموزش

نشان می‌دهد که در گروه آموزش، هر چه اندازه‌ی مدل از نظر تعداد ترکیبات منطقی و متغیرهای مشمول در این ترکیبات بیشتر می‌شود امتیاز مدل‌ها کمتر شده و برازش بهتر می‌شود که این امر حاکی از امکان بیش‌برازش شدن مدل است. مشاهده‌ی امتیازات گروه آزمون نیز بیش‌برازش شدن مدل‌های گروه آموزش را تایید می‌نماید، زیرا این امتیازات نشان‌گر این می‌باشد که از بین تمام مدل ها با اندازه‌های مختلف، مدلی با ۳ ترکیب منطقی متشکل از ۴ متغیر، کمترین امتیاز را نسبت به بقیه دارد. در نتیجه، انتخاب مدل منطقی با این اندازه مناسب داده‌ها بوده و این مدل بهترین برازش را برای متغیر پاسخ دارد. براساس این شکل، مشاهده می‌شود افزایش اندازه‌ی مدل و در نظر گرفتن تعداد زیاد متغیرهای پیش‌بین سبب بهبود مدل نشده، و فقط سبب بیش‌برازش شدن آن می‌گردد و بهینه‌ترین اندازه برای مدل مربوط به این داده‌ها مدلی با ۳ ترکیب متشکل از ۴ متغیر پیش‌بین است که کمترین آماره‌ی انحراف را در گروه آزمون دارد.

بعد از انتخاب اندازه‌ی مناسب مدل (مدل با ۳ ترکیب منطقی و ۴ متغیر)، یافته‌های جستجوی الگوریتم Annealing برای یافتن مدل رگرسیون لجستیک منطقی مناسب برای بیان ارتباط بین SNP ها و عوامل محیطی شامل: سن، جنس، تری‌گلیسرید بالا، فشار خون بالا، قند خون بالا، اندازه‌ی دور کمر بالا و سیگار با سطح پایین کلسترول ـ HDL به صورت زیر به‌دست آمد:

Logit (HDL ـ پایین کلسترول)=۱/۱۶+۰/۸۵۷(ApoE= ε3 بالا یا TG) +۱/۰۱(TG بالا) +۰/۳۴۳ (جنس=زن)

براساس مدل یاد شده، افرادی که دارای ترکیب منطقی (TG بالا یا ε3 =ApoE) هستند، یعنی الل پلی‌مورفیسم مربوط به ژنApoE آن‌ها به صورت ε3 است، یا تری‌گلیسرید بالایی دارند، نسبت به افراد دیگر شانس ۲/۳۵ برابری با فاصله اطمینان ۹۵٪ (۴/۲۵ و ۱/۳) برای داشتن سطح پایین کلسترول ـHDL دارند. هم‌چنین، داشتن تری‌گلیسرید بالا به تنهایی نیز قادر است شانس ابتلا به کلسترول ـ HDL پایین داشتن را ۲/۷۳ برابر افزایش دهد [فاصله‌ی اطمینان ۹۵٪: (۴/۵۳و ۱/۶۵)]. جنسیت نیز به عنوان یکی از عوامل تاثیرگذار بر کلسترول ـ HDL پایین در این مدل ظاهر شد، به طوری که زنان نسبت به مردان دارای شانس ۱/۴ برابری برای داشتن کلسترول ـ HDL پایین بودند. تاثیر این متغیر با فاصله اطمینان ۹۵٪ : (۲/۱۷ و ۰/۹۱) بر کلسترول ـ HDL پایین معنی‌دار نبود.

براساس مدل به دست آمده مقدار امتیاز این مدل که همان آماره‌ی انحراف مدل لجستیک است، ۵۰۳/۶۶ به دست آمد.

بحث

یافته‌های پژوهش حاضر بیان‌گر ارتباط یک پلی مورفیسم از مجموعه‌ی پلی‌مورفیسم‌های انتخابی با سطح پایین کلسترول ـ HDL است، به طوری‌که حضور الل ε3 ژن آپولیپوپروتئین E منجر به کاهش میزان کلسترول ـ HDL شده، و در نتیجه‌ی آن خطر بروز بیماری‌های قلبی ـ عروقی افزایش می‌یابد. بررسی حاضر وجود یک اثر متقابل بین این ژن و تری‌گلیسرید بالا را نشان می‌دهد، به طوری که براساس ترکیب منطقی " یا" به دست آمده، حضور الل ε3 یا داشتن تری‌گلیسرید بالا، یا حضور همزمان هر دو بر داشتن سطح پایین کلسترول ـ HDL تاثیر معنی‌دار دارد، در حالی‌که در پژوهش‌های قبلی، چنین تقابل‌هایی تشخیص داده نشده و اثر این پلی‌مورفیسم بر سطح کلسترول ـ HDL به صورت اثرات اصلی ارزیابی شده است.[۱۵،۱۸،۱۹] تفاوت در یافته‌های بررسی‌های مختلف نیز می‌تواند ناشی از تفاوت مدل‌های استفاده شده و تکنیک تحلیل دادها باشد، زیرا که روش‌های استفاده شده در بررسی‌های پیشین کمتر به مساله‌ی اثرات متقابل و اهمیت آن پرداخته‌اند و بیشتر تحلیل تک متغیری[i] در مورد اثرات اصلی متغیرها انجام گرفته است. استفاده از روش‌های رگرسیونی برای تحلیل همزمان متغیرها در کنار هم، به‌ویژه با امکان در نظر گرفتن برهمکنشی بین متغیرها می‌تواند منجر به یافته‌های دقیق‌تری گردد.

هم‌چنین، متغیر تری‌گلیسرید بالا به عنوان متغیر مهم و تاثیر گذار بر سطح کلسترول ـ HDL ظاهر شد که با نسبت شانس بالایی میزان کلسترول ـ HDL خون را تحت تاثیر قرار می‌دهد. این متغیر علاوه بر این‌که در حضور الل ε3 متغیر پاسخ را تحت تاثیر قرار می‌دهد، به تنهایی نیز اثر قابل توجهی بر پاسخ دارد، و لازم است پیشگیری و یا درمان تری‌گلیسرید بالا برای پیشگیری از سطح پایین کلسترول ـ HDL مورد توجه قرار گیرد.

i- Analysis

از سوی دیگر، متغیر جنس به عنوان عاملی موثر بر میزان کلسترول ـ HDL در مدل ظاهر شده، اما اثر متقابلی با سایر متغیرها ندارد. اگرچه اثر این متغیر از نظر آماری معنی‌دار نیست ولی نشان می‌دهد در بررسی عوامل موثر بر سطح کلسترول ـ HDL باید اثر عوامل مخدوش‌گر نیز در نظر گرفته شود.

یکی از هدف‌های بررسی‌های ژنتیکی، تشخیص اثرات اصلی و متقابل پلی‌مورفیسم‌ها بر بروز بیماری‌ها می‌باشد. از آنجا که گاهی اثرات تقابلی و برهمکنشی بین پلی‌مورفیسم‌ها تاثیرگذارتر از اثرات انفرادی آن‌ها می‌باشد، نیاز به روش‌هایی است که قادر به یافتن چنین اثرات متقابلی باشد. هم‌چنین، روش استفاده شده باید قادر به تعیین شدت اثر این تقابل‌ها نیز باشد. یکی از روش‌هایی که به تازگی معرفی شده و در بررسی‌های ژنتیکی کاربرد فراوان یافته، روش رگرسیون منطقی می‌باشد. این روش رگرسیونی، برای یافتن اثرات متقابل بین پلی‌مورفیسم‌ها و سایر عوامل خطر که به طور معمول با روش‌های مرسوم قادر به یافتن آن‌ها نیستیم، استفاده می‌شود. برتری رگرسیون منطقی نسبت به سایر روش‌های تجزیه و تحلیل متغیرهای دوحالتی مانند روش شبکه‌های عصبی مصنوعی و درخت تصمیم‌گیری، این است که یافته‌های رگرسیون منطقی به طور کامل به شکل یک مدل رگرسیون نوشته می‌شود، و در نتیجه امکان تفسیر ضرایب، انجام آزمون فرضیه در مورد ضرایب، و همچنین ارزیابی کفایت مدل با استفاده از تابع امتیاز آن مدل وجود دارد.[۲۰] این روش رگرسیونی به تازگی در بررسی‌های متعددی برای تحلیل داده‌ها استفاده گردیده از جمله در ایران، محرابی و همکاران در پژوهشی مدل رگرسیون منطقی را به منظور پیش‌بینی بروز دیابت ارایه داده‌اند.[۲۰] هم‌چنین، در پژوهشی که برای بررسی ارتباط پلی‌مورفیسم‌ها با بروز سرطان پستان انجام شده پلی‌مورفیسم‌های مرتبط با این سرطان و برهم‌کنش بین آن‌ها با روش رگرسیون منطقی شناسایی گردیده است.[۲۱] در بررسی دیگری اثر برخی پلی‌مورفیسم‌ها بر سطح کلسترول ـ LDL خون بررسی شده که در این پژوهش نیز اثرات متقابل بین این پلی‌مورفیسم‌ها با رگرسیون منطقی شناسایی شده‌اند.[۲۲]

در بررسی‌های ژنی، به دلیل اهمیت اثرات متقابل بر بروز بیماری‌ها، شناسایی این اثرات متقابل دارای اهمیت می‌باشد. رگرسیون منطقی روش مناسبی به منظور شناسایی و تعیین شدت اثر چنین اثرات متقابلی است. در بررسی حاضر نیز با استفاده از رگرسیون منطقی، اثرات متقابل موثر بین پلی‌مورفیسم‌ها بر سطح کلسترول ـ HDL با حضور عوامل مداخله‌گر بررسی گردید.

از محدودیت‌های پژوهش می‌توان بیان نمود که در وهله‌ی اول تعداد پلی‌مورفیسم‌های قابل اندازه‌گیری، در مقایسه با پلی‌مورفیسم‌های درگیر با کلسترول ـ HDL کم بود و نیز محدودیت بعدی در رابطه با جمع‌آوری سایر متغیرهای موثر بر سطح کلسترول ـ HDL بود.

سپاسگزاری: در پژوهش حاضر، از داده‌های طرح قند و لیپید تهران که توسط پژوهشکده‌ی علوم غدد درون‌ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی اجرا شده، استفاده گردید. پژوهش‌گران از تمام کسانی که در طراحی و جمع‌آوری داده‌های TLGS مشارکت داشتند نهایت قدردانی را به عمل می‌آورد. این مقاله، برگرفته از پایان‌نامه‌ی خانم پروین سربخش (دانشجوی دکتری آمار زیستی) است که در ضمن طرح مصوب پژوهشکده نیز می‌باشد.

References

1. Lucek PR, Ott J. Neural network analysis of complex traits. Genet Epidemiol 1997; 14: 1101-6.
2. Vermeulen SH, Den Heijer M, Sham P, Knight J. Application of multi-locus analytical methods to identify interacting loci in case-control studies. Ann Hum Genet 2007; 71: 689-700.
3. Ruczinski I, Kooperberg C, LeBlanc M. Logic Regression. Journal of Computational and Graphical Statistics 2003; 12: 475-511.
4. Goldbourt U, Yaari S, Medalie JH. Isolated Low HDL cholesterol as a risk factor for coronary heart disease mortality: a 21-year follow-up of 8000 men. Arterioscler Thromb Vasc Biol 1997; 17: 107-13.
5. Kim SM, Han JH, Park HS. Prevalence of low HDL-cholesterol levels and associated factors among Koreans. Circ J 2006; 70: 820-6.
6. Dwyer T, Calvert GD, Baghurst KI, Leitch DR. Diet, other lifestyle factors and HDL cholesterol in a population of Australian male service recruits. Am J Epidemiol 1981; 114: 683-96.
7. Brown CM, Rea TJ, Hamon SC, Hixson JE, Boerwinkle E, Clark AG, et al. The contribution of individual and pairwise combinations of SNPs in the APOA1 and APOC3 genes to interindividual HDL-C variability. J Mol Med (Berl) 2006; 84: 561-72.
8. McCarthy JJ, Lehner T, Reeves C, Moliterno DJ, Newby LK, Rogers WJ, et al. Association of genetic variants in the HDL receptor, SR-B1, with abnormal lipids in women with coronary artery disease. J Med Genet 2003; 40: 453-8.

9. Frikke-Schmidt R. Context-dependent and invariant associations between APOE genotype and levels of lipoproteins and risk of ischemic heart disease: a review. Scand J Clin Lab Invest Suppl 2000; 233: 3-25.
10. Azizi F, Salehi P, Etemadi A, Zahedi-Asl S. Prevalence of metabolic syndrome in an urban population: Tehran Lipid and Glucose Study. Diabetes Res Clin Pract 2003; 61: 29-37.
11. Daneshpour MS, Faam B, Hedayati M, Eshraghi P, Azizi F. ApoB (XbaI) polymorphism and lipid variation in Teharnian population. European Journal of Lipid Science and Technology 113: 436-40.
12. Daneshpour MS, Hedayati M, Azizi M. Hepatic Lipase C-514T polymorphism and its association with HDL-C level in Tehran. Kowsar Medical Journal 2005; 10: 135-42.[Farsi]
13. Daneshpour MS, Hedayati M, Azari F, Ghasemi F, Azizi F. Association between the cholesteryl ester transfer protein_TaqI polymorphism and low HDL-C concentration in Tehran population. Iranian Journal of Endocrinology and Metabolism 2004; Suppl 5: 355-61. [Farsi]
14. 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 Präventivmed 2002; 47: 408-26.
15. Daneshpour MS, Hedayati M, Eshraghi P, Azizi F. Association of Apo E gene polymorphism with HDL level in Tehranian population. European Journal of Lipid Science and Technology 112: 810-6.
16. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The 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). JAMA 2001; 285: 2486-97.
17. 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: 3141-421.
18. Zaman MM, Ikemoto S, Yoshiike N, Date C, Yokoyama T, Tanaka H. Association of apolipoprotein genetic polymorphisms with plasma cholesterol in a Japanese rural population: The Shibata Study. Arterioscler Thromb Vasc Biol 1997; 17: 3495-504.
19. Saidi S, Slamia LB, Ammou SB, Mahjoub T, Almawi WY. Association of apolipoprotein E gene polymerphism with ischemic stroke involving large-vessel disease and its relation to serum lipid levels. J Stroke Cerebrovasc Dis 2007; 16: 160-6.
20. Mehrabi Y, Sarbakhsh P, Khadem-Maboudi A, Hadaegh F. Prediction of Diabetes Using Logic Regression. Iranian Journal of Endocrinology and Metabolism 2010; 12: 16-24. [Farsi]
21. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. Biostatistics 2008; 9: 187-98.
22. Voora D, Reed CR, Zhai J, Salisbury BA, Shah SH, Ginsburg GS. Polymorphisms in ABCA1 predict statin mediated LDL cholesterol lowering and suggest an interaction with CETP. Circulation 2007; 116: II-178.

Original Article

Logic Regression Analysis for Finding Interaction Effects of Genes Polymorphisms and Other Risk Factors on Low HDL: Tehran Lipid and Glucose Study

Sarbakhsh P[1], Mehrabi Y[2], Daneshpour M[3], Zayeri F[1], Namdari M[1], Azizi F[4]

[1]Department of Biostatistics, Faculty of Paramedicine, & [2]Department of Epidemiology, Faculty of Public Health, & [3] Obesity Research Center, & [4] Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences,Tehran, I.R. Iran

e-mail: ymehrabi@gmail.com

Received: 07/02/2012 **Accepted**: 02/05/2012

Abstract

Introduction: Logic regression is a generalized regression method that can identify complex Boolean interactions of binary variables. This method has been successfully used for analyzing single-nucleotide polymorphism data, because in SNP association studies interactions are important. The aim of this study is to investigate the associations between some candidate gene polymorphisms and HDL concentration using Logic Regression. Materials and Methods: Subjects for this cross sectional study, 436 subjects (172 men and 264 women) aged≥20 with some polymorphisms, were randomly selected from among participants of the Tehran Lipid and Glucose Study (TLGS). Logic regression analysis was used to identify combinations of main genetic effects and interactions associated with HDL. Cross validation and randomization test were done to avoid over fitting of the models. Results: Cross validation test suggested that the Logic model with four Boolean combinations and four predictors was the best logic model, which after fitting, showed that individuals who carry Apoe SNP ε3 or have high TG have an odds ratio of 2.35 (CI 95%:1.3-4.25) for having low HDL compared to other subjects. Also subjects with high TG have odds ratio 2.73 (CI 95%: 1.65,4.53) for having low HDL. Conclusion: Results of this study shows that Logic Regression is a powerful method to determine the interaction effect between high TG and ApoE SNP for having low HDL.

Keywords: Interaction, Annealing algorithm, SNP, Logic regression, low HDL, TLGS