بررسی ریسک فاکتورهای سرطان پستان با استفاده از تحلیل مدل های درختی

امل ساکی[1] - ابراهیم حاجی زاده[2] - نجمه تهرانیان[3]

چکیده

زمینه و هدف: در بسیاری از تحقیقات پزشکی هدف تعیین ریسک فاکتورهای بیماری مورد نظر با توجه به اثرات متقابل آن‌ها و رده‌بندی بیماران بر اساس این ریسک فاکتورها است. برای این منظور از مدل‌های رگرسیون، تحلیل ممیزی و رده‌بندی استفاده می‌شود. حال آن‌که بررسی پیش‌فرض‌ها برای این مدل‌ها در داده‌های پزشکی اغلب بسیار مشکل خواهد بود. لذا باید از روش‌های جایگزین استفاده نمود. هدف از این مقاله ارائه ی روش رده‌بندی درختی برای بررسی ریسک فاکتورهای سرطان پستان است. این روش در سال 1984 توسط بریمن و همکاران تحت عنوان تحلیل رده‌بندی و رگرسیون درختی عمومیت یافت.

روش تحقیق: در این مطالعه از داده‌های یک مطالعه موردی شاهدی استفاده شده است که در آن گروه مورد شامل 312 خانم مبتلا به سرطان پستان زیر 40 سال با تشخیص قطعی بر مبنای آسیب‌شناسی و گروه شاهد شامل 312 خانم زیر 40 سال مراجعه‌کننده به بیمارستان امام خمینی تهران به علت بیماری غیرنئوپلاستیک و غیرهورمونی بودند. این داده‌ها با استفاده از مدل‌های رده‌بندی درختی و با استفاده از نرم‌افزار CART تحلیل شدند.

یافته‌ها: نتایج حاصل نشان داد که متغیرهایی مانند سابقه ی خانوادگی سرطان پستان، سابقه ی خانوادگی سرطان تخمدان، سابقه ی بیماری خوش‌خیم پستان، سن منارک، وضعیت قاعدگی، عدم فعالیت فیزیکی، قرار گرفتن در معرض حوادث استرس‌زا، عواملی مؤثر در ابتلا به سرطان هستند.

نتیجه گیری: تعیین ریسک فاکتورها در سرطان پستان اهمیت ویژه ای دارد. روش‌های رده‌بندی درختی به دلیل عدم نیاز به برقراری پیش فرض‌های سخت، ابزاری قدرتمند برای بررسی ریسک فاکتورهاست. همچنین به دلیل شهودی بودن مدل‌های درختی، تفسیر داده‌ها برای محققین ساده‌تر می‌باشد.

کلید واژه ها: تقسیم‌بندی بازگشتی داده‌ها؛ سرطان پستان؛ مدل رده‌بندی و رگرسیون درختی

افق دانش؛ فصلنامه ی دانشگاه علوم پزشکی و خدمات بهداشتی درمانی گناباد (دوره ی 17؛ شماره ی 1؛ بهار 1390)

دریافت: 1388/4/17 اصلاح نهایی: 1389/12/20 پذیرش: 1389/12/22

1- کارشناس ارشد آمار زیستی، گروه آمار زیستی، دانشگاه تربیت مدرس، تهران

2- **نویسنده ی مسؤول؛** دانشیار، گروه آمار زیستی، دانشگاه تربیت مدرس، تهران

آدرس: تهران- خیابان جلال آل احمد - پل نصر - دانشگاه تربیت مدرس- گروه آمار زیستی

تلفن: 021-82883810 **نمابر:** 021- 82884555 **پست الکترونیکی:** Hajizadeh@modares.ac.ir

3- مربی، گروه مامایی، دانشگاه تربیت مدرس، تهران

مقدمه

سرطان در بار جهانی بیماری در دهه‌های آینده، عاملی مهم و فزاینده خواهد بود و انتظار می‌رود تعداد موارد جدید سرطان، از 10 میلیون نفر در سال 2000 میلادی به 15 میلیون نفر در سال 2020 افزایش یابد که حدود 60 درصد این موارد جدید، در کشورهای کمتر توسعه یافته جهان ایجاد می‌شوند (1). سرطان پستان شایع‌ترین سرطان در زنان است و بعد از سرطان ریه، شایع‌ترین علت مرگ ناشی از سرطان در زنان می‌باشد. خطر ایجاد سرطان پستان در طول عمر زنان 12/5 درصد (یعنی یک مورد از هشت مورد) و خطر مرگ ناشی از سرطان پستان 3/6 درصد (یک مورد از بیست و هشت مورد) می باشد (2). در حدود 10 درصد زنان در آمریکا در مرحله‌ای از زندگی خود دچار سرطان پستان می‌شوند و این سرطان بعد از سرطان ریه دومین علت مرگ و میر ناشی از سرطان است. در حدود 5 درصد سرطان های پستان ارثی و 80 درصد تا 90 درصد اسپورادیک هستند و اگر چه در سنین بالای 50 سالگی شایع‌تر است ولی در هر سنی ممکن است رخ دهد (3). بر اساس پژوهش‌های انجام شده در سال‌های اخیر، سرطان پستان فراوان‌ترین نوع سرطان در میان زنان ایران به شمار می‌آید (4). مقایسه آمارهای جدید با گزارش‌های پیشین، نشانگر افزایش به نسبت سریع میزان بروز این بیماری در زنان کشورمان می‌باشد. در حال حاضر مرگ و میر ناشی از این بیماری کمتر از شیوع آن بوده و از هر 18-10زن سرانجام یک نفر فوت می کند. بنابراین سرطان پستان در سنین 50-35 سالگی شایع‌ترین علت مرگ زنان به حساب می‌آید (4). با توجه به اهمیت این موضوع و شیوع این بیماری، هدف ما در این مقاله تعیین ریسک فاکتورها با توجه به اثرات متقابل آن‌ها و رده‌بندی بیماران بر اساس این ریسک فاکتورها است. به دلیل محدودیت مدل‌های خطی و غیرخطی پارامتری مانند رگرسیون خطی و رگرسیون لجستیک در تعیین ریسک فاکتورها و همچنین تفسیر پیچیده این مدل‌ها برای محققین بالینی، در این مطالعه از روش رده‌بندی درختی برای بررسی ریسک فاکتورهای سرطان پستان استفاده شده است. تحلیل رده‌بندی درختی با استفاده از تقسیم‌بندی بازگشتی داده‌ها با استفاده از ضابطه ی جینی یک ساختار درخت تصمیم را می‌سازد و افراد را به گروه‌هایی با خصوصیات مشابه بالینی رده‌بندی می‌کند. کاربرد این روش‌ها بسیار وسیع و گسترده است و بیشترین کاربرد آن در حقیقات پزشکی برای آزمون‌های تشخیصی است (5). یکی از مزیت‌های این روش‌ها این است که نتایج به صورت یک نمودار درخت تصمیم نمایش داده می‌شوند که منجر به درک و تفسیر آسان آن‌ها برای محققان و پزشکان شده است (6). این مدل‌ها یک نوع مدل‌های پیش‌بینی بالینی هستند که با در نظر گرفتن اثر متقابل ویژگی‌های بالینی بیماران، نتایج بالینی آن‌ها را پیش بینی می‌کنند. این پیش‌بینی‌ها در تصمیم‌گیری‌های بالینی از جمله انتخاب بیماران برای درمان‌های مداخله‌ای و یا بهترین زمان ممکن برای شروع درمان و ... مؤثر هستند (7).

مدل‌های درختی که در رده‌بندی و رگرسیون درختی استفاده می‌شوند توسط مورگان و سونی‌کوئیست در سال 1963 برای بررسی اثرات متقابل متغیرها در داده‌های علوم اجتماعی پیشنهاد شده‌اند (8) و جنبه‌های نظری و کاربردی آن توسط بریمن و همکارانش در سال 1984 در رساله‌ای که در این مورد منتشر گردید، بسط و توسعه داده شد. به طور کلی روش‌های مبتنی بر مدل‌های خطی، فضای متغیرهای کمکی را به ناحیه‌های مجزا تقسیم می کند و داده‌ها را به گروه‌های متناظر تخصیص می دهد (7). این روش‌ها داده‌ها را به طور بازگشتی برای تعیین یا معرفی اثرات متقابل متغیرها و معرفی زیرگروه هایی از افراد با مشخصات دموگرافی و علایم بالینی مشابه برای تشخیص های پزشکی بعدی تقسیم می‌کند. با توجه به نوع مسئله، هدف اساسی در یک مطالعه ی مدل‌های رده‌بندی و رگرسیون درختی می تواند ایجاد یک رده‌بندی کننده دقیق و یا کشف یک ساختار پیش‌بینی‌کننده برای مسئله ی مورد نظر باشد. اگر هدف تعیین یک ساختار پیش بینی کننده باشد، آنگاه درک صحیح متغیرها و اثرات متقابل آن‌ها ضروری است. معمولاً در مسائل مختلف این دو هدف به موازات هم بررسی می‌شوند (9). مدل‌های رده بندی درختی نسبت به مدل لجستیک چندگانه دارای تفسیر ساده‌تری است. به همین دلیل این روش در مطالعات بالینی

کاربردی‌تر است، از طرف دیگر منطق اصلی داده‌ها به سادگی در درخت مشهود می‌باشد. به دلایل ذکر شده رده‌بندی و رگرسیون درختی توسعه ی زیادی یافته است (8).

روش تحقیق

جامعه ی مورد پژوهش شامل خانم‌های مبتلا به سرطان پستان بستری شده در بخش‌های 1و 3 کانسر و سانترال 1 بیمارستان امام خمینی و نیز بیماران مبتلا در 10 سال اخیر بودند که برای استخراج اطلاعات آن‌ها از پرونده‌های موجود در مرکز ثبت مدارک پزشکی استفاده شده است. نمونه‌ها شامل اطلاعات مربوط به 312 خانم بیمار مبتلا به سرطان پستان و 312 خانم مراجعه کننده جهت درمان بیماری‌های مختلف به غیر از بیماری نئوپلاستیک و هورمونی به این بیمارستان بودند. حجم نمونه بر اساس نسبت متغیر سابقه ی فامیلی سرطان پستان در مطالعه ی ابراهیمی و همکارانش (10) با انتخاب p_1=0/092 برای گروه مورد، p_2=0/024 برای گروه شاهد و با اختیار α=0/05 و توان آزمون 95 درصد، برآورد شده است.

$$n=\frac{(z_1-\frac{\alpha}{2}+Z_{1-\beta})^2[p_1(1-p_1)+p_2(1-p_2)]}{(p_1-p_2)^2}$$

$$n=\frac{(1.96+1.645)^2[0.092(1-0.092)+0.024(1-0.024)]}{(0.092-0.024)^2}=301$$

در این مطالعه اطلاعات لازم از طریق مصاحبه ی حضوری و تلفنی، اسناد و مدارک پزشکی مربوط به پرونده‌های بیماران جهت تکمیل اطلاعات و پرسشنامه (شامل سؤالات مربوط به فاکتورهای باروری، وضعیت تغذیه ای و ...) جمع آوری شده است.

متغیر پاسخ مورد نظر وجود بیماری سرطان پستان است که یک متغیر دو حالتی در نظر گرفته شده است. بنابراین مدل رده بندی درختی برای رده‌بندی افراد به دو رده ی سالم و بیمار مورد استفاده قرار می‌گیرد. در واقع در این مدل رده‌بندی با استفاده از ضابطه ی جینی که یک ضابطه ی نیکویی افراز است، افراد را به زیرگروه‌هایی با ویژگی‌های مشابه بالینی تقسیم‌بندی می‌کند. یکی از مزیت‌های این روش علاوه بر شهودی بودن گروه‌های مشابه بالینی، مشاهده اثرات متقابل بین متغیرهای پیش‌بین یا ریسک فاکتورهاست. زیرا متغیرهایی در مدل آشکار می‌شوند که با هم اثر متقابل داشته باشند. متغیرهای مورد استفاده برای رده‌بندی بیماران در این مدل به عنوان ریسک فاکتورها در نظر گرفته می‌شوند. در این مطالعه صحت مدل با استفاده از مساحت زیر منحنی ROC[1] اندازه‌گیری می‌شود که بر اساس حساسیت و ویژگی مدل مورد بحث، اجرا می‌شود. مقادیر بیش از 80 درصد مقدار مساحت زیر منحنی ROC نشان دهنده ی توان بالای مدل در رده‌بندی ممیزی و مقادیر بین 70 درصد تا 80 درصد بیانگر قابل قبول بودن مدل تشخیصی است (11). در این مطالعه سه مدل درختی مجزا مورد بررسی قرار گرفتند. مدل 1 برای افراد متاهل، مدل 2 برای افراد مجرد و مدل 3 به طور مشترک برای هر دو گروه ساخته شد. این سه مدل با معرفی افراد با ویژگی‌های مشابه بالینی و تعیین ریسک فاکتورهای مطرح بیماری مورد مطالعه، کمک شایانی به تشخیص بیماری خواهند کرد. سپس با استفاده از آزمون کای دو و آزمون دقیق فیشر ارتباط بین متغیرها مورد بررسی قرار گرفت. آنالیز این مطالعه با نرم افزار CART 6.0 و SPSS 16.0 انجام شد.

یافته ها

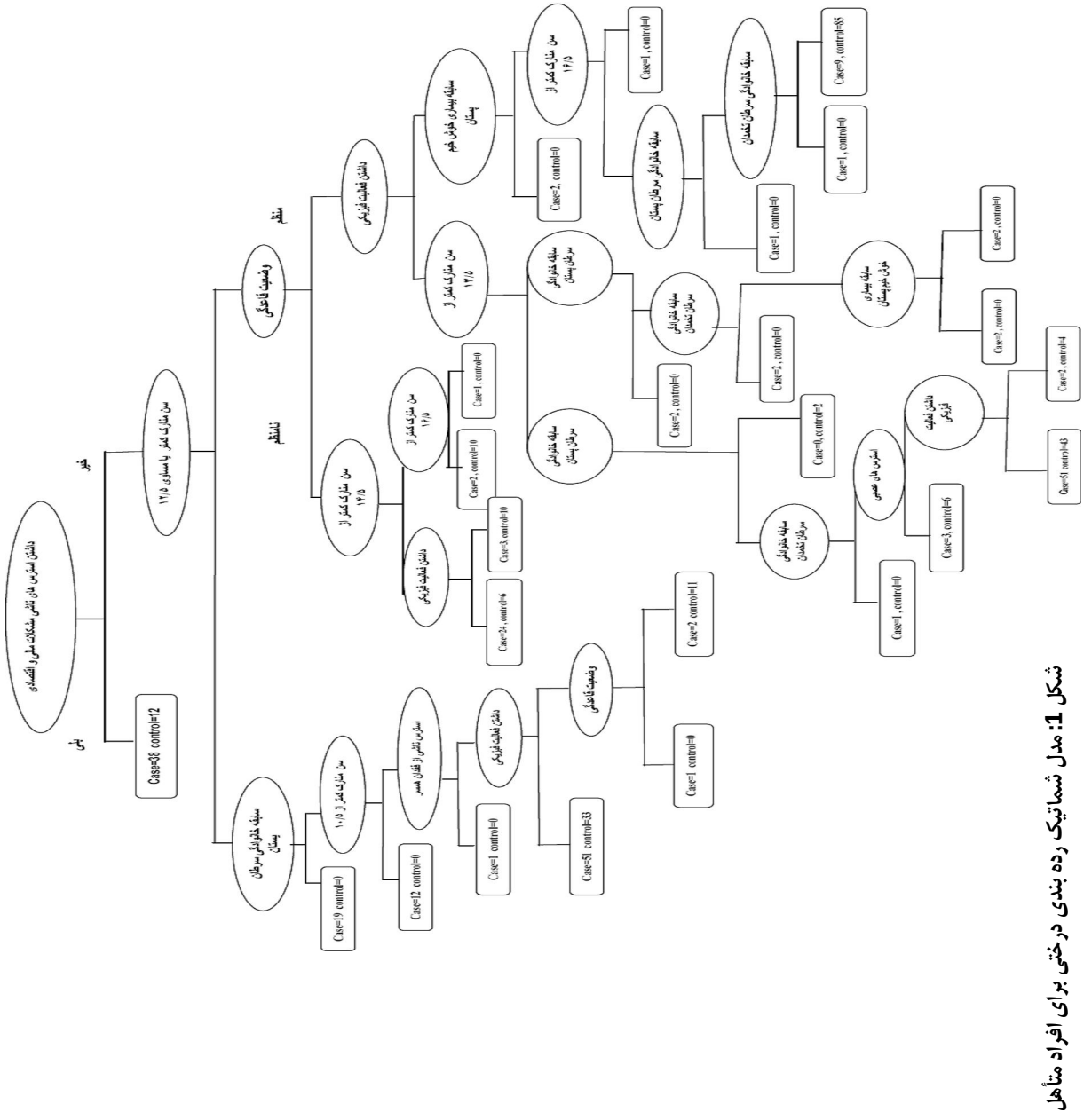
سه مدل رده‌بندی درختی با متغیرهای سابقه ی خانوادگی سرطان پستان، سابقه ی خانوادگی سرطان تخمدان، سابقه ی بیماری خوش خیم پستان، سن منارک، وضعیت قاعدگی، فعالیت فیزیکی، قرار گرفتن در معرض حوادث استرس‌زا که عبارتند از: مرگ همسر، مرگ فرزند، مرگ یکی از اعضای خانواده، بروز حوادث غیرمترقبه و متارکه ی همسر و عوامل عصبی ساخته شدند. دو مدل اول به طور مجزا روی گروه افراد متاهل و مجرد و مدل سوم روی هر دو گروه با هم اجرا شد.

در مدل‌های ارائه شده در این مطالعه، افراد بیمار و سالم بر اساس ضابطه ی افراز جینی انجام شده است. این ضابطه

1- Receiver Operator Curve

متغیرهایی را برای افراز انتخاب می‌کند که منجر بـه بهتـرین افراز ممکن روی داده‌ها می‌شوند. بـه ایـن ترتیـب ضـابطه ی جینی به عنوان ضابطه‌ای بـرای یـافتن بهتـرین متغیرهـا بـه عنوان عوامل مؤثر در افراز افـراد بـه زیرگـروه‌هـای مشـابه از لحاظ بالینی عمل می‌کند. بنابراین در این مطالعه متغیرهـای شناسایی شده به وسیله ی ضابطه ی جینی به عنوان ریسـک فاکتورهای سرطان پستان معرفی شدند.

در مـدل شـماتیک رده بنـدی درختـی کـه بـرای افـراد متاهل اجرا شده است، متغیرهای سابقه ی خانوادگی سـرطان پسـتان، سـابقه ی خـانوادگی سـرطان تخمـدان، سـابقه ی بیماری خوش‌خیم پستان، سـن منـارک، وضـعیت قاعـدگی، نداشتن فعالیت فیزیکی و اسـترس‌هـای ناشـی از متارکـه ی همسر به عنوان عوامل خطر مشاهده شدند. اندازه ی کـارایی این مدل در جداسازی افراد بیمار و سالم بـا مقـدار مسـاحت زیرمنحنی ROC نشان داده شده اسـت، کـه ایـن مقـدار در مدل (1) 83 درصد است (شکل 1).



شکل 1: مدل شماتیک رده بندی درختی برای افراد متأهل

همان طور که قبلاً اشاره شد، یکی از اهداف تحلیل مدل‌های رده‌بندی درختی تعیین زیرگروه‌هایی از افراد با سطح خطر مشابه است، لذا در مدل مورد بررسی نیز افراد با مشخصات زیر تحت عنوان زیرگروه‌های دارای خطر بالای ابتلا به سرطان پستان تشخیص داده شده‌اند.

1- افراد دارای سن منارک کمتر از 10/5سال

2- داشتن سن منارک کمتر از 12/5 سال به همراه هر یک از ویژگی های زیر:

داشتن سابقه ی خانوادگی سرطان پستان

وضعیت قاعدگی نامنظم

فشار روانی ناشی از متارکه ی همسر

نداشتن هیچ فعالیت فیزیکی

3- نداشتن فعالیت فیزیکی به همراه هر یک از ویژگی‌های بالینی زیر:

داشتن سابقه ی خانوادگی سرطان پستان

داشتن سابقه ی خانوادگی سرطان تخمدان

سابقه ی بیماری خوش خیم پستان

4- افرادی که در معرض استرس‌های ناشی از مشکلات مالی هستند، که برای بررسی بیشتر اثرات بالا، رابطه ی هر یک از این عوامل را با ابتلا به بیماری با استفاده از آزمون کای دو و آزمون دقیق فیشر آنالیز کرده و نتایج زیر حاصل شده است. در جداول 1 و2 نیز توزیع فراوانی متغیرهای معنادار، سابقه ی خانوادگی سرطان پستان و سرطان تخمدان مشاهده می‌شود که نشان‌دهنده ی رابطه ی مثبت بین سابقه ی خانوادگی سرطان و بالا بودن خطر ابتلا به سرطان پستان است.

جدول 1: توزیع فراوانی سابقه ی خانوادگی سرطان پستان در افراد متأهل

گروه	سابقه خانوادگی دارد					سابقه خانوادگی ندارد	جمع کل
	فامیل درجه ی 1	فامیل درجه ی 2	فامیل درجه ی 3	فامیل درجه ی 1و2	فامیل درجه ی 2و3		
مورد	10 (%4)	10 (%4)	6 (%2/4)	3 (%1/2)	1 (%0/4)	220 (%88)	250
شاهد	1 (%0/3)	2 (%0/7)	3 (%1/1)	0 (%0/0)	0 (%0/0)	277 (%97/9)	283
جمع کل	11 (%2/2)	12 (%2/3)	9 (%1/7)	3 (%0/6)	1 (%0/2)	497 (%93)	533

جدول 2: توزیع فراوانی سابقه ی خانوادگی سرطان تخمدان در افراد متأهل

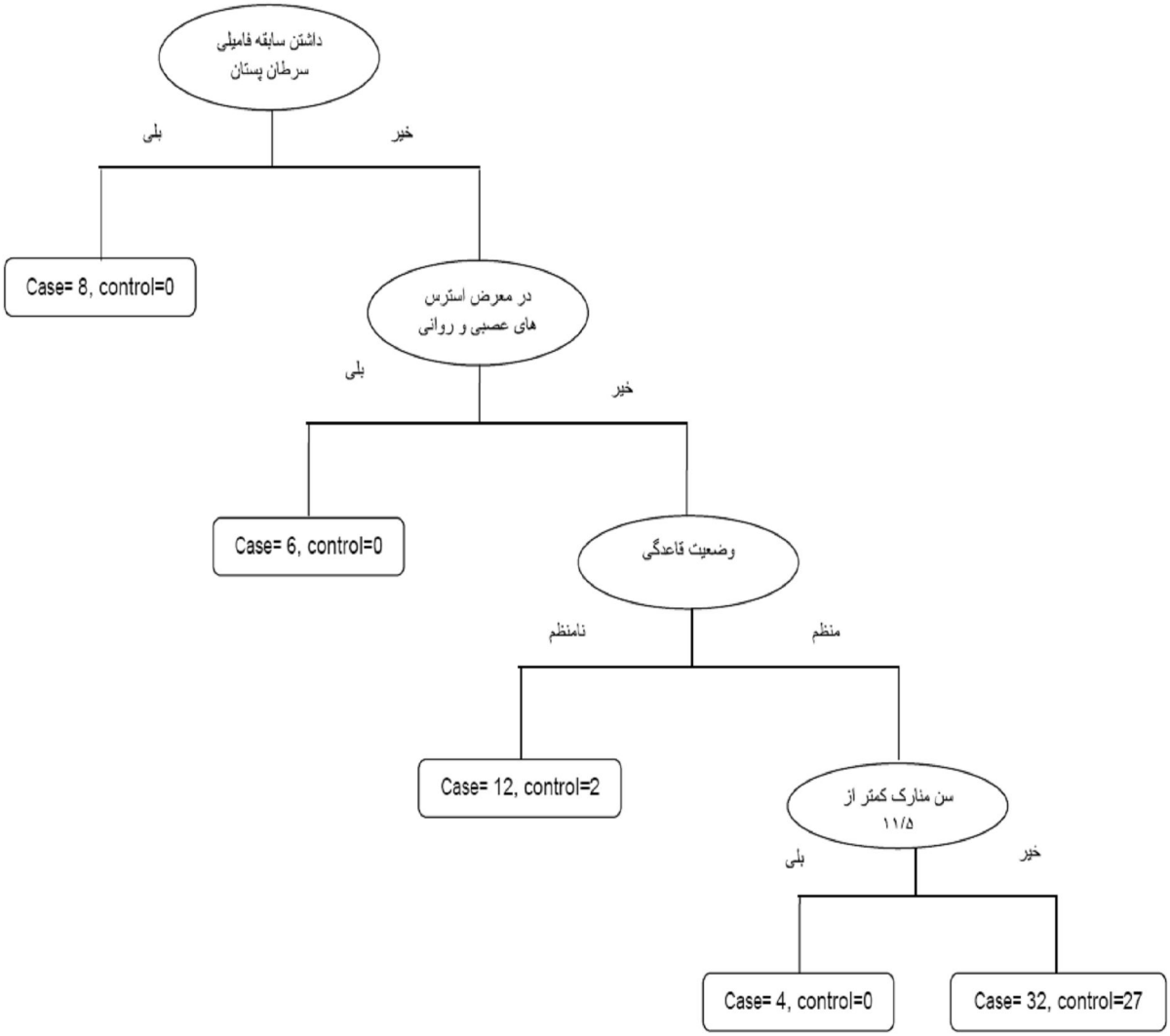
گروه	سابقه ی خانوادگی دارد		سابقه ی خانوادگی ندارد	جمع کل
	فامیل درجه ی1	فامیل درجه ی 2		
مورد	2 (%0/8)	3 (%1/2)	245 (%98)	250
شاهد	0 (%0/0)	0 (%0/0)	283 (%100)	283
جمع کل	2 (%0/4)	3 (%0/6)	528 (%99)	533

آزمون کای دو در سطح خطای 5 درصد نشان می‌دهد که بین سابقه ی خانوادگی سرطان پستان و ابتلا به بیماری سرطان رابطه وجود دارد ($p<0/001$). همچنین بین سابقه ی خانوادگی سرطان تخمدان و احتمال ابتلا به سرطان پستان رابطه معنی‌داری وجود دارد ($p=0/022$).

سن منارک، وضعیت قاعدگی و استرس‌های ناشی از مشکلات مالی نیز نشان‌دهنده ی رابطه ی قوی بین این عوامل و احتمال ابتلا به بیماری سرطان پستان است ($p<0/001$). در مورد دیگر عوامل مؤثر معرفی شده در مدل درختی نیز این آنالیز اجرا شد و نتایج به دست آمده تأیید کننده ی تحلیل مدل رده‌بندی درختی اجرا شده و ریسک فاکتورهای معرفی شده به وسیله ی مدل است.

در مدل رده بندی درختی برای افراد مجرد از بین متغیرهای اصلی که وارد مدل شده‌اند، فقط متغیرهای سابقه ی خانوادگی سرطان پستان، سن منارک، وضعیت قاعدگی و مشکلات عصبی در مدل مشاهده

می‌شوند که در احتمال بروز بیماری مؤثرند (شکل 2). مقدار مساحت زیر منحنی OC 72/2 درصد است. در بررسی بیشتر عوامل خطر در بین مجردها مشاهده می‌شود که 8 نفر دارای عامل سابقه ی خانوادگی سرطان پستان بودند که وضعیت سابقه ی خانوادگی آن‌ها در جدول 3 مشاهده می‌شود.



شکل 2: مدل رده بندی درختی برای افراد مجرد

جدول 3: توزیع فراوانی سابقه ی خانوادگی سرطان پستان در افراد مجرد

گروه	سابقه خانوادگی دارد			سابقه خانوادگی ندارد	جمع کل
	فامیل درجه ی 1	فامیل درجه ی 2	فامیل درجه ی 3		
مورد	4 (%6/5)	1 (%1/6)	3 (%4/8)	54 (%87/1)	62
شاهد	0 (%0/0)	0 (%0/0)	0 (%0/0)	29 (%100)	29
جمع کل	4 (%4/4)	4 (%4/4)	3 (%3/3)	83 (%87/9)	91

در مورد وضعیت قاعدگی نامنظم، 14 نفر دارای این ویژگی بودند که از این تعداد 12 نفر مبتلا به سرطان پستان هستند (جدول 4).

جدول 4: توزیع فراوانی وضعیت قاعدگی در افراد مجرد

گروه	قاعدگی منظم	قاعدگی نامنظم	جمع کل
مورد	47 (%79/7)	12 (%20/3)	59
شاهد	27 (%93)	2 (%7)	29
جمع کل	74 (%84)	14 (%16)	88

همچنین در مورد سن منارک 8 نفر دارای سن منارک کمتر از 11/5 سال هستند که همگی آن‌ها نیز در رده ی افراد مبتلا هستند (جدول 5). لذا در این مدل می‌توان یکی از زیرگروه‌های با ریسک بالای ابتلا به سرطان را افراد با سن منارک کمتر از 11/5 سال معرفی کرد.

جدول 5: توزیع فراوانی سن منارک در افراد مجرد

گروه ضعیت	سن کمتر از 11/5	سن بیشتر از 11/5	جمع کل
مورد	8 (%13/6)	51 (%86/4)	59
شاهد	0 (%0/0)	29 (%100)	29
جمع کل	8 (%9)	80 (%91)	88

مدل (3) برای افراد متاهل و مجرد با هم ساخته شده است. در این مدل پس از ورود همه ی متغیرهای مورد بحث، متغیرهای سابقه ی خانوادگی سرطان پستان، سن منارک، وضعیت قاعدگی و وضعیت تأهل به عنوان عوامل خطر در مدل قابل مشاهده بودند. مقدار مساحت زیرمنحنی ROC 80 درصد است. در این مدل نیز افراد با خطر ابتلای بالا به گروه‌های زیر تقسیم‌بندی شدند:

1- افراد دارای سن بالاتر از 39 سال

2- افراد دارای سن منارک کمتر از 12/5 و سابقه ی خانوادگی سرطان پستان

3- افراد دارای سن منارک کمتر از 10/5

4- افراد دارای وضعیت قاعدگی نامنظم

5- افراد مجرد

وضعیت تأهل به عنوان یک عامل مؤثر در مدل قابل مشاهده است.

بحث

با توجه به اهمیت موضوع مورد بحث، شناسایی ریسک فاکتورهای مربوط در بسیاری از مطالعات مورد توجه قرار گرفته است. از میان عواملی که در این مطالعه به عنوان عوامل تأثیرگذار آشکار شدند، متغیرهای سن پایین منارک و داشتن سابقه ی خانوادگی سرطان و وضعیت نامنظم قاعدگی و بیماری‌های خوش‌خیم پستان، عواملی هستند که در بیشتر مطالعات به تأثیرگذاری آن‌ها اشاره شده است. در مطالعه‌ای که توسط عطار پارسایی و همکارانش در مورد بررسی ارتباط مشخصات فردی، اجتماعی، سبک زندگی و عوامل تنش‌زا با سرطان پستان در زنان صورت گرفته به تأثیر فعالیت ورزشی در بیماری و هم چنین رابطه ی معنی‌دار بین سابقه ی خانوادگی درجه ی یک و بیماری خوش‌خیم پستان با ابتلا به سرطان پستان اشاره شده است. ولی در این مطالعه ارتباط معنی‌داری بین اولین سن قاعدگی و ابتلا به سرطان پستان مشاهده نشده است. همچنین در این مطالعه اکثریت افراد گروه مورد، تنش (در اثر مرگ همسر، مشکلات مالی، اختلافات خانوادگی و ...) زیادی را قبل از ابتلا تجربه کرده بودند که این رابطه معنی‌دار بوده است (12). در مطالعه ی دیگری سن منارک کمتر از 12 سال و مجرد بودن از عوامل مؤثر در بروز سرطان پستان معرفی شده‌اند (10). ولی در مطالعه‌ای که دکتر هلاکویی و همکارانش در استان مازندران انجام داده‌اند، ارتباط آماری معنی‌داری (در سطح اطمینان 95 درصد) بین ابتلا به سرطان پستان و متغیرهای سن اولین قاعدگی و سابقه ی قاعدگی نامنظم مشاهده نشده است (9). همچنین سابقه ی فامیلی سرطان پستان در مطالعات متفاوت نشان‌دهنده ی افزایش شانس ابتلا به سرطان پستان حدود 3-2 برابر است (9). در مطالعه ی دیگری که به بررسی عوامل مربوط به باروری بر خطر بروز سرطان پستان توسط یاوری و همکارانش اجرا شده است، خطر سرطان پستان در زنانی که قاعدگی آنان در سنین پایین (کمتر از 12 سال) شروع شده نسبت به آن‌هایی که در 15 سالگی یا دیرتر قاعده شده‌اند، اندکی بالاتر است، اما این اختلاف به لحاظ آماری معنی دار نیست (13). مطالعه ی جمشیدی ایوانکی نیز نشان‌دهنده ی رابطه ی معنی‌دار بین

سرطان پستان و سابقه ی بین فامیلی درجه ی اول است (p=0/0002) (14). در مطالعه ی حاضر بر اساس متغیرهای هورمونی سن پایین منارک و وضعیت نامنظم قاعدگی، داشتن سابقه ی خانوادگی سرطان (سرطان پستان و سرطان تخمدان) و بیماری‌های خوش‌خیم پستان و متغیرهای تنش‌زا؛ استرس ناشی از متارکه ی همسر و استرس‌های ناشی از مشکلات مالی و همچنین نداشتن فعالیت فیزیکی، افراد به زیر گروه‌هایی با ریسک بالای ابتلا مشابه رده‌بندی شدند. همچنین رابطه ی بین متغیرهای رده‌بندی کننده و سرطان پستان با آزمون کای دو مورد بررسی قرار گرفت که نتایج حاکی از معنادار بودن این رابطه بوده است، بنابراین در این مطالعه مدل درختی به خوبی ریسک فاکتورهای مهم را مشخص کرده است.

نتیجه گیری

مدل درختی با استفاده از متغیرهای مؤثر در ابتلا به بیماری، افراد را به زیرگروه‌هایی با ویژگی‌های مشابه بالینی رده‌بندی می‌کند که این روش رده‌بندی از نظر تشخیص بالینی بیماری بسیار حائز اهمیت است. در واقع این مدل‌ها همزمان با رده‌بندی بیماران، ریسک فاکتورها را تعیین می‌کنند. به عبارتی دیگر رده‌بندی بیماران بر اساس ریسک فاکتورهایی است که زیر گروه های همگن‌تر را ارائه می‌دهند.

تشکر و قدردانی

از حمایت و همکاری شورای پژوهشی دانشکده ی علوم پزشکی دانشگاه تربیت مدرس صمیمانه قدردانی می شود.

References:

1- National cancer registry report, 2004. Tehran, cancer administration. Non-communicable diseases sector, Iranian center for diseases control and prevention; 2005. [In Persian]

2- Akbarzade Pasha H, Akbarzade Pasha A. Obstetrics & Gynecology. 1st ed. Tehran: Golban Press; 2007: 890-905. [In Persian]

3- Gharekhani P, Sadatin A. Cardinal manifestation and management of diseases (CMMD). 3rd ed. Tehran: Nooredanesh Press; 2004. [In Persian]

4- Jami M, Tavassoli M, Hemmati S. Association of the length of CA dinucleotide repeat in the epidermal growth factor receptor with risk and age of breast cancer onset. Journal of Isfahan 2008; 88(26): 22-30. [In Persian]

5- Biswas A, Datta S, Fine J P, Segal M R. (eds). Statistical advances in the biomedical science. California: John Wiley & Sons, Inc; 2007: 265-285.

6- Breiman L, Friedman H J, Olshen A R, Stone J C. Classification and regression trees. New York: a division of wads worth Inc; 1984.

7- Siciliano R, Mola F. Multivariate data analysis and modeling through classification and regression trees. Comput Statis Data Analy; 2000: 285-301.

8- Clifton D S. Handbook of Statistics. V 24. (Classification and Regression Trees, Bagging, and Boosting). Elsevier Press; 2005.

9- Holakoee N K, Ardalan A, Mahmoodi M, Motevalian A, Yahyapoor Y. Investigation of risk factors of breast cancer patients in Mazandaran in 2004. Institute of Public Health Research. J School Pub H 2006; 1(4): 27-32.

10- Ebrahimi M, Vahdaninia M, Montazeri A. Investigation of reproductive risk factors in breast cancer patients. Payesh 2002; 3(1): 23-28. [In Persian]

11- Demaris A. Regression with Social Data. New Jersey: Wiley & Sons Inc; 2004: 247-277.

12- Atarparsaee F, Golchin M, Asvadi E. Investigation of relationship of individual and social characteristics, lifestyle and stressor factors with breast cancer in females. Medical Journal of Tabriz University of Medical Sciences & Health Services 2001; 50(35): 15-21. [In Persian]

13- Yavari P, Mosavizade A, Sadrolhefazi B, Khodabakhshi R, Madani H, Mehrabi Y. Effect of reproductive factors on the incidence risk of breast cancer. Iranian Journal of Epidemiology 2005; 2(1): 11-20. [In Persian]

14- Jamshidi Evanaki F. Relationship between breast cancer and first class familial background

in women suffered from breast cancer in hospitals of Tehran University of Medical Sciences. Hayat 2001; 14: 35-43. [In Persian]

15- Fiona M C. Classification trees for survival data with competing risks. [Thesis]. University of Pittsburgh; 2008.

16- Hyungjun Ch. Tree-structured regression modeling for censored data. [Thesis]. University of Wisconsin-Madison; 2002.

17- Dean L S A. Method for detecting optimal splits over time in survival analysis using tree–structured models. [Thesis]. University of Pittsburgh; 2007.

18- Segal M R. Regression trees for censored data. Biometrics 1988; 44(1): 35-47.

19- Ibrahim N A, Kudus A, Daud I, Abu Bakar M R. Decision tree for competing risks survival probability in breast cancer study. PWASET 2008; 28: 15-19.

20- Giudici P. Applied data mining. New York: John Wiley & Sons Inc; 2003.

21- Rojer R J, Geatz M W. Data mining. New York: Addison Wesley Press; 2003.

22- Keles S, Segal M R. Residual-based tree-structured survival analysis. Statistic in Medicine 2002; 21(2): 313-326.

23- Yu S A. Tree-structured survival model with incomplete and time-dependent covariates: Illustration using type1 diabetes data. [Thesis]. University of Pittsburgh; 2006.

24- Vladimir A V. Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. Annal Surg Oncol 2006; 14(1): 34-40.

25- Su X. Multivariate survival tree. [Thesis]. Davis: University of California; 2001.

Evaluating the Risk Factors of Breast Cancer Using the Analysis of Tree Models

Amal Saki[1], Ebrahim Hajizadeh[2] and Najme Tehranian[3]

Abstract

Background and Aim: In biomedical research, we are concerned with exploring the risk factors of disease and classifying the patients based on similarity of their responses. However, traditional methods need to consider the related assumptions that are difficult to establish in biomedical studies. In this study, an alternative analytical method was used for determining the risk factors of breast cancer and classifying patients into groups based on the similarity of their features. Advances in the practical and theoretical aspects of tree-based methods were developed by Breiman et al. (1984) in their monograph on classification and regression trees. Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures.

Materials and Methods: In this article, we used the data from a case-control study. The two groups included 628 women who were under 40 years old. The case group included women with positive breast cancer diagnosis, and the control group included the patients who referred to Emam Khomeiny hospital with non-hormone and non-neoplastic disease.

Results: Of covariate selected to build the classification tree methods, the following variables were determined as the risk factors for breast cancer: Family history breast cancer, and ovarian cancer, irregular menstruation, lack of physical activity, and low age of menarche.

Conclusion: The simplicity of result interaction in terms of clinical or other relevant patient characteristics made trees an appealing approach in clinical and epidemiologic investigation.

Keywords: Breast cancer, classification and regression tree, recursive partitioning data

Ofogh-e-Danesh. GMUHS Journal. 2011; Vol. 17, No. 2

1- MSc in Biostatistics, Faculty of Medicine, Tarbiat Modares University, Tehran Iran
2- **Corresponding Author:** Associate Professor, Faculty of Medicine, Tarbiat Modares University, Tehran, Iran.
Tel: +98 21 82883810 **Fax:** +98 21 82884555 **E-mail:** Hajizadeh@modares.ac.ir
3- MSc in Midwifery and Reproductive Health, Faculty of Medicine, Tarbiat Modares University, Tehran, Iran