مجله تخصصی اپیدمیولوژی ایران؛ ۱۳۹۵، دوره ۱۲، شماره ۲: صفحات ۵۷-۴۹.

مقاله پژوهشی

مقایسه دقت پیش‌بینی رگرسیون لجستیک و درخت رده‌بندی در تعیین عوامل خطر و پیش‌بینی ابتلا به سرطان پستان

فرید زایری[۱]، سید حسین سیدآقا[۲]، هاله آقامولایی[۳]، فرزانه برومند[۴]، پروین یاوری[۵]

[۱] دانشیار، گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۲] کارشناسی ارشد، گروه آمار زیستی، دانشکده پیراپزشکی، کمیته پژوهشی دانشجویان، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۳] کارشناسی ارشد، گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۴] کارشناسی ارشد، گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

[۵] استاد، گروه بهداشت و پزشکی اجتماعی، دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران

نویسنده رابط: سید حسین سید آقا: نشانی: تهران، میدان قدس، ابتدای خیابان دربند، دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی، گروه آمار زیستی، تلفن: ۲۲۷۱۸۵۳۱-۰۲۱،

پست الکترونیک: hosseinseyedagha@gmail.com

تاریخ دریافت: ۹۴/۸/۳۰؛ پذیرش: ۹۴/۱۱/۱۰

مقدمه و اهداف: سرطان پستان یکی از رایج‌ترین بدخیمی‌های زنان است که بعد از سرطان ریه بیشترین میزان مرگ را به خود اختصاص داده است. هدف این مطالعه، مقایسه دو مدل رگرسیون لجستیک و درخت رده‌بندی در تعیین عوامل مؤثر و پیش‌بینی ابتلا به سرطان پستان است.

روش کار: داده‌های مورداستفاده برگرفته از یک مطالعه مورد- شاهدی است که بر پایه اطلاعات بیمارستانی از ۳۰۳ بیمار مبتلابه سرطان پستان به‌عنوان مورد و همین تعداد افراد غیرمبتلا به این سرطان به‌عنوان شاهد به‌دست‌آمده است. ابتدا ۱۶ متغیر به‌عنوان عوامل خطر بالقوه در دو مدل درخت رده‌بندی و رگرسیون لجستیک وارد و نتایج حاصل با استفاده از سطح زیر منحنی مشخصه عملکرد (ROC) و مقدار حساسیت و ویژگی مقایسه شدند.

نتایج: از ۱۶ متغیر موردبررسی، پنج متغیر در مدل درخت رده‌بندی و پنج متغیر در مدل رگرسیون لجستیک معنی‌دار شدند. پیش‌بینی بر اساس این متغیرها منجر به حساسیت، ویژگی و سطح زیر منحنی ROC ۷۱ درصد، ۶۹ درصد و ۷۴/۷ درصد برای درخت رده‌بندی و۶۳/۳ درصد، ۶۸/۸ درصد و ۷۱/۱ درصد به‌ترتیب برای رگرسیون لجستیک گردید.

نتیجه‌گیری: با توجه به معیارهای به‌دست‌آمده، مدل درخت رده‌بندی از توانایی بالاتری نسبت به رگرسیون لجستیک در تفکیک بیماران از افراد سالم برخوردار بود. نتایج به‌دست‌آمده از مطالعه حاضر نشان داد که سه متغیر وضعیت یائسگی، تعداد افراد مبتلا به سرطان پستان در خانواده و سن مادر هنگام اولین تولد زنده در هر دو مدل رگرسیون لجستیک و درخت رده‌بندی به‌طور مشترک معنی‌دار بوده‌اند.

واژگان کلیدی: سرطان پستان، عوامل خطر، درخت رده‌بندی، رگرسیون لجستیک

مقدمه

اختلالات پستانی یکی از مهم‌ترین دلایل مراجعه زنان به پزشک است. به رشد و تقسیم خارج از کنترل سلول‌های یک بافت از بدن سرطان گویند. شیوه تقسیم سلول‌های سرطانی مشابه تقسیم سلول‌های طبیعی بدن است. سرطان پستان تنها یک بیماری نیست، بلکه مجموعه‌ای از بیماری‌ها است. این سرطان را می‌توان در سه مرحله بر اساس روند پیشرفت بیماری در نظر گرفت. توده‌های پستانی در دو گروه خوش‌خیم و بدخیم طبقه‌بندی می‌شوند و در زنان در تمام سنین شایع‌اند، اما در زنان جوان‌تر، معمولاً این توده‌ها بدخیم نیستند. مردان نیز می‌توانند به این سرطان مبتلا شوند هرچند که تعداد آن نادر است (۱،۲).

سرطان پستان یکی از رایج‌ترین بدخیمی‌ها در انسان و دومین علت مرگ ناشی از سرطان در زنان است؛ به‌طوری‌که در سال ۱۹۹۵ تعداد ۱۸۲هزار مورد جدید شناسایی‌شده و در سال ۲۰۱۲ تعداد برآورد شده از این سرطان برابر ۲۲۶۸۷۰ مورد بوده که ۲۹ درصد از سرطان‌های مربوط به زنان را شامل شده و بالاترین نرخ بروز را به خود اختصاص داده است (۵-۲). در ایران نیز ۲۱/۴ درصد از کل موارد گزارش‌شده سرطان متعلق به سرطان پستان است. میزان خام بروز سرطان پستان در ایران معادل ۲۲/۴ در هر ۱۰۰هزار زن برآورد شده و داده‌های موجود نشان‌دهنده روند

افزایشی این سرطان در ایران هستند (۶).

پس از سرطان ریه، این سرطان کشنده‌ترین نوع سرطان به شمار می‌آید. احتمال ابتلا به این بیماری با افزایش سن زنان افزایش می‌یابد، اکثر زنانی که به این بیماری مبتلا می‌شوند زنان بالای ۶۰ سال هستند. همچنین سابقه خانوادگی نیز باعث افزایش احتمال ابتلا به این بیماری می‌شود. علاوه بر این درصورتی‌که زنی دریکی از پستان‌ها مبتلابه این سرطان شود، احتمال ابتلا این سرطان در پستان دیگر افزایش می‌یابد (۷،۸). وجود ارتباط بین هورمون‌ها و دخالت آن‌ها در بیماری‌زایی سرطان پستان در بسیاری از مطالعات نشان داده‌شده است (۹). از دیگر عوامل مؤثر بر بروز سرطان پستان می‌توان به عواملی چون مواجهه با اشعه x، چاقی، بیماری‌های کبدی، مصرف داروهای حاوی استروژن و همین‌طور عوامل ژنتیکی اشاره کرد (۱۲-۱۰). همچنین، عوامل مرتبط با باروری نیز ازجمله عوامل مرتبط با سرطان گزارش شده‌اند (۱۳).

مطالعات نشان داده‌اند که خطر سرطان پستان در سال‌های اولیه پس از زایمان افزایش قابل‌توجهی داشته و پس از گذشت چند سال، به‌عنوان یک عامل محافظتی شناخته می‌شود (۱۴،۱۵). اگرچه سرطان پستان مرتبط با بارداری، در طول بارداری و یا حداکثر تا ۲ سال بعد از زایمان موردبررسی قرار می‌گیرد، برآوردی که از این دوره زمانی به‌دست‌آمده بین ۲ تا ۱۵ سال است (۱۶،۱۷). سابقه خانوادگی سرطان پستان یک عامل خطر شناخته‌شده برای این بیماری است به‌طوری‌که از آن برای تشخیص زنان در معرض خطر این سرطان استفاده می‌شود. البته نسبت خطر سرطان پستان برای زنان با سابقه خانوادگی این بیماری به‌طور واضح مشخص نشده است (۱۸،۱۹).

شناسایی الگو و طبقه‌بندی، ازجمله مهم‌ترین کاربردهای آمار در زمینه‌های مختلف است که وظیفه پیش‌بینی بر اساس واقعیات موجود پیش‌رو را از طریق روش‌هایی نظیر رده‌بندی، درخت‌های تصمیم، تحلیل ممیزی، رگرسیون لجستیک و غیره بر عهده دارد. در این میان رگرسیون لجستیک یکی از کاراترین مدل‌های آماری است که جهت مدل‌سازی و تحلیل رابطه بین یک پاسخ دوحالته با یک یا چند متغیر مستقل به کار می‌رود. ازجمله قابلیت‌های رگرسیون لجستیک، مقایسه احتمال قرار گرفتن هر یک از موارد تحت آزمایش در هر یک از سطوح متغیر وابسته است (۲۱،۲۰).

یکی دیگر از پرکاربردترین تحلیل‌ها، درخت رده‌بندی است که در آن، نمونه‌ها به نحوی دسته‌بندی می‌شوند که از ریشه به سمت پایین رشد می‌کنند. هر گره داخلی (برگ) با یک ویژگی مشخص می‌شود. به‌منظور رده‌بندی دقیق‌تر نمونه‌ها، هر گره به نحوی به دو زیر گره تقسیم می‌شود که نمونه‌های مشابه در زیر گره‌های یکسان و نمونه‌های متمایز، در زیر گره‌های متفاوت قرار گیرند. در داده‌کاوی از این روش به‌طور گسترده‌ای استفاده می‌شود. همچنین از درخت رده‌بندی در حل مسائلی که بتوان آن‌ها را به‌گونه‌ای مطرح نمود که پاسخ واحدی به‌صورت یک دسته یا کلاس ارائه دهند نیز استفاده می‌شود (۲۲-۲۰).

مطالعات زیادی درزمینه‌ی بررسی عوامل خطر سرطان پستان در ایران و جهان انجام‌شده و در هر مطالعه با توجه به تعداد افراد موردبررسی و نوع مطالعه، یک سری عوامل خطر شناسایی‌شده که تأثیر معنی‌داری بر ابتلا به این سرطان داشته‌اند. در برخی از این تحلیل‌ها از روش‌های کلاسیک و در برخی دیگر از روش‌های رده‌بندی مانند رگرسیون لجستیک، تحلیل ممیزی و داده‌کاوی استفاده‌شده اما توجه زیادی به درخت رده‌بندی نشده است. درواقع درخت رده‌بندی یک روش نسبتاً جدید و توانمند در تعیین عوامل مرتبط با پیامدهای مختلف نظیر سرطان‌ها است و تاکنون مطالعه مشابهی با استفاده از دو روش رگرسیون لجستیک و درخت رده‌بندی انجام‌نشده است. هدف از انجام مطالعه حاضر، بررسی عوامل خطر این سرطان و مقایسه قدرت پیش‌بینی دو روش تحلیلی رگرسیون لجستیک و درخت رده‌بندی است. در این مطالعه از داده‌های مربوط به مطالعه مورد- شاهدی سرطان پستان که در بخش انکولوژی بیمارستان شهدا در بازه زمانی یک‌ساله جمع‌آوری‌شده، استفاده‌شده است.

روش کار

داده‌های مورداستفاده در این مطالعه، متعلق به یک مطالعه مورد- شاهدی است که بر پایه اطلاعات بیمارستانی از ۳۰۳ بیمار مبتلابه سرطان پستان به‌عنوان مورد و ۳۰۳ فرد غیر مبتلابه این بیماری به‌عنوان شاهد به‌دست‌آمده است. در این مطالعه که در فاصله زمانی بهمن ۸۲ تا آذر ۸۳ انجام شده است، شناسایی گروه مورد[1] از طریق بخش انکولوژی بیمارستان شهدا و با تأیید هیستوپاتولوژی انجام شد و گروه شاهد[2] از بین مراجعین به درمانگاه‌های سرپایی یا بیماران بستری در سایر بخش‌های همین مرکز که فاقد سرطان پستان بودند و هیچ‌گونه سابقه‌ای از این بیماری در گذشته خود نداشتند انتخاب شدند. افراد گروه شاهد از نظر سن با گروه مورد با روش

[1] Case Group

[2] Control Group

همسان‌سازی، همسان شده و از تمامی آنان رضایت برای شرکت در مطالعه گرفته شد.

داده‌های این مطالعه به کمک مصاحبه حضوری و تکمیل پرسشنامه جمع‌آوری شدند. این اطلاعات شامل سن قاعدگی، سن اولین بارداری، سن مادر هنگام اولین تولد زنده، تعداد بارداری‌ها، تعداد تولدهای زنده، وضعیت یائسگی، علت یائسگی، سابقه بیماری‌های پستان، سابقه خانوادگی سرطان پستان، تعداد سرطان پستان در خانواده، سابقه سقط، سابقه شیردهی، رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی، وضعیت تأهل، سطح تحصیلات و قومیت بود.

در این مطالعه زنانی که در طول یک سال قبل از تاریخ مصاحبه خونریزی ماهیانه نداشتند به‌عنوان یائسه در نظر گرفته شدند. سنجش متغیرها در افراد مبتلا تا زمان تشخیص سرطان پستان و در افراد شاهد تا زمان انجام مصاحبه صورت گرفته است. به‌منظور جلوگیری از اریبی امتناع و یا به حداقل رساندن آن، از پرسشگران زن استفاده‌شده است. اطلاعات بیشتر درباره افراد نمونه و نحوه نمونه‌گیری در مقالات به چاپ رسیده قبلی از این داده‌ها قابل‌دسترسی است (۶،۲۳،۲۴).

در این مقاله از دو روش رگرسیون لجستیک و درخت تصمیم برای پیش‌بینی سرطان پستان استفاده شد و کیفیت مدل‌های برازش شده و تعیین توان پیش‌بینی صحیح آن‌ها، با استفاده از مقدار حساسیت و ویژگی و سطح زیر منحنی ROC موردسنجش و مقایسه قرار گرفت.

مدل کلی رگرسیون لجستیک به‌صورت $\mathrm{Log}\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum \beta X$ است که در آن X معرف متغیرهای مسضرایبβ ضرایب برآورد شده مدل و π احتمال ابتلا یا عدم ابتلا (بسته به هدف تحقیق) به بیماری است. درخت تصمیم تقسیم‌بندی داده‌ها به روش بازگشتی است، به‌طوری‌که در هر مرحله با استفاده از معیاری به نام ضابطه جینی، افراد را در گروه‌هایی با ویژگی‌های بالینی مشابه رده‌بندی می‌کند و درنهایت نتایج این رده‌بندی به‌صورت نمودار درخت رده‌بندی نمایش داده می‌شود. برای رشد درخت تصمیم، الگوریتم‌هایی ارائه‌شده که از مهم‌ترین آن‌ها می‌توان به الگوریتم‌های CART، CHAILD، CRUISE و QUEST اشاره کرد.

در تحقیق پیش رو، گسترش درخت تصمیم با الگوریتم CART مدنظر است. اساس کار این الگوریتم بدین‌صورت است که هر گره، بر اساس متغیر جداکننده‌ای که بهترین خلوص را ایجاد کند، به دو زیرگروه تقسیم می‌شود، اگر گره‌های ایجادشده به‌اندازه کافی یکدست و خالص باشند انشعاب آن‌ها متوقف‌شده و در حکم گره نهایی یا برگ خواهند بود، در غیر این صورت مجدداً به گره‌های دیگری منشعب خواهند شد، درواقع داده‌های هر ریزگره همگن‌تر از گره قبل هستند. به گره‌هایی که مابین گره ریشه و گره نهایی وجود دارند گره میانی گفته می‌شود (۲۵). این الگوریتم شامل ۲ بخش اصلی است: بخش رده‌بندی و بخش رگرسیون. اگر متغیر پاسخ دوحالته باشد، هدف تخصیص افراد به یکی از رده‌های متغیر پاسخ است که به آن رده‌بندی گویند؛ و اگر متغیر پاسخ پیوسته باشد، هدف ما پیش‌بینی متغیر پاسخ بر اساس متغیرهای پیشگوی مشاهده‌شده است که به آن درخت رگرسیون گویند (۲۶).

گره نهایی نشان می‌دهد که تقسیم‌های بیشتر بر اساس متغیرهای پیشگو، توانایی بیان واریانس کافی مربوط به توصیف متغیر پاسخ را ندارند. به کمک درخت تصمیم می‌توانیم متغیرهایی را بررسی کنیم که شاید به‌تنهایی متغیرهای مناسبی برای پیشگویی و یا رده‌بندی متغیر پاسخ نباشند، اما با قرار گرفتن در کنار متغیرهای دیگر باعث افزایش توان مدل و بهبود نتایج می‌شوند. مزیت الگوریتم CART این هست که به‌صورت ظاهری، تفسیر نتایج را با دقت زیاد آماری ترکیب کرده و طراحی مدلی پایا و مناسب را آسان می‌کند (۲۶).

در این مطالعه ابتدا ۱۶ متغیر که در مطالعات پیشین معنی‌دار شناخته‌شده بودند وارد مدل درخت رده‌بندی شدند. قبل از برازش مدل چند متغیره نهایی، ابتدا تحلیل تک متغیره برای هر یک از متغیرها صورت گرفت. پس‌ازآن تحلیل‌های چند متغیره با متغیرهای مختلف انجام‌شده و متغیرهایی که از نظر آماری تأثیری در افزایش توانایی مدل در پیشگویی افراد مستعد و یا شناسایی عوامل خطر نداشتند از مدل حذف شدند. در آخر مدل نهایی با ۷ متغیر برازش داده شد. این متغیرها عبارت بودند از: سن زن هنگام اولین تولد زنده، وضعیت یائسگی، تعداد سرطان پستان در خانواده، تعداد تولدهای زنده، سابقه سقط، مصرف قرص ضدبارداری و سن اولین قاعدگی. پس از برازش مدل فوق، پنج متغیر معنی‌دار شدند که عبارت بودند از: سن زن هنگام اولین تولد زنده، وضعیت یائسگی، تعداد سرطان پستان در خانواده، تعداد تولدهای زنده و سابقه سقط. همچنین متغیر وضعیت یائسگی به‌عنوان مهم‌ترین متغیر و ریشه درخت رده‌بندی انتخاب شد.

در مدل رگرسیون لجستیک نیز ابتدا ۱۶ متغیر را وارد مدل کردیم. پس از بررسی‌های انجام‌شده و حذف متغیرهای نامناسب از مدل، مدل نهایی با پنج متغیر برازش داده شد که هر پنج متغیر اثر

معنی‌داری را نشان داند. این متغیرها عبارت بودند از: سن زن هنگام اولین تولد زنده، وضعیت یائسگی، تعداد سرطان پستان در خانواده، رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی و سابقه مصرف قرص ضدبارداری.

در این تحقیق کارایی دو مدل رگرسیون لجستیک و درخت تصمیم در پیشگویی سرطان پستان بررسی شد.

یافته‌ها

میانگین (انحراف معیار) سن در دو گروه مورد و شاهد به ترتیب ۴۸/۸ (۹/۸) سال و ۵۰/۲ (۱۱/۱) سال با میانه ۴۸ سال و دامنه تغییرات ۲۴ تا ۸۴ سال بود. آمارهای توصیفی برای متغیرهای مختلف موردمطالعه به تفکیک دو گروه در جدول ۱ آمده است. بر این اساس، آزمون‌های تک‌متغیره مانند تی و کای دو نشان داد که دو گروه ازنظر سابقه سرطان پستان در خانواده، وضعیت یائسگی، رادیوگرافی قفسه سینه در فاصله سنی بلوغ تا ۳۰ سالگی و سن زن هنگام اولین تولد زنده تفاوت معنی‌داری داشته‌اند.

طبق جدول شماره ۲، نتایج حاصل از رگرسیون لجستیک نشان داد که متغیرهای وضعیت یائسگی، تعداد سرطان پستان در خانواده، سن زن هنگام اولین تولد زنده، رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی و مصرف قرص ضدبارداری اثر معنی‌داری بر ابتلا به سرطان پستان دارند. مقدار AOR (ضریب تعیین تعدیل‌شده) برای متغیرهای موردبررسی نشان داد که شانس ابتلا به سرطان پستان برای زنان یائسه سه برابر زنان غیر یائسه است، زنانی که قرص ضدبارداری مصرف می‌کنند، ۸۶٪ شانس بیشتری برای ابتلا به سرطان پستان دارند، زنانی که رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی انجام داده‌اند نیز ۴۶٪ شانس بیشتری برای ابتلا به این سرطان دارند. همچنین به ازای هرسال افزایش سن زن هنگام اولین تولد زنده، شانس ابتلا به سرطان پستان ۱۱ درصد افزایش می‌یابد.

در مدل رده‌بندی درختی نیز متغیرهای وضعیت یائسگی، تعداد سرطان پستان در خانواده، سابقه سقط‌جنین، تعداد تولدهای زنده و سن مادر هنگام اولین تولد زنده معنی‌دار شدند. مهم‌ترین متغیر، وضعیت یائسگی بود که به‌عنوان ریشه درخت رده‌بندی انتخاب شد. سپس دو متغیر سابقه سقط‌جنین و تعداد تولدهای زنده در رده دوم قرار گرفتند و در آخر نیز متغیرهای تعداد سرطان پستان در خانواده و سن مادر هنگام اولین تولد زنده وارد مدل شدند.

میزان حساسیت مدل رگرسیون لجستیک و درخت تصمیم به ترتیب برابر ۶۳/۳ درصد و ۷۱/۳ درصد و ویژگی آن‌ها به ترتیب برابر ۶۸/۸ درصد و ۶۹/۱ درصد محاسبه شد. به‌منظور مقایسه قدرت پیش‌بینی دو مدل، از سطح زیر منحنی ROC استفاده شد. این مقدار برای مدل رگرسیون لجستیک برابر ۷۱/۱ درصد و برای مدل درخت تصمیم برابر ۷۴/۷ درصد به دست آمد (جدول شماره ۳). بر اساس این نمودار، درخت تصمیم از قدرت پیش‌بینی بالاتری برخوردار بوده است. این دو مدل به ترتیب توانستند وضعیت ۶۶/۱ درصد و ۷۰ درصد از افراد را به‌درستی پیش‌بینی نمایند.

جدول شماره ۱ – مقایسه دو گروه مورد و شاهد ازنظر متغیرهای مختلف

متغیر	گروه		مقدار P
	مورد	شاهد	
سن زن در زمان مصاحبه	۴۸/۸±۹/۸*	۵۰/۲±۱۱/۱	۰/۰۹‡
سن زن هنگام اولین تولد زنده	۲۰/۷۹±۴/۹*	۱۹/۰۵±۴/۲	<۰/۰۰۱‡
سابقه سرطان پستان در خانواده	۴۴† (٪۱۴/۷)	۲۳ (٪۷/۶)	<۰/۰۰۱§
مصرف قرص ضد حاملگی	۱۶۹† (٪۶۱/۵)	۱۴۶ (٪۵۰/۹)	۰/۰۹§
وضعیت یائسگی	۲۱۱† (٪۷۶/۷)	۱۵۴ (٪۵۴)	<۰/۰۰۱§
رادیوگرافی قفسه سینه در فاصله سنی بلوغ تا ۳۰ سالگی	۸۱† (٪۳۰)	۶۲ (٪۲۱/۸)	<۰/۰۰۱§

* میانگین و انحراف استاندارد

† تعداد (درصد)

‡ نتیجه آزمون تی مستقل

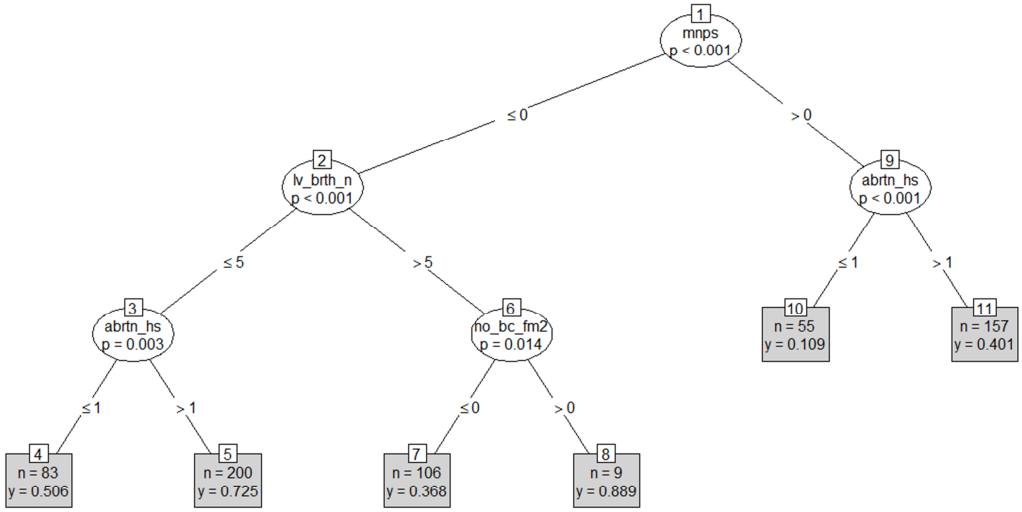
§ نتیجه آزمون کای دو

جدول شماره ۲ – نتایج مدل رگرسیون لجستیک چندگانه برای بررسی اثر متغیرهای مختلف بر سرطان پستان

متغیر	طبقه	ضریب تعیین تعدیل‌شده	فاصله اطمینان ۹۵٪
سن زن هنگام اولین تولد زنده		۱/۱۱	۱/۰۶ – ۱/۱۵
سابقه خانوادگی سرطان پستان		۱/۸۶	۱/۰۴ – ۳/۳۲
مصرف قرص ضد حاملگی	دارد	۱/۸۶	۱/۲۹ – ۲/۶۹
	ندارد		طبقه مرجع
وضعیت یائسگی	یائسه	۳/۰۲	۲/۰۵ – ۴/۴۴
	غیر یائسه		طبقه مرجع
رادیوگرافی قفسه سینه در فاصله سنی بلوغ تا ۳۰ سالگی	دارد	۱/۴۶	۱/۰۱ – ۲/۲۱
	ندارد		طبقه مرجع

جدول شماره ۳ – مقایسه نتایج رگرسیون لجستیک و درخت تصمیم در پیش‌بینی سرطان پستان

روش	حساسیت (درصد)	ویژگی (درصد)	سطح زیر منحنی ROC (درصد)
رگرسیون لجستیک	۶۳/۳	۶۸/۸	۷۱/۱
درخت تصمیم	۷۱/۳	۶۹/۱	۷۴/۷



شکل شماره ۱– نمودار درخت رده‌بندی برای اثر متغیرهای مختلف بر سرطان پستان

بحث

با توجه به افزایش روزافزون تعداد مبتلایان به سرطان پستان و شیوع بالای آن در بین زنان ایرانی، دست‌یابی به مدل‌هایی که با توان بالایی می‌توانند وجود این سرطان را در افراد پیش‌بینی نمایند، بسیار باارزش است. پزشکان همواره اهمیت زیادی برای تشخیص زودهنگام انواع سرطان‌ها قائل هستند، چراکه با تشخیص به‌موقع سرطان می‌توان از انتشار آن به سایر سلول‌ها و اندام‌های بدن جلوگیری کرده و حتی لوازم بهبودی کامل بیمار را نیز فراهم کرد.

در این مطالعه با توجه به مدل رده‌بندی درختی، متغیرهای وضعیت یائسگی، سابقه سقط، تعداد سرطان پستان در خانواده، تعداد تولدهای زنده و سن مادر هنگام اولین تولد زنده معنی‌دار شدند و متغیر وضعیت یائسگی به‌عنوان ریشه درخت در مدل رده‌بندی درختی انتخاب شد. در مدل رگرسیون لجستیک، متغیرهای تعداد سرطان پستان در خانواده، سن هنگام اولین تولد زنده، وضعیت یائسگی، رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی و درنهایت مصرف قرص ضدبارداری در سطح معنی‌داری ۰/۰۵ اثر معنی‌داری را نشان دادند.

متغیر سن مادر در هنگام اولین تولد زنده در مطالعه یاوری و همکاران معنی‌دار شناخته‌شده است (۶). در مطالعه‌ای که معتمد و همکاران بر روی۱۰۲ بیمار مبتلابه سرطان پستان و ۲۷۸ شاهد انجام دادند متغیرهای سن اولین ازدواج و سن اولین حاملگی کامل معنی‌دار شناخته شدند (۲۷). متغیر رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی در مطالعه یاوری و همکاران معنی‌دار شده است (۶).

اگرچه متغیر مصرف قرص ضدبارداری در مطالعه‌های یاوری و همکاران و کیهانیان و همکاران معنی‌دار شد اما در مطالعه‌هایی چون مطالعه مارک بانکز و همکاران، مطالعه تسارو و همکاران و یا مطالعه هم‌گروهی سلامت پرستاران که در بین سال‌های ۱۹۷۶ تا ۱۹۹۲ در آمریکا بر روی ۳۳۸۳ مورد سرطان پستان توسط هنکینسون و همکاران انجام شد، هیچ رابطه کلی معنی‌داری بین مدت‌زمان استفاده از قرص‌های ضدبارداری و ابتلا به سرطان پستان وجود ندارد. حتی برای زنانی که بیشتر از ۱۰ سال از این قرص‌ها استفاده کردند نیز رابطه معنی‌داری یافت نشد. البته این مطالعه بر روی زنان بالای ۴۰ سال انجام‌شده است (۳۰-۴،۶،۲۸). در مطالعه صداقت و همکاران نیز مصرف قرص‌های ضدبارداری به مدت بیش از ۴ سال به‌عنوان عامل خطر شناخته شد. در مطالعه کابات و همکاران، مصرف قرص ضدبارداری و طول مدت مصرف، به‌عنوان عوامل مؤثر بر این سرطان شناخته شدند (۳۱،۳۲). در مطالعه مانجوشا رأی و همکاران که در سال ۲۰۰۸ بر روی ۶۵ مورد مبتلابه سرطان و همین تعداد شاهد انجام شد، اثر متغیر مصرف قرص ضدبارداری معنی‌دار نشد (۳۳).

یاوری و همکاران، کیهانیان و همکاران، دسیلوا و همکاران و کابات و همکاران، اثر متغیر وضعیت یائسگی را معنی‌دار گزارش کردند (۶،۲۸،۳۱،۳۴). در مطالعه صداقت و همکاران، اثر متغیرهای سن یائسگی و سن در اولین قاعدگی معنی‌دار نشد (۳۲). در مطالعه بچر و همکاران، متغیر سن قاعدگی اثر معنی‌داری روی سرطان پستان نداشت (۱۸). در مطالعه گیلانی و همکاران در سال ۱۹۹۸ بر روی زنان پاکستانی، تحلیل‌ها ابتدا بر روی کل زنان و سپس برای زنان یائسه و غیر یائسه به‌طور جداگانه انجام شد. نتایج نشان داد که سن یائسگی بالای ۴۵ سال، یک عامل خطر مهم برای سرطان پستان است (۳۵). در مطالعه هلاکویی نایینی و همکاران که در سال ۱۳۸۳ بر روی ۲۵۰ نمونه مبتلابه سرطان پستان و ۵۰۰ شاهد همسان شده فردی انجام شد، وضعیت یائسگی معنی‌دار شد (۳۶). در مطالعه معتمد و همکاران اثر متغیرهای سن قاعدگی و سن یائسگی معنی‌دار نشد (۲۷). یوسا کیم در یک مطالعه مورد-شاهدی که بر روی ۴۸۱ مورد مبتلابه سرطان پستان و ۴۹۱ شاهد که ازنظر سنی همسان شده بودند انجام شد، نشان داد که سن قاعدگی و سن یائسگی اثر معنی‌داری بر ابتلا به سرطان پستان ندارند (۳۷). در مطالعه مانجوشا رأی و همکاران وضعیت یائسگی معنی‌دار نشد (۳۳).

در مطالعه‌های شریف زاده و همکاران، کیهانیان و همکاران، مورالس و همکاران و کابات و همکاران، متغیر سابقه خانوادگی سرطان پستان معنی‌دار شد (۳۸،۳۱،۲۸،۳۹). در مطالعه گیلانی و همکاران متغیرهای سابقه خانوادگی سرطان پستان و سابقه ازدواج فامیلی اثر معنی‌داری روی ابتلا به سرطان پستان در هر سه گروه دارند. همچنین متغیر تعداد سرطان پستان در خانواده نیز در مطالعه یاوری و همکاران معنی‌دار شد (۶،۳۵).

بچر و همکاران، دسیلوا و همکاران و کیهانیان و همکاران نشان دادند که رابطه معنی‌داری بین متغیر سابقه سقط و سرطان پستان وجود دارد (۱۸،۲۸،۳۴). در مطالعه هلاکویی نایینی و همکاران نیز متغیر سابقه سقط تحریکی به‌عنوان عامل خطر سرطان پستان شناخته شد. امینی ثانی و همکاران نیز در مطالعه خود که در سال ۱۳۸۰ بر روی زنان ساکن مشهد انجام شد، نشان دادند که متغیرهای سابقه سقط و دفعات سقط اثر معنی‌داری بر

ابتلا به سرطان پستان دارند و اگر سقط قبل از اولین حاملگی ترم باشد، خطر این سرطان افزایش می‌یابد (۴۰). یوسا کیم در مطالعه خود اثر متغیر سابقه سقط در اولین بارداری را معنی‌دار گزارش کرد (۳۷).

در مطالعه مورالس و همکاران، متغیرهای تعداد زایمان بالا، هیسترکتومی قبل از ۵۰ سالگی و دوره شیردهی طولانی‌مدت، از عوامل پیشگیری‌کننده سرطان پستان شناخته شدند (۳۹). در مطالعه دسیلوا و همکاران که بر روی زنان ۶۴-۳۰ ساله سریلانکا انجام شد، خطر ابتلا به سرطان پستان برای زنان با دوره شیردهی بیش از ۲ سال، به‌طور متوسط ۶۰ درصد کمتر از زنان با دوره شیردهی کمتر از ۲ سال به دست آمد (۳۴). در مطالعه کابات و همکاران تعداد زایمان‌های بالا رابطه معکوس معنی‌داری با خطر ابتلا به سرطان پستان نشان داد (۳۱). لوماچی و همکاران به کمک تحلیل تک متغیره نشان دادند که متغیرهای سن قاعدگی، تعداد زایمان‌ها و سن زن هنگام اولین تولد زنده تأثیر معنی‌داری روی سرطان پستان دارند (۴۱). سابقه سقط در اولین بارداری نیز در مطالعه یوسا کیم معنی‌دار شناخته شد (۳۷). متغیرهای سابقه سقط و تعداد فرزندان سالم در مطالعه مانجوشا رأی و همکاران معنی‌دار نشدند (۳۳).

پرس و ویلسون در سال ۱۹۷۸ به مقایسه دو روش رده‌بندی رگرسیون لجستیک و تحلیل ممیزی خطی با استفاده از اطلاعات مربوط به بیماران سرطان پستان در ایالات مختلف کلمبیا پرداخته‌اند. در این مطالعه نتایج متفاوتی از دو روش به دست آمد به‌طوری‌که در برخی ایالات درصد رده‌بندی برای رگرسیون لجستیک بیشتر بود و در برخی، درصد رده‌بندی برای تحلیل ممیزی (۲۰).

نتیجه‌گیری

در این مطالعه از دو روش درخت رده‌بندی و رگرسیون لجستیک برای بررسی عوامل خطر مرتبط با سرطان پستان و مقایسه قدرت پیش‌بینی آن‌ها استفاده شد. بر اساس معیارهای حساسیت و ویژگی، مدل رده‌بندی درختی از توانایی بالاتری نسبت به رگرسیون لجستیک برخوردار بود و بر اساس مساحت زیر منحنی ROC، هر دو مدل از قدرت پیش‌بینی یکسانی برخوردار بودند به‌طوری‌که سطح زیر منحنی ROC در مدل رده‌بندی درختی و رگرسیون لجستیک به ترتیب برابر ۶۹ درصد و ۶۸/۳ درصد به دست آمد.

ازجمله کاستی‌های این تحقیق می‌توان به این نکته اشاره کرد که متغیرهای معنی‌داری که در بسیاری از مطالعات داخلی و خارجی مورد ارزیابی قرار گرفتند، در این مطالعه جمع‌آوری نشدند. همچنین به دلیل اینکه داده‌ها مربوط به یک مرکز (بیمارستان شهدای تجریش واقع در شمال تهران) بودند، نتایج این مطالعه قابلیت تعمیم به کل کشور را ندارد.

تشخیص زودهنگام و به‌موقع یک بیماری، مخصوصاً انواع سرطان‌ها، در تمام دنیا و برای تمام پزشکان از اهمیت ویژه‌ای برخوردار است. با توجه به افزایش تعداد مبتلایان به سرطان پستان مخصوصاً در بین زنان ایرانی، تشخیص به‌موقع آن کمک شایانی به بهبود بیماران می‌کند. برخی از عوامل خطر موردبررسی در این مطالعه قابل‌کنترل بوده و برخی غیرقابل‌کنترل. برای مثال متغیرهای وضعیت یائسگی، سن اولین قاعدگی، تعداد سرطان پستان در خانواده ازجمله عوامل غیرقابل‌کنترل و متغیرهای سابقه سقط، مصرف قرص ضدبارداری، سن مادر هنگام اولین تولد زنده و رادیوگرافی قفسه سینه در فاصله زمانی سن بلوغ تا ۳۰ سالگی ازجمله عوامل قابل‌کنترل هستند. با توجه به نتایج مطالعه حاضر، توجه بیشتر به عوامل خطر قابل‌کنترل و بررسی دقیق‌تر این عوامل می‌تواند نقش بسزایی در کاهش خطر ابتلا به این سرطان داشته باشد.

تشکر و قدردانی

داده‌های این پژوهش با حمایت معاونت پژوهشی دانشکده پزشکی دانشگاه علوم پزشکی شهید بهشتی جمع‌آوری‌شده است. لذا نویسندگان از تمامی همکارانی که به نحوی در این مطالعه نقش داشته‌اند، تقدیر و تشکر می‌نمایند.

منابع

1. Ogden J. Understanding Breast Cancer [Internet]. 2004 [cited 2015 Jul 11]. 160.
2. Ghasemzadeh S, Khayat KM, Dadmanesh M, Safari A, Sahebi A. "Evaluation of Prevalence and Risk Factors of Asympyomatic Masses of Breast" in Women Visiting in Khanevadeh Hospital (Oct 2005-2006)).
3. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin. 2012; 62: 10–29.
4. Hankinson SE, Colditz GA, Manson JE, Willett WC, Hunter DJ, Stampfer MJ, et al. A prospective study of oral

contraceptive use and risk of breast cancer (Nurses' Health Study, United States). Cancer Causes Control. 1997 Jan; 8: 65–72.
5. Jangjoo A, Aliakbarian M, others. Seroma formation after breast cancer surgery: incidence and risk factors. Tehran Univ Med Sci [Internet]. 2009 [cited 2015 Oct 10]; 67.
6. Yavari P, Mosavizadeh MA, Sadrolhefaz i B, Khodabakhshi R, Madani H, Mehrabi Y. Reproductive Characteristics and the Risk of Breast Cancer: A Case-Control Study. Iran J Epidemiol. 2006 Feb 15; 1: 11–9.
7. Kwabi-Addo B, Lindstrom TL. Cancer Causes and Controversies: Understanding Risk Reduction and Prevention: Understanding Risk Reduction and Prevention. ABC-CLIO; 2011. 265.
8. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General [Internet]. Atlanta (GA): Centers for Disease Control and Prevention (US); 2014 [cited 2015 Jul 11].
9. Jl S, M S, La B. Estrogen receptors and breast cancer. Epidemiol Rev. 1985 Dec; 8: 42–59.
10. Friedman LS, Gayther SA, Kurosaki T, Gordon D, Noble B, Casey G, et al. Mutation analysis of BRCA1 and BRCA2 in a male breast cancer population. Am J Hum Genet. 1997 Feb; 60: 313–9.
11. Easton DF, Steele L, Fields P, Ormiston W, Averill D, Daly PA, et al. Cancer risks in two large breast cancer families linked to BRCA2 on chromosome 13q12-13. Am J Hum Genet. 1997 Jul; 61: 120–8.
12. Lynch HT, Watson P, Narod SA. The genetic epidemiology of male breast carcinoma. Cancer. 1999 Sep 1;86(5):744–6.
13. 13. Brinton LA, Schairer C, Hoover RN, Fraumeni JF. Menstrual Factors and Risk of Breast Cancer. Cancer Invest. 1988 Jan 1; 6: 245–54.
14. Lyons TR, Schedin PJ, Borges VF. Pregnancy and breast cancer: when they collide. J Mammary Gland Biol Neoplasia. 2009; 14: 87–98.
15. Lambe M, Hsieh C, Trichopoulos D, Ekbom A, Pavia M, Adami H-O. Transient increase in the risk of breast cancer after giving birth. N Engl J Med. 1994; 331: 5–9.
16. Hou N, Ogundiran T, Ojengbede O, Morhason-Bello I, Zheng Y, Fackenthal J, et al. Risk factors for pregnancy-associated breast cancer: a report from the Nigerian Breast Cancer Study. Ann Epidemiol. 2013; 23: 551–7.
17. Azim HA, Santoro L, Russell-Edu W, Pentheroudakis G, Pavlidis N, Peccatori FA. Prognosis of pregnancy-associated breast cancer: a meta-analysis of 30 studies. Cancer Treat Rev. 2012; 38: 834–42.
18. Becher H, Schmidt S, Chang-Claude J. Reproductive factors and familial predisposition for breast cancer by age 50 years. A case-control-family study for assessing main effects and possible gene–environment interaction. Int J Epidemiol. 2003 Feb 1; 32: 38–48.
19. Colditz GA, Willett WC, Hunter DJ, Stampfer MJ, Manson JE, Hennekens CH, et al. Family history, age, and risk of breast cancer: prospective data from the Nurses' Health Study. Jama. 1993; 270: 338–43.
20. 20. Press SJ, Wilson S. Choosing between Logistic Regression and Discriminant Analysis. J Am Stat Assoc. 1978 Dec 1; 73: 699–705.
21. Zandkarim E i, Afshari Safavi A. Comparison of artificial neural network predictive power with multiple logistic regressions to determine patients with and without diabetic retinopathy. Razi J Med Sci. 2014 Oct 15; 21: 79–90.
22. Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Methods in Prediction of Metabolic Syndrome. Iran J Endocrinol Metab. 2009 Apr 15; 11: 638–46.
23. Pourhosseingholi MA, Mehrabi Y, Alavi-Majd H, Yavari P. Using Latent Variables to Eliminate Multicollinearity Effect in A Logistic Regression on Risk Factors for Breast Cancer. Iran J Epidemiol. 2006 Feb 15; 1: 41–5.
24. Yavari P, Pourhosseingholi MA. Estimation of the Gene-Environment Interaction in Breast Cancer Patients. Iran J Epidemiol. 2006 Mar 15; 2: 49–52.
25. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1 edition. Boca Raton: Chapman and Hall/CRC; 1984. 368.
26. Morgan J. Classification and Regression Tree Analysis [Internet]. Technical Report Boston University School of Public Health; 2014 [cited 2015 Jul 4].
27. Motamed N, Hadi N, Taleie A. Evaluation of risk factors for breast cancer in women over 35 years Shiraz 1380. J Zanjan Univ Med Sci. 2004 Mar 1;12(46):25–33.
28. Keihanian S, Ghaffari F, Fotokian Z, Shoormig R, Saravi M. Risk factors of breast cancer in Ramsar and Tonekabon. J Qazvin Univ Med Sci. 2010 Jul 15; 14: 12–9.
29. Marchbanks PA, McDonald JA, Wilson HG, Folger SG, Mandel MG, Daling JR, et al. Oral Contraceptives and the Risk of Breast Cancer. N Engl J Med. 2002 Jun 27; 346: 2025–32.
30. Tessaro S, Béria JU, Tomasi E, Barros AJ. Oral contraceptive and breast cancer: a case-control study. Rev Saúde Pública. 2001 Feb; 35: 32–8.
31. Kabat GC, Jones JG, Olson N, Negassa A, Duggan C, Ginsberg M, et al. Risk factors for breast cancer in women biopsied for benign breast disease: A nested case-control study. Cancer Epidemiol. 2010 Feb; 34: 34–9.
32. Sedaghat M, Molavi Nojoumi M, Hosseini N. A case-control study on breast cancer risk factors in Iran. J Med Counc Islam Repub Iran. 1382 Jan 1; 21: 198–204.
33. Rai M, Pande A, Singh M, Rai A, Shukla HS. Assessment of epidemiological factors associated with breast cancer. Indian J Prev Soc Med. 2008; 39: 71–7.
34. De Silva M, Senarath U, Gunatilake M, Lokuhetty D. Prolonged breastfeeding reduces risk of breast cancer in Sri Lankan women: A case–control study. Cancer Epidemiol. 2010 Jun; 34: 267–73.
35. Gilani GM, Kamal S, Gilani SAM. Risk factors for breast cancer for women in Punjab, Pakistan:Results from a case-control study. Pak J Stat Oper Res. 2006 Jan 1; 2: 17–26.
36. Holakouie Naeini K, Ardalan A, Mahmoudi M, Motevallian A, Yahyapour Y. Risk factors for breast cancer in Mazandaran Province, 2004. J Sch Public Health Inst Public Health Res. 2006 Apr 15; 4: 27–36.
37. Kim YS. Risk Factors for Breast Cancer: A Case-Control Study. J Korean Breast Cancer Soc. 1998; 1: 109.
38. Sharif Zadeh GR, Hosseini M, Kermani T, Ataiee M, Akhbari SH. Breast cancer and the related factors: A case control study. J Birjand Univ Med Sci. 2011 Sep 1; 18: 191–9.
39. Morales L, Alvarez-Garriga C, Matta J, Ortiz C, Vergne Y, Vargas W, et al. Factors associated with breast cancer in Puerto Rican women. J Epidemiol Glob Health. 2013 Dec; 3: 205–15.
40. Nayereh AS, Seyed Morteza S, Ehdaei Vand F, Mardi A. Abortion and Breast Cancer Risk in Women in Mashhad: a case-control study. J Ardabil Univ Med Sci. 2003 Apr 15; 3: 7–12.
41. Lumachi F, Ermani M, Brandes AA, Basso U, Paris M, Basso SMM, et al. Breast cancer risk in healthy and symptomatic women: results of a multivariate analysis. A case–control study. Biomed Pharmacother. 2002 Oct; 56: 416–20.

Iranian Journal of Epidemiology 2016; 12(2): 49-57.
Original Article

Comparison of the Logistic Regression and Classification Tree Models in Determining the Risk Factors and Prediction of Breast Cancer

Zayeri F[1], Seyedagha SH[2], Aghamolaie H[3], Boroumand F[4], Yavari P[5]

1- Associate Professor, Department of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran
2- Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Students' Research Committee, Tehran, Iran
3- Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
4- Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
5- Department of Health and Community Medicine, Medical School, Shahid Beheshti University of Medical Sciences, Tehran, Iran
Corresponding author: Seyedagha SH, hosseinseyedagha@gmail.com

Background and Objectives: Breast cancer is one of the most common malignancies in women which accounts for the highest number of deaths after lung cancer. The aim of the current study was to compare the logistic regression and classification tree models in determining the risk factors and prediction of breast cancer.

Methods: We used from the data of a case-control study conducted on 303 patients with breast cancer and 303 controls. In the first step, we included 16 potential risk factors of breast cancer in both the logistic regression and classification tree models. Then, the area under the ROC curve (AUC), sensitivity, and specificity indexes were used for comparing these models.

Results: From 16 variables included in the models, 5 variables were statistically significant in both models. Sensitivity, specificity, and AUC was 71%, 69%, and 74.7% for the logistic regression and 63.3%, 68.8%, and 71.1% for the classification tree, respectively.

Conclusion: The obtained results suggest that the classification tree has more power for separating patients from healthy people. Menopausal status, number of breast cancer cases in the family, and maternal age at the first live birth were significant indicators in both models.

Keywords: Breast cancer, Risk factors, Classification tree, Logistic regression