دوماهنامه طبّ جنوب
پژوهشکده زیست-پزشکی خلیج فارس
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی بوشهر
سال هجدهم، شماره ۴، صفحه ۷۱۹ – ۷۱۱ (مهر و آبان ۱۳۹۴)

ISMJ 2015;18(4): 711-719

به کارگیری مدل مخاطره جمعی برای پیش‌بینی زمان بقای بیماران مبتلا به لنفوم منتشر سلول‌های بزرگ B و تعیین ژن‌های مؤثر بر آن با استفاده از داده‌های ریزآرایه

عارفه جعفرزاده کهنه‌لو [۱]، علیرضا سلطانیان [۲]، جلال پورالعجل [۲]، حسین محجوب [۳*]

[۱] گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، ایران
[۲] مرکز تحقیقات مدلسازی بیماری‌های غیرواگیر و گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، ایران
[۳] مرکز تحقیقات علوم بهداشتی و گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، ایران

(دریافت مقاله: ۹۳/۳/۱۱- پذیرش مقاله: ۹۳/۶/۹)

چکیده

زمینه: تحقیقات اخیر نشان داده است که بیان ژن‌های مؤثر بر بقای بیماران سرطانی، نقش مهمی را به‌عنوان عامل خطر یا عامل پیشگیری کننده بیماری بازی می‌کنند. مطالعه حاضر برای تعیین ژن‌های مؤثر بر بقای بیماران مبتلا به لنفوم منتشر سلول بزرگ B و پیش‌بینی زمان بقای آن‌ها با استفاده از این ژن‌ها طراحی شده است.

مواد و روش‌ها: مطالعه حاضر یک مطالعه مقطعی بوده که بر روی ۴۰ بیمار مبتلا به لنفوم منتشر سلول B انجام شده است. برای این بیماران ۲۰۴۲ بیان ژن اندازه‌گیری شده است، به‌منظور پیش‌بینی زمان بقای بیماران، ترکیب مدل بقای نیمه‌پارامتری جمعی با دو روش انتخاب ژن الستیک نت و لاسو به‌کار رفته است. مقایسه‌ی روش‌های ذکر شده با رسم سطح زیر منحنی مشخصه عملکرد و محاسبه میانه آن منحنی انجام گرفت.

یافته‌ها: براساس یافته‌های این مطالعه ده ژن مؤثر به‌وسیله روش الستیک نت- جمعی انتخاب شده است و روش لاسو- جمعی هفت ژن انتخاب کرده است. ژن **GENE3325X** زمان بقای بیماران را افزایش می‌دهد (P=۰/۰۰۶) و ژن‌های **GENE3980X** و **GENE377X** زمان بقای بیماران را کاهش می‌دهند (P=۰/۰۰۴). این سه ژن به‌عنوان مهم‌ترین ژن‌ها در هر دو روش انتخاب شده است.

نتیجه‌گیری: مطالعه حاضر نشان داد که روش الستیک نت کارائی بهتری از روش رایج لاسو در تعیین ژن‌های مؤثر بر زمان بقای بیماران دارد. علاوه‌بر این مشاهده شد که به‌کار بردن مدل بقای جمعی به جای مدل رایج کاکس و استفاده از داده‌های ریزآرایه راهکاری مفید و قابل استفاده برای پیش‌بینی زمان بقای بیماران مبتلا به سرطان است.

واژگان کلیدی: سرطان لنفوم، آنالیز بقا، انتخاب ژن، الستیک نت، لاسو

* همدان، خیابان مهدیه، دانشگاه علوم پزشکی همدان، دانشکده بهداشت، گروه آمارزیستی و اپیدمیولوژی

Email :mahjub@umsha.ac.ir

مقدمه

سرطان سیستم ایمنی بدن به دو دسته عمده لنفوم هوچکین (HL) و لنفوم غیرهوچکین (NHL) تقسیم می‌شوند (۱). لنفوم منتشر B سل بزرگ (DLBCL) رایج‌ترین زیرگروه NLH در بزرگسالان است، و در درجه اول افراد مسن‌تر تحت تأثیر آن قرار می‌گیرند (۲). DLBCL سرطان تهاجمی لنفوسیت پیشرفته سلول B است که هرساله حدود ۲۵۰۰۰ نفر بیمار جدید به این بیماری مبتلا می‌شوند (۳). فقط ۴۰ درصد از بیماران مبتلا به این بیماری به درمان‌های رایج پاسخ می‌دهند و زمان بقای طولانی‌تری دارند (۳). مطالعات مختلفی در زمینه بقای بیماران مبتلا به DLBCL انجام شده است که برخی از آن‌ها از فن‌آوری ریزآرایه بهره گرفته‌اند، استفاده از تکنیک‌های ریزآرایه الگوهای منظم بیان ژن را در سرطان سلول B می‌آزماید. سطوح مختلف بیان ژن منجر به نرخ متفاوت گسترش و وضعیت تومور و پاسخ بیمار می‌شود (۳).

مطالعات متفاوتی در مورد داده‌های مربوط به ۴۰ بیمار مبتلا به DLBCL انجام شده است، وجود تعداد بسیار زیادی ژن در مقابل حجم نمونه محدود از مشکلات تجزیه و تحلیل داده‌های بیان ژن است و برای آنالیز بقا ابتدا باید از روش‌های کاهش بعد استفاده کرد. به عنوان مثال باولستد (Bøvelstad) و همکاران چندین روش کاهش بعد رایج را براساس مدل کاکس انجام داده‌اند مانند رگرسیون مولفه‌های اصلی (PCR) و رگرسیون حداقل مربعات جزئی (۴). خوشحالی و همکاران با انجام مطالعه‌ای روی همان مجموعه داده از مدل کاکس برای پیش‌بینی زمان بقا و از رگرسیون ریج [1] برای کاهش بعد متغیرها استفاده کرده‌اند (۵). لی (Li) و همکاران ترکیب دو روش الگوریتم ژنتیک و k نزدیک‌ترین همسایه را برای تعیین ژن‌ها به‌کار بردند (۶). شانگ ما (shuangg Ma) و همکاران روش رگرسیون مؤلفه‌های اصلی را برای کاهش بعد متغیرهای مستقل به‌کار بردند و با استفاده از مدل بقای جمعی زمان بقای بیماران را پیش‌بینی کردند (۷). ژن‌هایی که توسط مطالعات مختلف انتخاب می‌شوند معمولاً همپوشانی کمی دارند و مطالعات نتیجه واحدی در مورد ژن‌های مؤثر نداشته‌اند.

هر کدام از روش‌های به‌کار رفته دارای مزایا و معایبی هستند، مثلاً برآورد پارامترهای حاصل از رگرسیون ریج معمولاً اریبی متمایل به پایین دارند و علاوه‌بر این روش رگرسیون ریج از نظر انتخاب متغیرهای مستقل موجود در مدل صرفه‌جو نیست (۸). در روش‌هایی که ابتدا ژن‌ها خوشه‌بندی می‌شوند و سپس مدل بقای کاکس برازش داده می‌شود، نتایج به انتخاب الگوریتمی که برای خوشه‌بندی استفاده می‌شود بسیار حساس هستند (۹). روش‌های دیگری مانند حداقل مربعات جزئی (۱۰ و ۱۱)، یا تجزیه و تحلیل مؤلفه‌های اصلی (۱۲)، به جای انتخاب متغیرهای اصلی ترکیب‌های خطی از ژن‌ها را انتخاب می‌کنند (۹). یکی از روش‌هایی که اخیراً برای داده‌های ریزآرایه استفاده می‌شود الستیک نت [2] است. در این روش تعداد متغیرهایی که انتخاب می‌شود با حجم نمونه محدود نشده است. مزیت دیگر این روش تشخیص مجموعه ژن‌های هم‌بسته است (۱۳).

مدل رگرسیونی کاکس رایج‌ترین مدل برای داده‌های زمان بقا است، اما این مدل همیشه برازش رضایت بخشی برای داده‌های ریزآرایه ندارد چون بررسی فرضیه متناسب بودن مخاطرات در داده‌های ریزآرایه به

[1] Ridge

[2] Elastic Net

علت تعداد بسیار زیاد متغیرها امکان‌پذیر نمی‌باشد (۱۴). مدل مخاطره جمعی یک جایگزین مناسب برای مدل کاکس است که نیاز به بررسی فرضیه متناسب بودن مخاطرات ندارد، با توجه به مطالب ذکر شده، در این مطالعه از روش جریمه‌ای الستیک نت تحت مدل بقای نیمه پارامتری جمعی برای انتخاب ژن‌های مؤثر بر بقای بیماران مبتلا به سرطان سلول B و پیش‌بینی زمان بقای آن‌ها استفاده شده است علاوه‌بر این برای مقایسه روش الستیک نت با روش‌های سابق روش لاسو به‌عنوان روش رایج کاهش بعد بررسی شده است.

مواد و روش‌ها

در مطالعه مقطعی حاضر از داده‌های DLBCL استفاده شده است که شامل ۴۰۲۶ بیان ژن است. داده‌ها در سایت http://llmpp.nih.gov قابل دسترس هستند (۳). این مجموعه داده شامل زمان بقای ۴۰ بیمار DLBCL است که در طول مدت پیگیری مطالعه ۲۲ نفر فوت شده‌اند (تقریباً ۴۵ درصد سانسور وجود دارد). برای انجام تجزیه و تحلیل آماری از ۲۰۴۲ بیان ژن استفاده شده است که اطلاعات کامل آن‌ها برای بیماران موجود بود.

برای پیش‌بینی زمان بقا، مدل بقای نیمه پارامتری جمعی استفاده شده است که تابع مخاطره مدل به صورت زیر می‌باشد:

$$h(t|Z) = h_{\cdot}(t) + \beta^T Z$$

در اینجا $h_{\cdot}(t)$ تابع مخاطره پایه نامعلوم است و $\beta=(\beta_1 \ldots \beta_p)^T$ ضرایب مستقل از زمان نامعلوم هستند. یکی از تکنیک‌های کاهش بعد متغیرهای مستقل استفاده از تابع جریمه لاسو است که به صورت زیر نوشته می‌شود:

$$L_{pen}(\beta;\lambda) = L(\beta) + \lambda\|\beta\|_1$$

در اینجا $L(\beta)$ تابع زیان و λ پارامتر غیرمنفی است، و $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$، این تابع جریمه امکان صفر برآورد شدن برخی از ضرایب را پدید می‌آورد (۱۵). روش دیگر انتخاب متغیر الستیک نت است، برای مدل مخاطره جمعی تابع الستیک نت به صورت زیر تعریف می‌شود:

$$L_{pen}(\beta;\lambda,\alpha) = L(\beta) + \lambda\|\beta\|_1 + \frac{1}{2}\lambda(1-\alpha)\|\beta\|_2^2$$

که α و λ پارامترهای غیرمنفی هستند و $\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$، در این روش تابع جریمه ترکیبی اکیداً محدب از نواحی تاوان در روش‌های لاسو و ریج است (۸).

ارزیابی قدرت پیش‌بینی دو مدل الستیک نت- جمعی و لاسو- جمعی با استفاده از منحنی مشخصه عملکرد انجام می‌گیرد، این منحنی حساسیت را در برابر تابعی از (ویژگی-۱) برای نقاط مختلف رسم می‌کند. هگرتی (Heagerty) و همکاران این روش را برای داده‌های زمان بقا بسط دادند (۱۳). سطح زیر منحنی مشخصه عملکرد یک معیار عددی برای ارزیابی عملکرد مدل در هر نقطه از زمان است (۱۶).

برای استفاده از روش‌های لاسو و الستیک نت ابتدا باید پارامترهای کنترل را برآورد کنیم، برای این منظور هشتاد درصد نمونه به تصادف انتخاب می‌شود و پارامترها با استفاده از روش اعتبارسنجی متقاطع با ۵ تاخوردگی (5-fold) محاسبه می‌شوند. جهت به‌دست آوردن نتایج معتبر در هر دو روش مرحله انتخاب ژن‌های مؤثر، صدبار تکرار شده است و ژن‌هایی انتخاب شده‌اند که در این صد بار تکرار فراوانی بیشتر از پانزده داشته‌اند. بیست درصد نمونه باقیمانده برای ارزیابی قدرت پیش‌بینی دو مدل الستیک نت- جمعی و لاسو- جمعی با استفاده از سطح زیر منحنی مشخصه عملکرد استفاده می‌شود. تجزیه و تحلیل

داده‌های بیان ژن با استفاده از نرم‌افزار R3.0.3 انجام شده است (http://www.r-project.org). برای برازش مدل جمعی از بسته نرم‌افزاری timereg و برای برآورد پارامترهای کنترل و تعیین ژن‌هایی با ضریب غیر صفر، از بسته نرم‌افزاری ahaz استفاده شده است. همچنین بسته نرم‌افزاری survAUC برای رسم سطح زیر منحنی مشخصه عملکرد و بسته نرم‌افزاری foreach برای تکرار مرحله انتخاب ژن به‌کار رفته است.

یافته‌ها

دامنه زمان بقا برای بیماران DLBCL بین ۱/۳ تا ۱۲۹/۹ ماه قرار دارد و میانه زمان بقا با استفاده از روش کاپلان مایر ۳۲/۵ ماه به‌دست آمده است. بعد از صدبار تکرار مرحله انتخاب ژن در هر روش، ژن‌هایی انتخاب شدند که بیش از پانزده بار ضریب غیرصفر برای آن‌ها برآورد شد.ده ژن به‌وسیله روش الستیک نت- جمعی انتخاب شده است و روش لاسو- جمعی تعداد هفت ژن انتخاب کرده است.

در هر دو روش ژن GENE3980X بیشترین فراوانی تکرار را داشته است به‌طوری که در روش الستیک نت و لاسو به‌ترتیب ۴۶ و ۴۴ بار این ژن انتخاب شده است. ژن‌های انتخاب شده و نتایج حاصل از تحلیل تک متغیره مدل بقای جمعی برای روش‌های الستیک نت و لاسو به ترتیب در جداول ۱ و ۲ آمده است.

جدول ۱) ژن‌های انتخاب شده مؤثر بر بقای بیماران DLBCL با استفاده از روش الستیک نت- جمعی

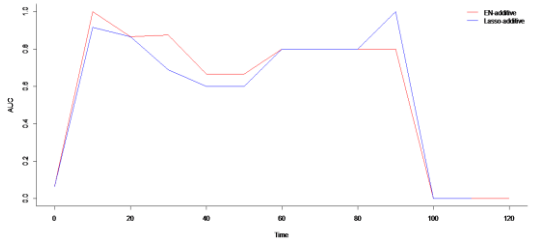
کد ژن	نام ژن (Clone)	ضریب	انحراف معیار	p-value	فراوانی در صدبار تکرار
GENE3980X	1356433	۰/۰۳۴	۰/۰۱۱	۰/۰۰۴	۴۶
GENE950X	199018	۰/۰۲۹	۰/۰۱۴	۰/۰۷۹	۴۲
GENE377X	1357489	۰/۰۲۴	۰/۰۰۹	۰/۰۰۴	۳۱
GENE3391X	824695	-۰/۰۰۵	۰/۰۰۲	۰/۰۳۹	۳۰
GENE3325X	1268870	-۰/۰۱۰	۰/۰۰۲	۰/۰۰۶	۲۹
GENE3973X	181998	۰/۰۲۰	۰/۰۰۷	۰/۰۲۴	۲۶
GENE3405X	1371030	۰/۰۳۴	۰/۰۱۸	۰/۰۵۳	۲۳
GENE172X	1339910	۰/۰۴۱	۰/۰۳۰	۰/۱۱۷	۱۷
GENE1186X	825976	-۰/۰۰۹	۰/۰۱۰	۰/۳۳۲	۱۵
GENE1683X	213797	-۰/۰۰۷	۰/۰۰۵	۰/۰۴۵	۱۵

جدول ۲) ژن‌های انتخاب شده موثر بر بقای بیماران DLBCL با استفاده از روش لاسو- جمعی

کد ژن	نام ژن (Clone)	ضریب	انحراف معیار	p-value	فراوانی در صدبار تکرار
GENE3980X	1356433	۰/۰۳۴	۰/۰۱۱	۰/۰۰۴	۴۴
GENE950X	199018	۰/۰۲۹	۰/۰۱۴	۰/۰۷۹	۳۴
GENE3973X	181998	۰/۰۲۰	۰/۰۰۷	۰/۰۲۴	۲۴
GENE377X	1357489	۰/۰۲۴	۰/۰۰۹	۰/۰۰۴	۲۳
GENE3325X	1268870	-۰/۰۱۰	۰/۰۰۲	۰/۰۰۶	۱۸
GENE1161X	1317052	-۰/۰۰۵	۰/۰۰۳	۰/۲۷۳	۱۵
GENE3045X	1240803	-۰/۰۳۳	۰/۰۱۲	۰/۰۱۳	۱۵
GENE3980X	1356433	۰/۰۳۴	۰/۰۱۱	۰/۰۰۴	۴۴
GENE950X	199018	۰/۰۲۹	۰/۰۱۴	۰/۰۷۹	۳۴
GENE3973X	181998	۰/۰۲۰	۰/۰۰۷	۰/۰۲۴	۲۴

ژن‌های GENE3325X، GENE3980X و GENE377X به‌عنوان مؤثرترین ژن‌ها در هر دو روش است. ضریب محاسبه شده توسط تابع مخاطره مدل نیمه پارامتری جمعی برای ژن‌های GENE3980X و GENE377X مثبت است، یعنی بیان این ژن‌ها باعث کاهش زمان بقای بیماران مبتلا به DLBCL می‌شود (P=۰/۰۰۴)، در حالی که بیان ژن GENE3325X باعث افزایش زمان بقای بیماران می‌شود (P=۰/۰۰۶).

نمودار ۱ سطح زیر منحنی مشخصه عملکرد را برای زمان‌های مختلف در دو روش نشان می‌دهد. میانه سطوح مختلف زیر منحنی مشخصه عملکرد، برای مدل الستیک نت- جمعی، ۰/۸۰ و برای مدل لاسو- جمعی، ۰/۷۴ به‌دست آمده است. سطح زیر نمودار ۱ برای محدوده زمان صفر تا بیشینه زمان بقا محاسبه شده است، مقدار این سطح برای مدل الستیک نت- جمعی، ۰/۸۸۱ و برای مدل لاسو- جمعی مقدار ۰/۷۹۷ محاسبه شده است. بنابراین در روش الستیک نت- جمعی سطح زیر منحنی مشخصه عملکرد بیشتر از روش لاسو- جمعی است.



نمودار ۱) سطح زیر منحنی مشخصه عملکرد برای دو مدل ایستیک نت- جمعی و لاسو- جمعی

جدول ۳ هفت مطالعه را نشان می‌دهد که روی داده‌های DLBCL انجام شده است. در این جدول سعی بر این شده است که ژن‌های مشترک انتخاب شده توسط این مطالعات گردآوری شود. مطالعاتی که تعداد ژن‌های منتخب بیشتری دارند برای پیش‌بینی زمان بقا ملزم هستند که روش کاهش بعد دیگری را نیز برای کاهش تعداد ژن‌های مؤثر به‌کار گیرند.

جدول ۳) بررسی ژن‌های مشترک و مهم در داده‌های DLBCL در هفت مطالعه

کد ژن	مطالعه							
	لی (۶)	هستی (۱۷)	اگولار (۱۸)	وانگ (۲۱)	ناجون شا (۹)	خوشحالی (۵)	نیل (۱۹)	مطالعه حاضر
GENE3332X		❋	❋					
GENE3429X	❋	❋						
GENE1296X	❋		❋					
GENE3328X		❋	❋	❋				
GENE3330X		❋	❋	❋				
GENE148X		❋			❋			
GENE3228X					❋	❋		
GENE3807X		❋				❋		
GENE3045X		❋					❋	❋
GENE1967X		❋					❋	
GENE3256X			❋				❋	
GENE3317X		❋				❋		
GENE3391X						❋		❋
GENE1161X						❋		❋
GENE3980X		❋						❋
GENE3325X	❋		❋					❋
کل تعداد ژن‌های انتخاب شده	۵۰	۲۳۳	۳۳	۶	۴	۱۶	۵۳	۱۲

بحث

یکی از راه‌های پیش‌بینی بقای بیماران مبتلا به سرطان استفاده از داده‌های بیان ژن و انتخاب ژن‌های مؤثر بر بقای بیماران است. مطالعه حاضر نشان داد که استفاده از مدل نیمه پارامتری جمعی و دو روش انتخاب ژن الستیک نت و لاسو به خوبی زمان بقای بیماران مبتلا به DLBCL را پیش‌بینی می‌کند.

با توجه به نتایج حاصل شده از سطح زیر منحنی مشخصه عملکرد در این مطالعه، مشاهده شد که مدل الستیک نت- جمعی برازش بهتری از مدل لاسو- جمعی دارد. در مطالعه‌ای که خوشحالی و همکاران بر روی همین مجموعه داده انجام دادند از روش رگرسیون ریج برای انتخاب ژن‌های موثر و مدل کاکس برای برآورد زمان بقای بیماران استفاده شده است. میانه نمودار مساحت سطح زیر منحنی مشخصه عملکرد برای ۱۶ ژن انتخاب شده توسط خوشحالی ۰/۶۰ و مساحت سطح زیر آن ۰/۶۱ محاسبه شد. این یافته‌ها بیانگر این هستند که برای مجموعه داده DLBCL هر دو روش الستیک نت- جمعی و لاسو- جمعی برازش بهتری نسبت به مدل کاکس مطالعه خوشحالی داشته‌اند. ژن GENE3391X که در این مطالعه توسط روش الستیک نت انتخاب شده است و ژن GENE1161X که توسط روش لاسو انتخاب شده است با ژن‌های انتخاب شده در مطالعه خوشحالی مشترک هستند (۵). در مطالعه‌ای که شانگ ما و همکاران روی داده‌های DLBCL انجام دادند ژن‌هایی را با استفاده از روش مؤلفه‌های اصلی انتخاب کردند. سطح زیر منحنی مشخصه عملکرد برای ده ژن که از بقیه معنی‌دارتر بودند محاسبه شد و مقدار آن تحت مدل بقای جمعی ۰/۸۵۱، و تحت مدل کاکس ۰/۶۵۱ به‌دست آمد (۷)، این مقادیر در مقایسه با سطح زیر منحنی مشخصه عملکرد در مدل الستیک نت- جمعی مطالعه حاضر، کمتر می‌باشند. نتایج حاصل نشان می‌دهد که روش الستیک نت- جمعی برازش بهتری از روش مولفه‌های اصلی برای داده‌هایDLBCL داشته است. در مطالعه‌ای که هستی و همکاران روی همان مجموعه داده انجام دادند از روش اصلاح ژن (Gene shaving) استفاده شد. آن‌ها خوشه‌های کوچکی از ژن‌ها را ایجاد کردند که زمان بقای بیماران را به خوبی پیش‌بینی می‌کرد، یکی از ژن‌های مؤثر انتخاب شده در آن مطالعه ژن GENE3980X بود که این ژن در مطالعه حاضر در هر دو روش انتخاب ژن به کار رفته بیشترین فراوانی را در صد بار تکرار داشته است (۱۷). ژن GENE3325X که توسط هر دو روش انتخاب شده با دو مطالعه جسوس و لی (Jesus & Li) همخوانی دارد (۶ و ۱۸). علاوه بر این ژن GENE3045X که توسط روش لاسو انتخاب شده است با مطالعه هستی و نیل (Hastie & Neill) مشترک است (۱۷ و ۱۹).

آنست (Annest) و همکاران از داده‌های DLBCL برای پیش‌بینی زمان بقای بیماران استفاده کردند و برای انتخاب زیرمجموعه‌ای از ژن‌های مؤثر الگوریتم تکراری بیز (BMA) را به‌کار بردند و ۲۵ ژن مؤثر را شناسایی کردند (۲۰). همچنین وانگ (Wang) و همکاران ۶ ژن تأثیرگذار را شناسایی نمودند (۲۱).

بیان ژن‌های مؤثر بر بقای بیماران نقش مهمی را به‌عنوان عامل خطر یا عامل پیشگیری کننده بازی می‌کند، بنابراین تعیین سطوح این بیان ژن‌ها در برنامه‌ریزی‌های اولیه پیشگیری تأثیرگذار است و می‌توان از آن‌ها به عنوان عامل پیش‌بینی کننده در برنامه‌ریزی‌های ثانویه استفاده کرد. پیشنهاد می‌شود برای تجزیه و تحلیل داده‌های ریزآرایه و پیش‌بینی زمان بقای بیماران در

مطالعات آتی، به جای رگرسیون کاکس از مدل‌های بقای دیگر مانند مدل جمعی استفاده شود و ژن‌های مشترک مطالعه انجام شده با مطالعات قبلی مورد بررسی بالینی قرار بگیرند.

نتایج به‌دست آمده با استفاده از سطح زیر منحنی مشخصه عملکرد نشان داد که روش الستیک نت کارائی بهتری از روش رایج لاسو در تعیین ژن‌های مؤثر دارد. علاوه‌بر این مقایسه مطالعه حاضر با مطالعات قبلی انجام شده روی مجموعه داده DLBCL نشان داد که مدل بقای جمعی برازش بهتری از مدل رایج کاکس برای این مجموعه داده دارد.

سپاس و قدردانی

مقاله حاضر برگرفته شده از پایان نامه کارشناسی ارشد در رشته آمار زیستی بوده و بدین جهت مراتب تشکر و قدردانی خود را از حمایت‌های معاونت پژوهشی دانشگاه علوم پزشکی همدان و مرکز تحقیقات علوم بهداشتی دانشکده بهداشت دانشگاه علوم پزشکی همدان اعلام می‌داریم.

References:

1.Cheng Y, Walkom E. Proposal for the inclusion of ifosfamide in the WHO model list of essential medicines. WHO, Geneva, Switzerland 2008.
2.Vose JM. Current approaches to the management of non-Hodgkin's lymphoma. Seminars in oncology; 1998. p. 483-91.
3.Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403: 503-11.
4.Bøvelstad HM, Borgan Ø. Assessment of evaluation criteria for survival prediction from genomic data. Biometr J 2011; 53: 202-16.
5.Khoshhali M, Mahjub H, Saidijam M, et al. Predicting the survival time for diffuse large B-cell lymphoma using microarray data. J Mol Genet Med 2012;6:287-92.
6.Li L, Weinberg CR, Darden TA, et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001; 17: 1131-42.
7.Ma S, Kosorok MR, Fine JP. Additive risk models for survival data with high-dimensional covariates. Biometrics 2006; 62: 202-10.
8.Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005; 67: 301-20.
9.Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics 2006; 22: 2262-8.
10.Park PJ, Tian L, Kohane IS. Linking gene expression data with patient survival times using partial least squares. Bioinformatics 2002; 18: S120-S7.
11.Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics 2002; 18: 1625-32.
12.Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 2004; 20: i208-15.
13.Engler D, Li Y. Survival analysis with high-dimensional covariates: an application in microarray studies. Stat Appl Genet Mol Biol 2009; 8: 1-22.
14.Ma S, Huang J, Shi M, et al. Semiparametric prognosis models in genomic studies. Brief Bioinform 2010; 11: 385-93.
15.Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997; 16: 385-95.
16.Marzban C. The ROC curve and the area under it as performance measures. Wea Forecasting 2004; 19: 1106-14.
17.Hastie T, Tibshirani R, Eisen MB, et al. Gene shaving'as a method for identifying distinct sets of genes with similar expression patterns.

Genome Biol 2000; 1: 1-0003.

18.Aguilar–Ruiz JS, Azuaje F, Riquelme JC. Knowledge discovery approaches to gene expression data interpretation. Available at: ftp://butler.hpl.hp.com/pub/toomany/paper06.pdf

19.O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC bioinformatics 2003; 4: 13.

20.Annest A, Bumgarner RE ,Raftery AE, et al. Iterative bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data. BMC bioinformatics 2009; 10: 72.

21.Wang Y, Tetko IV, Hall MA, et al. Gene selection from microarray data for cancer classification-a machine learning approach. Comput Biol Chem 2005; 29: 37-46.

ISMJ 2015; 18(4):711-719

Original Article

Applied the additive hazard model to predict the survival time of patient with diffuse large B- cell lymphoma and determine the effective genes, using microarray data

A. Jafarzadeh Kohneloo [1], AR. Soltanian [2], J. Poorolajal [2], H.Mahjub [3]*

[1] Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Iran
[2] Research Center for Modeling of Non-communicable Diseases and Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Iran
[3] Research Center for Health Sciences and Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Iran

(Received 1 Jun, 2014 Accepted 31 Aug, 2014)

Abstract

Background: Recent studies have shown that effective genes on survival time of cancer patients play an important role as a risk factor or preventive factor. Present study was designed to determine effective genes on survival time for diffuse large B-cell lymphoma patients and predict the survival time using these selected genes.

Materials & Methods: Present study is a cohort study was conducted on 40 patients with diffuse large B-cell lymphoma. For these patients, 2042 gene expression was measured. In order to predict the survival time, the composition of the semi-parametric additive survival model with two gene selection methods elastic net and lasso were used. Two methods were evaluated by plotting area under the ROC curve over time and calculating the integral of this curve.

Results: Based on our findings, the elastic net method identified 10 genes, and Lasso-Cox method identified 7 genes. GENE3325X increased the survival time (P=0.006), Whereas GENE3980X and GENE377X reduced the survival time (P=0.004). These three genes were selected as important genes in both methods.

Conclusion: This study showed that the elastic net method outperformed the common Lasso method in terms of predictive power. Moreover, apply the additive model instead Cox regression and using microarray data is usable way for predict the survival time of patients.

Key words: Lymphoma, Survival analysis, Variable selection, Elastic net, Lasso

* *Address for correspondence*: Department of Biostatistics & Epidemiology, School of Public Health, Hamadan Medical Sciences University, Mahdyeh street, Hamadan, Iran. Email: mahjub@umsha.ac.ir

Website: http://bpums.ac.ir
Journal Address: http://ismj.bpums.ac.ir