Modelo de decisión bayesiano para el diagnóstico de cáncer de mama

Quiroz R., Segundo U.; Rivas M., Belzaira

Recibido: 24-05-17 - Revisado: 11-06-17 - Aceptado: 11-09-16

Quiroz R., Segundo U.
Ingeniero de Sistemas.
Magister Scientiae en Estadística Aplicada y Computación. Doctor en Ciencias y Técnicas Estadísticas.
Universidad de Los Andes, Venezuela.
segquiroz@gmail.com

Rivas M., Belzaira
Licenciada en Estadística.
R&Q Asesorías. Empresa Personal dedicada a la asesoría y estudios estadísticos
belzaira@gmail.com

La presente investigación su objetivo fue, construir un modelo de decisión bayesiano automático para cuantificar el riesgo de cáncer de mama y evaluar las consecuencias de las alternativas del tratamiento, desde el punto de vista de las pérdidas y utilidades de todos los actores en la problemática de la salud del paciente (sistema sanitario, médicos y la propia paciente). El modelo incorpora los resultados de una mamografía, algunas variables históricas y una función de costos para clasificar una paciente en tres categorías mutuamente excluyentes: No Cáncer, Probablemente Cáncer y Si Cáncer. Para cuantificar el riesgo, se desarrolló un modelo de regresión binaria bayesiano con distribuciones a priori de Jeffreys y para la selección del modelo, se usó el Factor de Bayes Medio. El ajuste del modelo se realizó desde el punto de vista de las predicciones y de la clasificación, en vez del método clásico de estimación. La base de datos consta de 328 pacientes con 184 casos positivos, ésta fue suministrada por el Hospital Universitario de Granada, España. Se encontró que el diagnóstico puede cambiar drásticamente según cambie la función de pérdida y además que el modelo de predicción debe ser distinto dependiendo de la edad de la paciente, menor o mayor a 50 años.
Palabras clave: Modelo de decisión bayesiano; regresión binaria; distribución a priori de Jeffreys; cáncer de mama.

RESUMEN

An automatic Bayesian decision model was constructed to quantify breast cancer risk and evaluate the consequences of the treatment alternatives, from the point of view of the losses and utilities of all the actors in the patient's health problem (health system , doctors and the patient herself). The model incorporates the results of a mammogram, some historical variables and a cost function to classify a patient into three mutually exclusive categories: No Cancer, Probability of Cancer and Yes Cancer. To quantify the risk, a Bayesian binary regression model was developed with Jeffreys a priori distributions and the Bayes Mean Factor was used to select the model. The adjustment of the model was made from the point of view of predictions and classification, instead of the classical method of estimation. The database consists of 328 patients with 184 positive cases, this was provided by the University Hospital of Granada, Spain. It was found that the diagnosis can change drastically as the function of loss changes and also that the prediction model must be different depending on the age of the patient, younger or older than 50 years.
Keywords: Bayesian decision model; binary regression; a priori distribution of Jeffreys; breast cancer.

ABSTRACT

1. Introducción

El cáncer es una de las principales causas de morbilidad y mortalidad en todo el mundo, cada año se diagnostican alrededor de 14 millones de nuevos casos de cáncer y ocurren más de ocho millones de muertes relacionas con el cáncer. Advierte la Organización Mundial de la Salud (OMS), que los casos anuales de cáncer aumentarán a 22 millones en las próximas dos décadas. El cáncer de mama es el más común entre las mujeres y representa 16% de todos los cánceres femeninos. En los países en desarrollo, su incidencia ha aumentado y la mayoría (69%) de las defunciones por esa causa se registran en estos países (OMS, 2015). La bajas tasas de supervivencia observadas en los países poco desarrollados pueden explicarse principalmente por la falta de programas de detección precoz, que conlleva a que un alto porcentaje de mujeres acudan al médico con la enfermedad ya muy avanzada, pero también, por la falta de servicios adecuados de diagnóstico y tratamiento.

Sin duda, el diagnóstico precoz basado en un procedimiento de screening mamográfico, ha sido probado que es un procedimiento efectivo para el control del cáncer de mama, dado que efectivamente lo detecta en pacientes asintomáticos, lo cual, a su vez, puede

incrementar la probabilidad de curarlo o aumentar la esperanza y la calidad de vida de los pacientes. Según el National Center Institute (NCI) (2015), este método para detectar el cáncer mama ha demostrado que reduce la mortalidad por esta enfermedad entre mujeres de 39 a 69 años de edad, los exámenes de detección con mamografía se relacionaron con una disminución relativa mayor a 15 % en cuanto a la mortalidad por cáncer de mama (Nelson et al., 2009). Las mujeres con mayor riesgo de cáncer de mama, tienen altas probabilidades para la detección precoz y la prevención y, cuantificar la magnitud del riesgo de cáncer de mama, es un factor crucial en la optimización de beneficios médicos cuando se considera la eficacia de métodos de reducción de riesgo (Jeffrey A. et al., 2008).

Muchos programas de screening basan sus diagnósticos en la mamografía, un con-junto de radiografías de la mama, que sirven para detectar cambios en la mujer cuando aún los síntomas del cáncer de mama no se han producido. Por cada 1.000 mujeres que se examinan cada dos años y siguen un programa preventivo, se estima que se evitan entre siete y nueve muertes por cáncer de mama (Paci et al., 2014). También hay que reconocer que existen cánceres que no son visibles en la mamografía, excepcionalmente porque son invisibles, y con mayor frecuencia, porque se encuentran inmersos en mamas densas, con abundante tejido fibroglandular que dificulta o hace imposible su identificación por este procedimiento radiológico. Otro inconveniente que normalmente se presenta con el examen radiológico, es que puede ocurrir un diagnóstico en exceso. Según Paci et al. (2014), por cada 1.000 mujeres que se examinan cada dos años ocurren cuatro casos de diagnóstico en exceso. A pesar de estos inconvenientes, no se pone en duda la eficacia de la mamografía como procedimiento para detectar un carcinoma de mama clínicamente oculto.

Durante las últimas tres décadas, se han utilizado modelos matemáticos y modelos de asistencia computarizada que cuantifican los riesgos de sufrir cáncer de mama. Los profesionales de la medicina pueden incorporar estos modelos para hacer más eficaz los métodos disponibles para la prevención. Además, sería conveniente que cada paciente pueda conocer una estimación de la magnitud del riego de desarrollar cáncer de mama y que se le

brinde la oportunidad de considerar las opciones de minimizar el riesgo.

Las técnicas más usadas en el desarrollo de los modelos matemáticos para cuantificar el riesgo de cáncer de mama son: Análisis discriminante, regresión logística de Girón et al., (1998) y William E. et al. (2006) sistemas expertos, inteligencia artificial, de Wu et al. (1993), redes bayesianas de Burnside and R. (2000) y otros (Jeffrey A. et al., 2008 y Richard J et al., 2007). En la actualidad, ninguno de estos modelos goza de confianza en la comunidad médica que trata el cáncer de mama, aún existe incertidumbre sobre los factores de riesgo y no se han incorporado herramientas para la toma de decisiones, que consideren conjuntamente todos los elementos que en forma natural intervienen en una decisión: Las alternativas, las consecuencias, los actores (médico, paciente, sistema de salud) y los costos-utilidades.

En esta investigación, se construye un modelo de decisión bayesiano automático (objetivo), que utiliza una función de pérdida para incorporar en la decisión final las alternativas, las consecuencias, los actores y los costos-utilidades. La estimación de las probabilidades predictivas (el riesgo), que se usan en el diagnóstico se calculan a través de un modelo de regresión binaria bayesiano automático, con distribuciones a priori de Jeffreys sobre los parámetros del modelo. La base de datos y la información necesaria para construir el modelo, fue suministrada por el departamento de Radiología del Hospital Universitario de Granada, España y está compuesta de 328 pacientes de los cuales 184 son positivos (Quiroz S., 2002).

2. Modelo de regresión binaria

Un modelo bayesiano de respuesta binaria se define como:

$$p_i \equiv \Pr(Y_i = 1|\mathbf{x}_i, F, \boldsymbol{\beta}) = F\left(\mathbf{x}_i^t \boldsymbol{\beta}\right), \qquad i = 1,2,\dots,n, \tag{1}$$

Donde Y_i es una variable aleatoria (v.a.) observable, cuya distribución condicional sobre el parámetro es una Bernoulli(p_i), $\boldsymbol{\beta}=(\beta_1,\dots,\beta_k)^t$ es un vector k- dimensional de parámetros desconocidos, $\mathbf{x}_i^t=(x_{i1},\dots,x_{ik})$ es un vector k- dimensional de covariables observables, $F(\cdot)$ es una función de distribución sobre

P y n es el número de observaciones.

Para un conjunto de n v.a.'s observadas la verosimilitud de $\boldsymbol{\beta}$ viene dada por:

$$lik(\boldsymbol{\beta}|\mathbf{y},\mathbf{x},\mathrm{F}) = \prod_{i=1}^{n} F\left(\mathbf{x}_i^{\mathrm{t}}\boldsymbol{\beta}^{\mathbf{F}}\right)^{y_i}\left[1 - F\left(\mathbf{x}_i^{\mathrm{t}}\boldsymbol{\beta}^{\mathbf{F}}\right)\right]^{1-y_i} \quad (2)$$

Donde al conjunto de datos observados se denota por D=($\mathbf{y}$,$\mathbf{X}$), $\mathbf{y}$=$(y_1,\ldots,y_n)^{\boldsymbol{t}}$ es una muestra observada sobre la variable dependiente y $\mathbf{X}$=$(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\boldsymbol{t}}$ es una matriz de orden n×k con los correspondientes datos de las covariables.

Bajo ciertas condiciones de regularidad e independencia, se puede definir una distribución a priori de Jeffreys sobre

$$\pi^{J}(\boldsymbol{\beta}) = \frac{\det[\mathbf{Z}^t\mathbf{Z}]^{1/2}}{\pi^{k}} \prod_{j=1}^{k} \frac{\mathrm{f}(\mathbf{z}_i^t\boldsymbol{\beta})}{(\mathrm{F}(\mathbf{z}_i^t\boldsymbol{\beta})[1-\mathrm{F}(\mathbf{z}_i^t\boldsymbol{\beta})])^{1/2}}, \quad (3)$$

la cual es una distribución de probabilidad propia, véase Quiroz S. (2002), que depende de una matriz $\boldsymbol{Z}$ de valores de las variables explicativas y de la transformación $F(\cdot)$, pero no depende directamente de los valores observados de las variables aleatorias Yi. La matriz $\mathbf{Z} = (\mathbf{z}_1^t, \ldots, \mathbf{z}_k^t)$ es una submatriz de $\mathbf{X}$, no singular de orden k×k. La función $f(\cdot)$ es la correspondiente función de densidad de $F(\cdot)$, la cual define la forma de la transformación.

La distribución a posteriori de $\boldsymbol{\beta}$, también es una distribución propia y viene dada por:

$$\pi^{J}(\boldsymbol{\beta}|\mathbf{D},\mathrm{F}) \propto \prod_{j=1}^{k} \frac{\mathrm{f}\left(\mathbf{z}_i^t\boldsymbol{\beta}^{F}\right)}{\left(\mathrm{F}\left(\mathbf{z}_i^t\boldsymbol{\beta}^{F}\right)\left[1-\mathrm{F}\left(\mathbf{z}_i^t\boldsymbol{\beta}^{F}\right)\right]\right)^{\frac{1}{2}}} \times \prod_{i=1}^{n}\left(\mathrm{F}\left(\mathbf{x}_i^t\boldsymbol{\beta}^{F}\right)\left[1-\mathrm{F}\left(\mathbf{x}_i^t\boldsymbol{\beta}^{F}\right)\right]\right)^{1-y_i} \quad (4)$$

La densidad de una observación futura, y_{n+1}, condicional al vector de observaciones previas y, asumiendo conocido el vector de covariables y que y_{n+1} es estocásticamente independiente de $\mathbf{y}$, viene dada por:

$$m(y_{n+1}|\mathrm{x}_{n+1}) = \int \mathrm{F}(\mathbf{x}_{n+1}^t\boldsymbol{\beta})^{y_{n+1}}[1-\mathrm{F}(\mathbf{x}_{n+1}^t\boldsymbol{\beta})]^{1-y_{n+1}}\ \pi^{J}(\boldsymbol{\beta}|\mathbf{D},\mathrm{F})\, d\boldsymbol{\beta}. \quad (5)$$

Ninguna de las densidades definidas por (4) y (5) pueden obtenerse analíticamente. En su lugar usaremos los métodos

MCMC, véase por ejemplo Casella and Robert (1999), para calcular estas probabilidades.

La base de datos, **D**, que usaremos para calcular el modelo corresponde a información recogida durante el año 1995 por el Departamento de Radiología del Hospital Universitario de Granada, España, y está compuesta de 328 registros: 184 pacientes presentan cáncer de mama y 144 pacientes no presentan cáncer de mama.

La codificación de las variables se hizo de acuerdo a los factores de riesgo considerados por los médicos de este hospital: (i) Nódulos/Masa, (ii) Calcificaciones, (iii) Asimetrías, (iv) Cambios Arquitecturales. Además se consideraron cinco (5) variables históricas: Edad, antecedentes familiares, antecedentes personales, edad de embarazos y la lactancia. El resultado de una biopsia, realizado a cada paciente, es considerado la prueba infalible que valida el diagnóstico. En el modelo estas variables se denotan por:

$\boldsymbol{y}$: *Resultado de la biopsia*, $\boldsymbol{x_2}$: *Grupo de edad*, $\boldsymbol{x_3}$: *Antecedentes familiares*, $\boldsymbol{x_4}$: *Embarazos y lactancia*, $\boldsymbol{x_5}$: *Antecedentes personales*, $\boldsymbol{x_6}$: *Nódulos* , $\boldsymbol{x_7}$: *Micro-Calcificaciones*, $\boldsymbol{x_8}$: *Cambios arquitecturales*, $\boldsymbol{x_9}$: *Asimetrías*, con lo cual queda definido el vector de variables explicativas como: $\mathbf{x}_i = (x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)^t$.

3. Selección de modelos

La clase de modelos definida por (1) y (2) permite disponer con facilidad de una gran variedad de modelos alternativos con solo cambiar la forma de la distribución F(·). Para este estudio particular del cáncer de mama hemos considerado seis funciones *link*: *Logística (L), Probit (P), Cauchy (C), log-log complementario (O), t-Student con 2 g.l. (S_2)* y *t-Student con 4 g.l. (S_4)*. Denotemos por $M = \{M^L, M^P, M^C, M^O, M^{S_2}, M^{S_4}\}$ a este grupo de modelos formado por las seis subclases de modelos.

En una primera fase, corresponde seleccionar la función *link* que mejor ajusta a los datos y en la segunda fase, se trabaja el problema de selección de las variables del modelo. En ambos casos, usamos el factor de Bayes para comparar los modelos, asumiendo que todos los modelos son igualmente probables a priori.

La selección de la función link requiere de una distribución a priori sobre los parámetros de cada subclase de modelos. Tal y como sugiere Quiroz S. (2002)[sección 3.7] usaremos una misma

matriz $\boldsymbol{Z}$ para definir los modelos a comparar. Concretamente se usará en este caso particular la matriz

$$\boldsymbol{Z} = (\mathbf{x}_9, \mathbf{x}_{15}, \mathbf{x}_{36}, \mathbf{x}_{114}, \mathbf{x}_{146}, \mathbf{x}_{149}, \mathbf{x}_{250}, \mathbf{x}_{274}, \mathbf{x}_{316})^t$$

de orden $k + 1 = 9$ y det[$\boldsymbol{Z}$] = 72.9, la cual fue elegida en forma aleatoria de un conjunto de 5.000 matrices $\boldsymbol{Z}$ formadas a partir de los datos $\boldsymbol{D}$.

$$m(y|\boldsymbol{\beta}^F, M^F) = \int lik(\boldsymbol{\beta}^F|\mathbf{D}, \mathrm{F})\, \pi^J(\boldsymbol{\beta}^F|\mathbf{D}, \mathrm{F})\, d\boldsymbol{\beta}^F, \qquad (6)$$

Las marginales de los datos necesarias para el cálculo de las probabilidades a posteriori de los modelos, fueron estimadas por el método de Laplace, véase por ejemplo Tierney and Kadane (1986).

El cuadro 1, muestra las probabilidades a posteriori para las seis funciones link. Las funciones l*ink: Cauchy y log-log complementario* reciben muy poco soporte de los datos y está claro que podemos descartarlas. El resto de las funciones *link: Logística, Probit, t-Student con 2 g.l y t-Student con 4 g.l*, reciben el mismo soporte y pudiéramos elegir cualquiera de ellas. Sin embargo, el *link* logístico, ofrece grandes ventajas para el cálculo y la interpretación de sus coeficientes de regresión. Esto justifica la elección del *link logístico* para construir el modelo de predicción.

Para la selección de las variables a incluir en el modelo de predicción, se usará el Factor de Bayes Medio, una versión del Factor de Bayes intrínseco de Berger and Pericchi (1996), y el algoritmo para seleccionar variables, ambos propuestos por Quiroz S. (2002)[sección 3.9.2.].

Cuadro 1
Probabilidad a posteriori de M^F

Función *link*, *F*	***L***	***P***	***C***	***O***	S_2	S_4
$m(y\|M^F) \times 10^{-81}$	4.38	3.91	0.84	0.25	3.65	4.24
$\Pr(M^F\|\mathbf{D})$	.254	.226	.049	.015	.211	.245

Fuente: Elaboración propia.

Las matrices $\boldsymbol{Z}$ necesarias para calcular los Factores de Bayes Medio se escogieron en forma aleatoria de un total de 25.370.626.424.000 que resulta de combinar las filas distintas de la matriz **X** en *k* formas distintas.

El espacio de modelos de la clase logística está formado por $2^k = 256$ modelos, los cuales se asume son igualmente probables a priori. Para i=1,…25 y l=1,…,60, la marginal:

$$m_i(Z(l)) = \int lik(\mathrm{y}|\boldsymbol{\beta}_i)\, \pi_i{}^J(\boldsymbol{\beta}_i|Z(l))\, d\boldsymbol{\beta}_i,$$

fue estimada por el método de Laplace, con lo cual resultó un conjunto $\mathcal{B}$ de 8 modelos potencialmente aceptables, véase el cuadro 2. Debido a que $\mathbf{M}_2$ está anidado en cualquiera de los modelos restantes y $\Pr(M_2 \mid \mathbf{D}) > \Pr(M_j \mid \mathbf{D})$, para todo j=3,…,8, es natural reducir la búsqueda del modelo final sobre el conjunto de modelos $\mathcal{A}=\{M_1, M_2\}$.

Con las probabilidades a posteriori de los modelos en $\mathcal{A}$ actualizadas, resulta que $\Pr(M_1 \mid \mathbf{D})$=0.66 y $\Pr(M_2 \mid \mathbf{D})$=0.34. Esta evidencia aconseja considerar ambos modelos para hacer las predicciones. Naturalmente, usaremos la mixtura de $\mathbf{M}_1$ y $\mathbf{M}_2$ ponderados por $\Pr(M_1 \mid \mathbf{D})$ y $\Pr(M_2 \mid \mathbf{D})$, respectivamente.

El cuadro 3 muestra las probabilidades a posteriori de que un coeficiente de regresión no sea igual 0, $\Pr(\beta_j \neq 0 \mid \mathbf{D})$, j = 2,..,k, obtenidas al sumar las probabilidades de los modelos seleccionados que incluyen a ese coeficiente.

Cuadro 2
Modelos potencialmente aceptables (Conjunto $\mathcal{B}$)

Modelo	Covariables	$m(y\|M_j) \times 10^{-78}$	$\Pr(M_j\|\mathbf{D})$
* M1	1 2 5 6 7 8 9	9.82	.539
* M2	1 2 6 7 8 9	5.00	.276
M3	1 2 4 5 6 7 8 9	1.32	.072
M4	1 2 3 5 6 7 8 9	0.79	.044
M5	1 2 4 6 7 8 9	0.75	.042
M6	1 2 3 6 7 8 9	0.36	.019
M7	1 2 3 4 5 6 7 8 9	0.10	.005
M8	1 2 3 4 6 7 8 9	0.06	.003

Fuente: Elaboración propia

Cuadro 3
Pr ($\beta_j \neq 0 | D$)

j	Predictor	Probabilidad a post.
2	Grupo de edad	1.0
3	Antecedentes familiares	0.0
4	Embarazos y lactancia	0.0
5	Antecedentes personales	0.66
6	Nódulos	1.0
7	Microcalcificaciones	1.0
8	Cambios arquitecturales	1.0
9	Asimetrías	1.0

Fuente: Elaboración propia

Bajo este criterio, está claro que la evidencia de las variables x_3 y x_4, antecedentes familiares y embarazos y lactancia, para diagnosticar el cáncer de mama, es nula. Mientras que la evidencia es muy fuerte cuando relacionamos el cáncer de mama con el resto de las variables: $x_2, x_5, x_6, x_7, x_8, x_9$.

La conclusión es clara. Sin embargo, hay que recordar que estos cálculos se refieren a una muestra particular de pacientes. Por consiguiente, decir que la evidencia mostrada por estos datos sobre x_3 y x_4 para explicar el cáncer de mama es nula, no quiere decir que en otra muestra (por ejemplo, con equipos radiológicos de nueva generación, o en otro lugar, o en otro tiempo), estas variables van a tener el mismo comportamiento. Lo que si podemos afirmar es: Primero, para el caso particular considerado aquí se pueden ignorar esas dos variables y segundo, en futuras investigaciones no debe ignorarse este resultado.

4. Bondad de ajuste del modelo

Todos los cálculos para medir el ajuste del modelo $M = 0.66\, M_1 + 0.34\ M_2,$, fueron hechos mediante el método

de *Markov Chain Monte Carlo* (MCMC). Concretamente, en el cálculo de las probabilidades predictivas se usaron muestras de $\pi(\boldsymbol{\beta}_i|\mathbf{D}, M_i),\ i = 1,2$, generadas mediante el algoritmo híbrido Gibbs-Metropolis-Hastings. En todos los casos, el tamaño de la muestra es de 300 ($\mathbf{U} = 300$). Véase por ejemplo, Casella and Robert (1999).

Los datos para ajustar el modelo se obtuvieron de la manera tradicional, i.e. se dividió aleatoriamente la base de datos, **D**, en dos subconjuntos: un subconjunto de entrenamiento, $\mathbf{D}^e$, formado por 75% de los datos y un subconjunto de datos futuros $\mathbf{D}^f = \mathbf{D} \setminus \mathbf{D}^e$. El cuadro 4 muestra varias medidas comúnmente usadas para cuantificar el ajuste del un modelo a los datos. Veáse por ejemplo Raftery et al. (1997), Hosmer and Lemeshow (2000) y Quiroz S. (2002)[capítulo 4].

El valor de c^* es el punto de corte donde se minimiza la pérdida esperada con $l_{10} = l_{01}$. Con estos datos resultó que $0.69 < c^* < 0.72$, con lo cual el área bajo la curva ROC(α_{ROC}), área de la curva de calibración (α_C), la media del logaritmo de la probabilidad predictiva a posteriori (*LPPM*) y el porcentaje de aciertos, muestran la mixtura **M** es ligeramente superior a ambos M_1 y M_2. Además, siguiendo las indicaciones de Hosmer and Lemeshow (2000, p.163), estos valores indican que éste es un excelente ajuste y el modelo está bien calibrado.

Cuadro 4
Ajuste con $\mathbf{D}^f$ en $0.69 < c^* < 0.72$

Modelo	*LPPM*	α_{ROC}	α_C	$Se(c^*)$	$Es(c^*)$	Aciertos %
M_1	0.533	0.827	0.065	0.605	0.976	80.0
M_2	0.544	0.823	0.079	0.579	0.976	78.7
M	0.535	0.832	0.071	0.605	0.976	80.0

Fuente: Elaboración propia

5. Diagnóstico

La fase final del proceso de diagnóstico consiste en tomar una decisión que clasifique a una paciente en alguna de las tres siguientes categorías mutuamente excluyentes: *Cáncer, No Cáncer, y Probablemente Cáncer*. Estas tres categorías constituyen

el resultado de nuestro test de diagnóstico automático, los cuales denotamos por T^+, T^- y $T^\times$; respectivamente. El verdadero estado de la paciente: Cáncer o No Cáncer, los denotamos por C^+ y C^-; respectivamente. Las clasificaciones pueden variar, pero el objetivo de éstas es establecer el diagnóstico final de Cáncer o No Cáncer.

Las consecuencias de cada decisión dependen del verdadero estado de la paciente. En total existen seis situaciones distintas que pueden ser evaluadas en términos de pérdidas-utilidades y calidad de vida. Para una paciente cualquiera, se puede obtener una función de pérdida como que se muestra en el cuadro 5.

Cuadro 5
Función de pérdida

	T^-	T^+	$T^\times$
C^-	0	l_{01}	l_{02}
C^+	l_{10}	0	l_{12}

Fuente: Elaboración propia

Sin pérdida de generalidad se supone que $0 < l_{ij} \leq 1, i = 0,1, j = 0,1,2$, las cuales representan las pérdidas correspondientes al cruce de las categorías del test de diagnóstico y el verdadero estado de la paciente.

En este caso, los actores en la decisión final son tres: La paciente, los médicos tratantes y el sistema integral de salud. Es posible que las preferencias de unos y otros difieran un poco, con lo cual se obtienen distintas funciones de utilidad. Se puede contrastar las decisiones que se producen ante distintas preferencias o combinarlas y obtener una sola ponderando los tres puntos de vista.

Supongamos una nueva paciente con $\mathrm{x}_{n+1} = (x_2, x_5, x_6, x_7, x_8, x_9)$ La probabilidad condicional de que una nueva paciente tenga cáncer de mama, dada la información **D**, está dada por:

$$p_{n+1} \equiv m(y_{n+1} = 1|\mathrm{x}_{n+1}, \mathbf{D})$$

$$= 0.66\, m(y_{n+1} = 1|M_1, \mathrm{x}_{n+1}, \mathbf{D}) + 0.34\, m(y_{n+1} = 1|M_2, \mathrm{x}_{n+1}, \mathbf{D})$$

Y la regla de clasificación es la siguiente:

$$T(y_{n+1}, c) = \begin{cases} 0 & \text{Si } p_{n+1} \leq \mathrm{c}_I \\ 1 & \text{Si } p_{n+1} \geq \mathrm{c}_s \\ 2 & \text{En otro caso.} \end{cases} \quad (7)$$

Donde $c = (c_I, c_s), 0 < c_I \leq c_s \leq 1$, son los umbrales de clasificación inferior y superior, respectivamente.

Para las *n* observaciones en **D** y el par (c_I, c_s), podemos asociar la tabla de contingencia que muestra el cuadro 6.

Cuadro 6
Clasificación del modelo en $c = (c_I, c_s)$

	T^-	T^+	$T^\times$	Total
C^-	$n_{00}(c)$	$n_{01}(c)$	$n_{02}(c)$	$n_{0\cdot}$
C^+	$n_{10}(c)$	$n_{11}(c)$	$n_{12}(c)$	$n_{1\cdot}$
Total	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Fuente: Elaboración propia

Donde $n_{ij}(c), i = 1,2; j = 0,1,2;$ son los correspondientes números de observaciones que resultan del cruce de las categorías del test de diagnóstico con las categorías del verdadero estado de la paciente.

Denotemos por $\boldsymbol{\gamma}_0(c) = (\gamma_{00}(c), \gamma_{01}(c), \gamma_{02}(c))^t$, al vector $(\Pr(T^+ \mid C^-), (\Pr(T^+ \mid C^-), \Pr(T^\times | C^-))$ y por $\boldsymbol{\gamma}_1(c) = (\gamma_{10}(c), \gamma_{11}(c), \gamma_{12}(c))^t$, al vector $(\Pr(T^- | C^+), \Pr(T^+ | C^+), \Pr(T^\times | C^+)), \Pr(T^+ | C^+), \Pr(T^\times | C^+))$, los parámetros de las dos multinomiales asociadas a esta tabla de contingencia. Supongamos una distribución a priori de Jeffreys sobre $\boldsymbol{\gamma}_i, i = 0,1$; i.e., una distribución Dirichlet $\mathcal{D}_3\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$. Entonces, la distribución a posteriori sobre γ_i viene dada por:

$$\pi^J(\boldsymbol{\gamma}_i | \mathbf{D}) = \mathcal{D}_3\left(n_{i0}(c) + \tfrac{1}{2}, n_{i1}(c) + \tfrac{1}{2}, n_{i2}(c) + \tfrac{1}{2}\right), i = 0,1$$

De la pérdida total en $c = (c_I, c_s)$, $L(c) = n_{01}(c)l_{01} + n_{02}(c)l_{02} + n_{10}(c)l_{10} + n_{12}(c)l_{12}$, resulta que la pérdida esperada a posteriori viene dada por:

$$E[L(c)] = \frac{n_{0\cdot}}{\left(n_{0\cdot} + \frac{3}{2}\right)}\left(\left(\frac{1}{2} + n_{01}(c)\right) l_{01} + \left(\frac{1}{2} + n_{02}(c)\right) l_{02}\right) + \frac{n_{1\cdot}}{\left(n_{1\cdot} + \frac{3}{2}\right)}\left(\left(\frac{1}{2} + n_{10}(c)\right) l_{10} + \left(\frac{1}{2} + n_{12}(c)\right) l_{12}\right) \quad (8)$$

El punto $c^* = (c_l^*, c_s^*)$ que minimice (8), será el par de puntos de corte que definen el test de diagnóstico (7) óptimo.

Cualquier centro de salud dispone de la información necesaria para determinar una función de costos-utilidades que permita implementar este modelo. La base para definir dicha función son los costos asociados a los exámenes clínicos, medicamentos, hospitalización, atención médica, entre otros.

Y los beneficios están asociados a la calidad de vida de las pacientes, valoradas por los actores del problema, i.e. las pacientes, los médicos, los directores de los centros de salud y el personal paramédico. Lógicamente que no es fácil medir la calidad de vida para los distintos escenarios; no obstante, existen trabajos donde se estudian estos aspectos y que podemos tomar como referencia, véase por ejemplo, Parmigiani (1999), Paci et al. (2014), Arrospide et al. (2015).

6. Ejemplo de función de pérdida

Una función de pérdida particular puede ser construida basándose en entrevistas hechas a las pacientes, a los médicos, a los directores y administradores de los centros de salud, véase Arrospide et al. (2015). Es natural que en cada caso las pérdidas varíen, pero en la práctica, la mayor variación ocurre entre los pacientes. A manera de ejemplo, hemos construido tres funciones de pérdida basadas en las siguientes consideraciones:

(i) Asumimos que no existen costos asociados a las clasificaciones hechas correcta-mente. Es decir, las pérdidas por clasificar un verdadero negativo y un verdadero positivo son cero. Por lo tanto, l00 = 0 y l10 = 0, respectivamente.

(ii) Los costos asociados a una clasificación de *falso positivo*, se deben al costo de la biopsia y al costo de la calidad de vida de la paciente sometida a un severo examen extra.

(iii) Una clasificación de *Probablemente Cáncer* con *resultado negativo*, requiere de exámenes clínicos y revisiones médicas que resultan ser menos severas, en todos los sentidos, que el de una biopsia. Excepcionalmente, se recurre a la biopsia para confirmar el diagnóstico.

(iv) Los costos asociados a una clasificación *falso negativo*, se deben a tener que tratar a la paciente con un cáncer de mama en un estado avanzado, en vez de hacerlo en estados tempranos del desarrollo, cuando se supone existe un ahorro importante en el tratamiento. El tratamiento en estado avanzado del cáncer de mama, como en cualquier otra enfermedad incurable, conlleva entre otros costos los siguientes: Hospitalización, exámenes clínicos, medicamentos, pérdida de la calidad de vida, e incluso la misma muerte.

(v) Una clasificación *Probablemente Cáncer* con *resultado positivo*, comienza con exámenes clínicos y revisiones médicas no muy severas y siempre finaliza con una biopsia confirmatoria. La diferencia con una clasificación *verdadero positivo* son los exámenes y revisiones previas.

(vi) Por (ii) y (iii) asumimos que $l_{10} \geq l_{02}$.

(vii) Por (iv) asumimos que la mayor pérdida ocurre cuando se clasifica un *falso negativo*. Es decir $l_{10} = 1$.

(viii) Por (i) y (v) asumimos que $0 \leq l_{12} < l_{02}$.

En cualquier caso, una función de pérdida construida sobre estas consideraciones generales, verifica que:

$$(i) \quad c_I = \frac{l_{02}}{(l_{10} + l_{02} - l_{12})}$$

$$(ii) \quad c_S = \frac{(l_{01} - l_{02})}{(l_{01} + l_{12} - l_{02})} \qquad \text{sí } \frac{l_{02}}{(l_{10} + l_{02} - l_{12})} \leq c_M \leq \frac{(l_{01} - l_{02})}{(l_{01} + l_{12} - l_{02})},$$

$$(iii) \quad c_I = c_S = \frac{l_{01}}{(l_{10} + l_{01})} \qquad \text{en otro caso,}$$

donde $c_M = \frac{l_{01}}{(l_{10}+l_{01})}$

El cuadro 7 muestra la clasificación hecha por el modelo con tres funciones de pérdida construidas a manera de ejemplo. Allí puede verse que el diagnóstico depende, de forma crítica, del punto de vista de los costos, además del examen radiológico. La primera función de pérdida (costos), corresponde a un punto de vista que juzga como muy importante el gasto de los exámenes médicos. La segunda función de pérdida, muestra el efecto en la clasificación cuando el peso relativo de los exámenes médicos es menor que el del primer caso. La tercera función de pérdida, considera que el costo promedio para confirmar un *diagnóstico negativo*, es similar si se hace con una biopsia directamente o se hace con procedimientos menos agresivos.

Cuadro 7
Clasificación con diversas pérdidas ($\boldsymbol{n = 328},\ \boldsymbol{n_{0.} = 144, n_{1.} = 184}$)

l_{00}	l_{01}	l_{02}	l_{10}	l_{11}	l_{12}	c_I^*	c_S^*	Clasificación*
0	1.00	0.75	1	0	0.20	0.57	0.60	$\begin{matrix} & T^- & T^+ & T^\times \\ C^- & 109 & 28 & 7 \\ C^+ & 47 & 130 & 7 \end{matrix}$
0	0.75	0.50	1	0	0.10	0.30	0.66	$\begin{matrix} & T^- & T^+ & T^\times \\ C^- & 59 & 7 & 78 \\ C^+ & 18 & 101 & 65 \end{matrix}$
0	0.90	0.80	1	0	0.10	0.57	0.57	$\begin{matrix} & T^- & T^+ & T^\times \\ C^- & 109 & 35 & 0 \\ C^+ & 47 & 137 & 0 \end{matrix}$

Fuente: Elaboración propia

7. Análisis por grupos de edad

Está probado, que el diagnóstico precoz del cáncer de mama basado en un procedimiento de *screening* mamográfico es efectivo, sin embargo, existe controversia sobre la efectividad de este procedimiento para mujeres menores de 50 años, véase, por

ejemplo, Parmigiani (1999), Lee et al. (2010). Sin duda, la *prevalencia* de la enfermedad, junto con la *sensibilidad* y *especificidad* del test de diagnóstico, tienen mucho que ver con los beneficios que dejan los programas de *screening*. A menor prevalencia del cáncer de mama se necesitan pruebas de diagnóstico con mayor *sensibilidad* y *especificidad* que hagan eficiente el programa.

De algún modo, la controversia sugiere investigar la posibilidad de construir distintas pruebas para cada grupo de edad. Con esta idea, se repitió el cálculo y se construyó un modelo de respuesta binario, en forma separada, para dos grupos de edad. El Grupo 1, formado por 145 mujeres menores de 50 años de edad y que denotamos por $\mathbf{D}_{01}$ y el grupo 2, formado por las 183 mujeres restantes de un total de 328 mujeres. Así que $\mathbf{D}_2=\mathbf{D} \setminus \mathbf{D}_{01}$.

Un resumen de los resultados se muestra en los cuadros 8 y 9. La última columna de cada cuadro, muestra las probabilidades a posteriori de los modelos aceptados. Basándose en esta evidencia, podemos ver con claridad que las variables x_3, antecedentes familiares, x_4, embarazos y lactancia, y x_9, asimetrías, no son importantes para explicar la presencia de cáncer de mama en mujeres menores de 50 años. El comportamiento es distinto para el grupo de mujeres con 50 años o más, la evidencia muestra que las variables x_3, antecedentes familiares, x_4, embarazos y lactancia y x_5, antecedentes personales, no son importantes para explicar presencia de cáncer de mama en este caso.

Desafortunadamente la cantidad de datos y la composición en las submuestras ($\mathbf{D}$ y $\mathbf{D}_{01}$), no son suficientes para poder obtener pruebas de diagnóstico fiables. No obstante, estos resultados coinciden con otros estudios que evalúan la utilidad de los programas de screening en mujeres menores a 50 años, véase por ejemplo, Lee et al. (2010). Pero lo más importante, es que estos resultados sugieren que las variables regresoras son distintas según el grupo de edad y por lo tanto, indican la necesidad de desarrollar modelos distintos para cada grupo de edad, lo cual posiblemente incremente la precisión en el diagnóstico.

Cuadro 8
Edad < 50 años: Modelos potencialmente aceptables y modelos aceptados*

Modelo	Covariables								$\Pr(M_j \mid \mathbf{D}_{01})$	$\Pr(M_j \mid \mathbf{D}_{01})^*$
*M39	1			5	6	7	8		0.65	1
M49	1	3		5	6	7	8		0.15	
M62	1	3		5	6	7	8	9	0.08	
M57	1			5	6	7	8	9	0.07	
M53	1		4	5	6	7	8		.05	

Fuente: Elaboración propia

Cuadro 9
Edad ≥ 50 años: Modelos potencialmente aceptables y modelos aceptados*

Modelo	Covariables								$\Pr(M_j \mid \mathbf{D}_2)$	$\Pr(M_j \mid \mathbf{D}_2)^*$
*M16	1					6	7	8	9	0.47	0.65
*M5	1					6	7	8		0.25	0.35
M6	1					6	7		9	0.09	
M26	1				5	6	7	8	9	0.08	
M22	1			3		6	7	8	9	0.06	
M14	1				5	6	7	8		0.05	

Fuente: Elaboración propia

8. Discusión final

La dificultad del problema de la selección de modelos es bien conocida y más aún, cuando se trata de la construcción de un modelo de decisión que utiliza únicamente los métodos bayesianos automáticos (objetivos). Este trabajo muestra los métodos que permiten construir un modelo de decisión con estas características,

usando distribuciones a priori de Jeffreys y una versión del *Factor de Bayes Intrínseco* llamado *Factor de Bayes Medio*.

En particular, el modelo regresión binaria que se construyó consideró inicialmente seis funciones link (Logístico, Probit, Cauchy, log-log complementario, *t*-Student con 2 g.l. y *t*-Student con 4 g.l.) y ocho variables explicativas, lo cual indica trabajar con 1.536 posibles modelos. A través del Factor de Bayes Medio, se determinó que el modelo que mejor ajusta a los datos está conformado por una mixtura de dos modelos logísticos con seis variables explicativas solamente. Tomando en consideración varias medidas de la bondad del modelo, por ejemplo ROC = 0.832 y % de aciertos = 80.0, estas indican que la mixtura tiene un excelente ajuste y calibración.

También se encontró, que el diagnóstico puede cambiar drásticamente según cambie la función de pérdida y además se confirmó, que el modelo de predicción debe ser distinto si la paciente es menor o mayor a 50 años.

El modelo teórico desarrollado aquí es de carácter general y puede ser aplicado a otras enfermedades e infinidad de problemas de decisión que tiene esta misma estructura, varios potenciales factores de riesgo y varias alternativas de acción.

Una característica importante del modelo de decisión desarrollado, es que permite la posibilidad de clasificar a una paciente en tres categorías, en vez de dos que es la forma clásica: *Cáncer* o *No cáncer*. La tercera categoría, *Probablemente Cáncer*, indica que es necesario encontrar nuevas evidencias para poder establecer el diagnóstico final. Es fácil adaptar el modelo para que considere otras clasificaciones distintas, que permitan ahorrar recursos y molestias innecesarias a las pacientes al momento del diagnóstico, pero siempre con el objetivo de determinar del verdadero estado de la paciente.

Las funciones de pérdida que necesita el modelo parecen fáciles de construir, pero el escepticismo de los administradores de los centros de salud y del cuerpo médico tratante, dificulta la tarea de evaluar la verdadera utilidad de estos modelos. Es fundamental lograr la participación de la comunidad médica para poder desarrollar esta clase de modelos, una tarea no exenta de dificultades.

9. Referencias

Arrospide A., Soto M., Acaiturri T., López G., Abecita L., and Mar J. (2015). *Cost of breast cancer treatment by clinical stage in the Basque contry, Spain.* Revista Española Salud Publica, 89(1):93–97.

Berger J. and Pericchi L. (1996). *The intrinsic bayes factor for model selection and prediction*. Journal of the American Statistical Association, 91:109–122.

Burnside D., Rubin E. and Shachter R. (2000). *A bayesian network for mammography. Management Science and Engineering.* Stanford University. USA.

Casella G. and Robert C. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Girón F., Martínez M., and Moreno E. (1998). *Automatic diagnostic of breast cancer: A case study. Applied Decision Analysis,* Edited by F. Javier Girón. (Kluwer Academic Publishers), pages 101–113.

Hosmer W. and Lemeshow S. (2000). *Applied Logistic Regression*. New York: John Wiley and Sons Ltd.

Jeffrey T., Steven R., Smith-Bindman R., Ichikawa, Barlow William E. and Kerlikowske K. (2008). *Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model.* Annals of Internal Medicine, 148(5):337–347.

Lee C, Dershaw D, Kopans D, Evans P, Monsees B, Monticciolo D, Brenner R, Bassett L, Berg W, Feig S, Hendrick E, Mendelson E, D'Orsiç C, Sickles E, and Burhenne L. (2010). *Breast cancer screening with imaging: recommendations from the society of breast imaging and the acr on the use of mammography, breast mri, breast ultrasound, and other technologies for the detection of clinically occult breast cancer.* Journal of the American College of Radiologiy, 7(1):18–27.

NCI. (2015) *Screening tests that have been show to reduce cancer deaths.* Technical report, National Cancer Institute, USA.

Nelson H, Tyne K, Naik A, Bougatsos C, Chan B, and Humphrey L. (2009). *Screening for breast cancer: an update for the u.s. preventive services task force.* Annals of Internal Medicine, 151(10):727–737.

Organización Mundial de la Salud. (2015). *Carga mundial de*

morbilidad. Technical report, Organización Mundial de la Salud.

Paci E, Broeders M, Hofvind S, Puliti D, Duffy S and EUROSCREEN Working Group. (2014). *European breast cancer service screening outcomes: a first balance sheet of the benefits and harms.* Cancer Epidemiol Biomarkers Prev., 23(7):1159–1163.

Parmigiani G. (1999). *Decision models in screening for breast cancer. Bayesian Statistics 6*, Edited by Bernardo, J., Berger, J., Dawid, A. and Smith, A. (Clarendon Press Oxford), 6:525–546.

Quiroz S. (2002). *Modelos de Respuesta Binaria Bayesianas.* Tesis Doctoral, Universidad de Granada, Facultad de Ciencias, Departamento de Estadística e Investigación de Operaciones, Granada.

Raftery A, Madigan D, and Hoeting. (1997). *Bayesian model averaging for linear regres-sion models*. Journal of the American Statistical Association, 92:179–191.

Richard S, Norman B, Rowan R, Cummings S, Cuzick J, Dowsett M, Easton D, Forbes J, Key T, Han-kinson S, Howell A, and Ingle J. (2007). *Critical assessment of new risk factors for breast cancer: considerations for development of an improved risk prediction model*. Endocrine-Related Canc, 14:169–187.

Tierney L and Kadane J. (1986). *Accurate approximations for posterior moments and marginal densities.* Journal of the American Statistical Association, 81:82–86.

William B, White E, Ballard-Barbash R, Vacek P, Titus-Ernstoff, Carney P, Jeffrey T, Buist D, Geller B, Rosenberg R, Bonnie Y, and Kerlikowske K. (2006). *Prospective breast cancer risk prediction model for women undergoing screening mammography.* Journal of the National Cancer Institute, 98(17):1204–1214.

Wu Y, Giger M, Doi K, Vyborny C, and Schmidt R. (1993). *Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer*. Radiology, 187:81–87.