Psicothema
ISSN: 0214-9915
psicothema@cop.es
Universidad de Oviedo
España

Richard´s, María Marta; Solanas, Antonio; Ledesma, Rubén D.; Introzzi, Isabel M.; López Ramón, María Fernanda

Técnicas estadísticas de clasificación: un estudio comparativo y aplicado

Psicothema, vol. 20, núm. 4, 2008, pp. 863-871
Universidad de Oviedo
Oviedo, España

Psicothema 2008. Vol. 20, nº 4, pp. 863-871
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Técnicas estadísticas de clasificación: un estudio comparativo y aplicado

María Marta Richard's, Antonio Solanas*, Rubén D. Ledesma, Isabel M. Introzzi y María Fernanda López Ramón
Universidad Nacional de Mar del Plata y * Universidad de Barcelona

El propósito de este trabajo es evaluar y comparar tres técnicas estadísticas de clasificación —regresión logística, análisis discriminante y árboles de clasificación— para identificar las características de personalidad asociadas al riesgo de padecimiento de episodios isquémicos cardiovasculares agudos (EICA). Se utilizaron la tasa de error y el índice C de discriminación como criterios para comparar los resultados obtenidos con las distintas técnicas. El tamaño de la muestra, compuesta por mujeres y hombres con edades comprendidas entre 36 y 80 años, fue igual a 313 participantes, quienes fueron divididos en dos grupos: clínico (n= 143) y control (n= 170), equiparándose por sexo, edad, nivel socioeconómico y nivel educacional. El análisis de los resultados posibilitó la elección de los árboles de clasificación como la técnica más adecuada, ya que aporta un porcentaje de predicción más elevado, un tratamiento simple de los datos y una significativa interpretación clínica. Los resultados posibilitaron reducir a 9, de las 22 originales, las escalas de personalidad asociadas a una mayor probabilidad de padecer EICA y generar un modelo empírico de siete perfiles de personalidad asociados al aumento de la probabilidad de EICA y cinco perfiles de personalidad vinculados a la ausencia de patología.

Classification statistical techniques: An applied and comparative study. The aim of this article is to assess and compare three classification statistical techniques —logistic regression, discriminant analysis and classification trees— to identify the personality characteristics associated with the risk of suffering from ischemic cardiovascular acute episodes (ICAE). The sample comprised 313 participants, men and women, aged from 36 to 80. Participants were divided into two groups: a clinical group of patients (n= 143) who were diagnosed as suffering from ICAE, and a control group (n= 170). Both groups were equated in gender, age, socio-economic and educational level. In view of the comparative study of the analytical procedures, we recommend classification trees as the best choice, as it was the most accurate for the individuals in the clinical group, a simple data analysis and a meaningful clinical interpretation. The predictive validity analysis of the MCMI-II allowed the construction of a reduced version made up of 9 personality scales from the 22 scales in the original version. Thus, we could identify the patients with a higher probability of suffering from ICAE, and additionally, generate an empirical model comprising seven and five personality profiles associated, respectively, with the increase and the decrease of the probability of suffering from ICAE.

De acuerdo con la literatura científica, las enfermedades cardiovasculares son la principal causa de muerte y morbilidad severa en la mayoría de los países desarrollados y, en particular, el infarto agudo de miocardio se encuentra entre las principales causas de muerte entre la población adulta del mundo occidental (*American Heart Association*, 1996). La enfermedad coronaria, también conocida como cardiopatía isquémica, es el proceso patológico que implica déficit de riego sanguíneo en sectores del músculo cardíaco (miocardio) en forma permanente o transitoria. Este déficit circulatorio, que pone en juego diversos mecanismos de compensación, dispara síntomas y puede concluir en la muerte celular de porciones del miocardio (isquemia). En el presente trabajo se considera el grupo de los episodios isquémicos cardiovasculares agudos (EICA), englobando en éstos a la angina inestable y al infarto agudo de miocardio, de acuerdo con la propuesta de Marso, Griffin y Topol (2002).

El modelo médico ya no resulta suficiente para el abordaje de las enfermedades cardiovasculares, razón por la cual actualmente se suelen incluir aspectos psicológicos como factores de riesgo de padecimiento de EICA. En la actualidad se observa que la literatura científica suele destacar que los factores psicológicos tienen un papel importante en el impacto, curso y tratamiento de las enfermedades cardiovasculares. Numerosos estudios han encontrado evidencias sobre la relación entre depresión, ansiedad y otras variables psicológicas con el incremento de la mortalidad y morbilidad cardiaca en pacientes con patologías isquémicas (Buceta y Bueno, 1996; Denollet et al., 1996; Welin, Lappas y Wilhelmsen, 2000; Yusuf et al., 2004; Zellweger, Osterwalder, Langewitz y Pfisterer, 2004). Sin embargo, no existe acuerdo sobre cuáles serían los rasgos, perfiles o patrones de personalidad que se asocian a los episodios cardiovasculares. Para identificar los mencionados perfiles es frecuente utilizar distintas técnicas estadísticas de clasi-

Fecha recepción: 31-10-07 • Fecha aceptación: 23-3-08
Correspondencia: María Marta Richard's
Facultad de Psicología
Universidad Nacional de Mar del Plata
7600 Mar del Plata - Buenos Aires (Argentina)
E-mail: mariamartarichards@gmail.com

ficación, como son la regresión logística (RL) y el análisis discriminante (AD).

La RL no requiere supuesto alguno sobre la distribución de los datos y, por tanto, resulta más robusta que el AD si la variable dependiente es binaria (Pohar, Blas y Turk, 2004). El modelo de la RL permite estimar la probabilidad de un suceso en función de un conjunto de variables de predicción, que pueden ser cualitativas o cuantitativas, siendo el objetivo de esta técnica hallar el modelo conjunto más parsimonioso para describir las relaciones existentes entre la variable de respuesta y las variables de predicción. La interpretación de la ecuación de regresión logística es relativamente simple y puede usarse como una función para clasificar la pertenencia a un grupo u otro. Además, el resultado permite valorar la calidad de la clasificación en términos de sensibilidad y especificad, refiriéndose estos dos conceptos a la capacidad de una prueba diagnóstica para detectar correctamente a las personas que realmente padecen una patología y a aquellas que ciertamente no la sufren, respectivamente. No obstante, la RL también presenta ciertos inconvenientes, entre los cuales cabe mencionar los efectos negativos de la multicolinealidad sobre las estimaciones de los errores estándar de los parámetros del modelo (McGee, Reed y Yano, 1984; Pohar et al., 2004). También deberíamos añadir ciertas consideraciones críticas, como las que se refieren a la utilización de la RL para seleccionar grupos de variables de predicción y evaluar la importancia relativa de cada variable de predicción. En esta línea, la RL por pasos hacia delante o hacia atrás produce resultados discrepantes, tanto cuando se emplean diferentes alternativas dentro de este método, como cuando se trabaja con distintas muestras de la misma población (Silva y Barroso, 2001).

Si las variables explicativas son numéricas, también puede recurrirse al AD para clasificar casos en categorías. Este método estadístico requiere un conjunto de variables discriminantes y una variable de agrupación. El objetivo básico cuando se utiliza esta técnica es obtener una función discriminante tal que permita la mejor clasificación de los individuos en los grupos especificados. Esto último acostumbra a evaluarse mediante tablas de confusión y así determinar la tasa de aciertos o errores. No obstante, cabe destacar que, a pesar de la existencia de varios estudios en el ámbito de la personalidad que utilizan esta técnica (Abbate-Daga, Amianto, Rogna y Fassino, 2007; Freeman, Hayes, Kuch y Taub, 2007; Rice y Ashby, 2007), su uso requiere el cumplimiento de ciertos supuestos. Estos supuestos son, en general, difíciles de aceptar en la práctica clínica. Así, puesto que los contrastes de decisión requieren una distribución normal multivariante de las variables de predicción, que la variabilidad de éstas sea homogénea y que sus distribuciones no sean extremadamente asimétricas, algunos autores han sostenido que el AD es más robusto para variables con más de dos o tres categorías (Pitarque, Ruiz y Roy, 2000; Pohar et al., 2004; Worth y Cronin, 2003).

Para identificar perfiles clínicos, los *árboles de clasificación* (AC) o *árboles de decisión* pueden mejorar la descripción clínica de los patrones de personalidad asociados a las distintas psicopatologías. Brevemente, los árboles de clasificación permiten asignar a los individuos de la muestra a las distintas categorías o valores de una variable objetivo o, si se prefiere, obtener segmentos a partir de un conjunto de variables de clasificación, pudiendo ser estas últimas variables de tipo categórico o numérico. Los distintos algoritmos disponibles proceden realizando divisiones en secuencia de los individuos de la muestra. Cada una de estas particiones puede ser dividida hasta que se cumple un criterio de parada del algoritmo. A fin de ilustrar más claramente cómo proceden este tipo de algoritmos, en la figura 1 se muestra un árbol de clasificación. Algunos de los algoritmos se fundamentan en criterios estadísticos para realizar las particiones, mientras otros se basan en la optimización de alguna función objetivo, como puede ser la función de entropía o, complementariamente, la función de información. Debido a la breve descripción aquí realizada, se sugiere la lectura de alguno de los distintos manuales donde se explican extensamente los árboles de clasificación (Berry, 1997; Breiman, Friedman, Olshen y Stone, 1998; Picón, Varela y Lèvy, 2004; Román y Lévy, 2003).

Los árboles de clasificación expresan la información en términos directamente inteligibles para los psicólogos clínicos y resultan, al mismo tiempo, simples de interpretar —las ramas del árbol de clasificación emulan en parte el proceso humano de decisión y, además, es posible obtener la precisión de cada regla de asignación a las clases—. Los AC también permiten reducir el número de variables y detectar relaciones no lineales. Se trata de un enfoque más flexible y menos restrictivo que la RL y el AD, posibilitando que los individuos sean asignados a distintas clases a partir de un conjunto de variables de clasificación. Los árboles de clasificación no suponen modelos estadísticos a priori y posibilitan hallar patrones o perfiles, siendo común obtener una medida sobre la calidad del método de clasificación por medio de la tasa de aciertos (Picón y Varela, 2000).

Este trabajo se propone evaluar y comparar la precisión alcanzada mediante tres técnicas estadísticas de clasificación —RL, AD y AC— en un estudio aplicado, a fin de identificar las características de personalidad asociadas al riesgo de padecimiento de episodios isquémicos cardiovasculares agudos. Se pretende mostrar a los psicólogos clínicos la utilidad práctica que pueden tener los árboles de clasificación.

Método

En el contexto de aplicación del presente estudio, la variable de respuesta es categórica binaria; por tanto, el problema se plantea en términos de clasificación o pertenencia de los participantes a los grupos clínico o control. En otros términos, ¿en qué medida podemos clasificar correctamente a los participantes en el estudio a partir de sus características de personalidad, que han sido evaluadas con un instrumento de medida? Desde el punto de vista estadístico, este problema puede abordarse mediante diferentes técnicas, como son la RL, el AD y los AC.

Participantes

Todos los participantes asignados al grupo clínico fueron diagnosticados por el Servicio de Cardiología del Hospital Interzonal General de Agudos (HIGA) de la ciudad de Mar del Plata, Argentina. Como criterio de inclusión en el grupo clínico era preciso que a la persona, en el proceso diagnóstico, se le hubiera detectado una angina inestable y/o un infarto agudo de miocardio, siempre de acuerdo con los criterios de Marso et al. (2002). Se excluyeron todos aquellos pacientes internados por otras patologías cardiovasculares (por ejemplo, valvulares y arritmias) y aquellos que no pudieron ser entrevistados a causa de su estado (por ejemplo, precisaron asistencia respiratoria o manifestaron algún tipo de conmoción), además de quienes se negaron a responder a las preguntas. Una vez se dispuso de la muestra inicial para el grupo clí-

nico, se descartaron del análisis de datos 17 individuos, un 11.8% de ese grupo, por presentar protocolos inválidos en función de los criterios establecidos por los índices de validez del *Millon Clinical Multiaxial Inventory* (MCMI-II, Millon, 1999).

Finalmente, la muestra quedó compuesta por 313 participantes adultos, todos ellos sin antecedentes de patologías psiquiátricas y con un nivel de comprensión equivalente al nivel de 8 años de escolarización. La muestra estaba formada por dos grupos: 1) *grupo clínico* (n= 143), compuesto de forma incidental por varones (72.7%) y mujeres (27.3%) que habían sufrido episodios agudos isquémicos cardiovasculares, con edades comprendidas entre 31 y 80 años (*M*= 55.95; *DS*= 10.53); y 2) *grupo control* (n= 170), constituido por varones (66.5%) y mujeres (33.5%) cuyas edades variaban entre 31 y 72 años (*M*= 55.40; *DS*= 9.40). Los participantes asignados al grupo control, que fueron seleccionados de forma intencional, no debían tener antecedente alguno de enfermedad cardiovascular y fueron equiparados con respecto a los participantes del grupo clínico atendiendo a la edad, el género, el estado civil, la posición socioeconómica y el nivel de formación.

Instrumentos

La información sociodemográfica se obtuvo a través de una entrevista estructurada con el único objeto de formar un grupo de control comparable al grupo clínico para las posibles variables de confundido que han sido mencionadas (edad, género, estado civil,

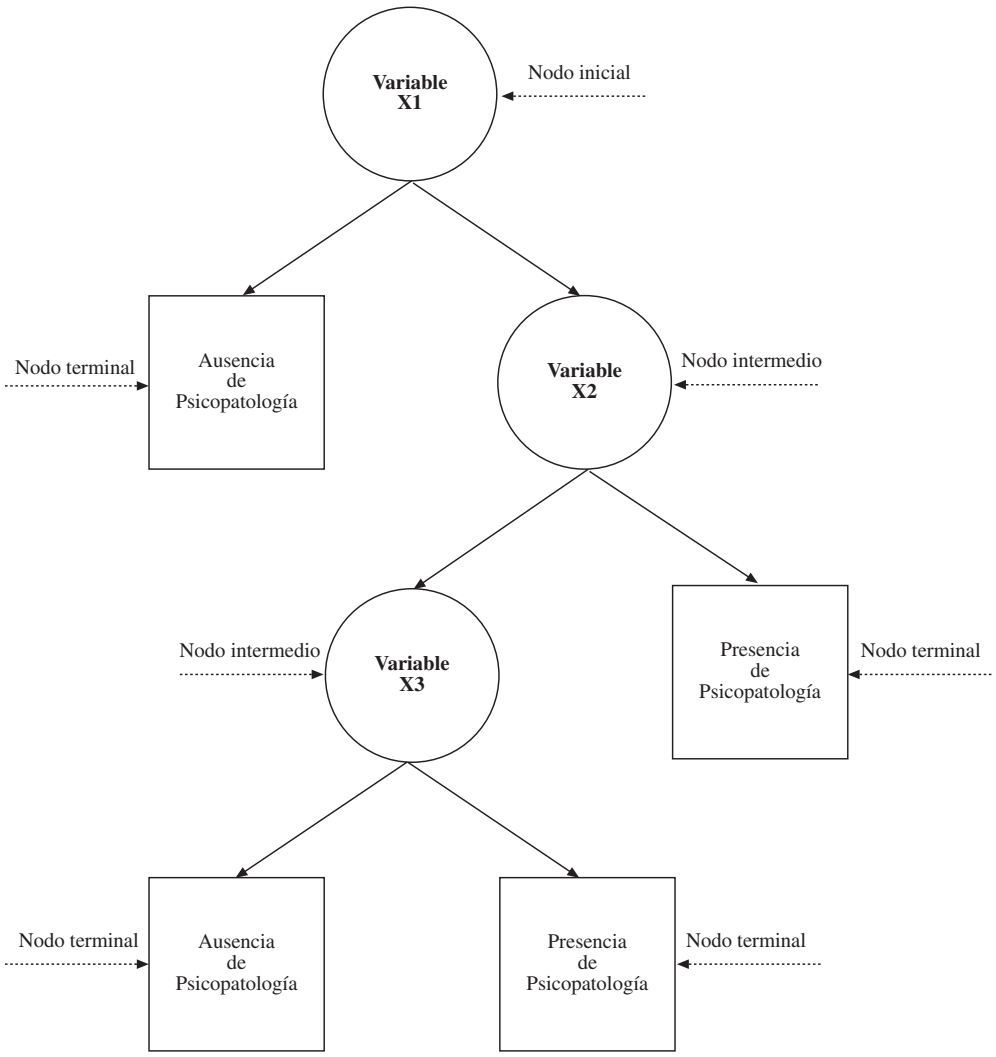

Figura 1. *Representación gráfica de un árbol de clasificación. A partir del nodo inicial, definido por la variable que mejor diferencia o discrimina entre los participantes que padecen y los que no sufren una determinada psicopatología, se identifican sucesivas particiones o ramas de los datos de la muestra. Los nodos intermedios están asociados con otras variables de clasificación, además de la inicial, que resultan discriminantes entre ambos grupos de individuos. Los nodos terminales expresan el grupo de pertenencia de los individuos de la muestra. Nótese que, a partir de las diferentes ramas y los criterios de división de éstas, se pueden establecer los perfiles para diferenciar entre aquellos participantes que han sido asignados al grupo de personas que padece una determinada psicopatología y los individuos que no la sufren. Es decir, se pueden extraer un conjunto de reglas de inducción que pueden ser utilizadas, siempre que éstas resulten suficientemente precisas, para, en una aplicación futura, reconocer si una persona se halla o no en un grupo de riesgo para padecer una determinada psicopatología. A cada regla de inducción se le asocia un valor de propensión, es decir, una estimación empírica de la precisión de la regla*

posición socioeconómica y nivel de formación). La presencia y clasificación de rasgos y trastornos de la personalidad fue evaluada con el *MCMI-II*, segunda versión y adaptación española de *TEA* (Millon, 1999), por considerarlo un instrumento que permite establecer, con precisión y economía, un continuo entre aspectos funcionales y disfuncionales de la personalidad. Una de las críticas más comunes de la clasificación de los trastornos de personalidad que aparece en el *DSM-IV* (APA, 1995) es la ausencia de un modelo teórico unificado. En este sentido, la propuesta de Millon (1999) solventa este problema. El *MCMI-II* consta de 175 reactivos con un formato binario de respuesta (Verdadero/Falso) y se administra en un tiempo relativamente breve (entre 20 y 40 minutos). Las respuestas dicotómicas son posteriormente transformadas, obteniéndose las puntuaciones ponderadas (PP) para cada una de las escalas, realizándose un ajuste final de las puntuaciones de acuerdo a los criterios establecidos por Millon (1999). En el presente estudio se utilizaron las siguientes 22 escalas del *MCMI-II*: 10 escalas básicas de personalidad (Esquizoide, Fóbica, Dependiente, Histriónica, Narcisista, Antisocial, Agresivo-sádica, Compulsiva, Pasivo-agresiva y Autodestructiva; 3 escalas de personalidad patológica (Esquizotípica, Límite y Paranoide); 6 síndromes clínicos de gravedad moderada (Ansiedad, Histeriforme, Hipomanía, Distimia, Abuso de alcohol y Abuso de drogas); y 3 síndromes clínicos de gravedad severa (Pensamiento psicótico, Depresión mayor y Trastorno delirante).

Procedimiento

Todos los participantes fueron evaluados individualmente. En cuanto al grupo clínico, los participantes se evaluaron, siempre y cuando las condiciones lo permitieran, durante los dos días posteriores a su ingreso en la unidad coronaria del hospital ya mencionado. Con cada participante se explicitó y acordó el propósito del trabajo, así como la confidencialidad de las respuestas. Para controlar la posible incidencia de los distintos niveles educativos, los reactivos del instrumento fueron leídos por los administradores. A cada participante se le solicitó que emitiera su respuesta (Verdadero/Falso) para cada unos de los 175 reactivos del *MCMI-II*. Se les indicó que se les iban a leer una serie de frases usuales que las personas suelen utilizar para describirse a sí mismas, solicitándoles que fueran lo más sinceros posibles, que respondieran a todas las frases a pesar de que no estuvieran totalmente seguros y que no había límite de tiempo para responder, aunque lo mejor es que lo hicieran con rapidez. Seguidamente se les leyeron dos ejemplos del cuadernillo de administración para asegurarse de que comprendieran correctamente la consigna. El mismo procedimiento se utilizó con los participantes del grupo control, pero procurándose obtener dos grupos homogéneos, para lo cual se emparejaron las variables referidas al género, la edad, el nivel educativo, la posición socioeconómica y el estado civil.

Análisis de datos

Una vez obtenidos los datos de la muestra se eliminaron los casos atípicos a través de métodos gráficos exploratorios de visualización con el Programa *ViSta* (Ledesma, Molina, Young y Valero-Mora, 2007; Young, Ledesma, Molina, Valero y Llorens, 2001) y se calcularon las puntuaciones directas (PD) de todas las escalas, excepto para «V» (Validez) y «X» (Sinceridad). Se realizaron los ajustes correspondientes y, finalmente, se calcularon las Tasas Base (TB). Estas puntuaciones *TB* se obtienen, mediante el uso de puntuaciones de corte estimadas para optimizar las clasificaciones diagnósticas correctas (maximizando los verdaderos positivos y minimizando los falsos positivos), a partir de datos conocidos de prevalencia para los diferentes trastornos. El número de verdaderos positivos es la cantidad de personas que, padeciendo éstas realmente una patología, son detectadas por un protocolo de diagnóstico y, por tanto, son aciertos del sistema clasificador. En cuanto al número de falsos positivos, éste se corresponde con errores del sistema de clasificación, pues, no sufriendo las personas la patología, éstas son clasificadas como enfermas al utilizar el protocolo de diagnóstico. Si se considera ahora sólo aquellas personas que realmente no padecen la patología, se definen de forma análoga los falsos negativos, que son errores del sistema clasificador, y los verdaderos negativos, que son aciertos del sistema de clasificación. La cantidad de verdaderos y falsos positivos, junto al número de verdaderos y falsos negativos, permiten obtener los valores para los coeficientes de sensibilidad y especificidad.

Se optó por el uso de las puntuaciones directas (PD) de las escalas del *MCMI-II*.

Las razones de esta elección pueden resumirse en algunos puntos clave: la escasez de datos de prevalencia regionales, el hecho de que las puntuaciones *TB* se hallan vinculadas a la prevalencia de características en la población psiquiátrica, el impacto clínico que posibilita comparar las puntuaciones entre las mismas escalas para los diferentes grupos (clínico y control) en función de una única escala de medida, las limitaciones de la muestra utilizada (tamaño moderado y mejorable representatividad) y la existencia de una población de referencia en lugar de los baremos (grupo control).

Se realizó un análisis de datos mediante la RL con el objetivo de, a partir de sus puntuaciones directas en las distintas escalas de personalidad, estimar la probabilidad de pertenencia de cada participante a cada uno de los grupos (clínico y control). El método de selección algorítmica de modelos utilizado fue *Paso a paso hacia atrás*. Se ha seleccionado dicho método de introducción de las variables por resultar más robusto que el método *Paso a paso hacia adelante,* ya que una vez eliminada una variable no significativa ésta nunca vuelve a aparecer en la ecuación, ni tampoco ningún modelo alternativo que la contenga (Montero, 2004). También se ejecutó un AD con el objetivo de utilizar los valores observados de las variables independientes para realizar las clasificaciones. Fue preciso estimar la matriz de variancias-covariancias de forma separada en el AD, pues se rechazó la hipótesis nula de identidad de las matrices de variancias-covariancias para ambas poblaciones de referencia. El método de selección de variables se llevó a cabo mediante el mismo procedimiento utilizado en la RL.

Finalmente, se utilizó la técnica de AC para asignar los participantes a los grupos en función de las variables de personalidad. Con la finalidad de generar un árbol de decisión para clasificar a los participantes, se llevó a cabo un análisis mediante el algoritmo C5.0, incluido en la aplicación informática *Clementine V6.0* (SPSS, 2001). El conjunto de las variables de entrada para el sistema de clasificación estaba compuesto por las PD de las 22 escalas clínicas anteriormente mencionadas. La variable de salida era haber padecido o no un episodio isquémico cardiovascular agudo. Se estableció que ningún nodo terminal podía tener un número de individuos inferior al 3% del total de casos de la muestra, quedando el número mínimo para cada nodo igual a 9. Mencionamos que

para elaborar los árboles de clasificación utilizamos el algoritmo C5.0 (Kantarazic, 2003), que no requiere supuestos sobre las variables aleatorias y, por tanto, no se fundamenta en la inferencia estadística. Se trata de un algoritmo en el cual las distintas divisiones se realizan identificando en cada paso las variables de clasificación que minimizan la función de entropía, es decir, permite obtener la máxima información. En otros términos, el algoritmo C5.0 opera a partir de la reducción de la incertidumbre o, si se prefiere, identificando las variables de clasificación que discriminan mejor entre los individuos que poseen y los que no poseen un determinado atributo.

Una vez obtenido un árbol de clasificación existe la posibilidad de realizar la denominada *poda*. Los métodos de poda pretenden alcanzar árboles de clasificación en los cuales no existan más particiones que las necesarias para lograr un nivel adecuado de correctos clasificados. No se realizó poda alguna por distintos motivos. Primero, se requiere llevar a cabo una validación cruzada en el proceso de poda (Picón, Varela y Lévy, 2004) y no se disponía de muestra suficiente para utilizar una parte de ésta para extraer el árbol de clasificación, mientras la otra se utilizaba como conjunto de datos para la prueba. Segundo, se desaconseja llevar a cabo una poda manual, una vez obtenido el árbol de decisión (Picón, Varela y Lévy, 2004). Tercero, la poda se realiza, en general, tras obtener una gran cantidad de nodos terminales (Breiman, Friedman, Olshen y Stone, 1998), situación que se evitó en el análisis realizado al especificar un valor suficientemente restrictivo para el criterio de parada del algoritmo. Cuarto, no es frecuente realizar un proceso similar a una poda cuando se utiliza RL y AD, razón por la cual parecía preferible mantener unas condiciones equivalentes en los tres tipos de análisis realizados. Quinto, resultaban de difícil justificación las razones para determinar el número mínimo y relevante de perfiles clínicos necesarios para obtener un compromiso entre complejidad de la solución, significación clínica y tasa de correctos clasificados.

Existen otros algoritmos que posibilitan obtener árboles de clasificación (véase Picón, Varela y Lévy, 2004; Román y Lévy, 2003). Algunos de estos algoritmos se descartaron porque sólo permiten trabajar con variables de clasificación categóricas, condición que no cumplían las escalas de medida utilizadas en el presente estudio. De haberse utilizado, se hubiera requerido obtener arbitrariamente valores discretos para las variables de clasificación. Otros algoritmos utilizan técnicas de inferencia estadística para obtener las divisiones y tampoco fueron considerados porque implican aceptar determinados supuestos sobre las variables aleatorias. Al respecto se consideró que la distribución normal no se halla habitualmente para las escalas clínicas e, incluso, en el ámbito de las consideradas medidas psicológicas (Micceri, 1989). Además, el *t*-test para datos independientes, y por ende el análisis de la variancia para un factor con dos categorías y grupos al azar, no resulta robusto para variables psicométricas con distribución marcadamente asimétrica, siempre que los grupos no tengan idéntico tamaño o el tamaño de la muestra sea suficientemente elevado (Sawilowsky y Blair, 1992). Adviértase que el problema de la desigualdad de los tamaños de los grupos puede ir acentuándose en el proceso de particiones sucesivas. A partir de los resultados de las dos investigaciones mencionadas se decidió no utilizar ningún algoritmo que utilizara una prueba estadística, como las que se fundamentan en el estadístico *F*, donde incluso existe el problema adicional de la heterogeneidad de las variancias de las poblaciones.

También podría haberse utilizado algún algoritmo propio de las redes neuronales artificiales (véase Picón, Varela y Lévy, 2004) para llevar a cabo el análisis numérico en lugar de obtener árboles de clasificación. Ahora bien, si bien las redes neuronales artificiales nos hubieran permitido conocer la contribución o peso de cada variable de clasificación en la solución final, es decir, en la asignación de los participantes a uno u otro grupo, éstas no permiten determinar los perfiles en la forma simple e inteligible que facilitan los AC. Debido a que el presente estudio está dirigido a mostrar a los psicólogos clínicos la utilidad de un sistema de asignación de los individuos, que además permite establecer los perfiles asociados a las psicopatologías y se caracteriza por su fácil comunicabilidad e interpretación clínica, se decidió utilizar árboles de clasificación en lugar de redes neuronales artificiales. En cualquier caso, debemos remarcar que, si el objetivo del estudio hubiera sido determinar las contribuciones de cada variable de clasificación, las redes neuronales artificiales hubieran sido utilizadas para analizar los datos.

De acuerdo con Pohar et al. (2004), se utilizó, además del criterio de la tasa de aciertos, el índice *C* de discriminación entre poblaciones (Harrell y Lee, 1985), pues se complementa la referida tasa con un índice de precisión. El índice *C* es una medida cuyo valor 1 indica una discriminación perfecta entre las poblaciones, mientras el valor 0.5 se corresponde con la precisión esperada de la clasificación que se obtendría mediante un sistema de asignación aleatorio.

Resultados

La tabla 1 muestra los resultados para la clasificación obtenida mediante la RL. Por una parte, el modelo más parsimonioso (Paso 12) incluye, en orden decreciente, las siguientes escalas del *MCMI-II*: *Pensamiento psicótico*, *Histriónica*, *Agresivo-sádica*, *Paranoide*, *Dependiente*, *Evitativa*, *Antisocial*, *Ansiedad* y *Abuso de drogas*. Por otra parte, en el Paso 12 se alcanza el 76% global de correctos clasificados, existiendo un 82.4% de verdaderos negativos para el grupo control y un 68.5% de verdaderos positivos para el grupo clínico. Respecto al índice de discriminación *C*, éste toma un valor igual a 0.7.

En cuanto al AD, mediante el estadístico *M* de *Box* se rechazó la hipótesis nula de que las matrices de variancia-covariancia son idénticas en ambas poblaciones, por lo que se procedió a realizar el AD con estimación separada de la matriz de variancias-covariancias. La *Lambda (λ) de Wilks* tomó un valor igual a 0.595 (χ^2= 159.469; p<0.001), resultado que muestra que las puntuaciones

Tabla 1
Tabla de clasificación para la regresión logística binaria, método de selección por pasos hacia atrás. Clasificados correctamente el 76,0% de los participantes

Valores observados			Valores de predicción		
			Grupo		
			Control	Clínico	Porcentaje correctos clasificados
Paso 12	Grupo	Control	140	30	82.4
		Clínico	45	98	68.5
	% total				76.0

medias de las escalas de personalidad seleccionadas en el último paso del AD (*Distimia*, *Dependiente*, *Narcisista*, *Histeriforme*, *Pasivo-agresiva*, *Ansiedad* y *Abuso de drogas*) presentan diferencias significativas entre los grupos clínico y control. La función discriminante canónica da cuenta de una apreciable parte de la variabilidad inicial, pues presenta un autovalor de 0.680 y una correlación canónica de 0.636. La función discriminante muestra, en los centroides de los grupos, una clara discriminación (-0.754 para el grupo control y 0.896 para el grupo clínico). En la tabla 2 se presentan los coeficientes estandarizados de la función discriminante canónica, los coeficientes de estructura, los estadísticos *lambda de Wilks* y los valores del estadístico *F* con sus correspondientes niveles de significación.

En la tabla 3 se muestran los resultados de la clasificación obtenida con el AD para la estimación separada de la matriz de variancias-covariancias. El 78.9% de los casos totales fueron correctamente clasificados, existiendo un 77.6% de verdaderos negativos y un 80.4% de verdaderos positivos para los grupos control y clínico, respectivamente. El índice de discriminación *C* es aproximadamente igual a 0.87, muy superior al obtenido para la RL.

A partir de los resultados alcanzados con la técnica de AC se observa que, del total de las 22 escalas de personalidad que fueron inicialmente consideradas, sólo 9 de aquellas se encuentran asociadas con los perfiles de riesgo de EICA, a saber: *Distimia, Abuso de alcohol, Dependiente, Ansiedad, Pensamiento psicótico, Compulsiva, Agresivo-sádica, Pasivo-agresiva* y *Abuso de drogas.* Aunque un análisis detallado se expone en un estudio reciente (Richard's y Solanas, en prensa), en la figura 2 se muestra el árbol de clasificación obtenido en el presente estudio. A partir de la información disponible en la tabla 4 pueden obtenerse distintos indicadores sobre la precisión del sistema clasificador, como son la sensibilidad y la especificidad, entre otros. Los valores que toman la sensibilidad y la especificidad de este último sistema clasificador son 0,895 y 0,858, respectivamente. Estos últimos valores, que son superiores a los obtenidos mediante las técnicas RL y AD, nos indican, en términos relativos, que se espera una menor cantidad de falsos negativos frente a los falsos positivos, lo que, desde un punto de vista aplicado, es un resultado preferible que el contrario. Puede apreciarse que, para el conjunto de individuos, la fiabilidad del sistema de clasificación no difiere en exceso dependiendo del grupo, existiendo una diferencia en torno al 4% a favor del grupo clínico. En cuanto al porcentaje global de participantes correctamente clasificados, se obtuvo un valor igual a 87.5%, un valor apreciablemente superior al logrado con las otras dos técnicas. Respecto del índice de discriminación *C*, éste toma un valor aproximadamente igual a 0.88, marcadamente superior que el obtenido para la RL y escasamente mayor que el alcanzado en el AD.

Tabla 2
Análisis discriminante, mediante el método de selección por pasos hacia atrás, con estimación separada de la matriz de variancias-covariancias para la identificación de escalas de la personalidad asociadas a la probabilidad de padecimiento de EICA

Predictores escalas MCMI-II	Coeficientes estandarizados	Coeficientes estructura	*Lambda* de *Wilks*	Estadístico *F* y nivel de significación
Distimia	0.662	0.723	0.636	20.692 p<0.001
Dependiente	0.508	0.464	0.649	27.654 p<0.001
Narcisista	0.394	0.239	0.616	10.719 p<0.001
Histeriforme	0.592	0.231	0.641	23.480 p<0.001
Pasivo-agresiva	-0.594	0.173	0.632	18.855 p<0.001
Ansiedad	0.481	0.681	0.612	8.474 p<0.001
Abuso de drogas	0.332	0.311	0.606	5.264 p<0.001

Tabla 3
Resultados de la clasificación del análisis discriminante con estimación separada de la matriz de variancias-covariancias. Clasificados correctamente el 78,9% de los participantes

		Grupo de clasificación		
		Clínico	*Control*	Total
Realidad	*Grupo clínico*	115 (80.4%)*	28 (19.6%)	143
	Grupo control	38 (22.4%)	132 (77.6%)*	170

* Los porcentajes de la diagonal principal corresponden a la sensibilidad y la especificidad

Discusión y conclusiones

El objetivo principal de este estudio consistió en evaluar y comparar tres técnicas estadísticas de clasificación —regresión logística, análisis discriminante y árboles de clasificación— para identificar las características de personalidad asociadas al riesgo de padecer EICA. Además de mejorar los resultados obtenidos mediante dos técnicas de clasificación, RL y AD, la técnica de AC nos posibilitó obtener información adicional sobre las asociaciones entre los perfiles de personalidad y el aumento de la probabilidad de padecer EICA.

Los resultados muestran que los AC conducen a mayores tasas de correctos clasificados, ya sea considerando los porcentajes globales o los correspondientes a la sensibilidad y la especificidad. Además, el AC obtenido destaca por los elevados porcentajes de correctos clasificados alcanzados, tanto para el grupo clínico como para el grupo control. Estos resultados a favor de los AC también se han hallado mediante el índice *C*, que es una cuantificación de la discriminación.

En cuanto a las variables de personalidad asociadas al aumento de la probabilidad de padecer EICA, las escalas *Ansiedad*,

Tabla 4
Tabla de confusión de la precisión global del sistema clasificador o conjunto de reglas de decisión para el conjunto de individuos de la muestra, mediante la técnica de árboles de clasificación. Clasificados correctamente el 87,5% de los participantes

		Grupo de clasificación		
		Clínico	*Control*	Total
Realidad	*Grupo clínico*	128 (89.51%)*	15 (10.49%)	143
	Grupo control	24 (14.12%)	146 (85.88%)*	170

* Los porcentajes de la diagonal principal corresponden a la sensibilidad y la especificidad

Dependiente y *Abuso de drogas* aparecen en las tres técnicas utilizadas, aportando evidencia empírica a favor de la asociación con las dimensiones de la personalidad Tipo D (Denollet et al., 1996; Denollet, Vaes, Dirk y Brutsaert, 2000). La razón por la cual no se obtuvieron en la solución final las mismas variables en las tres técnicas se debe a los algoritmos propios de cada procedimiento.

La técnica de AC es un enfoque más flexible y menos restrictivo que la RL y el AD, puesto que es un método exploratorio que busca patrones sin requerir supuestos sobre la distribución de los datos. Además, los AC utilizan algoritmos que posibilitan detectar relaciones no lineales, siendo este hecho una de las posibles razones de que, en este estudio aplicado, se hayan encontrado mejores tasas de clasificación para el AC. Ahora bien, también debe men-

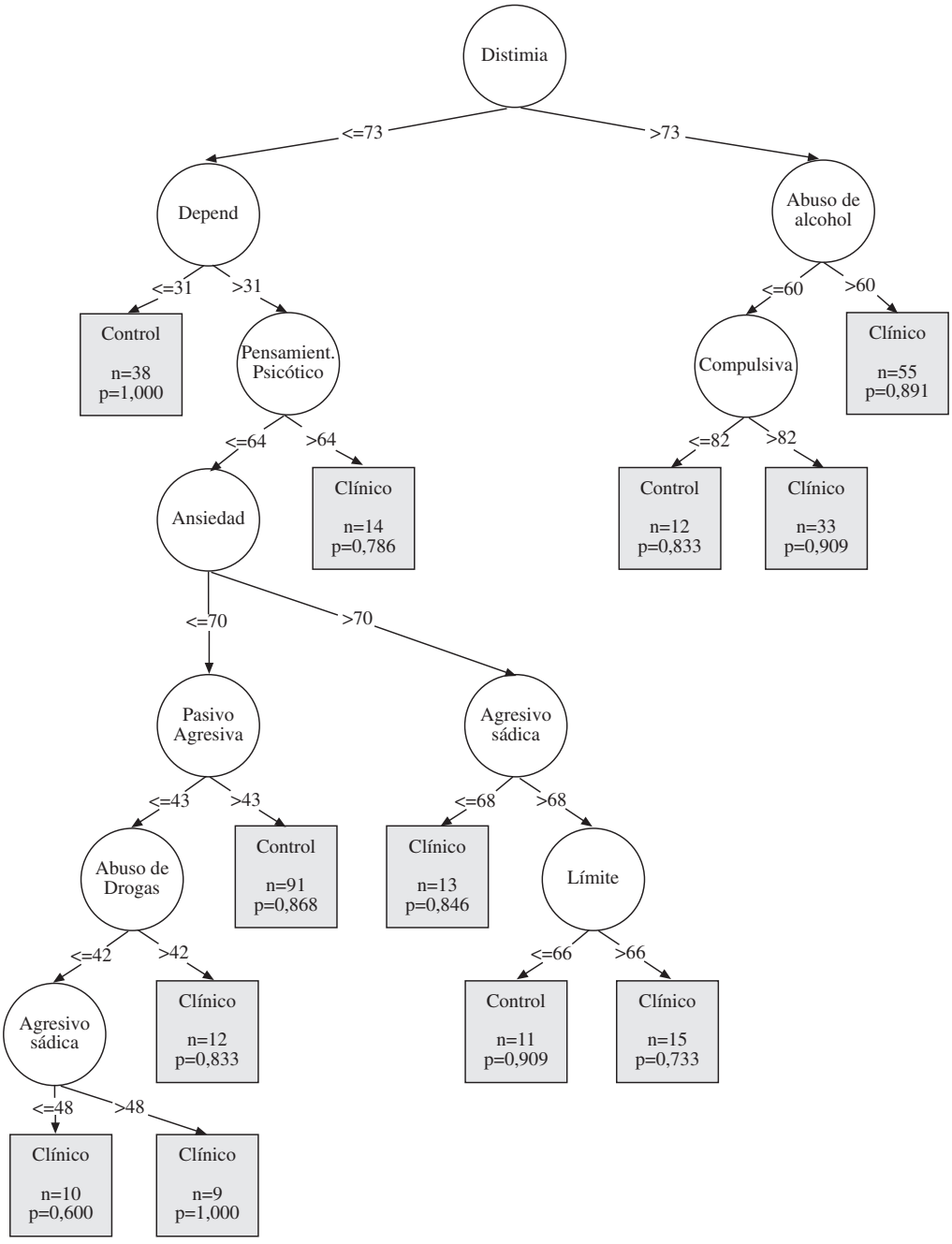

Figura 2. *El árbol de clasificación muestra los nodos terminales, representados mediante cuadrados sombreados, y los nodos intermedios y el nodo inicial, simbolizados por círculos. En cada nodo terminal se indica el valor de propensión, p, para la regla inducida y el número de individuos de la muestra, n, que cumplían cada una de las reglas. Desde un punto de vista clínico, aquellas reglas inducidas con un mayor valor de propensión y un elevado número de individuos resultan especialmente relevantes, aunque eso no implica que otros perfiles menos frecuentes no tengan interés alguno. Los valores de propensión pueden entenderse como una cuantificación de la precisión de la regla inducida*

cionarse una limitación del AC utilizado en la presente investigación, pues se trata de una técnica descriptiva y no posibilita realizar inferencias estadísticas. Por otro lado, una ventaja comparativa del AC que se ha utilizado es que no requiere supuestos, como ocurre en el RL y AD, y, por tanto, resulta de mayor aplicabilidad en el ámbito clínico.

Otra ventaja de los AC es que las reglas se expresan de una forma inteligible, lo cual facilita su comunicabilidad y utilización en el ámbito aplicado. Desde un punto de vista teórico, la técnica de AC permitió diferenciar cinco perfiles distintivos del grupo control, siendo éstos coincidentes con los resultados de los análisis de RL y AD, ya que también se observa el peso de las puntuaciones elevadas en la escala *Pasivo-agresiva* y de las puntuaciones bajas en la escala *Límite*; además, permitió identificar siete perfiles de personalidad vinculados al aumento de la probabilidad de padecer EICA (Richard's y Solanas, en prensa).

Entendemos que las conclusiones propuestas en este estudio deben interpretarse con precaución. El tipo de diseño y la naturaleza de la muestra clínica que se han utilizado implican algunas restricciones para la generalización y la aplicación de los resultados. Las investigaciones futuras podrían considerar muestras mayores y diferenciadas por tipo de síndromes coronarios. Además, sería importante implementar estudios nacionales con el objeto de extender el alcance de los resultados y de las técnicas. Así, al disponer de una muestra mayor, podría hacerse una validación cruzada. Cabe añadir que, para realizar una comparación sistemática entre las tres técnicas, debe llevarse a cabo un estudio de simulación. Aunque nuestro estudio no pretendía tratar ese objetivo, es conveniente tener en consideración esta última observación para establecer claramente las limitaciones del presente trabajo y, así, evitar posibles generalizaciones que no están fundamentadas en los resultados aquí presentados.

En resumen, el presente trabajo muestra que los AC podrían ser de utilidad en el campo aplicado, tanto por las tasas de aciertos como por el índice de discriminación obtenidos, además de la comunicabilidad de sus resultados y la flexibilidad de los requisitos para su utilización.

Agradecimientos

Los autores agradecen los comentarios y sugerencias realizadas por dos revisores anónimos sobre una primera versión del manuscrito. Sin duda, sus observaciones permitieron mejorar el manuscrito inicial. La investigación incluida en este trabajo ha sido financiada parcialmente por el Consejo Nacional de Investigaciones Científicas y Técnicas de la República Argentina (CONICET).

Referencias

Abbate-Daga, G., Amianto, F., Rogna, L., y Fassino, S. (2007). Do anorectic men share personality traits with opiate dependent men?: A case-control study. *Addictive Behaviors, 32*, 170-174.

American Heart Association Journal Report (1996). *Being, chronically "blue" raises risk of heart attack, all cause mortality. NR-96-4416,* (Circ/Barefoot).

APA (1995). American Psychiatric Association. *Manual diagnóstico y estadístico de los trastornos mentales. DSM-IV.* Barcelona: Masson (orig. 1994).

Berry, M.J.A. (1997). *Data mining techniques for marketing, sales and customer support*. New Cork: John Wiley & Sons.

Breiman, L., Friedman, J.H., Olshen, R.A., y Stone, C.J. (1998). *Classification and regression trees*. Boca Raton: Chapman & Hall.

Buceta, J.M., y Bueno, A.M. (1996). *Tratamiento psicológico de hábitos y enfermedades*. Madrid: Pirámide.

Denollet, J., Sys, S.U., Stroobant, N., Rombouts, H., Gillebert, T.C., y Brutsaert, D.L. (1996). Personality as independent predictor of long-term mortality in patients with coronary heart disease. *The Lancet, 347,* 417-421.

Denollet, J., Vaes, J., Dirk, L., y Brutsaert, D.L. (2000). Inadequate response to treatment in coronary heart disease: Adverse effects of type D personality and younger age on 5-year prognosis and quality of life. *Circulation, 102*, 630-635.

Freeman, M.S., Hayes, B.G., Kuch, T.H., y Taub, G. (2007). Personality: A predictor of theoretical orientation of students enrolled in a counseling theories course. *Counselor Education and Supervision, 46*, 254-265.

Hair, J.F., Anderson, R., Tatham, R.L., y Black, W.C. (1999). *Análisis multivariante.* Madrid: Prentice Hall.

Harrell, F.E., y Lee, K.L. (1985). A comparison of the discrimination of discriminant analysis and logistic regresión under multivariate normality. En P.K. Sen (Ed.): *Biostatistics in biomedical: Public Health and Enviromental Sciences* (pp. 333-343). North-Holland: Elsevier Science Publishers.

Kantarazic, M. (2003). *Data mining. Concepts, models, methods and algorithms.* New York: John Wiley & Sons.

Ledesma, R., Molina, G., Young, F.W., y Valero-Mora, P. (2007). Desarrollo de técnicas de visualización multiple en el programa ViSta: ejemplo de aplicación al análisis de componentes principales. *Psicothema*, *19*, 497-505.

Marso, S., Griffin, B., y Topol, E. (2002). *Cardiología*. Madrid: Marban.

McGee, D.L., Reed, D., y Yano, K. (1984). The results of logistic analysis when the variables are highly correlated. *American Journal of Epidemiology, 37*, 713-719.

Micceri, T. (1989). The unicorn, the normal curve and other improbable criatures. *Psychological Bulletin*, *105*, 156-166.

Millon, T. (1999). *Inventario Clínico Multiaxial de Millon-II [MCMI-II]* Madrid: TEA Publicaciones de Psicología Aplicada.

Millon, T., y Davis, R. (2000). *Personality disorders in modern life.* New York: Wiley & Sons.

Montero, L. (2004). *Open course ware*. Universidad Politécnica de Catalunya. Recuperado el 13/02/08, de Internet: http://emd.upc.edu/gestor/index.php?idcentre=270&id_assig=270-6-28008&idtit=6&propia=yes

Picón, E., y Varela, J. (2000). Segmentando mercados con análisis conjunto. Una aplicación al sector turístico. *Psicothema, 12*, 453-458.

Picón, E., Varela, J., y Lévy, J.P. (2004). *Segmentación de mercados. Aspectos estratégicos y metodológicos.* Madrid: Pearson Educación, S.A.

Pohar, M., Blas, M., y Turk, S. (2004). Comparison of logistic regresion and linear discriminant analysis: A simulation study. *Metodoloskizvezki, 1*, 143-161.

Rice, K.G., y Ashby, J.S. (2007). An efficient method for classifying perfectionists. *Journal of Counseling Psychology, 54*, 72-85.

Richard's, M.M., y Solanas, A. (en prensa). Millon's Personality Model and ischemic cardiac-vascular acute episodes: Profiles of risk in a decision tree. *International Journal of Clinical and Health Psychology.*

Román, M.V., y Lévy, J.P. (2003). Clasificación y segmentación jerárquica. En J.P. Lévy y J. Varela (Eds.): *Análisis multivariable para las ciencias sociales* (pp. 567-630). Madrid: Pearson-Prentice Hall.

Sawilowsky, S.S., y Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, *111*, 352-360.

Silva, L.C., y Barroso Ultra, I.M. (2001). Selección algorítmica de modelos en las aplicaciones biomédicas de la regresión múltiple. *Medicina Clínica, 116*, 741-745.

SPSS Inc. (2001). *Clementine 6.0 User's Guide*. Chicago, IL: Author.

Welin, C., Lappas, G., y Wilhelmsen, L. (2000). Independent importance of psychosocial factors for prognosis after myocardial infarction. *Journal Internal of Medicine, 247*, 629-639.

Worth, A.P., y Cronin, M.T.D. (2003). The use of discriminant analysis, logistic regresión and classification tree analysis in the development of classification models for human health effects. *Theochem, 622*, 97-111.

Young, F., Ledesma, R., Molina, G., Valero, P., y Llorens, A. (2001). ViSta «The Visual Statistics System». *Metodología de Encuestas, 3*, 127-133.

Yusuf, S., Hawken, S.P., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., y Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (The interheart study): Case-control study. *The Lancet*, *364*, 937-952.

Zellweger, M.J., Osterwalder, R.H., Langewitz, W., y Pfisterer, M.E. (2004). Coronary artery disease and depression. *European Heart Journal, 25*, 3-9.