FACULTAD DE MEDICINA
DEPARTAMENTO DE MEDICINA

UN MODELO MULTIVARIANTE EN EL CARCINOMA DE COLON ESPORÁDICO: SU UTILIZACIÓN COMO INSTRUMENTO DE ESTUDIO DEL SESGO DE BERKSON EN LOS CENTROS DE ATENCION PRIMARIA.

Doctorando: José María Villadiego Sánchez

Director: Manuel Ortega Calvo.

- b1763141 5
- 011528394
: 24121666

Anexo 15

AUTORIZACIÓN DE LA PRESENTACIÓN DE LA TESIS DOCTORAL

El Departamento de **MEDICINA**

con fecha **15-05-2006**, utilizando el procedimiento acordado por su Consejo a los efectos de garantizar la calidad de la Tesis Doctoral(1) , previo informe del Director/es de la misma(2), ha acordado proceder a su presentación.

APELLIDOS Y NOMBRE DEL AUTOR

VILLADIEGO SÁNCHEZ, JOSÉ MARÍA

PROGRAMA DE DOCTORADO CURSADO

"PEDIATRÍA"

TÍTULO DE LA TESIS

"UN MODELO MULTIVARIANTE EN EL CARCINOMA DE COLON ESPORÁDICO: SU UTILIZACIÓN COMO INSTRUMENTO DE ESTUDIO DEL SESGO DE BERKSON EN LOS CENTROS DE ATENCIÓN PRIMARÍA".

DIRECTOR/ES

DR. D. MANUEL ORTEGA CALVO

TUTOR:

DR. D. CARLOS MARTÍNEZ MANZANARES

Sevilla, 15 de mayo de 2006

UNIVERSIDAD DE SEVILLA
Departamento de Medicina
DIRECCIÓN
Prof. Dr. R. Pérez Cano

Fdo.: Ramón Pérez Cano

EXCMO. SR. PRESIDENTE DE LA COMISIÓN DE DOCTORADO.

(1) se adjunta, en su caso, la documentación que haya generado el proceso de evaluación previa de la calidad de la tesis (art. 11.4 R.D. 56/2005)

(2) se adjunta informe previo del Director/es de la Tesis Doctoral

" Vosotros sois la sal de la tierra....Vosotros sois la luz del mundo "

(Mateo 5, 13-16)

Al concluir esta tesis quiero dedicársela con mi más profundo agradecimiento a mis padres y mi hermana por todo lo que me han dado a lo largo de mi vida sin pedir nada a cambio, porque lo que soy se lo debo a ellos. De forma muy especial a mis amigos que siempre han caminado junto a mi ayudándome a levantar en los momentos difíciles. A Manolo Ortega por su cariño y perseverancia, pero sobre todo por su profunda confianza en mi. Por ultimo quiero entregarle todo este esfuerzo a Silvia que me ha estado acompañando y motivando para seguir adelante durante todo este tiempo.

Sevilla, Otoño de 2006

AGRADECIMIENTOS

Al Profesor Dr. Ramón Pérez Cano (Dpto. de Medicina) por el estímulo intelectual y humano que ha mostrado hacia mí y hacia mi director.

Al Profesor Dr. Carlos Martínez Manzanares (Dpto. de Medicina) por su labor de tutoría hacía mí y hacia mi director.

Al Dr. Manuel Ortega Calvo, Médico de Familia del Centro de Salud Esperanza Macarena (Sevilla) por su labor de Dirección de esta Tesis y por su amistad y cariño durante todos estos años. Mi especial agradecimiento por los valores humanos y profesionales que me ha transmitido.

Al Dr. Aurelio Cayuela Domínguez, Epidemiólogo Clínico de la Unidad de Documentación e Investigación de los Hospitales Virgen del Rocío de Sevilla por su labor de estímulo, magisterio y consultoría científica de esta Tesis.

Al Profesor Dr. Emilio Sánchez-Cantalejo por la claridad pedagógica con que imparte sus cursos de análisis multivariante en la Escuela Andaluza de Salud Pública (Granada).

A la Dra. Maria Isabel Fernández Fernández, Médico de Familia en el Centro de Salud de Camas por su labor de consultoría científica y estadística de esta Tesis.

Al Profesor Dr. Rafael Pino Mejías del Departamento de Estadística de la Universidad de Sevilla por su labor de análisis de datos mediante "bootstrap".

A la Dra. Maria de los Ángeles Tarilonte, Médico de Familia del Centro de Salud de Camas (Sevilla), por su inestimable colaboración en la selección y recogida de controles.

Al Dr. Victoriano Macías Pérez, Médico General, por su labor de selección y recogida de controles en el Centro de Salud Huerta del Rey (Sevilla).

A los Dres. José Luís Arias Jiménez , Oscar Aramburu Bodas y Román Cerro González, Facultativos Especialistas de Área en el Servicio de Medicina

Interna del Hospital Universitario Virgen Macarena (Profesor Pérez Cano) de Sevilla, por su labor de selección de controles hospitalarios.

A los Dres José Manuel Muriel Benítez y Angel González Manero Médicos de Familia del Centro de Salud Mérida Norte (Badajoz) .

Por su inestimable colaboración y ayuda a mi amiga la Dra. Beatriz Pascual de la Pisa Médico de Familia y responsable de la Unidad de Investigación del Distrito Sanitario Aljarafe (Sevilla)

A las Dras. Ana Marcos (Médico Residente de Neurología),Yesenia Tordecilla (Médico Becario de Medicina Interna), Prado Salamanca (Médico Residente de Medicina Interna), Rosa Romero (Médico Residente de Alergia) y Rosa Castillo (Médico Residente de Farmacología Clínica) por su labor de recogida de controles hospitalarios en la Planta 5ª B de Medicina Interna del Hospital Virgen Macarena de Sevilla (segundo semestre del año 2003).

Al Profesor Dr. Juan Polo Padillo por las facilidades que nos dio en la recogida de casos retrospectivos en el Hospital Universitario Virgen Macarena de Sevilla.

A la Dra. Martín Blanco responsable de la Unidad de Documentación Clínica del Hospital Juan Ramón Jiménez de Huelva por su labor de recogida de datos.

INDICE

Indice de **Figuras.**

Indice de **Tablas.**

1. INTRODUCCION

El curso que puede llevar una investigación es inesperado en muchas ocasiones. Esto es lo que nos ha ocurrido durante el desarrollo de esta Tesis Doctoral, que comenzó siendo un trabajo sobre el sesgo de Berkson[1] y termina siendo esencialmente un profundo ejercicio de validación interna.

La mayoría de nosotros sobreestimamos la importancia del azar en comparación con el sesgo cuando interpretamos los datos. En esencia podríamos decir que: "*Si p es menor de 0,001, un poco de sesgo no nos hace demasiado daño*" [2]. Si embargo, si los datos se reúnen con algún tipo de vicio no identificado, ninguna elegancia estadística nos salvará del ridículo.

Como manifestaba Johnson[3] *"el estudio bien diseñado y llevado a cabo cuidadosamente suele deparar resultados que son evidentes incluso sin un análisis formal, y por otra parte, si existen fallos importantes en su diseño o en la ejecución, el análisis más sofisticado no nos resultará de ninguna utilidad"*. Aunque sea verdad quizás exprese en nuestra opinión, una perspectiva algo extremista.

1.1. OBSERVACION CIENTIFICA.

Toda la materia que conforma la ciencia empírica está construida sobre las observaciones. Aunque nuestras teorías posean intuición epistemológica o seriedad matemática, sólo podrán entrar en lo que es el dominio de la ciencia estándar cuando sean probadas, contrastadas y confrontadas con la realidad externa por medio de las observaciones[4] . La observación posee por lo tanto un papel central y constitutivo dentro de la ciencia empírica y por lo tanto de la epidemiología. Pero...,¿Qué es la observación?

En la mayoría de las lenguas occidentales, es el sintagma sustantivo que corresponde al verbo ***"observar".*** Tal acción a su vez posee dos significados básicos: el de mantener o practicar una costumbre o una prohibición y el de mirar atentamente. Estos dos significados se contemplan también en el griego clásico "*teréo, téresis*", en latín "*ob-servare, ob-servatio*", en alemán "*be-ob-achten, Be-ob-achtung*" (esta última con raíz latina). A nosotros nos interesa el segundo significado. El **modo de percepción atento, deliberado y explícitamente cognitivo** que subyace en este concepto[5].

Sin profundizar en los matices personales o impersonales que puedan tener las observaciones científicas[4, 5] nosotros queremos destacar la importancia de la tecnología informática a la hora del estudio de las observaciones realizadas. Desde la detección de señales, hasta la recogida, el análisis, la selección de los datos y la interpretación de los resultados, los ordenadores actúan como extensiones artificiales de nuestras mentes y de nuestros sentidos. Esto ocurre en campos como la física de partículas de alta energía o la astronomía[4] en los que el "software" que se emplea posee muchos requisitos de automatismo.

Sin embargo, en la investigación estadística y epidemiológica , el automatismo del software es algo con lo que hay que tener cuidado. Sin ir más lejos, nos podemos encontrar con los problemas metodológicos del "stepwise" (selección mecánica de variables hacia atrás y hacia delante según criterios matemáticos preestablecidos) sobre los que siempre tiene que predominar la mente experta del investigador [6 7]. En primer lugar siempre hemos de ***"dejar hablar a los datos"*** [6 8] .

Vamos a exponer el substrato teórico de este trabajo de investigación epidemiológica y estadística que no habría podido realizarse sin el desarrollo

actual de la tecnología informática, pero que explora también las imperfecciones y los lugares hasta donde todavía ésta no llega. Una investigación que comenzando en los entresijos del sesgo de Berkson nos va a llevar al mundo virtual de la validación interna mediante el "bootstrappping".

1.2 SESGO. DEFINICION Y TIPOS.

Hace algunos años se definía el sesgo como un fenómeno que podía ocurrir en cualquiera de las etapas del estudio epidemiológico y que tendía a producir resultados o conclusiones que diferían sistemáticamente de la realidad. Se podía producir durante el diseño, la ejecución, el análisis y la interpretación, distorsionando o falseando los resultados en uno u otro sentido[9].

La definición más actual identifica el sesgo con la falta de validez interna o con la asociación incorrecta de la exposición y del efecto en una población diana[10,11]. En contraste la validación externa conlleva la capacidad de generalización de los resultados en otras poblaciones. La presencia de validez externa no asegura la presencia de la validez interna. *Desde un punto de vista conceptual, el sesgo debe de distinguirse del error aleatorio y de la falta de precisión* [10]. En otras ocasiones el sesgo se identifica con el mecanismo por el que falta validez interna en una investigación epidemiológica[10, 11].

Los sesgos pueden clasificarse según la dirección del cambio que producen en el parámetro estudiado, por ejemplo en la odds ratio (OR). El sesgo negativo o inclinado hacia el valor nulo arroja resultados más bajos de los reales y cercanos a cero para la OR. En caso contrario se producen resultados elevados y alejados del valor nulo que tampoco son los reales. El máximo en estas circunstancias es el denominado sesgo inverso (switch-over bias) que transforma una asociación positiva o negativa en su contraria[10, 12].

Se suelen distinguir tres grandes grupos de sesgos, el de selección, el de información y el de confusión, aunque existen muchos más [9, 10].

1.2.1 SESGO DE SELECCION EN ESTUDIOS DE COHORTE.

El sesgo de selección es insidioso, difícil de medir y casi inevitable a pesar de todos los medios de los que podamos disponer. Puede ocurrir como una consecuencia natural del diseño (por ejemplo con la utilización de los pacientes de una consulta externa de un hospital como base para el muestreo). También puede tener efecto durante la fase de realización del protocolo del estudio (por ejemplo mediante la violación no aleatoria del proyecto o mediante la censura no aleatoria). Finalmente puede ocurrir también durante la fase de análisis (análisis parcial de los datos de un ensayo clínico aleatorizado como si hubiera un solo grupo de tratamiento por ejemplo) [9].

A pesar de ser un tema suficientemente tratado en los libros de texto clásicos de epidemiología y estadística [9], las revistas médicas siguen aceptando trabajos que contienen un nivel de sesgo de selección que hipoteca las conclusiones .De esta forma el investigador en activo recibe un mensaje incongruente, advertencias de tipo metodológico desde el ámbito académico por una parte y permisividad de los comités de revisión de los originales en las revistas de impacto por otra.

1.2.2. DEFINICIÓN Y AMPLITUD DEL SESGO DE SELECCIÓN.

La representatividad de una muestra es un problema muy discutido[13] incluso a nivel de filosofía de la ciencia[14]. Hill define ***una muestra "seleccionada"*** como *aquella que no es representativa del universo del cual forma parte*[15]. Cuando se sale de la seguridad que proporciona la aleatorización o el análisis completo, la muestra estará conformada básicamente por la dinámica de los mecanismos de inclusión[9].

¿Podemos identificar una muestra seleccionada? A esta pregunta no se puede responder con un sí o con un no. Se puede saber el grado de selectividad. En algunos casos el grado de selección es obvio, por ejemplo cuando se trabaja con una muestra de pacientes hospitalizados. Con algunas excepciones, los pacientes hospitalizados no son representativos de la población afecta de esa enfermedad [1,16] . La llegada de un paciente a un hospital depende de muchos factores. La forma que tienen los médicos de familia de enviar sus pacientes, la auto-selección de los mismos o el nivel socio-económico entre otros.

En el lado opuesto del espectro, la representatividad de una muestra recogida en un área geográfica bien definida o de un registro completo de pacientes con una enfermedad determinada, puede ser muy alta [9, 13].

Aunque se diseñen covariables específicas en el modelo predictivo para estudiar el grado de representatividad de la muestra con respecto de la población, la medición de las diferencias es muy difícil[9]. No existen escalas cuantitativas a este respecto. Se podría sugerir que *la evaluación cualitativa del proceso de diseño* podría arrojar más luz que la medición de variables que muestren aspectos parciales del problema [9]. Incluso sería más interesante señalar la selección potencial en base a un tipo de diseño que la comparación de las características de la muestra con respecto a la población de donde procede [9].

¿ Es el sesgo de selección un factor que relacione la validez interna y la validez externa ? Rothman [17] señala que: *"La validez de una investigación se puede separar en dos componentes: la forma en que la validez de las inferencias obtenidas se adecuan a los propios sujetos del estudio (validez interna) y la forma en que las conclusiones obtenidas se adecuan a sujetos que están fuera de la investigación (validez externa o capacidad de generalización de los resultados). A partir de este esquema, la validez interna es necesariamente un prerrequisito para la validez externa* ".

Sin embargo, la validez interna no necesariamente siempre implica validez externa. Los resultados obtenidos de una muestra "seleccionada" no pueden inferirse al exterior a menos que se haga una descripción minuciosa de cómo la selección afecta a la muestra.

En el diseño de un estudio de prevalencia son muchos los factores que pueden afectar al proceso de selección: la falta de recursos económicos en los sistemas sanitarios no universales, la localización geográfica de unos servicios médicos adecuados, el nivel de severidad de la alteración clínica, problemas sociales, problemas de desempleo, creencias religiosas, etc. Esto ocurre de una forma más ostensible cuando se analiza una enfermedad no aceptada socialmente en una determinada comunidad [9] con implicaciones en la estructura antropológica.

En los estudios de la historia natural de la enfermedad existen diferencias según que las conclusiones hayan sido extraídas de una serie de casos clínicos

referenciados a una consulta o se extraigan de un análisis con base comunitaria, sobre todo en lo que se refiere a la magnitud del riesgo estimado[18].

Como conclusión a este capítulo podríamos referir que no existen estudios observacionales o experimentales perfectos en epidemiología y que *la mejor forma de evitar la selección es mediante un diseño adecuado* [9] lo cual provoca un nivel de validez interna riguroso [10, 11]. En todo caso se puede realizar un estudio previo de los medios con los que se cuenta y del nivel de sesgo de selección aceptable a la hora de la validez externa de los resultados.

1.3 EL SESGO DE BERKSON EN LOS ESTUDIOS DE CASOS Y CONTROLES

Como hemos referido antes, el sesgo de selección se produce cuando en el proceso de recogida de los participantes en el estudio operan factores que provocan una falta de representatividad en la muestra poblacional. Esta falta de representatividad puede ser de los casos, de los controles o de ambos. Entre los sesgos de selección destaca el que fuera descrito por Joseph Berkson[1] que se produce cuando los datos se obtienen entre pacientes hospitalizados en los que la combinación de enfermedad y del factor de exposición no suele estar bien representada.

El fenómeno conocido como ***sesgo de Berkson*** está basado en una serie de hechos que a pesar de que haya pasado el tiempo se siguen manteniendo casi como un dogma en los estudios epidemiológicos[19]. La existencia de este tipo de sesgo se puede demostrar de forma empírica [19, 20] y la estructura teórica del problema estadístico se puede discutir también en profundidad [21, 22].

En la publicación original de Berkson, se postuló que la frecuencia relativa de enfermedades y de factores de exposición etiológicos de una población hospitalaria, estarán sesgados cuando se compara con la población comunitaria[1]. Su ocurrencia se debe a las diferentes verosimilitudes o probabilidades de hospitalización que presentan los pacientes con diferentes enfermedades, con diferentes factores de exposición y con sus combinaciones.

Nomenclatura y símbolos

Las observaciones que vamos a exponer están referidas a personas que tengan o no tengan una de estas tres características principales:

a) la exposición a un agente etiológico sospechoso,
b) la enfermedad principal supuestamente causada por el agente
c) una condición control específica comparativa o varias condiciones.

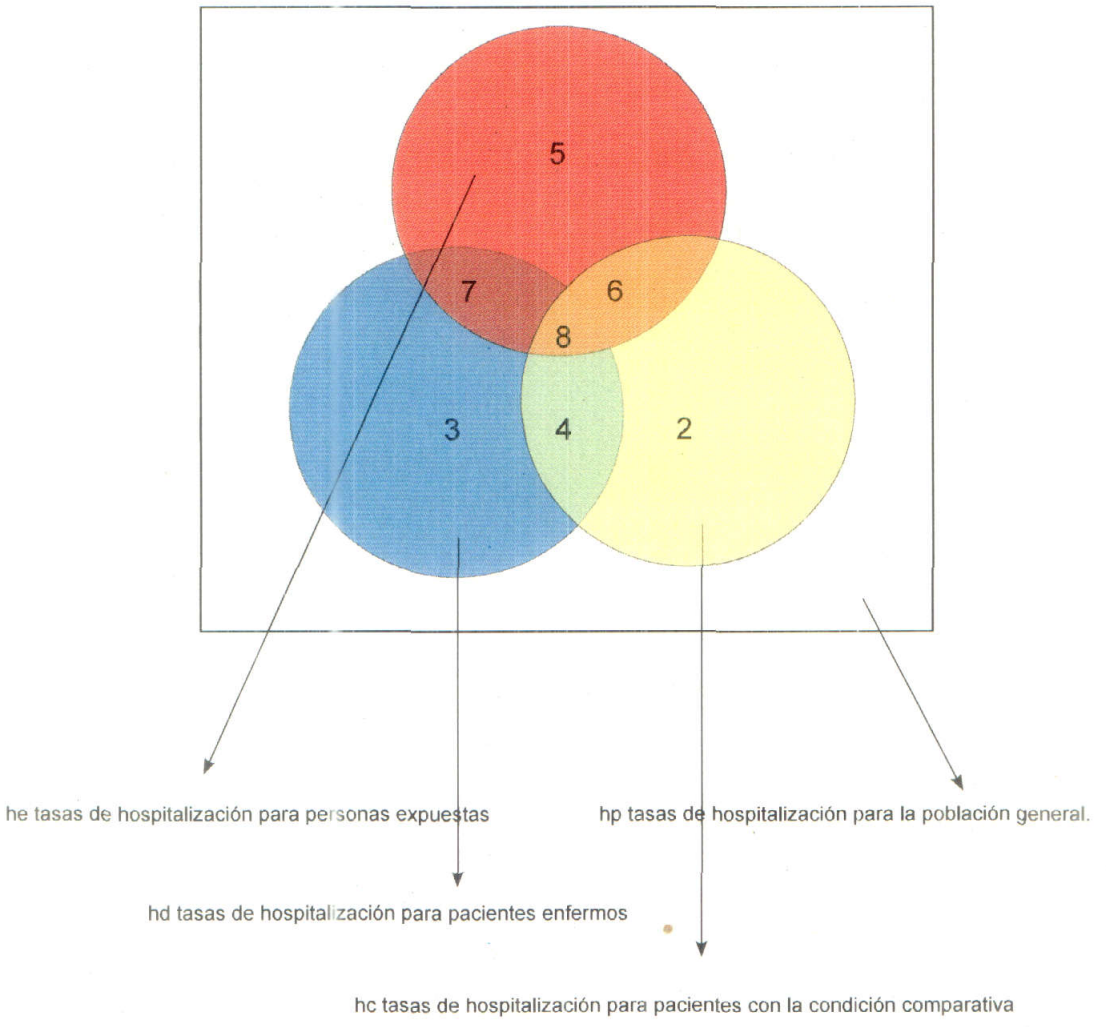
Los subgrupos formados por la superposición de estas tres características y la existencia de otras personas que no tengan ninguno de los fenómenos citados, se pueden identificar con dos conjuntos alternativos de símbolos y anotaciones.

La relación entre las tres características principales (exposición, enfermedad principal y condición de comparación) están ilustradas por un diagrama de Venn (Tabla y Figura nº 1).

Los símbolos "e" y "1-e" son utilizados para representar las proporciones de personas que han sido expuestas y no expuestas respectivamente. La enfermedad principal ocurre con una tasa p_2 de personas expuestas y p_1 en las personas no expuestas. La condición comparativa ocurre con una tasa de p_c.

Sector	Expuestos	Enfermos	Comparacion	Proporción básica	Factor hospitalizac.
1	No	No	No	$(1-e)(1-p_1)(1-p_c)$	h_p
2	No	No	Si	$(1-e)(1-p_1)p_c$	h_c,h_p
3	No	Si	No	$(1-e)p_1(1-p_c)$	h_d,h_p
4	No	Si	Si	$(1-e)p_1\ p_c$	h_c,h_d,h_p
5	Si	No	No	$e(1-p_2)(1-p_c)$	h_e,h_p
6	Si	No	Si	$e(1-p_2)p_c$	h_c,h_e,h_p
7	Si	Si	No	$ep_2(1-p_c)$	h_d,h_e,h_p
8	Si	Si	Si	$ep_2\ p_c$	h_c,h_d,h_e,h_p

Tabla y Figura nº 1.(Diagrama de Venn)



La relación entre p_c (probabilidad de la condición comparativa o de control) y las otras dos características (exposición y la enfermedad principal) es un aspecto muy importante en las asunciones del modelo aceptado clásicamente.

En algunos análisis [22,23] del sesgo de Berkson en los estudios casos y controles, el grupo de comparación puede ser escogido de cualquier manera por el investigador. El análisis está entonces en relación con la manera en la cual la odds - ratio de los casos-controles sustituye a la razón de riesgo (el riesgo relativo) de los estudios de cohorte.

Para permitir a la odds ratio de los estudios casos y controles actuar como expresión de la razón de riesgo verdadero, el grupo de comparación no debería estar esencialmente afectado por la exposición o por la falta de exposición, y debería ser también independiente de la enfermedad principal que se estudia en los grupos de casos.

De acuerdo con los símbolos citados y los conceptos de un estudio de cohorte convencional de relación etiológica, el investigador podría seguir dos grupos de personas e identificaría eventualmente cuatro grupos teniendo las proporciones siguientes:

Expuestos y enfermos; ep_2
Expuestos y no enfermos; $e\ (1 - p_2)$
No expuestos y enfermos; $(1 - e)\ p_1$
No expuestos y no enfermos $(1 - e)\ (1 - p_1)$.

El riesgo de enfermedad sería,

En los expuestos: $ep_2\ /\ [ep_2 + e\ (1 - p_2)] = p_2$
En los no expuestos $(1 - e)p_1\ /\ [(1 - e)p_1 + (1 - e)\ (1-p_1)] = p_1$

El riesgo relativo (la razón de riesgo) sería: $p_2\ /\ p_1$.

Cuando esta misma relación se examina en un estudio de tipo caso control retrospectivo, en el que alguno o todos los grupos investigados, pueden ser pacientes hospitalizados, los tamaños de los grupos originales, pueden estar alterados por diferentes tasas de hospitalización, como describió primitivamente Berkson[1]. Para proporcionar una anotación algebraica en su tratamiento, nosotros usamos cuatro símbolos adicionales:

- ✓ $\mathbf{h_e}$: tasas de hospitalización para personas expuestas;
- ✓ $\mathbf{h_d}$: tasas de hospitalización para pacientes enfermos;
- ✓ $\mathbf{h_c}$: tasas de hospitalización para pacientes con la condición comparativa
- ✓ $\mathbf{h_p}$: tasas de hospitalización para la población general, sin que se hayan tenido en cuenta ninguno de los atributos anteriormente citados (Figura nº 1).

El modelo algebraico que se va a discutir a continuación, como el que utilizó Berkson sobre este tema en su día[1], *depende de la asunción de que los factores conducentes a la hospitalización, actúan de forma independiente*. Como veremos más adelante en el texto, esta asunción pudiera no ser totalmente correcta. Por ejemplo, en el modelo desarrollado por Peritz [22] no se hace ninguna asunción sobre los factores conducentes a la hospitalización. De todas formas, la premisa de independencia se usará para permitir el desarrollo del modelo algebraico.

Las tasas de hospitalización para las personas que tienen más de un problema de los factores citados pueden ser definidas como una ecuación de álgebra de Boole:

$$P(AUB) = P(A) + P(B) - P(A \cap B)$$

Por lo tanto, los factores del ámbito general y de la enfermedad producirían ($h_p + h_d - h_p \cdot h_d$), como una tasa de hospitalización para la población enferma. Por ejemplo, si h_p es 0.04 y h_d es igual a 0.2, 40 de cada mil personas enfermas serán hospitalizadas por razones de su entorno general. De las 960 personas restantes, 192 serán hospitalizadas por la enfermedad. El número total de personas hospitalizadas serán por lo tanto 40 + 192 igual a 232, que es la tasa 0.232. Esta misma tasa podría haber sido calculada desde la fórmula siguiente como:

$$0.2 + 0.04 - 0.2 \times 0.04 = 0.24 - 0.008 = 0.232$$

Cuando más de dos factores afectan a la hospitalización, la expresión queda de esta forma:

$$h_p + h_d - h_p \cdot h_d = 1 - h_d \cdot 1 - h_p$$

Este tipo de notación para la tasa de hospitalización, está usada en la figura 1 y su leyenda asociada para identificar las proporciones de personas que debieran ser estudiadas en un diseño caso-control. La figura 1 indica las proporciones que existen para los grupos en cada sector de la población general y también las

tasas de hospitalización para cada uno de los ocho sectores.

Cálculos en los estudios de casos y controles

En los estudios retrospectivos de tipo casos y controles, la razón de riesgo verdadera está estimada a partir de la odds ratio. Si los grupos no estuvieran afectados por la hospitalización, las proporciones en la tabla tetracórica serían:

	ENFERMOS	NO ENFERMOS
EXPUESTOS	ep_2	$e(1-p_2)$
NO EXPUESTOS	$(1-e)p_1$	$(1-e)(1-p_1)$

La razón del producto cruzado es la odds ratio (OR), Con una serie de arreglos algebraicos de los términos e y (1 – e), la odds ratio sería:

$$\frac{p_2}{p_1} \cdot \frac{1-p_1}{1-p_2}$$

Como señaló Cornfield (9) por vez primera, cuando p_1 y p_2 son suficientemente pequeñas, del orden de 0.05 o menores, los valores de $1\text{-}p_1$ y $1\text{-}p_2$ y su cociente $1\text{-}p_1 / 1\text{-}p_2$ se aproximarán a la unidad. La razón de riesgo de p_2 / p_1 será por lo tanto, aproximadamente la odds ratio. En la discusión posterior, nosotros mantendremos la asunción de que $1\text{-}p_1 / 1\text{-} p_2$ es aproximadamente igual a 1. Nos centraremos en la forma en que p_2 / p_1 está afectada por los factores relacionados con las diferentes tasas de hospitalización.

En muchos estudios retrospectivos de tipo caso control, el grupo de casos ha sido escogido de pacientes tratados en un hospital. El grupo control se escoge de alguna de estas tres diferentes formas.

1. De pacientes hospitalizados sin la enfermedad principal;
2. De pacientes hospitalizados que tienen la condición de comparación
3. De personas de la comunidad sin la enfermedad principal.

ODDS RATIO EN DIFERENTES DISEÑOS DE CASOS Y CONTROLES.

Situación I.

Los controles son pacientes hospitalizados sin la enfermedad principal. En esta situación y también en las dos que siguen, los casos son pacientes hospitalizados con la enfermedad principal.

Expuestos y enfermos: sectores 7 y 8 de la Figura nº 1.
No expuestos y enfermos: sectores 3 y 4 de la Figura nº 1.

El grupo control en esta situación tiene la siguientes fuentes y proporciones:

Expuestos y no enfermos, sectores 5 y 6 de la Figura nº 1.
No expuestos y no enfermos, sectores 1 y 2 de la Figura nº 1.

La odds de exposición en el grupo control será aproximadamente:

$$\frac{p_2}{p_1} \times \frac{(1- h_d\ h_e\ h_p)\ (1-h_p)}{(1-h_d\ h_p)\ (1-h_e\ h_p)}$$

Con los cambios algebraicos adecuados, esta expresión se transforma en la formula:

$$\frac{p_2}{p_1} \times \left[1 - \frac{h_d\ h_e\ (1- h_p)}{(1 - h_d\ h_p)\ (1-h_e\ h_p)}\right]$$

El término de la derecha que está entre corchetes, será siempre inferior que 1 al menos que h_e, h_d o h_p sean iguales a 0. Estas dos últimas circunstancias son poco realistas, porque si h_d es igual a 0, ningún caso los enfermos estarán en el hospital y si $1- h_p$ es igual a 0, casi todo el mundo en la población general está hospitalizada. Consecuentemente, la OR en esta situación, siempre infraestimará la razón de riesgo, al menos que h_e sea igual a 0, cuando la exposición no tenga efectos sobre la hospitalización.

Situación II.

Los controles son pacientes hospitalizados con la condición de comparación. En esta circunstancia, se debe tomar una decisión estadística sobre el estatus de los pacientes que tienen la enfermedad y la condición de control. Si tales pacientes, son contemplados como enfermos (que es lo más frecuente), los dos grupos enfermos son los mismos que la situación 1. Los dos grupos controles hospitalizados vienen de los sectores 6 y 2 de la Figura 1, y contendrán las proporciones siguientes:

Expuestos y no enfermos: $e\,(1 - p_2)\,p_c\,(1- h_c\,h_e\,h_p)$
No expuestos y no enfermos: $(1 – e)\,(1 – p_1)\,p_c\,(1 – h_c\,h_p)$

La OR se obtendrá mediante la fórmula siguiente:

$$\frac{p_2}{p_1} \times \left[\frac{(1- h_d\,h_e\,h_p)\,(1 – h_c\,h_p)}{(1-h_d\,h_p)\,(1-h_c\,h_e\,h_p)} \right]$$

En esta expresión, que es similar a la de Walter[23], el término que está a la derecha, será igual a 1 para producir una OR no sesgada, si h_e es igual a 0, o h_c es igual a h_d, de tal forma que la enfermedad y las condiciones de comparación tengan similares tasas de hospitalización. De otra forma, la OR, dará una estimación sesgada de la razón de riesgo. La dirección y magnitud de sesgo dependerá de

los tamaños relativos de h_c y h_d. El sesgo producido por una falta de igualdad en h_c y h_d no se afecta por la clasificación asignada a pacientes que tengan la enfermedad y la condición de control.

Situación III.

Los controles son pacientes comunitarios sin la enfermedad principal. En esta situación los pacientes controles no están hospitalizados, y se distribuirán como se muestra en los sectores 1, 2, 5 y 6 de la Figura nº 1. Las proporciones en los grupos no enfermos serán como siguen:

Expuestos y no enfermos, sectores 5 y 6,

$$e\ (1 - p_2)\ (1 - p_c) + e\ (1 - p_2)\ p_c = e\ (1 - p_2)$$

No expuestos y no enfermos, sectores 1 y 2,

$$(1- e)\ (1 - p_1)\ p_c + (1 - e)\ (1 - p1)\ (1 - p_c) = (1 - e)\ (1 - p_1)$$

La odds ratio será:

$$\frac{ep_2\ (1 - h_d\ h_e\ h_p)}{(1-e)\ p1\ (1- h_d\ h_p)} \times \frac{(1-e)\ (1-p_1)}{e\ (1-p_2)}$$

Para evitar sesgos, el grupo control comunitario debería ser elegido de las personas que tienen la condición de comparación, aunque identificar a tales personas podría ser especialmente difícil en la investigación comunitaria. Asumiendo que tal tipo de personas pudieran encontrarse, todavía no se resolvería el problema del sesgo. Sus proporciones relativas en los grupos que tengan una condición de comparación serán

Pacientes expuestos: $e\ (1 - p_2) \times p_c$

Pacientes no expuestos $(1 - e)\ (1 - p1) \times p_c$

La razón de exposición en el grupo control será entonces: $e(1 - p_2) / (1 - e)(1 - p_1)$ y la OR seguirá teniendo sesgo. Debido a que la situación III produce el inverso de lo que ocurre en la situación I, los resultados mostrarán que si h_e no es igual a 0, no se puede obtener una OR totalmente exenta de sesgo mediante la selección del grupo control solamente de personas que carecen de la enfermedad principal.

La razón de riesgo estará falsamente disminuida si los controles son pacientes hospitalizados, y falsamente elevada si los controles vienen de la comunidad. Consecuentemente cuando h_e no sea igual a 0, la mejor esperanza para la obtención de una OR no sesgada (si los casos están hospitalizados) es elegir un grupo de comparación hospitalizado que tenga una condición para la cual h_c sea casi igual que h_d.

Otras Situaciones.

Aunque es más difícil de llevar a cabo, la comunidad podría ser el único escenario para la obtención de los casos y de los controles. En esas circunstancias las odds ratios carecerían prácticamente de sesgo si las proporciones de pacientes hospitalizados fuera prácticamente despreciable en todos los grupos. Por lo tanto si despreciamos a las personas hospitalizadas, la odds de exposición en los casos será:

$$[ep_2(1-p_c) + ep_2p_c] / [(1-e)p_1(1-p_c) + (1-e)p_1p_c] = ep_2 / (1-e)p_1.$$

Si los controles son individuos que no tienen la enfermedad principal, su odds de exposición será: $e(1-p_2) / (1-e)(1-p_1)$.

Cuando se dividen entre sí estas dos odds, la odds ratio obtenida no estará sesgada. Si los controles son personas que presentan la condición de comparación, su odds de exposición será: $e(1-p_2)p_c / (1-e)(1-p_1)p_c = e(1-p_2) / (1-e)(1-p_1)$. Y la odds ratio obtenida de esta forma también estará libre de sesgo.

En síntesis, la odds ratio permanecerá sin sesgo si los casos y los controles son recogidos de la comunidad. . De una forma parecida, Peritz [22] demostró que los "emigrantes" o "nómadas" (es decir los pacientes hospitalizados) no crean un desequilibrio especial en la composición de las poblaciones que dejan atrás.

Concretando, hemos profundizado en el sesgo de Berkson con el objeto de:

✓ Indicar la relación entre la razón de riesgo para expuestos y no expuestos en un estudio de cohorte y la odds ratio en un estudio de casos y controles. Hemos utilizado un modelo algebraico que es relativamente fácil de entender para el clínico [19].

✓ El modelo ha sido desarrollado específicamente para tres tipos de muestreo que son los más utilizados en los estudios casos control restropectivos. Hemos considerado la situación en las cuales los casos son todos pacientes hospitalizados, con los grupos controles viniendo de:

a. Pacientes hospitalizados con otras enfermedades
b. Pacientes hospitalizados con una condición comparativa específica
c. Personas de la comunidad y no hospitalizadas.

✓ El modelo tiene en cuenta a través del factor h_p, la hospitalización que ocurre en la población general por razones no asociadas con la enfermedad, la condición control o la exposición.

✓ La OR no estará sesgada si h_e es igual a 0, es decir, si la exposición a los agentes etiológicos imputados en la hipótesis, no tienen efectos como un factor separado que conduzca a la hospitalización.

Efectos de la exposición sobre la hospitalización.

La premisa de que la exposición no tiene impacto en la hospitalización, es decir que h_e es igual a 0, raramente se podrá mantener de una forma realista. Por ejemplo, el uso de ciertas sustancias farmacéuticas tales como los estrógenos, están regularmente asociados con efectos colaterales, como el sangrado que puede provocar hospitalización. Casi cualquier agente farmacéutico estará

asociado con un tipo de supervisión médica aumentada que puede conducir a la detección de cualquier dolencia que pudiera escaparse en otras circunstancias.

Incluso con los agentes no farmacéuticos, como es el tabaco, se pudiera provocar un efecto colateral (por ejemplo la tos) que conduciría a un aumento en la probabilidad de hospitalización del paciente y a la detección de enfermedades que no hubieran sido diagnosticadas de otro modo. Consecuentemente para la generalidad de los agentes etiológicos, lo más conservador es asumir que h_e es mayor que 0. Asumiendo esto, se puede llegar a las siguientes conclusiones:

I. El sesgo de Berkson podrá evitarse en los estudios de caso control restropectivos, si los casos y los controles son escogidos de la comunidad.

II. Si los casos están escogidos en una población hospitalizada, los efectos matemáticos del sesgo de Berkson, no serán eliminados mediante la selección del grupo control de la comunidad. De hecho, esta elección, elevará falsamente la OR. De forma inversa, los efectos matemáticos serán falsamente menores, disminuirá la OR, si el grupo control está escogido de todos los pacientes hospitalizados.

III. Con un grupo de casos hospitalizados, la mejor esperanza de evitar el sesgo de Berkson, es coger un grupo control hospitalizado, teniendo una condición de comparación para la cual la tasa de hospitalización iguala a la tasa análoga en el grupo de casos enfermos. A menos de que estas tasas sean iguales, la OR se verá sesgada hacia arriba, si h_d es mayor que h_c o hacia abajo si h_c es menor que h_d.

Problemas en el modelo matemático

Aunque el papel de "h_e" (tasa de hospitalización para personas expuestas) es fundamental en la producción del sesgo probabilístico descrito por Berkson, aunque la magnitud del sesgo dependa de la razón de hospitalización por diversos factores relacionados tanto con la exposición como con la enfermedad, estas tasas se producen a partir de decisiones de los médicos, y del tipo de paciente que envían a un hospital en donde finalmente se transforman en casos o en controles.

El modelo matemático para la evaluación de los efectos de estas decisiones clínicas depende de que las tasas de envío al hospital *puedan ser sumadas como probabilidades independientes*, pero esta asunción sobre la independencia no es totalmente congruente con la realidad. En la práctica, la concurrencia de dos causas potenciales de hospitalización afecta a los médicos examinadores de una forma especial aumentando la verosimilitud de hospitalización para el paciente que posea ambos procesos. En realidad este tipo de paciente es hospitalizado casi siempre.

Al examinar las relaciones, entre las tasas de hospitalizaciones esperadas para las dos condiciones coincidentes, Roberts y colaboradores [24] encontraron solamente una correlación débil. Concluyeron que el problema tenía dos componentes: el sesgo probabilístico descrito por Berkson y un tiempo de selección clínico separado que afecta a la verosimilitud para personas con dos o más condiciones clínicas o exposiciones.

En lugar de contemplar la selección clínica y las uniones probabilísticas, los investigadores podrían identificar eventos específicos y decisiones que conduzcan a la hospitalización. Los componentes más lógicos y convincentes de estos fenómenos son la supervisión rutinaria de los pacientes, los signos clínicos producidos por la enfermedad y por el agente de exposición y la consecuente orden e interpretación de este diagnóstico. Estos fenómenos determinan si el paciente será o no será hospitalizado.

1.4 DISCREPANCIAS EN LOS RESULTADOS DE LOS ESTUDIOS DE AMBITO HOSPITALARIO Y COMUNITARIO CUANDO SE ANALIZA UNA MISMA PREGUNTA DE INVESTIGACIÓN.

Por el carácter marcadamente inductivista de la investigación biomédica no basta con la realización de un solo estudio para responder una pregunta de investigación. Son necesarios varios trabajos para valorar cualquier tipo de asociación. Con frecuencia esos estudios están practicados en diferentes ámbitos, por ejemplo, en el hospital y en la comunidad [25]. En muchas ocasiones los resultados de los trabajos realizados a nivel comunitario no son consistentes con los realizados a nivel hospitalario según el tercer principio de Hill [26] (principio de especificidad entre una causa y un efecto: relación biunívoca exclusiva).

En el año 1946, Joseph Berkson[1] puso de manifiesto por primera vez que

los resultados que se observaban en una investigación realizada en el hospital no se correspondían con lo que ocurría realmente en la población de referencia, dando origen al sesgo que lleva su nombre y que hemos tratado anteriormente en la introducción de esta Tesis.

Se pueden enumerar las siguientes causas de discrepancia o falta de consistencia entre los estudios realizados en el hospital y los realizados en la comunidad.

1.- **Particularidades de la investigación hospitalaria**.

- Proceso de selección: Es fundamental tener en cuenta el proceso por el que un paciente acude al hospital a la hora de sacar conclusiones sobre la información obtenida.

- Recogida de datos – Historia Clínica: El escollo principal está constituido por los datos que faltan o no constan. Un dato que no consta puede deberse a varios hechos. A que no se haya valorado, a que habiendo sido medido haya dado un resultado negativo o incluso a que habiendo dado un resultado positivo no se encuentre reflejado en la historia. Los resultados pueden cambiar dependiendo de las asunciones que se hagan.

Los estudios hospitalarios pueden tener una tendencia a utilizar más los datos de la historia clínica, mientras que los estudios comunitarios necesariamente han de utilizar datos de entrevista.

El otro gran problema que presenta la historia clínica es la falta de uniformidad (estandarización) en las pruebas realizadas o en la anotación de los resultados y la falta de criterios establecidos a la hora del redondeo de los datos numéricos[25]

2.- **Discrepancias según la línea de investigación**.

Las preguntas de investigación se pueden agrupar de una forma sencilla y simplificando las ideas de Hulley y Cummins[27] dentro de las cinco líneas siguientes: frecuencia, diagnóstico, etiología, pronóstico y tratamiento-prevención.

Estimación de la frecuencia de un proceso.

Los centros asistenciales concentran enfermos y los profesionales sanitarios se ven tentados a utilizar estos datos como los que expresan la magnitud de un proceso. Esto sería válido tan sólo cuando los "elegidos" por el hospital sean los mismos que los afectados, circunstancia que raramente ocurre en la realidad. Cobran en este sentido un interés especial aquellos procesos en los que no todos los afectados se sienten enfermos y en los que no todos los afectados buscan asistencia.

También debe de cuestionarse la frecuencia de un proceso que se estima con los datos de centros de referencia [25].

Valoración del diagnóstico.

La mayor parte de la investigación en este apartado se realiza en instituciones asistenciales. Las pruebas de cribado constituyen una excepción. En ese caso es muy importante el espectro de la enfermedad en los casos avanzados que afecta la sensibilidad del cribado y el grupo de referencia que afecta a la especificidad[25]. La prevalencia de la enfermedad afecta a todos los parámetros de la prueba de cribado pero especialmente a los valores predictivos[28]

Valoración de la etiología.

El primer factor a tener en cuenta es el sesgo de Berkson[1] que hemos tratado con anterioridad. Otros factores que pueden afectar a la investigación etiológica son:

- **La frecuencia de la exposición influye en el valor de la fuerza de la asociación.** En el estudio de la etiología o relación de causalidad no se puede olvidar en ningún momento que la estimación de la frecuencia de la exposición ha de ser válida, ya que según el modelo de Rothman [28] ésta influye en el valor de la fuerza de asociación. Este parámetro en fácil que cambie según el grado de referencia que ocupe un centro. También puede ocurrir si la exposición es una intervención sanitaria que puede cambiar según el grado de aplicación en ciertas instituciones o sectores de la población.

- Sesgo de detección. El sesgo de detección se produce cuando el efecto se detecta más en el grupo de los expuestos que en el de los no expuestos y su consecuencia es una sobreestimación de la fuerza de asociación. A priori el sesgo de detección debe de ser más frecuente en las investigaciones hospitalarias porque tienen la mayoría de herramientas diagnósticas mucho más accesibles.

- Sesgo de inclusión. El sesgo de inclusión se produce en los estudios de casos y controles cuando en el grupo de controles se incluyen individuos con procesos que mantienen relación con la exposición y produce una subestimación de la magnitud de asociación. También es más probable en los casos-control hospitalarios.

- Sesgo protopático. El sesgo protopático se produce cuando los estadios iniciales del efecto, normalmente subclínicos, condicionan un cambio en el nivel de exposición. Como consecuencia se tiende a sobreestimar la fuerza de asociación Este error se puede presentar en cualquier tipo de investigación, pero su probabilidad aumenta cuando la enfermedad está más evolucionada que es cuando se suele diagnosticar en el hospital. El sesgo protopático ha sido descrito por Thijs y colaboradores[29] al estudiar la asociación del consumo de alcohol y la existencia de litiasis biliar.

- Sesgo por indicación (confusión por indicación). Se puede confundir con frecuencia con el sesgo protopático. El sesgo por indicación se presenta cuando se analizan medicaciones o intervenciones como consecuencia de procesos asociados con el efecto (la indicación de intervención es un auténtico factor de confusión) o también como consecuencia de la gravedad del proceso.

Estos errores también pueden presentarse en investigaciones de tipo comunitario.

- *Valoración del pronóstico*. Las inconsistencias que se pueden encontrar entre los diferentes estudios radican en el tipo de centros en donde se realizan (en su posición de referencia dentro del sistema sanitario), el patrón de remisión al centro y la selección en función del estadio de la enfermedad.

- *Valoración del tratamiento-prevención*. La mayoría de los estudios sobre tratamiento suelen ser institucionales mientras que los de prevención suelen ser

de ámbito comunitario. Las técnicas de aleatorización y enmascaramiento reducen la probabilidad de la existencia de sesgos y le dan el carácter de experimentales. Los problemas pueden venir a la hora de la evaluación de la validez externa.

1.5 DISCREPANCIAS ENTRE LA INVESTIGACIÓN HOSPITALARIA Y COMUNITARIA. EN LA LITERATURA BIOMEDICA.

Delgado Rodríguez ha intentado responder a esta pregunta[25] mediante una búsqueda en Pubmed. No existe un término MeSH (Medical Subjects Headings) específico para diferencias o discrepancias por lo que tuvo que utilizar varios términos equivalentes en lengua inglesa con la estrategia siguiente: " (differen* OR diverse OR divergen* OR disparate OR inconsisten* OR incoherenc* OR incongruous* OR discrepan* OR discordan* AND community stud* AND hospital stud*)" . La búsqueda se realizó en todos los campos y para el período 1981-2001. Tan sólo obtuvo tres citas, de las que sólo una comparaba estudios comunitarios con estudios hospitalarios y estaba referida al papel del alcohol en el cáncer de mama[30]

En una segunda fase de su estrategia valoró la frecuencia de mención de algunos de los errores comentados líneas arriba buscando de forma independiente los términos siguientes para el mismo período temporal : " detection bias "(104 citas) , "inclusion bias " (3 citas de la que tan sólo una era relevante) y "protopathic bias OR confounding by indication OR indication bias " (56 citas). El riesgo relativo entre el hecho de ser un estudio hospitalario y tratar el sesgo de detección con respecto al total de trabajos consultados[25] fue del 2,71 (IC al 95%, 1,69-4,37). Esto no significa que en el proceso de revisión sea más frecuente el sesgo de detección en los estudios hospitalarios , sino que lo mencionan más porque los autores conocen su amenaza. La asociación entre estudio hospitalario y mención del sesgo protopático o del sesgo por indicación no fue significativa estadísticamente (Riesgo Relativo = 1,76 , IC del 95% , 0,90-3,42).

La valoración científica de estos errores tiene la dificultad de **no tener un criterio de verdad**, es decir, cuando se analizan las discrepancias entre estudios realizados en el ámbito hospitalario y comunitario surge la pregunta inevitable: *¿Cuáles son los realmente ciertos*? Se podría extraer la conclusión de que, acorde con los ejemplos descritos, la validez radica en los estudios

comunitarios, no sometidos al proceso de selección particular que suele existir en los realizados en torno a las instituciones cerradas.

Sería un tanto arriesgado asumir esto pues , en primer lugar, los ejemplos propuestos [25] no provienen de una búsqueda sistemática . Un método adecuado para conocer realmente las discrepancias sería el metaanálisis. En segundo lugar , no siempre se puede penalizar a los estudios hospitalarios frente a los comunitarios porque hay otras situaciones en las que los estudios hospitalarios no proporcionan la misma información que los asentados en la comunidad y no se encuentra una razón definida para estas diferencias.

Según lo anterior, ante la ausencia de un criterio de verdad concluyente, no deberían de minusvalorarse en teoría los estudios hospitalarios frente a los comunitarios.No obstante, habrá de aceptarse que en base a las razones metodológicas antes expuestas, los estudios hospitalarios ofrecen resultados inferiores en calidad a los producidos en otros ámbitos [25].

Pudiera parecer que la discordancia entre los estudios de casos y controles hospitalarios y los comunitarios sea la norma en la literatura biomédica, cuando esto no es así. Existen muchos ejemplos en los que no se han encontrado diferencias substanciales entre ellos[25] ni tampoco con los estudios de cohortes. A continuación exponemos una serie de **problemas frecuentes** en la investigación hospitalaria y las recomendaciones a seguir en cada caso [25].

A. Sesgo de referencia / remisión especialmente importante en los centros terciarios. -Solución: Analizar la remisión y organizar el proceso en estadios.

B. Uso de la historia clínica. -Solución: Reflexionar de la forma en que está completada. Tratar la información que "no consta" de forma separada. Lo mejor es que la recogida de datos no sea retrospectiva.

C. Los sesgos más frecuentes en los estudios casos-controles hospitalarios (de base secundaria) son los de Berkson y de inclusión. - S o l u c i ó n : Identificación correcta de la base del estudio y utilización de criterios de selección adecuados.

D. Sesgo de detección en las enfermedades de largo período de latencia. -

Solución: Analizar como se realiza el diagnóstico del efecto y comprobar que no se practica con más frecuencia en los expuestos.

E. Sesgo Protopático. -Solución: Pensar siempre en el caso de enfermedades con fases prolongadas de latencia y seleccionar si es posible casos sin clínica.

F. Confusión por indicación. -Solución: Analizar las razones que motivan la aplicación de una intervención y tratarlas en el análisis. Realizar análisis en función del tiempo transcurrido entre la intervención y el efecto.

G. Falta de validez externa. -Solución: No hay. Si la población estudiada no representa a la que se pretende aplicar los resultados hay que hacer otro estudio. Los procesos de aleatorización y enmascaramiento no controlan este problema.

El sesgo de referencia consiste en que la población que atiende el centro no representa lo que sucede en la colectividad. Este error tiene trascendencia cuando se intenta establecer la frecuencia de la enfermedad. Conlleva también una estimación distorsionada de la frecuencia de exposición generalmente aumentándola. Si la exposición está aumentada, la consecuencia bajo el modelo de Rothman , es que aumenta el valor del Riesgo Relativo (RR). Esto explicaría porqué en muchos metaanálisis se aprecia que el valor del RR que dan los estudios hospitalarios es superior al de los estudios de la colectividad [25].

Es necesario mencionar los problemas que plantea la historia clínica como herramienta única para la obtención de la información. Es recomendable en estos casos tratar los "no consta" como tales y no asumir que sean "no expuestos". Si la frecuencia de los "no consta" es alta nunca se podrá estar seguro de que la asociación encontrada sea real. Los problemas de falta de uniformidad introducen un error de mala clasificación en principio no diferencial, que si la exposición tiene más de dos niveles , conviene recordar que puede sesgar el valor de la asociación en cualquier sentido [31]. Lo mejor por lo tanto es la recogida prospectiva de la información en donde se minimizan las pérdidas y se mantiene el principio de uniformidad.

Se puede concluir que para realizar una investigación hospitalaria que

responda correctamente ante un problema que surge en la comunidad, aparte de una recogida de datos apropiada y uniforme , es conveniente plantear el marco de población en la que se origina, el proceso por el que un individuo llega a una institución , la forma en que se realiza el diagnóstico (influencias de la exposición u otros procesos relacionados con la exposición) y si las intervenciones se aplican como consecuencia de otros procesos relacionados con el efecto o por el propio efecto en sí.

1.6. LA VALIDACIÓN DE MODELOS PRONÓSTICOS.

El análisis mediante regresión se utiliza en muchas ocasiones para desarrollar modelos estadísticos predictivos de un resultado a partir de una o más variables explicativas [32] En medicina, los modelos de regresión referentes a una característica del enfermo se denominan *modelos pronósticos*. Su intención fundamental es estudiar el proceso de enfermar mediante la determinación de cuáles son las variables que están relacionadas con el devenir de la enfermedad.

Debe quedar claro desde el principio que un modelo no tiene valor clínico a menos que sea capaz de predecir un hecho o una característica con cierto grado de éxito. *La utilidad viene determinada más* ***por la forma práctica en que el modelo se comporta*** *que por la cantidad de ceros que pueda haber en los valores de sus "p".*

Hay dos formas en la que un modelo puede tener utilidad. La primera, es que permita una clasificación de los pacientes en dos o más grupos de pronósticos diferentes. Estos esquemas clasificatorios se pueden manejar en el diseño de una conducta terapéutica o para evitar pruebas o estudios innecesarios. En segundo lugar, un modelo se puede utilizar para estimar el pronóstico de los pacientes a nivel individual. Los dos abordajes difieren bastante.

Por ejemplo, con un modelo excelente se puede distinguir fácilmente entre los pacientes con riesgo alto y los pacientes con riesgo bajo, pero puede resultar difícil a nivel individual informar del tiempo de supervivencia esperado con un intervalo de confianza del 95%. Esta distinción de lo que es aplicable al grupo y lo que es a nivel individual es a veces difícil de comprender.

Un aspecto básico de la predicción es considerar si el modelo construido a partir del análisis de una serie de datos originales es transportable a pacientes similares en un lugar distinto. Este concepto se conoce como ***capacidad de generalización o validación***. El modelo que supere las pruebas se dice que está validado.

¿QUE QUIERE DECIR VALIDACION?

La idea de validar un modelo pronóstico es la de demostrar que el modelo trabaja de forma correcta cuando se le aplica a pacientes diferentes de aquellos de los que derivó. Existen varias formas de aproximación al problema del funcionamiento (*exactitud de predicción*) de un modelo pronóstico. Por ejemplo, la comparación de la tasa de eventos observados y predichos para los grupos de pacientes anteriormente definidos (*calibración*) o para pacientes individuales (*niveles de exactitud*), y medidas que distinguen entre pacientes que experimentan o no experimentan el evento de interés (*discriminación*). Altman y Royston [32] piensan que no es correcto decir que se haya validado un modelo sino que **es más adecuado afirmar que se ha estudiado su funcionamiento.**

Otra acepción semántica del término validación es la que se deriva del concepto psicométrico de validez., usado ampliamente en los estudios de medida. La validez es la propiedad de un método de medición para medir lo que realmente intenta medir [33]. En este caso se utiliza una aproximación de tipo correlacional, juzgando de la forma más eficiente posible la variación intra-sujeto en relación con la variación inter-sujeto. En los trabajos bio-médicos prima más el criterio de calidad en los pronósticos a nivel individual o para grupos de pacientes.

En un mundo ideal, nosotros podríamos crear un modelo con datos recogidos de un estudio de factores pronósticos bien diseñado con un número adecuado de pacientes. Esperaríamos que el modelo tuviera toda la información pronóstica de los datos y que además ignorara todas las *variables "ruidosas"*. Imaginaríamos también que el modelo tuviera todas las variables correctas (incluyendo quizás alguna interacción) y que hubiéramos determinado correctamente el efecto funcional de cada variable contínua en la variable dependiente. La validación de un modelo como el que acabamos de describir, tendría en cuenta una muestra

adecuada de pacientes nuevos y la medida de todas las variables pronosticas contenidas en el mismo con las mismas técnicas de laboratorio.

¿VALORABLE O SIMPLEMENTE VALIDO?

En la práctica ocurre en ocasiones que un modelo validado desde un punto de vista estadístico, puede que no tenga ninguna importancia clínica. Si la ***información pronóstica intrínseca*** es débil, las predicciones aunque no estén sesgadas, no permitirán separar a pacientes en grupos pronósticos de utilidad clínica. En contraste, si la información pronóstica intrínseca es sólida, incluso en un modelo sesgado se podrán extraer grupos clínicos de utilidad.

El éxito del desarrollo de un modelo depende de varias circunstancias:

- La exactitud potencial en el pronóstico, que presumiblemente es desconocida.
- La información pronóstica intrínseca en las variables sometidas a estudio, lo cual depende de varios factores como por ejemplo la fisiopatología de la enfermedad.
- El proceso de medida que convierte la información intrínseca en dígitos, siendo algunas medidas más dóciles que otras en este sentido.
- La exactitud en que el modelo convierte los dígitos de las mediciones en predicciones.

De lo expuesto surgen dos preguntas:

a. ¿Con las variables introducidas y medidas, es el modelo obtenido el mejor que se puede obtener?
b. ¿Puede el modelo predecir con suficiente exactitud para el propósito que se ha creado?

La primera pregunta implica claramente un propósito de tipo estadístico mientras que la segunda implica uno de tipo clínico. Un modelo puede fallar porque no sea válido desde un punto de vista estadístico (para lo cual muchas veces hay digamos "compostura") o porque la información pronóstica intrínseca sea débil (lo cual no se puede "arreglar").Aún más, un mismo modelo puede fallar para un criterio clínico pero ser aceptable para otro. Todas estas consideraciones nos

hacen poder definir dos tipos de modelos validados.

TIPOS DE MODELOS DESPUES DE SER VALIDADOS.

En cierta forma el juicio sobre la validez depende de las circunstancias. Por ejemplo, si el análisis de un modelo arroja que la discrepancia entre unas tasas de recaída del 90% (pronosticadas) y las del 75% (observadas) pueden ser aceptables si el propósito es identificar subgrupos pronósticos suficientemente separados. Sin embargo si el propósito primitivo es el identificar pacientes con una tasa de recaída probable del 85%, el modelo no sería considerado válido para ese propósito.

Se pueden proponer por lo tanto dos tipos de modelos:

- **Un modelo validado estadísticamente**, que pueda superar todas las pruebas, incluyendo los tests de bondad del ajuste en los datos originales y los de predicción no sesgada en los datos nuevos.
- **Un modelo validado clínicamente**, que pueda "funcionar" de manera satisfactoria sobre los datos nuevos de acuerdo con nuevos criterios estadísticos.

Según esto, un modelo clínicamente válido puede no estar validado estadísticamente (porque por ejemplo sus predicciones estén sesgadas o porque no cumpla las pruebas del bondad del ajuste) y un modelo estadísticamente válido puede que no sea relevante desde un punto de vista clínico (porque por ejemplo la información pronóstica intrínseca sea demasiado débil).

Generalmente es más difícil obtener un modelo validado estadísticamente que uno validado clínicamente, porque es especialmente difícil superar el problema del optimismo pronóstico y del sesgo a la hora de la construcción del modelo. Sin embargo, en muchas ocasiones un modelo validado clínicamente probablemente sea más útil que uno validado estadísticamente.

NECESIDAD DE VALIDAR UN MODELO PRONÓSTICO

La necesidad fundamental es que existan pruebas de que el modelo

realiza el trabajo para el que fue diseñado. Pueden presentarse tres razones interrelacionadas por las que un modelo pronóstico no "funcione" bien.

I. Por deficiencias en los métodos de modelización estándar: Tienen que ver con mucha frecuencia con aspectos dependientes de los datos utilizados. En ocasiones existen muchas variables candidatas a un posible análisis con lo que es difícil aplicar el **principio de parsimonia**. Los aspectos dependientes de los datos provienen de la selección de variables. La forma de selección "stepwise" en ocasiones no prospera hacia el mejor modelo predictivo porque es una forma de selección informática de tipo puramente automático. Otro tipo de construcción se basa en seleccionar aquel modelo que optimiza una medida de bondad del ajuste que está penalizada a su vez cada vez que se introduce una nueva variable (**criterio de información de Akaike**). Este método puede tener también inconvenientes serios como el de omitir predictoras importantes [32]. Otra forma más reciente para tratar de solucionar estos problemas son **los árboles de regresión** (CART) y **las redes neurales**.

II. Por deficiencias en el diseño de los estudios pronósticos: Existen varias causas de debilidad en la fase de diseño de un estudio pronóstico que pueden conllevar la obtención de resultados demasiado optimistas. Entre ellas están:

a) Falta de criterios claros de inclusión y exclusión,
b) Existencia de muchos pacientes excluidos a través de datos perdidos ("missing data") que puede que no se hayan perdido de forma aleatoria,
c) Existencia de razonamientos poco claros a la hora de elegir un tratamiento,
d) Manejo de un tamaño de muestra inadecuado.

La definición de las características de la muestra es también fundamental para el clínico que desea saber si un modelo es realmente válido para un paciente en particular. Los problemas derivados de la selección de variables dependiente de los datos se exacerban cuando existe un tamaño de muestra pequeño [34]. Con una muestra pequeña habrá una razón baja de **señales / ruido** *("signal–to–noise ratio"),* aumentando el riesgo de selección de variables poco relevantes y el de no inclusión de variables importantes. En este sentido es muy importante el concepto de eventos de interés por variable[34].En los análisis mediante regresión logística

se consideran necesarios un mínimo de 10 eventos de interés por variable para el buen funcionamiento del modelo.

III. Por la obtención de modelos que no se puedan transportar: A pesar de que la metodología empleada en la confección sea impecable, puede que no sea transportable a otros lugares. La falta de similitud entre los pacientes de dos centros diferentes puede que sea su causa. Es lo que se denomina "**case-mix**". No obstante si un modelo contiene todas las variables importantes puede que funcione bien a pesar de las diferencias en el "case-mix". Lo malo es que el investigador casi nunca se sabe cuáles son las variables realmente importantes.

1.7 COMO VALIDAR UN MODELO.

Hay una serie de consideraciones a la hora de realizar la validación correcta de un modelo. Las enumeramos a continuación:

I. El diseño del estudio de validación
II. La medición de la información pronóstica intrínseca.
III. La comparación de las predicciones con las observaciones.
IV. La cuantificación del funcionamiento del modelo.
V. La especificación de un funcionamiento adecuado.

Vamos a tratar a continuación todos y cada uno de estos aspectos.

I. DISEÑO DEL ESTUDIO DE VALIDACIÓN.

Las siguientes estrategias de validación están ordenadas de menor a mayor grado de rigor.

a.- Validación Interna La forma común de analizar el comportamiento de un modelo en el futuro es hacer una división de los datos originales, es lo que se denomina validación cruzada. La forma de cómo dividir los datos es algo que los diversos autores no están muy de acuerdo en clarificar. La división al azar nos conduce a un contexto en el que existe la misma posibilidad de variación

que en el principio por lo que se considera un método débil. Una prueba más consistente es la división de los datos de manera no aleatoria, por ejemplo según períodos de tiempo diferentes. Otra forma de abordaje de este problema es el ***bootstrapping*** y la técnica de dejar uno fuera ***("leave one out").***Mediante ellas se pueden evaluar muchos grupos de datos, pero no dejan de ser procedimientos de validación interna.

b.- Validación Temporal Consiste en la evaluación del modelo en pacientes posteriores a los primitivos y recogidos en el mismo centro. Se trata de una técnica prospectiva, independiente de los datos originales y del proceso de evaluación de la bondad del ajuste del modelo original.

c.- Validación Externa. Ninguna de las dos anteriores abogan por la generalización del modelo Es absolutamente deseable evaluar el modelo con nuevos datos recogidos de una población de pacientes adecuada y en otro centro. Aspectos muy importantes de este apartado son la forma en que la muestra se selecciona y el tamaño de la misma.

II. MEDICION DE LA INFORMACIÓN PRONÓSTICA INTRÍNSECA.

Hemos de asumir que los pronósticos se construyen como *probabilidades* de que un hecho ocurra y que éstas van acopladas de forma explícita o implícita a un punto temporal. Las probabilidades se obtienen como resultados de un modelo pronóstico, el cual puede ser lo más sencillo posible, por ejemplo dos grupos de pacientes definidos por una variable pronóstica binaria o de lo más complejo cuando el índice pronóstico se define como una función lineal o no lineal de muchas predictoras con factores de constricción *("shrinkage")* aplicados a los coeficientes de regresión [34].

En este apartado no es necesaria la distinción entre modelos de supervivencia (regresión de Cox) y modelos de regresión binaria (regresión logística) porque los modelos de supervivencia pueden generar probabilidades de predicción en cualquier punto temporal dentro del período de seguimiento del estudio [32].

Intuitivamente, la idea de información pronóstica es sencilla y está relacionada con la magnitud o intervalo de las probabilidades pronosticadas. Por

ejemplo, en un análisis no ajustado por otros factores , la probabilidad estimada de supervivencia a los tres años siguientes del tratamiento del cáncer de mama con metástasis linfáticas locales (uno a tres nódulos) puede ser del 90% mientras que para aquellos casos con diez o más nódulos está alrededor del 60%. En contraste, los resultados para mujeres pre y post-menopaúsicas oscila entre el 84% y el 82%. La información pronóstica que reside en el número de nódulos linfáticos afectados es mucho mayor que la que tiene el estado menopaúsico, porque la amplitud de probabilidades es 0.3 contra 0.02.

Sin embargo esta concepción tan elemental tiene sus dificultades. La amplitud de las probabilidades depende de lo finamente que esté graduado el índice pronóstico. A más finura en la graduación mayor será la amplitud de las probabilidades. También depende de la prevalencia del evento en estudio. En los análisis de supervivencia la amplitud de las probabilidades aumenta cuando se alarga el tiempo de seguimiento. También le afecta el optimismo con que el modelo estima las probabilidades, lo cual a su vez está influenciado por el tamaño muestral y por la forma en que se haya construido el modelo original (por ejemplo con el método "stepwise").

La información pronóstica intrínseca de un modelo multivariante constituye aún una pregunta abierta [32].

PSEP

Supongamos que la variable dependiente sea la muerte en un período de tiempo dado después de la medición de una serie de factores pronósticos. Supongamos también que el esquema de clasificación pronóstica se ha obtenido directamente de predictoras individuales por medio de un índice pronóstico multifactorial o por algún otro procedimiento como los árboles de clasificación y de regresión (CART) o por la opinión experta de un especialista. Todo lo que se necesita es que cualquier paciente pueda ser clasificado en uno de los dos o más grupos pronósticos y que se puedan identificar los grupos con el mejor y el peor pronóstico. Entonces tendremos:

- P_{peor} = Probabilidad pronosticada de morir para un paciente en el grupo con peor pronóstico.

- P_{mejor} = Probabilidad pronosticada de morir para un paciente en el grupo con mejor pronóstico.

De esta forma podemos medir la información pronóstica predicha por el modelo mediante la resta de esos dos valores.

$$PSEP = P_{peor} - P_{mejor}$$

El PSEP se considera como un índice simple de ***separación*** entre el peor y el mejor pronóstico de un modelo multivariante. Cuando sólo existan dos grupos, los valores de P_{peor} y P_{mejor} se pueden identificar con los valores predictivos negativo (VPN) y positivo (VPP) de una prueba diagnóstica porque

$$P_{peor} = VPP$$

$$P_{mejor} = 1 - VPN$$

Por lo que

$$PSEP = VPP + VPN - 1$$

III. COMPARAR LAS PREDICCIONES CON LAS OBSERVACIONES.

Con las definiciones expuestas más arriba, la evaluación consiste en la comparación de los valores observados y los pronosticados, lo cual a su vez es un aspecto de la calibración del modelo.

Supongamos en un ejemplo hipotético que se ha pronosticado en tres grupos de pacientes una probabilidad de supervivencia a los tres años del 90, 60 y 30% sobre la base de un pequeño estudio inicial para encontrar un índice pronóstico de muerte en una enfermedad X. Entonces la $P_{peor} = 0,7$, la $P_{mejor} = 0,1$ y la PSEP = 0,6.Supongamos que un investigador diferente lleva a cabo un estudio de validación de forma adecuada y encuentra unas probabilidades de supervivencia respectivas del 70, 60 y 50%, todas ellas diferentes a las del estudio inicial. En este caso, la PSEP = 0,2, de lo cual se concluiría en primer lugar el considerable optimismo de la separación pronóstica inicial aún teniendo una relación directa con el período medido de tiempo.

Hay evidencia de que el índice pronóstico "funciona" al menos indicando una diferencia en las tasas de supervivencia. Sin embargo, si en los pacientes con tan sólo un 30% de probabilidad de supervivencia a los tres años estuviera justificado el administrar un tratamiento agresivo con efectos colaterales peligrosos y desagradables, esa decisión sería insostenible en aquellos que tienen un 50% de probabilidad de supervivencia, invalidando el modelo como instrumento de decisión terapéutica.

Si las probabilidades obtenidas hubieran sido, por ejemplo, de 65, 68 y 62%, respectivamente, se podría concluir que el modelo no era válido porque las probabilidades son aproximadamente las mismas y tienen una relación directa con la cantidad de tiempo analizada.

IV. CUANTIFICACION DEL FUNCIONAMIENTO DE UN MODELO.

Parece claro que la validación no debe de ser determinada tan sólo por criterios puramente estadísticos, sino que debe descansar también en los fines clínicos que nos hayamos propuesto. No obstante, los tecnicismos estadísticos no debemos olvidarlos.

Una forma de medir los valores pronosticados y los observados es el ***índice de Brier*** [35] que es una función de pérdida cuadrática definida como la diferencia de medias al cuadrado entre los valores obtenidos en la muestra de validación y los correspondientes valores pronosticados por el modelo. El índice de Brier posee unas propiedades matemáticas agradables digamos pero tiene también un inconveniente , carece de una interpretación obvia y sólo en términos

generales se puede afirmar que cuanto más valor tenga , peor es la calidad de la predicción. Una estimación estadística menos elaborada pero más interpretable es la diferencia de probabilidades observadas y pronosticadas a nivel de grupo (PSEP) .En el análisis real de los datos se debe de utilizar más de una medida.

ESPECIFICACION PREVIA DE FUNCIONALIDAD.

Los estudios pronósticos se pueden clasificar en dos grandes grupos: **pragmáticos** *y* **explicativos**[36]..Los pragmáticos tienen que ver básicamente con intenciones o supuestos de contexto clínico. La idea es prejuzgar la calidad de la predicción a partir de un modelo pronóstico haciendo que sea o no sea aceptable.

En este caso hay una idea clara en términos cuantitativos, por ejemplo sobre el efecto de un tratamiento en el ámbito de un ensayo clínico. Si el objetivo es la identificación de los pacientes que tengan un 80% de probabilidad de supervivencia a los tres años, un estudio de validación que muestre una probabilidad de supervivencia del 60% será rechazado para el propósito original aún teniendo una información pronóstica intrínseca muy sólida.

Los estudios explicativos se preocupan más de la comprensión científica y de la generación de hipótesis para la respuesta de preguntas tales como ¿Qué factores son importantes para la predicción del curso de una enfermedad? ¿Se puede discriminar de una forma reproducible entre un buen o un mal pronóstico de una enfermedad? En este caso no hay un objetivo pragmático por lo que se pueden analizar aspectos generales de tipo cuantitativo y cualitativo como estos:

- ¿Son las mismas variables importantes todavía?
- ¿Es correcta la forma en que trabaja el modelo funcionalmente?
- ¿Los coeficientes de regresión estimados son compatibles?
- ¿Qué bondad de ajuste tiene para los nuevos datos?
- ¿Se preserva el orden correcto en los grupos pronósticos?
- ¿La tasa de eventos entre los diversos grupos pronósticos son significativamente diferentes?

Es muy difícil juzgar si son importantes las mismas variables porque las

diferencias en la distribución de las predictoras en las muestras original y en la de validación puede afectar a la precisión con que estimen los coeficientes de regresión. Esto se puede acentuar más si las variables se seleccionan por el método "stepwise".

Se pueden aplicar también los conceptos de PSEP, P_{peor} y P_{mejor} en el grupo original y en el validatorio.

El estudio de validación representa solamente una parte del proceso científico de entendimiento de una enfermedad. Es muy útil la especificación previa del funcionamiento de un modelo. No hay que olvidar que **un aspecto de la validación es proporcionar una estimación no sesgada del error de predicción del modelo**. Miller y colaboradores sugirieron que [37] es más interesante pensar en la validación en este sentido más que en la capacidad del modelo de pasar las pruebas o no pasarlas.

Por lo tanto [32], nos debemos centrar en las medidas que cuantifican el funcionamiento de un modelo y aceptar que la decisión final sobre su validez requiere juicios clínicos y es dependiente del contexto. ***La estadística sola no puede determinar la validez clínica***.

PERSPECTIVAS GENERALES

Después de haber obtenido lo datos nuevos y haber seguido el proceso de validación es necesario que nos hagamos la pregunta siguiente:

¿Cual es real y definitivamente el modelo final?

En la realidad[32], ***una validez perfecta es casi imposible de obtener***. Van Houwelingen y Thorogood adoptaron este punto de vista y propusieron la actualización del modelo original a la luz de los datos nuevos[38] en el análisis de supervivencia. Su actualización se basaba en una menor calibración del modelo original en sus índices pronósticos más que en una reconstrucción completa.

Pueden existir muchos otros tipos de fallos en la validación y en algunas ocasiones el modelo original presenta tantas imperfecciones que es imposible

salvarlo [32]. Si el "case-mix" en la muestra validatoria difiere mucho del de la muestra primitiva, el modelo puede fallar sin ningún tipo de remisión. También es verdad que se puede mejorar introduciendo una o varias nuevas variables que relacionen los case-mix diferentes. Por ejemplo, el rango de edad puede que varíe mucho entre la muestra original y la validatoria por lo que sería posible que desapareciera como factor pronóstico.

Otro aspecto es la ***simplicidad/complejidad*** del modelo final y su efecto sobre la transportabilidad. La "navaja de Occam "sugiere preferencia por modelos pequeños más que por los grandes si "existe" realmente el modelo más pequeño. Para dirimir tal existencia hay criterios estadísticos que no son como hemos referido antes los definitivos. La existencia del modelo real más pequeño no está resuelta aún y puede que tenga una respuesta particular en cada investigación [32, 38]

La creación de grupos pronósticos es otra de las finalidades de este tipo de investigaciones. Altman y Royston sugieren que la creación de grupos pronósticos se haga mejor por criterios casi puramente clínicos y no por criterios estadísticos, especialmente en el caso de los estudios pragmáticos.Algunas técnicas como los árboles de regresión y clasificación (CART) y algunos modelos pronósticos simples con pocas combinaciones de valores pronósticos generan grupos de forma directa. Sin embargo la técnica CART no está diseñada de forma especifica para generar grupos pronósticos con utilidad clínica[32].

Cuantificando el funcionamiento del modelo. se pueden distinguir principalmente dos aspectos en este tema.

- La cantidad de información pronóstica que se relacione con una potencial utilidad clínica , y

- Si el modelo funciona de la misma forma con los datos nuevos que con los de la muestra original de donde nació.

Existen varias formas de medir la cantidad de información pronóstica. Un tipo de medida general aplicable a cualquier modelo es el método de Kullback-Leibler [39] que tiende a infinito a medida que aumenta la capacidad predictora. El PSEP es otro método [32]

Sobre los datos en los que se creó el modelo, el PSEP no es una medida de la cantidad de información pronóstica sino de la habilidad estimada del modelo para separar individuos en grupos pronósticos. Cuando se haga el análisis de un grupo de datos independiente se verá si esa estimación fue razonable. Cuando existen más de dos grupos pronósticos, es necesario asegurarse que exista un número aceptable de individuos en los grupos extremos para que existan suficientes eventos a la hora de calcular la P_{peor} y la P_{mejor}. Si no fuera así, sería mejor unir algunos grupos.

En los estudios de supervivencia, la PSEP es una función del tiempo de seguimiento. El clínico tendrá en su mente uno o varios tiempos de supervivencia y la PSEP se estimará a partir de la diferencia entre las curvas de supervivencia de Kaplan-Maier de los grupos pronósticos extremos en cada punto temporal [32].

La PSEP es una diferencia entre dos proporciones, también conocida como diferencia de riesgos en el contexto de un ensayo clínico aleatorizado. El inverso de la diferencia de riesgo es el *"número necesario de tratar"*, una medida de amplia utilización en el ensayo clínico y en la Medicina Basada en la Evidencia. Representa el número de pacientes que necesitan recibir un nuevo tratamiento para prevenir la producción de un efecto adverso adicional [39]. Por analogía, el recíproco del PSEP representa el número extra de pacientes en el grupo de más riesgo necesarios para generar un evento adverso adicional [32].

Se podría decir finalmente, que ***un modelo pronóstico multivariante debiera de ser creíble desde un punto de vista clínico, debiera tener generabilidad*** (es decir que pueda ser validado en otro lugar) y ***debiera demostrar su efectividad clínica***, en el sentido de que proporcione información adicional valiosa [32]

1.8 ¿COMO PODEMOS MEDIR LA GENERABILIDAD DE LA INFORMACIÓN PRONÓSTICA?

La **"generabilidad"** de un modelo pronóstico es lo que se conoce más comúnmente como ***validez externa***[40] Vamos a definir las relaciones entre generabilidad y exactitud y entre sus componentes. Trataremos el sistema pronóstico (por ejemplo, un modelo multivariante) como si fuera una *"caja negra"* de la que se analizan sus posibilidades.

EXACTITUD Y GENERABILIDAD.

Se trata de dos conceptos relacionados entre sí. La ***exactitud*** es el grado en el que las predicciones coinciden con los resultados y la ***generabilidad*** (validez externa) es la habilidad del sistema pronóstico para proporcionar predicciones exactas en una muestra diferente de pacientes.

COMPONENTES DE LA EXACTITUD.

Una serie de predicciones numéricas pueden ser inexactas de dos modos. La probabilidad pronosticada puede ser demasiado alta o demasiado baja (un error en la calibración) o puede también que la ordenación relativa del riesgo individual esté fuera de lugar (un error en la discriminación).

- ***Calibración*** Si la predicción se utiliza a nivel individual para aconsejar a un paciente, la exactitud de la probabilidad numérica (calibración) es un aspecto importante a tener en cuenta. En la investigación de servicios sanitarios es importante la calibración, por ejemplo en trabajos sobre mortalidad hospitalaria observada y esperada.
- ***Discriminación.*** Si por el contrario lo que se intenta conseguir es la estratificación de los pacientes en estadios de severidad de una enfermedad para la comparación de tratamientos diferentes, el aspecto más importante es la discriminación, que intenta estratificarlos de forma correcta en orden de riesgo.

La calibración y la discriminación se pueden medir de diferentes formas La calibración se mide mediante curvas que contengan los valores pronosticados frente a los valores observados[40]. La discriminación se suele medir mediante curvas de tipo ROC ("Area under the Receiver – Operating Characteristics")[41]. El área por debajo de la curva oscila entre 0,5 (no existe discriminación) y la unidad (discriminación perfecta) y refleja la probabilidad de todas las parejas posibles de pacientes en la que uno vive y otro muere, otorgando un riesgo algo más alto al que muere que al que sobrevive.

El área por debajo de una curva ROC se puede calcular directamente a partir de una tabla de resultados observados y pronosticados[40] .Se puede calcular también a partir de estimaciones pronósticas de tipo continuo (con o sin observaciones censuradas) mediante el estadístico C [42].

COMPONENTES DE LA GENERABILIDAD.

Si un modelo pronóstico multivariante que esté casi perfectamente calibrado y que además sea discriminatorio, no es capaz de predecir resultados fuera de su muestra original, no sirve. Para que un modelo pronóstico sea generalizable, la exactitud (es decir, la calibración y la discriminación) debe de ser reproducible y transportable.

- ***Reproducibilidad.*** Si el modelo refleja una exactitud parecida o igual a la obtenida en la muestra original en otra muestra proveniente de la misma población, se dice entonces que es reproducible. Una prueba que evalúe la reproducibilidad mide el grado en el cual el modelo se ajusta a los patrones reales de los datos mejor que al *ruido aleatorio*[40]. A medida que el número de eventos de interés por variable sea menor (sobreoptimización) el modelo tendrá más en cuenta el ruido aleatorio y será menos generalizable[40]. Los métodos para la evaluación de la reproducibilidad se basan en las técnicas de remuestreo, tales como el "bootstrap", que miden el grado de sobreoptimización. *El "bootstrap" puede medir errores en la discriminación y en la calibración, siendo de una especial utilidad en los casos en los que el modelo se desarrolla primariamente en una muestra pequeña*[40]

- ***Transportabilidad***. Un modelo puede que sea reproducible (es decir que funcione bien cuando se le aplique el "bootstrap") pero que se degrade de forma notable cuando se aplique a una o varias muestras siguientes de pacientes o individuos a causa de infraoptimización

La infraoptimización ocurre cuando una o varias predictoras faltan en el modelo final. Por ejemplo, un modelo para la neoplasia de mama que omita la presencia o ausencia de metástasis puede que "funcione" muy bien con pacientes sin metástasis pero puede que "decaiga " lamentablemente en una muestra más heterogénea de la misma enfermedad. El "bootstrapping" de la muestra primitiva puede que no sea capaz de detectar este problema, porque sea homogénea con respecto a la enfermedad metastásica. Pero por otra parte, la omisión de la variable que contenga información sobre la existencia o no metástasis es un error obvio.

La transportabilidad necesita que el modelo haga predicciones con exactitud en una muestra extraída de una población diferente a la primitiva pero con características similares. Los modelos con infraoptimización pueden mostrar una buena reproducibilidad pero no una adecuada transportabilidad.

Vamos a analizar a continuación cinco ***componentes de la transportabilidad*** que van a servir también para su medición objetiva.

- ***Transportabilidad Histórica***: requiere que el modelo mantenga un nivel de exactitud aceptable cuando se aplique a una cohorte en un tiempo histórico distinto. Es muy importante cuando la severidad del padecimiento analizado pueda cambiar a lo largo del tiempo. Esto puede ocurrir sobre todo cuando el diagnóstico se puede realizar en una etapa temprana de la enfermedad.

- ***Transportabilidad Geográfica***: requiere que se mantenga la exactitud del modelo pronóstico cuando se aplica a muestras de otras localizaciones.

- ***Transportabilidad Metodológica***: requiere que el modelo pronóstico mantenga un nivel de exactitud aceptable cuando se aplica a datos que se han recogido por métodos alternativos al que primitivamente generó el modelo. Charlson y colaboradores han propuesto que este suele ser el problema más frecuente cuando un modelo falla en su generabilidad[43]. Diferencias en la forma de definir las variables o de recoger los datos pueden originar diferencias substanciales en el funcionamiento del modelo. Si otros investigadores diferentes no pueden aplicar el modelo con exactitud se debe de dudar seriamente de su generabilidad.

- ***Transportabilidad de Espectro***: se refiere a la habilidad del modelo para realizar predicciones calibradas y discriminatorias en pacientes que se encuentren, en términos generales, en fases más avanzadas o menos avanzadas de la enfermedad en estudio o también que presenten una enfermedad parcialmente diferente. El nivel de severidad de la enfermedad en estudio está en relación con la frecuencia de presentación de la misma en la muestra. Un modelo pronóstico "perfecto" será capaz de manejar pacientes de acuerdo a su riesgo de muerte sin alterar su calibración en muestras con una frecuencia de presentación diferentes. La discriminación también se puede alterar, sobre todo cuando el modelo se ha

desarrollado en una muestra que contiene, por ejemplo, muchos pacientes con un estadio intermedio de la enfermedad, pero pocos en estadios más avanzados o menos avanzados.

- ***Transportabilidad con Período de Seguimiento***: requiere que el modelo mantenga su exactitud cuando las predicciones se realicen en períodos de seguimiento cortos o largos. Este tipo de transportabilidad está muy ligada a la transportabilidad de espectro[40] .Cambios en el espectro de la enfermedad y en el período de seguimiento pueden afectar a la prevalencia de presentación y por lo tanto alterar la calibración del modelo.

JERARQUIA DE LA VALIDACION PRONOSTICA.

Justice y colaboradores[40] propusieron una ***estructura jerárquica*** de cinco ***niveles*** para la validación externa de un modelo pronóstico. Cada nivel refleja los tipos de generabilidad que se han analizado para el modelo en cuestión de una forma acumulada y el grado de exactitud (calibración y discriminación) exhibido en estas pruebas.

Un modelo pronóstico jamás podrá ser totalmente validado de manera que el investigador nunca podrá estar absolutamente seguro de poder aplicarlo al siguiente paciente que se nos presente en la consulta, en un pase de sala o en una sesión clínica[40]. Si muestra niveles de exactitud aceptables en escenarios diferentes al primitivo, es más probable que se comporte en de la misma forma en cualquier entorno nuevo en donde se le analice. El esquema jerárquico que se presenta a continuación será por lo tanto ***un proceso iterativo y acumulativo***[40]. No tiene en cuenta la calidad intrínseca de cada trabajo de investigación.

- ***Nivel 0.Validación Interna***: Su objetivo consiste en analizar la exactitud del modelo en la muestra original en donde se desarrolló. Se emplean técnicas de exclusión de datos (muestra obtenida mediante corte aleatorio) o técnicas de remuestreo ("bootstrapping") mediante las cuales se estudia el grado de sobreoptimización del modelo[40,43]. En cierta forma se evalúa nada más que la reproducibilidad. No obstante la validez interna se considera un prerrequisito para la validez externa.

- ***Nivel 1.Validación Prospectiva***: Su objetivo es analizar la exactitud del modelo en datos muestrales recogidos después del desarrollo primitivo. Se lleva a cabo por lo general por el mismo investigador y en la misma institución sanitaria. Se evalúa básicamente la reproducibilidad y la susceptibilidad del modelo a cambios mínimos en un tiempo diferente (transportabilidad histórica). No se suele medir la transportabilidad metodológica ni la geográfica.

- ***Nivel 2.Validación Independiente***: Se analiza la exactitud del modelo en datos que hayan recogido investigadores diferentes y en lugares diferentes. Las muestras también se recogen en un tiempo histórico diferente. Este nivel de validación es importante porque los investigadores independientes probablemente tengan una "idiosincrasia" distinta a la hora de seleccionar la muestra y de recoger los datos.

- ***Nivel 3.Validación en Varios Lugares***: Mide de forma directa la transportabilidad geográfica y puede dar una idea de la transportabilidad metodológica. Se debe de informar siempre que se realice pues también da idea de la transportabilidad de espectro porque la expresividad clínica de la enfermedad puede ser diferente dependiendo del lugar.

- ***Nivel 4.Validaciones Independientes Múltiples***: Se mide la exactitud del modelo pronóstico cuando lo utilizan varios investigadores en lugares diferentes. El grado de transportabilidad metodológica del que informa es mayor que en el caso de la validación independiente única, aunque se lleve a cabo ésta en lugares diferentes.

- ***Nivel 5.Validaciones Independientes Múltiples con Períodos de Seguimiento Variables***: Si un modelo pronóstico es capaz de conservar el grado de exactitud durante varios períodos de seguimiento con varios investigadores y en lugares diferentes, posee más generabilidad que el medido sólo en el nivel 4 [40] Aunque es un método complejo, si se pueden realizar tablas de vida en vez de un punto de corte único a los 5 ó 10 años, es posible calcular la discriminación y la calibración del modelo en cualquier punto intermedio de interés.

Estos cinco niveles jerárquicos creemos que son muy útiles a la hora de discutir un trabajo de investigación validatorio[40]

1.9. ASPECTOS ESTADÍSTICOS DE LA VALIDACIÓN DE MODELOS PRONÓSTICOS.

Si se violan las *asunciones de tipo matemático* puede que un modelo resulte inadecuado para su utilización pronóstica. Por ejemplo, si no existe linealidad cuando se emplea un modelo lineal, o si se omiten variables predictoras de importancia, o si existe una elevada frecuencia de datos perdidos, o si se usan métodos de imputación que no sean los adecuados o si se produce una sobreoptimización, siendo esto último más frecuente en muestras pequeñas[44].

LOS PASOS PRELIMINARES: VALORES PERDIDOS E INTERACCIONES

Los métodos más simples para la *imputación* de valores perdidos incluyen el uso de la mediana, la media aritmética o la moda. Sin embargo esta forma de trabajar comporta mucho sesgo. Son mejores los *métodos de imputación* a variables predictoras basados en modelos de regresión. Tener en cuenta las *posibles interacciones* es también un dato muy importante a tener en cuenta. Por ejemplo:

1. Interacciones entre el tratamiento y la severidad de la enfermedad que está siendo tratada. Los pacientes con poca evolución tienen menos oportunidad de recibir un beneficio.

2. Interacciones que afecten a la edad y a los factores de riesgo. *Los individuos más ancianos son los que potencialmente se afectan menos por los factores de riesgo*. Han sido lo suficientemente fuertes como para sobrevivir hasta su edad estando presentes los factores de riesgo.
3. Interacciones que afecten a la edad y al tipo de enfermedad. Algunas enfermedades son incurables y tienen el mismo pronóstico independientemente de la edad del individuo.

4. Interacciones entre una medida y el estado del individuo en el momento de esa medida. Por ejemplo, la medición de la función ventricular izquierda en reposo puede ser menos predictiva que la medición realizada en estrés.

5. Interacciones entre la calidad y la cantidad de los síntomas.

6. Interacciones entre las variables que presenten las probabilidades más significativas de sus OR en un modelo confeccionado mediante regresión logística.

La optimización final del modelo debe de tener en cuenta todos estos aspectos de posibles interacciones de variables [44].

VERIFICACION DE LAS ASUNCIONES DEL MODELO. LINEALIDAD.

En su forma más simple, todos los modelos comunes de regresión asumen que para una cierta escala de Y, cada variable predictora X está linealmente relacionada con Y. En el modelo de regresión logística con respuesta binaria [44, 34], la asunción inicial es que X está relacionada linealmente con el logaritmo natural de la odds de la respuesta, o sea con:

$$(\log [P / (1 - P)]$$

Donde P es la probabilidad de respuesta, para pacientes agrupados por valores de X. Una manera de medir la asunción lineal es expansionar los valores de X elevándolos al cuadrado e introduciéndolos en el modelo.

ASUNCION DE ADITIVIDAD.

Otra asunción necesaria en la mayoría de los modelos de regresión es *la aditividad de efectos de las predictoras en ausencia de interacción*. Si existiera interacción se deben de introducir en el modelo mediante términos de productos cruzados.

La búsqueda de interacciones se puede hacer de varias formas. Por ejemplo en un modelo con las variables predictoras edad, sexo y dosis se pueden ensayar todas las interacciones que afecten a la edad, al sexo y a la dosis de forma combinada. También se pueden realizar interacciones de segundo orden. Si después de todas estas maniobras no se encuentran significaciones, es lógico pensar que no haya interacciones en el modelo [44].

CUANTIFICACION DE LA EXACTITUD PREDICTIVA

La medición de la exactitud puede tener las utilidades siguientes:

- Para cuantificar la utilidad de una variable o de un modelo completo para predecir o para despistar sujetos con un elevado riesgo de la enfermedad sometida a estudio. A menudo nos hemos de conformar con designar un modelo como "*mínimamente aceptable*" desde un punto de vista estadístico y en otros muchos casos solo es posible juzgar la exactitud de un modelo con respecto a la de otro [44].

- Para comprobar si un modelo obtenido está **sobreoptimizado** (con "ruido" de optimización resultante de unos coeficientes de regresión inestables) o **infraoptimizado** (por una especificación no apropiada del modelo, por omisión de predictoras o por infraoptimización estadística).

- Para comparar métodos o modelos competitivos.

En el caso de que la variable respuesta sea de tipo continuo y que se haya medido de manera completa (y no como en el caso de las mediciones censuradas determinadas por el cese del seguimiento antes de que todos los sujetos hayan desarrollado la variable de salida o de interés) la medida más frecuente de la exactitud predictiva es el ***error cuadrático esperado*** de la estimación ("expected squared error") [44].

Existen también otros dos términos utilizados en estos casos: ***la calibración y la discriminación***. Ambos han sido definidos en los capítulos anteriores de esta Tesis. Cuando la variable respuesta es dicotómica y las predicciones se expresan como probabilidades de que ocurra un evento, la calibración y la discriminación son más útiles que el error cuadrático [44].

Una forma de medir la calibración de las probabilidades es hacer grupos de individuos y verificar si existe sesgo comparando las respuestas predichas y las observadas .Otra forma es la utilización de un ***"alisador"*** ("smoother") [44].

CONSTRICCION ("SHRINKAGE")

La constricción es el aplanamiento de la curva de datos (predichos u observados) causado por sobreoptimización. Es un concepto relacionado con la regresión a la media [34]. Se puede calcular la magnitud de constricción presente en un modelo mediante validación externa o la magnitud probable del mismo mediante "bootstrapping" o por técnicas de validación cruzada [44].

INDICE DE DISCRIMINACIÓN GENERAL.

La discriminación se puede definir de manera más unívoca que la calibración. Se puede medir con una medida de la correlación sin el requisito de la formación de subgrupos o del alisamiento.

El ***índice c*** (de "concordance") es una medida ampliamente utilizada de la discriminación predictiva tanto en variables continuas, dicotómicas u ordinales. Está relacionado con el rango de correlación entre las observaciones predichas y las observadas[44]. *Se define como la proporción de todos los pares de pacientes en los cuales las predicciones y las frecuencias observadas son concordantes.*

Para la predicción de variables binarias (presencia o ausencia de enfermedad por ejemplo), el índice **c** se reduce a la proporción de todos los pares de pacientes, unos con y otros sin enfermedad, en los cuales el paciente portador de la enfermedad tuvo la probabilidad predicha más alta de enfermedad. En este caso el índice c es esencialmente el estadístico de Wilcoxon - Mann - Whitney para la comparación de predicciones y es idéntico al área bajo una curva ROC [44]

. METODOS DE VALIDACIÓN DE MODELOS.

La sobreoptimización del modelo primitivo es la causa más frecuente de que un modelo clínico falle a la hora de la validación externa [44]. *Los métodos estadísticos principales para medir la exactitud interna del modelo, cuando*

aplica a una población diferente de la primitiva, son: ***el "corte" de datos ("data splitting")*** [45] ***, la validación cruzada*** [46] ***y el "bootstrapping"*** [46, 47]

En el *"corte" de datos* se selecciona al azar una parte de los datos (por ejemplo los 2/3) y se realiza con ellos todo la labor analítica del modelo (transformación de datos, selección de variables por el método "stepwise", prueba de la posibles interacciones, estimación de los coeficientes de regresión ...[6]. El modelo así obtenido se "congela" y se aplica al resto de la información desechada en la primera selección para computar los estadísticos de calibración, el índice c , etc.

La *validación cruzada* es un "corte" de datos repetido. Para obtener una estimación de cierta exactitud al utilizar este método se necesita evaluar al menos unos 200 modelos[46]. El tamaño muestral ha de ser mayor que para el "corte" de datos. La validación cruzada reduce la variabilidad al no detenerse sobre un solo corte de datos [44, 45, 46] Efron demostró que la validación cruzada es ineficiente debido al alto grado de variación de la exactitud cuando se repite el proceso completo. Con el "corte de datos" los índices de exactitud varían aún más al realizar varios "cortes" [44, 46].

El ***"bootstrapping"*** es el método alternativo para la validación interna de un modelo predictivo [44,46,47]. *Se basa en el muestreo por reemplazamiento de los hallazgos originales*. Posee la ventaja añadida de que se emplean todos los datos para el desarrollo del modelo, no desperdiciándose nada de la información disponible[44,46,47] .Los datos recogidos son demasiado "preciosos" como para desecharlos activamente dentro de una investigación [44].

La ***investigación de la estabilidad*** de un modelo de regresión obtenido, es un aspecto fundamental dentro del análisis de datos [48]. El énfasis más importante ha de ponerse en la selección de las covariables (variables explicativas) que ejerzan influencia en la variable respuesta, por lo tanto el término estabilidad comporta un significado de repetitividad [49]. En este aspecto, el "bootstrapping" es una técnica alternativa al corte de datos ("data splitting") y al "jackknife"[48]. La inestabilidad del método de selección denominado "stepwise" ha sido demostrada por diversos autores [48, 50 .51, 52].

EL "BOOTSTRAP" COMO ESTRATEGIA DE SELECCIÓN.

Descrito por Efron [46,47,53] , consiste en la creación un número elevado de "repeticiones" submuestrales tomadas de la muestra original y tratarlas como "independientes". En cada repetición se intenta identificar la variable que tenga influencia en la variable respuesta. Para cada repetición de "bootstrap" se utiliza un método de selección para identificar las variables "significativas". En el caso más simple, cada variable posee un indicador para la inclusión o la exclusión del modelo seleccionado.

Las variables con un valor pronóstico importante se seleccionarán en la mayoría de las repeticiones "bootstrap". Se asume que cada repetición, al ser una muestra aleatoria de la muestra original, reflejará la estructura subyacente de los datos[8]. Por lo tanto, ***el porcentaje de inclusiones de una variable en el modelo "bootstrap" será un criterio de la importancia pronóstica de esa variable*** [48].

Haciendo una descripción algo más profunda del método, podemos comentar el denominado *"hipermodelo"* [48] en el cual se supone que los coeficientes de regresión siguen una cierta distribución con una punta en el cero. Esto significa, desde un punto de vista matemático, que cada componente del vector de los coeficientes de regresión se asume que son idénticos a cero con una probabilidad positiva o que provienen de una distribución común de forma independiente. Algunos autores consideran al hipermodelo como un punto de partida para el desarrollo de estrategias de tipo Bayesiano para la selección de variables [44,46].

Otro aspecto, digno de ser tenido en cuenta, tanto por sus implicaciones prácticas como por el entorno teórico del problema, es que cuando hay sólo una variable para ser incluida en el modelo, existe una relación directa entre el nivel de selección y la fracción de inclusión en el "bootstrap" [48]. La precisión de esta estimación puede ser la que deseemos tan sólo aumentando el número de repeticiones por lo que podemos igualar la fracción de inclusión en el "bootstrap" con el poder muestral [48].

Finalmente también es de importancia la decisión de cual es el punto de corte del grado porcentual de inclusión en el modelo para distinguir las variables que han de quedarse y las que han de salir. Obviamente depende de los límites que hayamos impuesto a la investigación que estemos realizando. Por ejemplo, elegiremos un punto de corte muy alto si lo que queremos distinguir son factores

de pronóstico muy fuertes, incluyendo tan sólo las variables que se repitan un porcentaje de veces muy alto[48] . El ***"bootstrap"*** es quizás el ***método más eficiente para la validación interna*** de un modelo pronóstico [44,47].

1.10. *APRETARSE BIEN LOS CORDONES DE LAS BOTAS Y ... ¡SALTAR!*

El escritor alemán Rudolph Raspe publicó en el año 1785 "Las narraciones del Barón de Munchuasen y sus maravillosos viajes y campañas en Rusia". En esa obra refería un episodio en que el protagonista lograba salir de un lodazal en donde se hallaba retenido con su caballo al "apretarse los cordones de sus botas y saltar hacia fuera..." ("pulling his bootstraps up ..."). El significado exacto de "bootstraps" es el de unos cordones enlazados una o dos veces en la parte superior de la bota de montar a caballo [54]. El contenido de esa imagen literaria sustenta el significado del término inglés "bootstrapping" que no sólo se utiliza en estadística sino también en informática, en economía, en biología, en electrónica e incluso en lingüística[55].

El ***núcleo semántico común*** para todos estos usos es el de obtener una acción compleja mediante la realización de un gesto sencillo (un individuo y su caballo pueden dar un gran salto después de que tan sólo el jinete se haya tirado de los cordones de las botas).

Como hemos referido anteriormente, el "bootstrapping" es un método estadístico diseñado para la evaluación de la distribución muestral de un estimador mediante remuestreo con reemplazamiento[55]. Fue "inventado" por Bradley Efron [55,56,57,58,59] y desarrollado posteriormente por él mismo junto a Tibshirani[60] Realizar un "bootstrap" es crear una serie de muestras "posibles" a partir de la real. La técnica que se emplea es el ***remuestreo con reemplazamiento***.

Llevando a cabo el análisis estadístico de todas las muestras posibles (**"población virtual"**) se pueden extraer conclusiones de cómo se distribuye el estimador obtenido a partir de la muestra original. Mientras que el objetivo de la validación cruzada era la verificación de la replicabilidad de los resultados y el del "jacknife" era la detección de los valores sobresalientes ("outliers"), Efron diseñó el "bootstrap" con el objetivo de realizar inferencias[55].

En el "bootstrap", la muestra original se puede duplicar tantas veces

como lo permita la capacidad del ordenador y del programa utilizado. La muestra expandida se trata como una "población virtual" extrayéndose nuevas muestras a partir de ella y observando el comportamiento del estimador dentro de estas. La fuente para el remuestreo en el "boostrap" es obviamente mucho mayor que en la validación cruzada y en el "jacknife".

Otra diferencia estriba en que el "bootstrapping" emplea remuestreo con reemplazamiento que es más exacto en términos de probabilidad de simulación. Otra ventaja es que en la validación cruzada y en el "jacknife" el tamaño de la submuestra es menor que el de la muestra original, mientras que cuando utilizamos "bootstrapping" cada submuestra tiene el mismo tamaño que la muestra original. Por lo tanto el bootstrap posee la ventaja de modelizar los impactos que se pueden producir sobre el tamaño de muestra verdadero [61], [62]

VENTAJAS DE LAS TÉCNICAS DE REMUESTREO.

En primer lugar hay ventajas de tipo empírico. Los procedimientos basados en el muestreo convencional se fundamentan en distribuciones teóricas que necesitan unas fuertes asunciones previas al análisis tanto en la muestra como en la población. Existen también ventajas de tipo conceptual, al ser el remuestreo un método limpio y claro que no necesita de un soporte matemático sofisticado [61].

Las asunciones previas que necesitan las distribuciones estadísticas teóricas se suelen cumplir sólo en los tamaños de muestra grandes. Cuando el tamaño de muestra es pequeño y no satisface las asunciones paramétricas, el remuestreo es un recurso bastante útil [56, 61]. Sin embargo, Good[63] estableció que el test de permutación esta sujeto todavía al *problema de Behrens-Fisher*, en el cual la estimación se considera problemática cuando la varianza poblacional es desconocida. Para ser correctos, los tests de permutación asumen todavía varianzas iguales que es también una premisa requerida en las pruebas clásicas de muestreo[61]. El problema de Behrens-Fisher consiste en el estudio de la igualdad de las medias de dos distribuciones normales que no tienen la misma varianza. Se han propuesto varios estadísticos de prueba no siendo ninguno completamente satisfactorio.

Los procedimientos tradicionales requieren de muestreo aleatorio para

validar las inferencias desde la muestra a la población general. Edgington[64] demostró que el remuestreo es válido para cualquier tipo de datos, tanto aleatorios como no aleatorios. Lunneborg[65] sugirió que aunque la utilización de muestras no aleatorias en el remuestreo puede que no conduzca a una conclusión de tipo inferencial, al menos el trabajar con muestras no aleatorias nos puede informar más acerca de las características locales de los datos y de la estabilidad de los resultados.

Si disponemos de una muestra pequeña, aunque su estructura cumpla los requerimientos paramétricos, puede que carezca del suficiente poder. El "bootstrapping" puede tratar una muestra pequeña como una población virtual para generar así más observaciones. No existe un tamaño de muestra mínimo en remuestreo. El caso extremo puede ser el de un tamaño de muestra de dos observaciones. Esta técnica de análisis puede generar múltiples muestras con un tamaño igual a 2. Como hemos referido antes, el propósito del remuestreo es ***la simulación de probabilidad.***

Los procedimientos clásicos de muestreo no informan a los investigadores de que probabilidad tienen los resultados de ser replicados [61]. Los estudios repetidos mediante remuestreo con validación cruzada o con "bootstrapping"pueden utilizarse como **replicaciones internas** [54,56].Las replicaciones son esenciales para un método estadístico tan clásico como es la regresión múltiple [61].

CRÍTICAS AL REMUESTREO.

Algunos metodólogos son escépticos con el remuestreo por las razones que a continuación exponemos. Se puede aducir que mediante estas técnicas se utilizan los mismos números una y otra vez hasta conseguir una respuesta que no se puede conseguir de ninguna otra forma[61]. En cierta manera se esta asumiendo algo que puede estar oculto y que se descubra después. Cualquier método que empleemos está construido sobre algún tipo de asunción previa y requiere "un salto cualitativo de fe" en mayor o menor grado. De hecho los métodos tradicionales necesitan más asunciones previas que los métodos de remuestreo.

Otros autores discuten sobre la capacidad de generalización del remuestreo al

estar basado este último en el análisis de una sola muestra [66]. En este sentido cabe decir que Fan y Wang[62] establecieron que el estudio de la estabilidad de los resultados del test es descriptivo y no inferencial para el estudio de la naturaleza.

Bosch[61] describió que los intervalos de confianza obtenidos mediante "bootstrapping" simple estarán siempre sesgados a pesar de que este sesgo disminuya con el tamaño de muestra. Si la muestra proviene de una población normal, el sesgo en la magnitud del intervalo de confianza es al menos n / n-1, en donde "n" es el tamaño muestral. Se puede disminuir ese sesgo realizando "bootstraps" más complejos.

Otros autores han señalado que si los datos recogidos están sesgados, el remuestreo se encargará de repetir y magnificar el mismo error. Rodgers[67] admitió que efectivamente la magnificación de los aspectos más inusuales de la muestra es uno de los problemas más importantes para la validez de las conclusiones de la aplicación del remuestreo. También es cierto que si el investigador descubre que sus datos están sesgados es porque conoce los atributos de la población de la que proviene En términos generales, la población es finita en tamaño y desconocida en su distribución, por lo que es realmente difícil asegurar que los datos muestrales obtenidos sean malos[61]. En todo caso, si los datos estuvieran sesgados, los procedimientos de muestreo clásico tendrían el mismo problema. Mientras que las replicaciones del remuestreo podrían aliviar en cierta forma este hecho, el muestreo clásico no podría aportar remedios.

Otro aspecto que algunos cuestionan, es la exactitud de las estimaciones por remuestreo. Si no se realizan el suficiente número de ensayos experimentales, el remuestreo puede ser menos exacto que los métodos paramétricos de muestreo convencional. Parece que este no es un argumento convincente cuando el poder de los computadores actuales los hace capaces de realizar miles de millones de simulaciones [61]. Por ejemplo, mediante el StatXact, un programa para tests exactos, el investigador puede configurar el remuestreo utilizando un máximo de memoria RAM para 1000000000 muestras sin límite temporal.

En todo caso, tanto los investigadores que están de acuerdo como aquellos que critican las técnicas de remuestreo tienen algo de razón [61]. En los métodos clásicos y en el remuestreo existen pros y contras. Lo adecuado

de la metodología a emplear depende mucho de la situación. Por ejemplo, Noreen [68] apuntó que si la población cumplía las asunciones que se derivan de la distribución muestral, el método a emplear debe de estar basado en las pruebas paramétricas convencionales y no en ninguna otra. Las técnicas de remuestreo como el bootstrapping han supuesto en la investigación estadística y epidemiológica un avance parecido al que existió en la biología molecular con el desarrollo de las técnicas de ADN recombinante mediante las cuales se podían estudiar en el laboratorio secuencias de ADN no existentes en la realidad[69]

1.11 CARCINOMA DE COLON ESPORADICO, LIPIDOS PLASMÁTICOS Y MARCADORES TUMORALES.

Desde el trabajo clásico de Rose y cols[70] sobre las relaciones del colesterol plasmático y la neoplasia maligna de colon han existido múltiples referencias bibliográficas a favor y en contra de esta asociación[71,72,73,74,75,76,77,78,79,80,81]. En la actualidad no se puede afirmar una relación clara de asociación entre la aparición de un carcinoma colorrectal esporádico (CCRE) y la disminución del colesterol plasmático o de algunas de sus fracciones, tampoco se han discriminado grupos diferenciados de pacientes con CCRE y la existencia del referido marcador lipídico. Cuando hablamos de grupos diferenciados nos referimos tanto a características genéticas como a clínicas.

Sobre los marcadores tumorales sí que existe una abundante bibliografía a favor de su valor pronóstico tanto en estadíos pre-clínicos como en fase ter apéutica[82,83,84,85.,86] El antígeno cárcinoembrionario **(CEA)** es una glicoproteína presente en el plasma en muy pequeñas cantidades (del orden de nanogramos) que se eleva ante la existencia de adenocarcinomas ocultos. Está bien descrita su utilidad en el carcinoma colorrectal[82,83] tanto en su fase diagnóstica como en el seguimiento clínico[84]. Los grandes fumadores también pueden poseerlo elevado [82,83,84.]

El **CA 19.9** es un antígeno asociado a tumor que se haya presente en tejidos que contengan mucina o en la circulación y que se localiza en el antígeno sializado del grupo sanguíneo Lewis A[85,87.] Los individuos con genotipo Lewis a-b no pueden sintetizar este antígeno (un 5% aproximado de la población general). Fue utilizado en primer lugar para el diagnóstico y seguimiento del

carcinoma de páncreas pero ha demostrado también su utilidad en el CCRE[85]. También se han podido observar valores elevados en casos de carcinoma de estómago, carcinoma de la vesícula y/o del tracto biliar y de hepatomas. No se le ha considerado hasta ahora como un instrumento válido para el despistaje de CCRE por su baja sensibilidad.

Nosotros hemos publicado un trabajo[87] sobre las relaciones que pudieran existir entre ambos tipos de substancias **en el momento de la aparición clínica de CCRE** Mediante un diseño de casos y controles y aplicando regresión logística multivariante obtuvimos un modelo predictor-explicativo con seis variables dependientes: edad del paciente en años, colesterol total (CT), colesterol HDL (C-HDL), colesterol VLDL (C-VLDL), fosfatasa alcalina (FA) y el marcador tumoral CA 19.9 (CA 19.9). Este original fué consecuencia de otro publicado también por nosotros en la misma línea[71], siendo ambos favorecedores de la hipótesis asociativa entre la disminución de lípidos plasmáticos y la aparición de un CCRE. En ningún momento presumimos una asociación causal.

Basándonos en el estadiaje clásico de Dukes[88] y realizando mediciones pre-quirúrgicas para evitar el estrés metabólico de la intervención[89] obtuvimos las seis variables predictoras antes referidas que son las que intentamos validar en esta Tesis. Las técnicas de medición fueron las estándar[90]. No se encontraron diferencias significativas entre casos y controles para las variables plasmáticas que definen el estado nutricional[91] (proteínas totales, albúmina, inmunoglobulinas, complemento, hemoglobina, patrones corpusculares).

La hipocolesterolemia, no obstante, se considera un marcador de desnutrición[92,93] debido a su disminución de síntesis hepática y a la disminución de secreción de lipoproteínas en individuos con desnutrición severa[93]. Otros factores metabólicos no directamente relacionados con el estado nutricional y que pudieran contribuir a la hipocolesterolemia son el hipercatabolismo de partículas lipoproteicas, una malabsorción intestinal o la extravasación de lipoproteínas séricas hacia el espacio extravascular[94].

Es interesante reseñar que si la disminución de CT y de HDL se debiera tan sólo a causas nutricionales en los pacientes con CCRE o con lesiones preneoplásicas[71,75,81] ésta se produciría más rápidamente que la disminución de la albúmina y de las otras fracciones proteicas al menos en nuestra experiencia[71].

De ahí el interés de esa posible asociación para un diagnóstico pre-clínico de la neoplasia por ejemplo en atención primaria[71,87]

Ante la existencia de información bibliográfica tanto a favor[71,75,81,95] como en contra[78,79,96] del espectro lipídico como factor predictivo en el momento biológico del diagnóstico de un CCRE, nosotros hemos persistido en nuestro esfuerzo investigador estudiándolo analíticamente junto al CA 19.9 y la fosfatasa alcalina (FA)[84,86] tratando de validar el modelo obtenido anteriormente.

2. PREGUNTAS DE INVESTIGACION

Nuestro interés en esta Tesis ha sido **en primer lugar** el intentar validar el modelo obtenido anteriormente[87] con el nuevo tamaño de muestra. Nuestra primera pregunta ha sido por lo tanto *¿Sigue funcionando nuestro modelo multivariante con los nuevos casos y controles?* o mejor dicho *¿Qué grado de funcionalidad presenta?* En su versión primitiva contenía seis variables predictoras (edad en años, colesterol total, HDL-colesterol, VLDL colesterol, fosfatasa alcalina y el marcador tumoral CA 19.9) siendo la variable resultado el ser caso de carcinoma colorrectal esporádico o control[87]

En segundo lugar nuestra intención fue *comparar el sesgo de selección de la submuestra hospitalaria con el de la submuestra recogida en atención primaria* por médicos de familia primordialmente. Por naturaleza, el sesgo de selección hospitalario más clásico es el sesgo de Berkson[1] descrito hace muchos años por este autor y aceptado desde entonces como un pilar de la epidemiología contemporánea[10,11]

Nuestra intención no era evidentemente suplantarlo de semejante estatus, era tan solo someterlo a una crítica constructiva y arrojar sobre él si fuera posible alguna sombra de duda con las características metodológicas de nuestro trabajo. Nuestra segunda pregunta ha sido por todo ello *¿Existe diferencia entre el modelo de regresión logística construido con controles hospitalarios con aquél construido con controles de atención primaria?*

Según el trabajo de Feinstein et al[19] debiera de existir diferencia si funciona el sesgo de Berkson en nuestra investigación pues los diseños casos-control que toman controles hospitalarios tienden a disminuir la OR mientras que en aquellos en los que se toman los controles de la comunidad ocurre lo contrario[19] . Un modelo mixto debería de poseer unas OR equilibradas, al menos desde un punto de vista estadístico.

Contábamos de partida con los datos del trabajo anterior[87] que constaba de 53 casos de neoplasia colorrectal esporádica y de 40 controles hospitalarios. A ellos y durante la presente investigación se han unido más de trescientos nuevos registros entre casos y controles tanto de atención primaria como hospitalarios. A los controles se les ha ido recogiendo las variables incluidas en el modelo multivariante primitivo así como también su origen hospitalario o de atención primaria.

Los casos que engrosaban los modelos comparativos eran los mismos en ambos pues la información sobre ellos tuvo que recogerse siempre a nivel hospitalario. Aunque también nos planteamos la posibilidad de hacer comparaciones con casos diferentes para los controles hospitalarios y los de primaria pero con un tamaño muestral menor obviamente.

Con posterioridad apareció **una tercera pregunta de investigación** que por su importancia no hemos querido desgajar de esta Tesis. El problema de la inducción (validación externa) de los resultados de los observados en una muestra a los que en realidad concurren en el universo está prologado por una buena validación interna. Es por ello que hemos decidido analizar esa vertiente con una técnica estadística sofisticada como es el "bootstrapping" basada en el remuestreo probabilístico[53]. Por lo tanto nuestra tercera pregunta ha sido *¿Tienen realmente validez interna nuestros resultados?*

Con esta última cerramos el círculo inquisitivo de esta Tesis, pues si la definición más actualizada de sesgo es la falta de validez interna[10,11] lo estaríamos explorando entonces de dos formas, una mediante el análisis de OR en dos diseños multivariantes diferentes (sesgo de Berkson[19] y otra mediante un riguroso estudio de validez interna (otros tipos de sesgo de selección)[10]

El ***orden lógico*** de esta serie de cuestiones sería:

1. Validación interna de los resultados muestrales ("bootstrapping").

2. Validación externa del modelo emanado de nuestro trabajo anterior con el nuevo tamaño muestral (regresión logística multivariante de la información "cruda").

3. Comparación del modelo logístico construido con controles de atención primaria con aquel otro construido con controles hospitalarios para discernir si existe sesgo de Berkson en nuestros resultados siguiendo las conclusiones de Feinstein et al[19]

3. METODOLOGIA

3.1 DISEÑO EPIDEMIOLÓGICO.

El estudio se ha construido en base a un diseño ***casos - control no pareado***, intentando alcanzar la proporción 1:3, para mejorar la eficiencia del estudio primitivo cuya proporción no llegaba a la 1:1[87]

3.2 RECOGIDA DE LA INFORMACIÓN : CRITERIOS.

Los casos y controles nuevos se han ido recogiendo durante un período de aproximadamente tres años en diferentes hospitales y centro de salud de Andalucía Occidental y Extremadura tanto urbanos como rurales. Los investigadores que han recogido la información en los centros de salud (*controles*) han sido médicos de familia con formación MIR y un médico general sin ella pero con una amplia experiencia en atención primaria. Todos tenían un *desarrollo comunitario* de más de tres años sobre sus cupos respectivos.

Los investigadores que han recogido la información a nivel hospitalario (*casos y controles*) han sido especialistas y residentes en medicina interna, neurología, inmunología y alergia y farmacología clínica y también médicos de familia en formación. A todos ellos se les explicaron los objetivos del trabajo y se les facilitaron hojas de registro que contenían los criterios de inclusión y de exclusión. Los investigadores principales (IP) han sido dos médicos de familia con formación MIR.

Los controles tanto en primaria como a nivel hospitalario han sido reunidos de forma prospectiva. Tan sólo un control hospitalario (CSVR) fue retrospectivo. Los controles de primaria se han recogido en los centros de salud siguientes: **Pilas** (Sevilla-Rural), **Camas** (Sevilla-Rural), **Huerta del Rey** (Sevilla-Urbano) y **Mérida** (Badajoz-Rural).

Los casos pertenecientes a esta nueva muestra se han recogido de forma retrospectiva en los archivos de los **Hospitales Virgen Macarena, Virgen del Rocío de Sevilla, Hospital General de Mérida** y también en el **Hospital Juan Ramón Jiménez de Huelva**, consultando las historias clínicas elegidas de forma aleatoria entre los últimos cinco años (2000-2004). Se ha intentado evitar en lo posible la "información perdida". La muestra "primitiva" constaba de 40 controles hospitalarios y 53 casos que fueron recogidos todos en su día de forma prospectiva[87].

Los criterios de inclusión y de exclusión utilizados en esta parte del estudio han sido los mismos que para la primera parte de esta investigación[71]. Los **criterios diagnósticos de inclusión** para los **casos** fueron la endoscopia y la biopsia ; **los de exclusión**: la existencia de metástasis a distancia , un trastorno severo del metabolismo lipídico , la coexistencia con otra neoplasia, los síndrome polipoideos hereditarios , el cáncer colorrectal hereditario sin poliposis, la enfermedad inflamatoria intestinal, las neoplasias no epiteliales y las inmunodeficiencias. En consecuencia todos los casos estaban como máximo en el estadio II A de Dukes[88]

Para los **controles**, el **criterio de inclusión** fue la ausencia de CCRE. Los **criterios de exclusión** fueron: cualquier tipo de enfermedad neoplásica maligna, la existencia de lesión pre-maligna colorrectal, un trastorno severo del metabolismo lipídico y las inmunodeficiencias. No se realizó enema opaco ni colonoscopia en los controles. A los dos años de la selección de los controles en atención primaria se realizó un rastreo telefónico por si alguno hubiera desarrollado un CCRE.

3.3 TÉCNICAS BIOQUÍMICAS.

El colesterol total se midió mediante el sistema RA TECNICON. El HDL colesterol fue medido por el método del reactivo precipitante. En la muestra primitiva el LDL colesterol se calculó mediante la formula de Friedwald [LDL = CT – HDL – TG / 5]. Las VLDL también se calcularon mediante la fórmula de Friedwald [VLDL = TG / 5]. Los TG se determinaron mediante mediante el test enzimático colorimétrico consistente en la hidrólisis enzimática de los TG y la medición posterior del glicerol mediante colorimetría[90]

El CA 19.9 (antígeno carbohidratado del grupo sanguíneo Lewis dializado) se determinó también mediante una técnica de "sándwich" semejante a la utilizada en la medición del CEA[82,83,84,85,86]

3.4 LIMITES TEMPORALES.

La información recogida en esta Tesis comenzó a ser recopilada en 1992 y ha terminado en el 2003.

3.5 TAMAÑO MUESTRAL

El tamaño muestral definitivo (n=401) se ha obtenido uniendo la muestra primitiva con la muestra multicéntrica "nueva" recogida en este trabajo.

3.6 CONTROL DE CALIDAD

Se ha llevado a cabo por dos investigadores diferentes (IP) con especial interés en los registros recogidos por los diferentes colaboradores, apreciando la exactitud de la información vertida en el paquete de datos. Fruto de ese examen cualitativo **se desecharon** un total de 9 controles y 3 casos en el momento final de la construcción del paquete. La causa fundamental fue la falta de cumplimiento de criterios de inclusión.

3.7 CONSTRUCCION DEL PAQUETE DE DATOS.

Se realizó el ensamblaje del paquete de datos anterior con un total de 93 registros (53 casos y 40 controles) en formato DBase IV al paquete de datos nuevo en formato EXCEL y con un total de 308 registros. El paquete en formato EXCEL se transformó en formato SPSS para su ulterior análisis estadístico, realizándose también controles de calidad en este paso.

3.8 ANALISIS ESTADISTICO DESCRIPTIVO.

Se realizó un estudio descriptivo inicial sobre el conjunto de registros resultante para obtener medidas de centralización y de dispersión. Se despistaron valores sobresalientes ("outliers") como parte final del control de calidad de los datos, teniéndolos presentes pero sin que constituyeran un criterio de exclusión. Se llevó a cabo un estudio de normalidad de los valores observados tanto en los controles como en los casos mediante la prueba de Kolmogorov-Smirnov[97] (SPSS. Versión 12.0)[98,99]

3.9 ANALISIS ESTADISTICO ANALITICO.

Se realizó un estudio bivariante mediante la prueba de U-Mann-Whitney[97] Se llevó a cabo un estudio de regresión logística (RL) no condicionada multivariante

partiendo del modelo obtenido en nuestro estudio anterior[87] con la variable ser caso o control como dependiente y las variables edad en años (EDAD), colesterol total (CT), fracción HDL (HDL), fracción VLDL (VLDL), fosfatasa alcalina (FA) y el marcador CA 19.9 (CA 19.9) como predictoras. Un primer análisis se llevó cabo sobre el paquete de datos digamos ***"crudo".*** La selección de variables fue siempre hacia atrás ("backward").

3.10 EVALUACIÓN DE LA INFORMACION PERDIDA ("MISSING VALUES").

En las variables en las que la información perdida superase el 20% decidimos imputar valores mediante el Programa SPSS (interpolación lineal).

3.11 EVALUACIÓN DE LA LINEALIDAD.

Previamente a la aplicación de la RL como prueba de análisis estadístico era necesario estudiar la relación lineal entre las variables. Primero se llevó a cabo una exploración visual de las nubes de puntos bivariantes y después se realizó un test de correlación de Spearman (prueba no paramétrica)[99]

3.12 ESTUDIO EPIDEMIOLÓGICO DE LA VALIDEZ EXTERNA.

Con la aplicación de la RL no condicionada al nuevo paquete de datos obtenido en esta Tesis, se intentaba "repetir" el análisis observacional de nuestro estudio anterior[87], para intentar validarlo con arreglo a los criterios de Justice et al[40], expuestos en la introducción de esta Tesis.

3.13 SESGO DE BERKSON.

Para apreciar si el sesgo de Berkson influía o no en nuestros resultados diseñamos un estudio doble con RL, construyendo primero un modelo a base de los controles recogidos en atención primaria y los casos y segundo construyendo otro a base de los controles recogidos en los hospitales y los mismos casos. Después se compararían ambos modelos. Si existe sesgo de Berkson y siguiendo las ideas de Feinstein et al[19]., los controles recogidos en primaria tienden a elevar las OR de forma estructural en los diseños de casos y controles.. La separación entre controles de primaria y de hospital se llevó a cabo mediante la creación de una variable filtro con SPSS.

3.14 BOOTSTRAPPING.

Mediante el programa F se aplicó el siguiente ***algoritmo informático***:

1. Generar 2000 muestras "bootstrap" (muestras virtuales).

2. Para cada muestra, construir un modelo de RL mediante selección hacia atrás y calcular el área bajo la *curva ROC.*

3. Resumen de los 2000 coeficientes[50,58.]

A partir de este algoritmo se recogieron las variables que se habían mostrado más pronósticas en las 2000 muestras "bootstrap".[47,52,53,56]

4. RESULTADOS

Estadística Descriptiva de los Datos Crudos.

Tabla nº 1. Variable CASO

		Frequencia	Porcentaje	
Valido	control	275	68,6	68,6
	caso	126	31,4	100,0
	Total	401	100,0	

Tabla nº 2.

Indices Generales

Relación CASO / CONTROL	1 / 2,18
Relación PROSPECTIVO / REPROSPECTIVO	4,41 / 1

Tabla nº 3. Variable Sexo

		Frequencia	Porcentaje	Porcentaje Válido
Valido	mujer	213	53,1	53,1
	varón	188	46,9	46,9
	Total	401	100,0	100,0

Tabla nº 4. Variables Centro de Referencia y Caso.

	Centro							Total
	1	2	3	4	5	6	7	
Var. control caso	60	36	14	114	32	18	1	275
Caso	0	12	0	64	5	0	45	126
Total	60	48	14	178	37	18	46	401

Centros de Referencia de la información: 1.- Centro de Salud de Pilas (Sevilla), 2.- Centro de Salud de Mérida (controles) , Hospital General de Mérida (casos) (Badajoz). 3.- Centro de Salud de Camas (Sevilla). 4.- Hospital Universitario Virgen Macarena (Sevilla).(HUVM) 5.- Hospital Juan Ramón Jiménez (Huelva) (casos y controles). 6.- Centro de Salud Huerta del Rey (Sevilla). 7.- Ciudad Sanitaria Virgen del Rocío (Sevilla)(CSVR).

NOTA: El Centro de Salud de Pilas tenía como centro hospitalario de referencia a CSVR , el Centro de Salud de Camas al HUVM y el Centro de Salud de Huerta del Rey a CSVR y HUVM.

Tabla nº 5. Desglose de la Tabla nº 4 en Controles de Primaria y de Hospital.

Tipo			centro							Total
			1	2	3	4	5	6	7	
control hospital	CASO	control				114	32		1	147
	Total					114	32		1	147
control primaria	CASO	control	60	36	14			18		128
	Total		60	36	14			18		128
caso	CASO	caso		12		64	5		45	126
	Total			12		64	5		45	126

Los códigos de la variable CENTRO corresponden con los expuestos en la tabla anterior.

Tabla nº 6. Estimadores de Centralización y de Dispersión de Variables Contínuas.

	N	Mínimo	Máximo	Media	Error Estándar de Media	Desviación Stándar
EDAD	401	24	94	63,42	,744	14,890
CT	399	81	313	197,70	2,140	42,737
HDL	346	17	176	45,65	,925	17,202
LDL	191	38	235	131,68	2,508	34,656
VLDL	228	10	216	54,90	3,617	54,623
TG	264	25	566	121,21	4,276	69,482
FA	357	29	500	154,67	4,292	81,104
CA19_9	380	,1	162,0	19,938	1,2423	24,2168

Tabla nº 7. Análisis Descriptivo de Valores Perdidos en la información cruda.

	CASO	Casos					
		Válidos		Perdidos		Total	
		N	Porcentaje	N	Porcentaje	N	Porcentaje
EDAD	control	275	100,0%	0	,0%	275	100,0%
	caso	126	100,0%	0	,0%	126	100,0%
CT	control	274	99,6%	1	,4%	275	100,0%
	caso	125	99,2%	1	,8%	126	100,0%
HDL	control	262	95,3%	13	4,7%	275	100,0%
	caso	84	66,7%	42	33,3%	126	100,0%
LDL	control	134	48,7%	141	51,3%	275	100,0%
	caso	57	45,2%	69	54,8%	126	100,0%
VLDL	control	163	59,3%	112	40,7%	275	100,0%
	caso	65	51,6%	61	48,4%	126	100,0%
TG	control	142	51,6%	133	48,4%	275	100,0%
	caso	122	96,8%	4	3,2%	126	100,0%
FA	control	234	85,1%	41	14,9%	275	100,0%
	caso	123	97,6%	3	2,4%	126	100,0%
CA19_9	control	266	96,7%	9	3,3%	275	100,0%
	caso	114	90,5%	12	9,5%	126	100,0%

Tabla nº 8. Percentil
Bisagras de Tukey

	CASO	25	50	75
EDAD	control	52,00	63,00	75,00
	caso	59,00	68,00	76,00
CT	control	181,00	203,50	236,00
	caso	156,00	177,00	203,00
HDL	control	35,00	45,00	57,00
	caso	32,00	34,00	44,00
LDL	control	117,00	136,00	159,00
	caso	101,00	111,00	121,00
VLDL	control	22,00	31,00	107,50
	caso	16,00	20,00	27,00
TG	control	85,00	110,00	147,00
	caso	79,00	100,00	132,00
FA	control	92,00	146,00	202,00
	caso	98,00	150,00	203,50
CA19_9	control	5,300	10,200	19,000
	caso	7,700	20,800	45,100

Tabla nº 9. Análisis de la Normalidad de las Variables Continuas.
Prueba de Kolmogorov-Smirnov para una muestra

	N	Parámetros normales(a,b)		Z de Kolmogorov-Smirnov	Sig. asintót. (bilateral)
		Media	Desviación típica		
EDAD	401	63,42	14,890	1,496	,023
CT	399	197,70	42,737	,649	,794
HDL	346	45,65	17,202	2,227	,000
LDL	191	131,68	34,656	1,075	,198
VLDL	228	54,90	54,623	4,664	,000
TG	264	121,21	69,482	2,543	,000
FA	357	154,67	81,104	1,146	,145
CA19_9	380	19,938	24,2168	4,035	,000

a La distribución de contraste es la Normal.
b Se han calculado a partir de los datos.

Tabla nº 10. Imputación de Valores Perdidos mediante Interpolación Lineal en las Variables Lipídicas (HDL, LDL y VLDL)

Variable resultante	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
HDL_1	55	1	401	401	LINT(HDL)

Variable resultante	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
LDL_1	191	18	399	382	LINT(LDL)

Variable resultante	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
VLDL_1	172	1	400	400	LINT(VLDL)

Tabla nº 11. Estadística Descriptiva de Variables Lipídicas con Valores Imputados.

	N	Minimum	Maximum	Mean	Std. Deviation
LINT(LDL) LDL_1	382	38,0	235,0	132,304	34,7017
LINT(VLDL) VLDL_1	400	10,0	216,0	56,570	52,1448
LINT(HDL) HDL_1	401	17,0	176,0	46,091	17,5727

Análisis Bivariante con Valores Crudos.

Tabla nº 12. Tabla de Contingencia: CASO – SEXO.

		Sexo		Total
		mujer	varón	
CASO	control	152	123	275
	caso	61	65	126
Total		213	188	401

Tabla nº 13. Prueba de Chi-Cuadrado.

	Valor	Grados de libertad	Significación Asintótica (dos colas)	Significación Exacta (dos colas)	Significación Exacta (una cola)
Pearson Chi-Cuadrado	1,633(a)	1	,201		
Likelihood Ratio	1,631	1	,202		
Test Exacto de Fisher				,236	,121
Casos Válidos	401				

a Ninguna celda mostró un valor esperado menor de 5.

Tabla nº 14. Centro / Sexo

		Sexo		Total
		mujer	varón	
centro	1	34	26	60
	2	30	18	48
	3	6	8	14
	4	84	94	178
	5	23	14	37
	6	12	6	18
	7	24	22	46
Total		213	188	401

Código de Centro en Tabla nº 4

Tabla nº 15. Test de Chi-Cuadrado (centro/sexo)

	Valor	Grados de Libertad	Significación Asintótica (dos colas)
Pearson Chi-Cuadrado	7,661(a)	6	,264
Likelihood Ratio	7,728	6	,259
N de Casos Válidos	401		

a. Ninguna celda tiene un valor esperado menor de 5.

Tabla nº 16. Estadísticos de contraste(a) para comparación de variables contínuas, según sean casos o controles.

	U de Mann-Whitney	Sig. asintót. (bilateral)
EDAD	14285,500	,005
CT	10815,500	,000
HDL	7097,000	,000
LDL	1835,000	,000
VLDL	2881,500	,000
TG	7779,500	,154
FA	14037,000	,702
CA19_9	10417,500	,000

a Variable de agrupación: CASO

CASO		CT	LDL	TG	HDL	VLDL	EDAD	CA19_9	FA
Control	Media	205,57	140,07	127,12	47,64	62,52	61,96	15,221	154,32
	N	274	134	142	262	163	275	266	234
Caso	Media	180,45	111,96	114,34	39,44	35,78	66,60	30,946	155,33
	N	125	57	122	84	65	126	114	123
Total	Media	197,70	131,68	121,21	45,65	54,90	63,42	19,938	154,67
	N	399	191	264	346	228	401	380	357

Figura nº 2.
Análisis de Linealidad.
Distribución Visual de los Valores de las Variables.

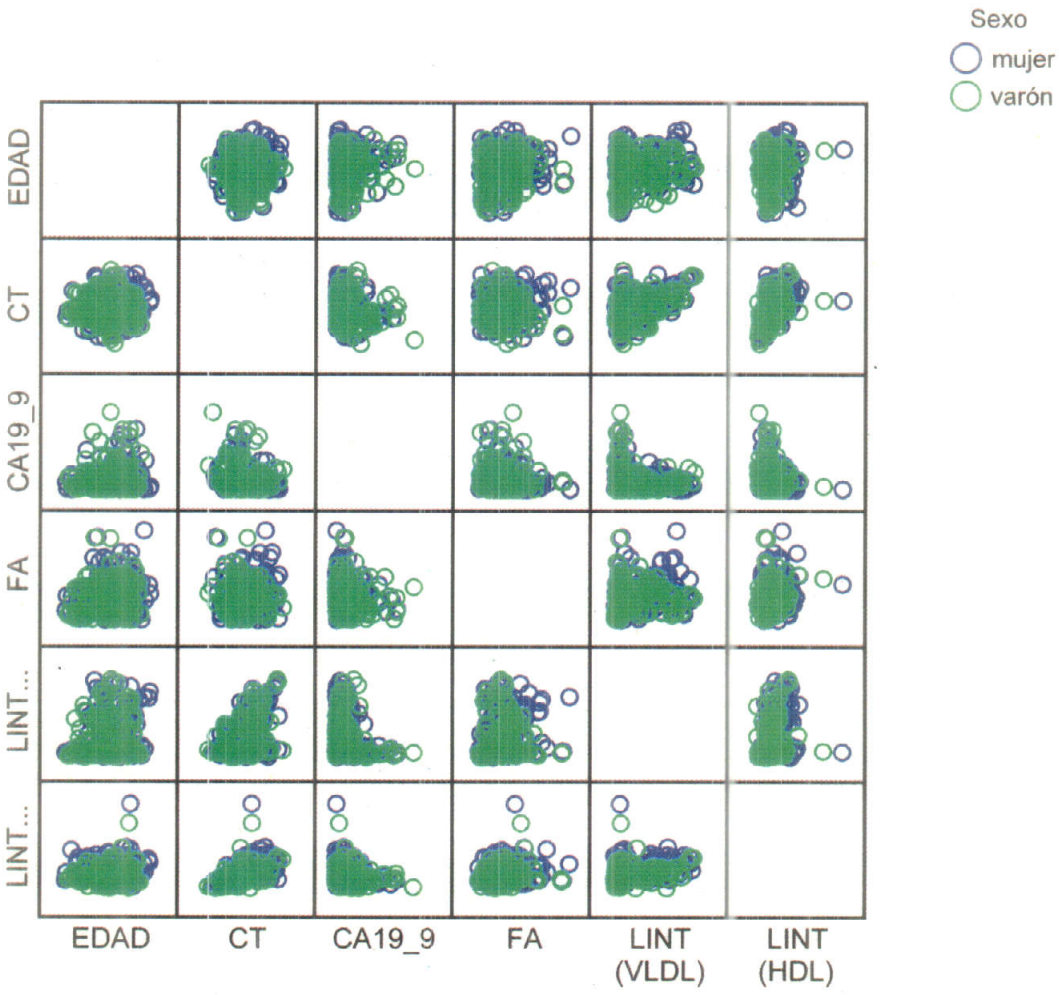


Tabla nº 17. Coeficientes de Correlación de Spearman.

			EDAD	CT	FA	CA19_9	LINT(VLDL)	LINT(HDL)
Spearman's rho	EDAD	Correlation Coefficient	1,000	,016	,166(**)	,033	,026	,044
		Sig. (2-tailed)	.	,747	,002	,519	,604	,374
		N	401	399	357	380	400	401
	CT	Correlation Coefficient	,016	1,000	,077	-,160(**)	,508(**)	,510(**)
		Sig. (2-tailed)	,747	.	,146	,002	,000	,000
		N	399	399	356	378	398	399
	FA	Correlation Coefficient	,166(**)	,077	1,000	-,149(**)	,234(**)	,301(**)
		Sig. (2-tailed)	,002	,146	.	,006	,000	,000
		N	357	356	357	338	356	357
	CA19_9	Correlation Coefficient	,033	-,160(**)	-,149(**)	1,000	-,172(**)	-,268(**)
		Sig. (2-tailed)	,519	,002	,006	.	,001	,000
		N	380	378	338	380	379	380
	LINT(VLDL)	Correlation Coefficient	,026	,508(**)	,234(**)	-,172(**)	1,000	,367(**)
		Sig. (2-tailed)	,604	,000	,000	,001	.	,000
		N	400	398	356	379	400	400
	LINT(HDL)	Correlation Coefficient	,044	,510(**)	,301(**)	-,268(**)	,367(**)	1,000
		Sig. (2-tailed)	,374	,000	,000	,000	,000	.
		N	401	399	357	380	400	401

** La correlación es significativa al 0.01 (dos colas)

Análisis Multivariante mediante Regresión Logística con Valores Crudos.

Codificación de la Variable Dependiente.

Valor Original	Valor Interno
Control	0
Caso	1

Tabla nº 18. Clasificación(a)

Observado			Pronosticado		
			CASO		Porcentaje correcto
			control	caso	
Paso 1	CASO	Control	251	14	94,7
		Caso	81	32	28,3
	Porcentaje global				74,9

a El valor de corte es ,500

Tabla nº 19. Resumen del Proceso en el Modelo Final con Valores Crudos.

Casos no ponderados		N	Porcentaje
Casos seleccionados	Incluidos	378	94,3
	Casos perdidos	23	5,7
	Total	401	100,0
Casos no seleccionados		0	,0
Total		401	100,0

Tabla nº 20. Modelo Final Ajustado con Valores Crudos.

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1(a)	EDAD	,020	,008	5,563	1	,018	1,020	1,003	1,037
	CT	-,014	,003	19,695	1	,000	,986	,980	,992
	CA19_9	,023	,005	17,946	1	,000	1,023	1,012	1,034
	Constante	,073	,785	,009	1	,926	1,076		

a Variable(s) introducida(s) en el paso 1: EDAD, CT, CA19_9.

Análisis Multivariante mediante Regresión Logística con Valores Imputados.

Tabla nº 21. Modelo de valores crudos con variable LDL_1

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDAD	,025	,009	7,626	1	,006	1,025
	CT	-,013	,008	2,759	1	,097	,987
	CA19_9	,037	,007	26,465	1	,000	1,038
	LDL_1	-,003	,009	,111	1	,739	,997
	Constant	-,117	,897	,017	1	,896	,889

a Variable(s) entered on step 1: EDAD, CT, CA19_9, LDL_1.

Tabla nº 22. Modelo de valores crudos con variable VLDL_1

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDAD	,021	,008	6,065	1	,014	1,021
	CT	-,012	,003	12,712	1	,000	,988
	CA19_9	,021	,005	15,777	1	,000	1,022
	VLDL_1	-,004	,003	1,853	1	,173	,996
	Constant	-,075	,792	,009	1	,925	,928

a Variable(s) entered on step 1: EDAD, CT, CA19_9, VLDL_1.

Tabla nº 23. Modelo de valores crudos con variable HDL_1

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	EDAD	,020	,008	5,525	1	,019	1,020
	CT	-,014	,003	16,490	1	,000	,986
	CA19_9	,023	,005	17,357	1	,000	1,023
	HDL_1	,001	,008	,013	1	,909	1,001
	Constant	,065	,788	,007	1	,935	1,067

a Variable(s) entered on step 1: EDAD, CT, CA19_9, HDL_1.

Análisis Multivariante mediante Regresión Logística con Interacción.

Tabla nº 24. Variables.

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	EDAD	,049	,013	14,486	1	,000	1,051	1,024	1,078
	CA19_9	,129	,036	12,687	1	,000	1,138	1,060	1,222
	CT	-,013	,003	18,409	1	,000	,987	,981	,993
	CA19.9xEDAD	-,002	,001	9,391	1	,002	,998	,997	,999
	Constant	-2,035	1,064	3,656	1	,056	,131		

a Variable(s) entered on step 1: EDAD, CA19_9, CT, CA19.9xEDAD.

Tabla nº 25. Clasificacion (a)

Observed			Predicted		
			CASO		Percentage Correct
			control	caso	
Step 1	CASO	control	250	15	Esp. 94,3
		Caso	78	35	Sens. 31,0
	Overall Percentage				75,4

a El valor de corte es 0,5 .

Especificidad y Sensibilidad del Modelo Ajustado con Interacción.

Análisis Multivariante mediante Regresión Logística
Controles recogidos en Atención Primaria.

Tabla nº 26. Resumen del procesamiento de los casos

Casos no ponderados.		N	Percent
Selected Cases	Included in Analysis	241	60,1
	Missing Cases	13	3,2
	Total	254	63,3
Unselected Cases		147	36,7
Total		401	100,0

Tabla nº 27. Clasificación. Sensibilidad.Especificidad.

Observed			Predicted					
			Selected Cases(a)			Unselected Cases(b,c)		
			CASO		Porcentaje Correcto	CASO		Porcentaje Correcto
			control	caso		control	caso	
Step 1	CASO	control	105	23	82,0	83	54	60,6
		caso	41	72	63,7	0	0	.
	Overall Percentage				73,4			60,6

a Selected cases tipo NE -1

b Unselected cases tipo EQ -1

c Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

d The cut value is ,500

Tabla nº 28. Modelo de Regresión Logística para Controles de Atención Primaria.

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	EDAD	,035	,010	11,420	1	,001	1,036	1,015	1,057
	CT	-,017	,004	17,891	1	,000	,983	,975	,991
	CA19_9	,045	,010	19,738	1	,000	1,046	1,026	1,067
	Constant	,267	,973	,075	1	,784	1,306		

a Variable(s) introducida(s) en el paso 1: EDAD, CT, CA19_9.

Análisis Multivariante mediante Regresión Logística con los Controles recogidos en los Hospitales.

Tabla nº 29. Resumen del procesamiento de los casos

Casos no ponderados.		N	Percent
Selected Cases	Included in Analysis	250	62,3
	Missing Cases	23	5,7
	Total	273	68,1
Unselected Cases		128	31,9
Total		401	100,0

a Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos

Tabla nº 30. Clasificación. Sensibilidad.Especificidad.

Observed			Predicted					
			Selected Cases(a)			Unselected Cases(b)		
			CASO		Percentage Correct	CASO		Percentage Correct
			control	caso		control	caso	
Step 1	CASO	control	105	32	76,6	113	15	88,3
		caso	57	56	49,6	0	0	.
	Overall Percentage				64,4			88,3

a Selected cases tipo NE 0

b Unselected cases tipo EQ 0

c El valor de corte es ,500

Tabla nº 31. Modelo de Regresión Logística para Controles de Hospital.

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	EDAD	,013	,009	1,888	1	,169	1,013	,994	1,032
	CT	-,012	,003	12,475	1	,000	,988	,982	,995
	CA19_9	,015	,005	7,393	1	,007	1,015	1,004	1,026
	Constant	,892	,887	1,011	1	,315	2,440		

a Variable(s) entered on step 1: EDAD, CT, CA19_9, CA19.9xEDAD.

a Variable(s) introducida(s) en el paso 1: EDAD, CT, CA19_9.

Interacción en los modelos construidos con controles Atención Primaria y de Hospital.

Tabla nº 32. Resumen del proceso de Registros.

Casos no ponderados		N	Percent
Selected Cases	Included in Analysis	241	60,1
	Missing Cases	13	3,2
	Total	254	63,3
Unselected Cases		147	36,7
Total		401	100,0

a If weight is in effect, see classification table for the total number of cases.

Tabla nº 33. Interacción en Modelo construido con Controles de Primaria.

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	EDAD	,062	,016	15,036	1	,000	1,064	1,031	1,098
	CT	-,018	,004	18,494	1	,000	,982	,974	,990
	CA19_9	,172	,057	8,987	1	,003	1,188	1,061	1,329
	CA19.9xEDAD	-,002	,001	5,632	1	,018	,998	,997	1,000
	Constant	-1,467	1,267	1,341	1	,247	,231		

a Variable(s) entered on step 1: EDAD, CT, CA19_9, CA19.9xEDAD.

Tabla nº 34. Resumen del proceso de Registros.

Casos no ponderados		N	Percent
Selected Cases	Included in Analysis	250	62,3
	Missing Cases	23	5,7
	Total	273	68,1
Unselected Cases		128	31,9
Total		401	100,0

Tabla nº 35. Interacción en Modelo construido con Controles de Hospital.

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	EDAD	,039	,014	7,470	1	,006	1,040	1,011	1,070
	CT	-,011	,003	11,153	1	,001	,989	,982	,995
	CA19_9	,100	,036	7,507	1	,006	1,105	1,029	1,187
	CA19.9xEDAD	-,001	,001	5,899	1	,015	,999	,998	1,000
	Constant	-1,034	1,189	,756	1	,384	,356		

Tabla nº 36. "Bootstrapping" del Modelo sobre 2000 Muestras Virtuales.

```
Coeficientes:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.679851   1.462390   2.516  0.01186 *
EDAD         0.038496   0.012338   3.120  0.00181 **
CT          -0.061438   0.021325  -2.881  0.00396 **
TG           0.002551   0.003497   0.729  0.46574
FA           0.005029   0.002957   1.701  0.08899 .
CA19_9       0.065791   0.015593   4.219 2.45e-05 ***
LDL_1        0.016314   0.023151   0.705  0.48102
VLDL_1       0.019089   0.005893   3.239  0.00120 **
HDL_1        0.019786   0.014507   1.364  0.17259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura nº 3.
2000 valores bootstrap de Area bajo la Curva ROC.

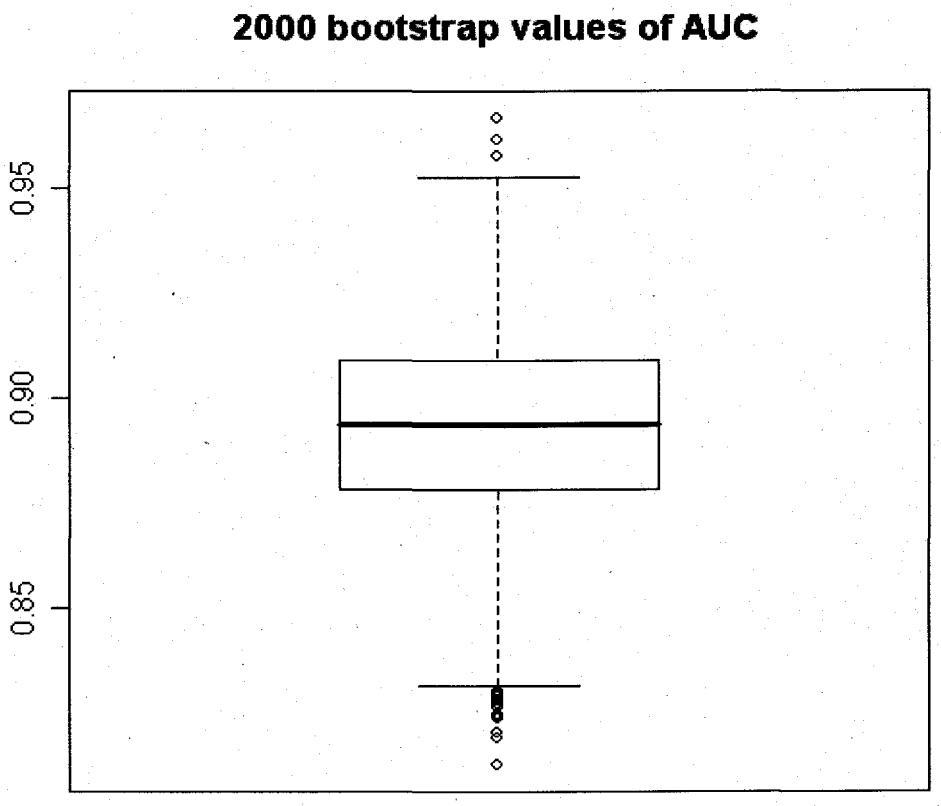


Figura nº 4.
2000 coeficientes bootstrap.

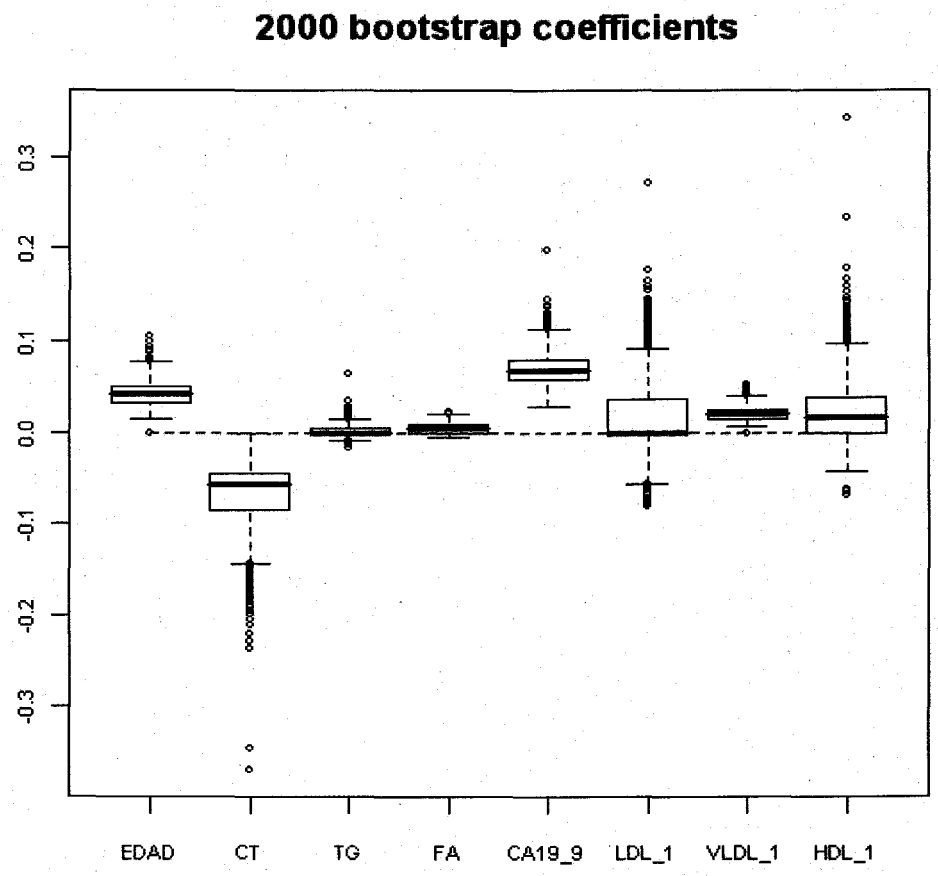


Figura nº 5.
Sensibilidad y Especificidad de los diferentes Modelos Multivariantes (R.L.)

Variables.	S	E
EDAD,CT y CA 19.9 con todos los controles	28,3%	94,7%
EDAD,CT Y CA 19.9 con controles de primaria	63,7%	82%
EDAD,CT y CA 19.9 con controles de hospital	49,6%	76,6%
EDAD,CT,CA 19.9 y (EDAD x CA 19.9) con todos los controles	31%	94,3%

Figura nº 6.
Análisis del Tamaño Muestral con respecto a la utilización de Regresión Logística No Condicionada.

	n	E.I.V. previos a RL	E.I.V. definitivos
Valores Crudos	378	18,8	37,6
Con controles de primaria	241	18,8	37,6
Con controles de hospital	250	18,8	37,6
Con variable de interacción Valores crudos	378	9,41	28,25

E.I.V= eventos de interés por variable = nº de casos / nº de variables

5. DISCUSION

Hemos realizado una investigación para estudiar la funcionalidad de un modelo multivariante explicativo del diagnóstico del CCE en términos de validez. El modelo original está publicado[87] y contiene seis variables, las mismas que hemos manejado para esta Tesis. El tamaño de la muestra primitiva fué de 93 elementos (53 casos y 40 controles).

El nuevo tamaño de muestra ha sido de 401 elementos repartidos en 126 casos y 275 controles (Tabla nº1). El diseño fue *no pareado*. Un total de 308 elementos pertenecen a la fase validatoria del trabajo. La relación caso / control ha sido por lo tanto de **1 / 2,18.**

Origen Geográfico y Límites Temporales.

La muestra primitiva fue recogida íntegramente en el Hospital Universitario Virgen Macarena de Sevilla (HUVM) entre los años 1992 a 1995 de una forma prospectiva. A este centro pertenecen también 11 casos y 74 controles de la muestra validatoria. Los casos nuevos han sido recopilados en forma retrospectiva del archivo general de historias clínicas respetando siempre los criterios de inclusión (período 2000-2003) y los controles nuevos se han recogido de forma prospectiva en el Servicio de Medicina Interna (Prof. Pérez Cano) durante el año 2003.

Entre los años 2001 a 2003 se recogieron de forma prospectiva un total de 60 controles en el Centro de Salud de Pilas (Sevilla).Durante ese mismo período de tiempo se han recogido 36 controles (prospectivos) en el Centro de Salud de Mérida y 12 casos (retrospectivos) en el Hospital General de la misma ciudad En el 2003 se recogieron los 14 controles del Centro de Salud de Camas (prospectivos) Durante el 2003 también se recogieron los 32 controles y los cinco casos del Hospital Juan Ramón Jiménez de Huelva (prospectivos) .Desde el 2002 al 2003 se recogieron los 18 controles del Centro de Salud Huerta del Rey de Sevilla (prospectivos). Y finalmente, también en 2003 se recogieron los 45 casos (retrospectivos) y el control (retrospectivo) de CSVR (Tablas 4 y 5). Por lo tanto, los límites temporales de nuestra recogida de datos han sido desde 1992 hasta 2003. Durante todo este tiempo se han respetado escrupulosamente los criterios de inclusión y exclusión expuestos en el capítulo de Metodología de esta Tesis.

La relación general de elementos prospectivos / retrospectivos de nuestra investigación ha sido de **4,41 / 1**. Cada centro de salud y cada hospital estaban conectados entre sí de tal forma que los usuarios de los primeros eran ingresados en los segundos cuando tenían algún problema de salud que así lo necesitara.

Investigadores.

El total de personas que han colaborado en la recogida de información de esta Tesis los podemos clasificar como sigue: cinco médicos de familia formados por el programa nacional de especialidades, un medico general, dos médicos internistas hospitalarios formados también por la vía MIR, residentes en Medicina de Familia, Cirugía General y Digestiva, Medicina Interna, Neurología, Inmunología y Alergia y Farmacología Clínica. Su trabajo confiere *transportabilidad metodológica* a nuestros resultados.

Descriptiva.

Los resultados descriptivos del paquete de datos completo (401 registros) están expuestos en la sección correspondiente de esta Tesis (Tablas nº 6-8). Cabe destacar de entre ellos las medias aritméticas de los casos que son más bajas que las de los controles en cuanto a las variables lipídicas se refiere (Tabla nº 16), excepto para los triglicéridos.

Los límites temporales de la recogida de datos han sido amplios. Una forma de intentar equilibrar el sesgo de este hecho era estandarizar estadísticamente las frecuencias observadas, sin embargo ello conllevaría invariablemente una tendencia a la centralización, y siendo este trabajo un intento de validar un modelo construido mediante RL la estandarización habría hipotecado nuestros resultados.

Análisis de Normalidad.

Tras la aplicación de la prueba de Kolmogorov-Smirnov se pudieron considerar como normales las variables: CT (n = 399) , LDL (n = 191) y FA (n = 357) (Tabla nº 9). El resto de variables poseían una distribución no normal[97].

Valores Perdidos

Una de las consecuencias más trascendentales del análisis descriptivo de los datos fue el elevado porcentaje de información perdida existente sobre todo en las variables lipídicas (Tabla nº 7) y dentro de ellas, en la lipoproteína de baja densidad LDL con un 52,4 % de pérdidas y en la lipoproteína de muy baja densidad (VLDL) con un 43,1%. Para intentar soslayar esté déficit construimos variables con *valores imputados.*

Entre las diversas formas de imputación que ofrece SPSS (versión 12.0) nos inclinamos por la *interpolación lineal* porque nos pareció la más idónea para la estructura de nuestros datos[8]. Aplicamos esta técnica a las variables LDL, VLDL y HDL de la forma que está expuesta en la sección correspondiente del epígrafe de Resultados. Para la LDL se repusieron un total de 191 registros, para la VLDL un total de 172 registros y para la HDL un total de 55 registros (Tabla nº 10).

En la fase validatoria no se aplicó la fórmula de Friedwald[90] para la obtención de valores perdidos de las variables LDL ni VLDL. No queríamos agregar linealidad artificial a los valores reales observados.

Comparaciones entre Variables Cualitativas.

Contemplado el paquete de datos en su totalidad no hubo diferencia significativa en la distribución por sexo entre los casos y los controles (Chi-Cuadrado de Pearson p = 0,20 y Test Exacto de Fisher p = 0,23) (Tabla nº 13). Tampoco la hubo en la distribución por sexo y centros de referencia (Chi-Cuadrado de Pearson p = 0,26 con 6 grados de libertad) (Tablas nº 14 y 15). Creemos que estos resultados conceden representatividad a la muestra.

Comparación entre Variables Cuantitativas.

Al no poseer la mitad de las predictoras del modelo primitivo una distribución normal nos decidimos por un método de comparación no paramétrica a la hora del análisis bivariante. Aplicando la prueba de U-Mann-Whitney se observaron diferencias significativas entre los casos y los controles en todas las variables del modelo primitivo excepto en la fosfatasa alcalina (FA) (Tabla nº 16).

Análisis de Linealidad.

Previamente a la aplicación de la RL como técnica multivariante es necesario hacer un estudio de la asociación lineal de la distribución de valores[6].Se realizó primero un análisis visual de las nubes de puntos (Figura nº 2) y después se aplicó el coeficiente de correlación de Spearman al tener la mayoría de las variables una distribución no normal De las 30 correlaciones posibles (Tabla nº 17), 18 mostraron valores significativos. De ellas, 12 las ocasionaban variables con valores imputados. Decidimos que era correcto aplicar la RL. (6,44)

Análisis Multivariante con Valores Crudos.

Aplicamos la RL no condicionada para intentar obtener un modelo ajustado con el nuevo tamaño muestral mediante la técnica de selección de variables hacia atrás. Como ya hemos referido anteriormente el programa empleado fue el SPSS. Pudimos obtener un modelo ajustado que contenía la variable EDAD, la variable CA19.9 y la variable CT (colesterol total). Habían podido entrar tres de las seis variables primitivas (Tabla nº 19).

Para nosotros ese era un resultado bastante aceptable porque conservaba la mitad de predictoras y porque era "coherente" con la realidad clínica y biológica. La **edad** y el **CA 19.9** se mostraban "favorecedoras" de la condición de ser caso con unas OR respectivas de 1.02 y de 1.23 y la tasa de colesterol total adoptaba un sentido contrario con una OR de 0.98 , alcanzando todas ellas la significación estadística ($p<0,05$). Se puede afirmar que este modelo ajustado y definitivo ha presentado una nivel de validez de 4 sobre 5 según los criterios de Justice[40] pues contiene validaciones independientes múltiples.

La sensibilidad (28,3%) y la especificidad (94,7%) estaban descompensadas y no eran muy útiles clínicamente (Tabla nº 18 y Figura nº 5) pero esto último no le restaba valor como resultado estadístico.

El valor clínico y biológico del CA 19.9 es un hecho constatado anteriormente en la bibliografía[86,87]. Su elevación es mucho más frecuente en procesos malignos que benignos sobre todo neoplasias de páncreas, colon-recto, pulmón, hígado y ovario[100]. Es útil también para el pronóstico de pacientes con neoplasia de colon.

Puede elevarse en procesos no oncológicos[101] y se ha descrito en la literatura una relación inversa del CA 19.9 con el calcio sérico en un estudio casos-control apareado con pacientes afectos de carcinoma de colon[102].Una evidencia muy interesante para el control del sesgo de clasificación en esta Tesis, ha sido el artículo de Varol et all[103], en donde se demuestra la normalidad del Ca 19.9 en paciente con insuficiencia cardiaca crónica., aunque la media aritmética del grupo de pacientes fuera más alta que la del grupo control ($p<0,001$). En otras publicaciones, el CA 19.9 no ha mostrado tanta capacidad diagnóstica para el CCRE cuando se le intenta incluir en modelos multivariantes[104]

En esta investigación no se maneja el antígeno cárcinoembrionario (CEA) como variable explicativa porque no formaba parte del modelo primitivo. Se están describiendo otros marcadores tumorales de utilidad en combinación con el CEA y con el CA 19.9, como por ejemplo el Factor de Células Madre (Stem Cell Factor) y la Interleukina 3[105].

El valor clínico y biológico de la tasa de lípidos plasmáticos en el carcinoma de colon es otro hecho que viene siendo también refrendado por las evidencias Eichholzer et al han publicado unos resultados sobre el valor predictivo de la tasa de colesterol en paciente con diversas neoplasias, entre ellas el cáncer de colon[81] Todo ello en el contexto de seguimiento del estudio prospectivo de Basilea.

Otro trabajo muy interesante sobre la importancia de los niveles plasmáticos en el CCRE es el de Notarnicola et al[106].En él se encuentra una asociación entre la capacidad de desarrollar metástasis y los niveles elevados de CT y LDL en pacientes afectos de CCRE. Estos hallazgos son congruentes con nuestros resultados porque un criterio de selección de los casos ha sido el que no hubieran desarrollado metástasis a distancia (Estadío de Dukes IIA como máximo). Nuestros casos tienden a presentar los lípidos bajos. Preferimos la clasificación de Dukes[88] a la Astler-Coller por el largo período de recogida de datos de nuestra investigación.

Ese mismo grupo ha publicado también unos hallazgos muy sugestivos sobre los cambios enzimáticos en la vía del mevalonato en pacientes con CCRE dependiendo de la localización del tumor en el tracto del intestino grueso[107]

La utilización de la RL para los estudios observacionales sigue siendo refrendada por la bibliografía más actualizada. Mostrando resultados similares si se compara con los índices de propensión[108] o con las redes neurales artificiales[109]. La utilización de un método de selección manual de variables es un hecho también cada vez más constatado en la bibliografía , sobre todo si el modelo multivariante se complementa después con un "bootstrapping" como era nuestro caso[110].

Análisis Multivariante diferenciando los controles según su origen

Cuando se generaron dos modelos diferentes mediante RL , el primero confeccionado a base de controles recogidos en atención primaria junto con los casos, y el segundo confeccionado con los controles recogidos en los hospitales y los mismos casos , el primero arrojó valores significativos en las tres predictoras estudiadas mientras que el segundo tan sólo los arrojó en dos. El primero era más eficiente y tenía unas OR más altas que el segundo tal como predecía Feinstein en su trabajo sobre el sesgo de Berkson[19] (Tablas nº 28 y 31). La sensibilidad y la especificidad también fueron más compensadas (Tabla nº 28 y Figura nº 5) que las del modelo confeccionado con todos los valores. .

Análisis Multivariante con valores imputados.

La substitución de las variables con un alto porcentaje de información perdida (LDL,VLDL y HDL) por las correspondientes con valores imputados mediante interpolación lineal (LDL_1,VLDL_1 y HDL_1) no permitió la agregación de nuevas predictoras al modelo ajustado con anterioridad que contenía a la EDAD, el CA19.9 y el CT (Tablas nº 21-23).

Interacciones del Modelo creado con todos los controles

La exploración de interacciones mostró un resultado significativo en la interacción entre EDAD y CA 19.9 ($p<0,005$) (Tabla nº 24) . Se potenciaba el efecto favorecedor de la condición de caso de ambas variables a nivel individual (OR de 1,051 para EDAD y OR de 1,138 para CA 19.9) respecto al modelo obtenido con valores crudos , pero la OR de la variable interacción (EDAD x

CA 19.9) tenía un sentido contrario. A pesar de que la especificidad se elevó al 94,3%, la sensibilidad bajaba hasta el 31% con lo que la utilidad práctica de este modelo no era la adecuada (Tabla nº 24 y Figura nº 5). No se obtuvieron otras interacciones significativas ni con los valores crudos ni con valores imputados.

El análisis epidemiológico intenta ir más allá del modelización matemática. El concepto de *interacción estadística* vinculado a la arbitrariedad en la elección del modelo, precisa de una interpretación epidemiológica porque no descansa en un fundamento teórico explícito. Si no se hace una inferencia que vaya más allá del modelo, un concepto puramente estadístico de la interacción no podría contribuir al análisis epidemiológico de una forma que tuviese sentido[111] Para nosotros tiene importancia epidemiológica esta interacción aunque la variable producto muestre una OR de diferente sentido a las de las variables individuales.

Greenland[111] proporciona una explicación al hecho de que, como en nuestro modelo de interacción, el coeficiente de la variable producto sea diferente a los de las variables contempladas individualmente. El coeficiente de la variable interacción refleja solamente el balance neto entre los diferentes tipos de respuesta implicados en la interacción. Un coeficiente > 0 implica solamente que las respuestas sinergísticas son más frecuentes que las antagonistas y que las respuestas competitivas pero que esta últimas estén ausentes. Un coeficiente < 0 implica solamente que las respuestas antagonistas y competitivas son más frecuentes que las sinergísticas, pero no que estas últimas no existan. Un coeficiente = 0 implica que las respuestas sinergísticas están equilibradas con las respuestas antagonistas y competitivas, pero no que las interacciones estén ausentes.

Interacciones en los modelos construidos con controles de Atención Primaria y de Hospital.

Ensayando la misma interacción (EDAD x CA 19.9) existente en el modelo generado con todos los controles se obtuvieron valores significativos en ambos casos (Tablas nº 33 y 35) aunque el límite superior del IC alcanzaba la unidad, con lo cual la utilidad clínica era mínima.

"Bootstrapping"

Mediante el programa S se confeccionaron 2000 muestras "bootstrap" a partir de los datos reales con las seis variables del modelo primitivo. Las variables HDL,VLDL y LDL en sus versiones con valores imputados HDL-1. VLDL-1 y LDL-1. Se estudiaron los coeficientes obtenidos mediante RL no condicionada con el método de selección de variables hacia atrás. La EDAD, el CT y el CA 19.9 fueron también las variables que arrojaron valores significativos (Tabla nº 36 y Figura nº 4). Los resultados obtenidos bajo la curva ROC fueron también muy interesantes, la figura de caja muestra la mayoría de valores alrededor de 0,9 (Figura nº 3).

Estos hallazgos conceden en primer lugar un alto grado de validez interna a nuestro trabajo y dan solidez a nuestras observaciones. Aunque el "bootstrapping" no sea una técnica de medición de la validez externa, sí que lo es de la validez interna que es previa en términos epidemiológicos [47,48,51,52,56,57]

Tamaño Muestral.

Vergouwe et al[112] proponen un mínimo de 200 elementos muestrales (100 eventos y 100 no eventos) como tamaño efectivo en estudios que tengan como objetivo una validación externa y empleen la RL como método estadístico predictivo. En su análisis han empleado un entorno convencional del 80% de poder y del 5% de nivel de significación. Ellos mismos refieren también que algunas hipótesis específicas pueden requerir tamaños mayores. Como hemos referido antes en esta discusión, nuestro diseño constaba de un total de 126 casos y 275 controles y la intención en el empleo de la RL era más bien explicativa que predictiva.

Cepeda et al han publicado una simulación muy buena[113] en donde mantienen que los índices de propensión eran menos sesgados, más robustos y más precisos que la RL en los diseños en que existían hasta siete eventos por variable confusora. Mientras que la RL se mostraba más eficiente cuando los eventos de interés superaban los siete por variable. Los índices de propensión se postulan como un método de control de la confusión en los estudios observacionales[114] mediante el cual se podrían obtener las mismas ventajas que la asignación aleatoria otorga a los ensayos clínicos. Sin

embargo esto último dista mucho de ser una realidad meridiana[115,116]

En la Figura nº 6 se pueden observar el número de eventos de interés por variable con el que se realizaba los diferentes análisis contenidos en esta Tesis. Todos están por encima de los 10 eventos de interés por variable[34] exigidos por la literatura más aceptada en este campo excepto uno, el de interacciones del modelo con todos los valores. Según Hsieh[117] trabajando con un tamaño de muestra de 378 se pueden discernir unas OR de 1,7-1,8 y con un tamaño de muestra de 241 unas OR de aproximadamente 2.

6. CONCLUSIONES

Tras la obtención y posterior análisis de una muestra de 401 elementos, constituida por 126 casos de CCRE, obtenidos en varios hospitales y seleccionados mediante los criterios indicados anteriormente, y por 275 controles no pareados con el fin de validar un modelo multivariante definido en una investigación anterior[87] y tratando de observar si el sesgo de Berkson influía o no en nuestros resultados, hemos llegado a las siguientes conclusiones:

1. La aplicación de RL no condicionada a los datos obtenidos en el estudio de validación ha mostrado la permanencia de tres de las seis variables que componían el modelo primitivo: la EDAD (OR = 1,02; I.C. al 95%: 1,003-1,037), el CT (OR = 0,986; I.C. al 95%: 0.980-0.992) y el marcador tumoral CA19.9 (OR = 1,023; I.C. al 95% 1,012-1,034) (Tabla nº 20). Este es el modelo más equilibrado al estar construido con controles de primaria y de hospital[19], favoreciendo además una interacción (EDAD x CA 19.9) (Tabla nº 24).

2. Nosotros estimamos que semejante resultado proporciona un grado de funcionalidad y de validez externa bastante aceptable en términos epidemiológicos a nuestro modelo primitivo[87]. Siguiendo los criterios de Justice et. al[40], ésta Tesis presenta un nivel de validez externa de 4 sobre 5 pues contiene validaciones independientes múltiples, manteniendo la mitad de las variables primitivas.

3. El modelo elaborado tiene transportabilidad geográfica (ha surgido del primitivo después de recoger datos en varios lugares) y también transportabilidad metodológica (ha sido realizado sobre información recogida por varios investigadores y en diferentes períodos de tiempo)[40,43]. Por otro lado adolece de transportabilidad de espectro (sólo se estudian casos hasta el estadío IIA de Dukes) y de seguimiento.

4. Al dividir la muestra por los controles recogidos en atención primaria y por aquellos otros obtenidos en los hospitales, el modelo de RL multivariante construido con los primeros es más eficiente que el construido con los segundos, elevando las OR, la significación, la especificidad y la sensibilidad del modelo general obtenido con todos los valores crudos.

5. La conclusión anterior nos conduce a afirmar que además de por razones puramente estadísticas, el sesgo de Berkson *sí* ha influido en nuestros resultados de validación, siguiendo la hipótesis operativa de Feinstein et al[19] quienes afirman que los controles obtenidos en primaria tienden a elevar estructuralmente las OR en un diseño como el nuestro.

6. El modelo multivariante de RL construido con los casos y los controles hospitalarios mantiene la significación de las variables CT y CA 19.9 con unas OR de sentido contrario (Tabla nº 31). Esta información nos permite reforzar la plausibilidad biológica y clínica de nuestro modelo, ya que en estas circunstancias es más probable que se controle el sesgo de Berkson[19]. ($h_e \neq 0$; hc $\approx$ hd)

7. La aplicación de la técnica del "bootstrapping" mediante el análisis de 2000 muestras virtuales obtenidas a partir de nuestros datos, ha revelado unos resultados idénticos a los obtenidos mediante RL no condicionada, siendo de nuevo las variables EDAD, CT y CA 19.9 las que se benefician de unos coeficientes más significativos que además están refrendados en el área bajo la curva ROC. (Figura nº 3).

8. La conclusión anterior nos permite afirmar que nuestra investigación presenta un elevado grado de validez interna la cual, en términos epidemiológicos, es previa a la validez externa. Este grado de validez interna se obtiene tras la imputación de valores a las variables con un mayor porcentaje de información perdida y con su posterior análisis mediante "bootstrapping".

7. BIBLIOGRAFIA

[1] Berkson, J.: Limitations of the application of the four-fold table analysis to hospital data. Biometrics Bulletin.1946;2:47-53.

[2] Fletcher, R. H., Fletcher, S.W., Wagner, E.H. : Epidemiología Clínica. Aspectos Fundamentales. Masson-Williams & Wilkins España S.A. Barcelona. 2ª edición. 1998. p.:193.

[3] Johnson, A.F. : Beneath the technological fix. Outliers and probability statements. J.Chron.Dis.1985;38:957-961.

[4] Mosterín.J. : Technology-mediated observation. Techné. Journal of the Society for Philosophy and Technology. Virginia Polytechnic Institute and State University.1998.vol.4n2. http://scholar.lib.ut.edu/ejournals/SPT/v4n2/MOSTERIN. htlm. Visitado el 16 de Octubre de 2004.

[5] Torretti, R. :Observation. The British Journal for the Philosophy of Science. 1986;37: 1-23.

[6] Sánchez-Cantalejo Ramírez, E.: Regresión Logística en Salud Pública. Publicaciones de la Escuela Andaluza de Salud Pública. Granada. Serie Monografía nº 26. 2000.

[7] Steyerberg, E. W., Eijkemans , M.J., Habbema , J.D. : Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J. Clin. Epidemiol. 1999;52 : 935-942.

[8] Barroso Utra , I. Mª. , Cañizares Pérez , M., Lera Márquez , L. : Influencia de la estructura de los datos en la selección de los métodos de análisis estadísticos. Revista Española de Salud Pública.2002;76:95-103.

[9] Ellenberg,J.H.: Cohort studies. Selection bias in observational and experimental studies. Statistics in Medicine. 1994; 13: 557-567.

[10] Delgado-Rodriguez M, Llorca J : Bias. J Epidemiol Community Health. 2004; 58 : 635 – 641.

[11] Rothman K., Greenland S. : Modern epidemiology. 2nd ed. Boston: Lippicontt-Raven, 1998.

[12] Kleinbaum DG, Kupper LL, Morgenstern H. : Epidemiologic research. Belmont, CA: Lifetime Learning Publications, 1982.

[13] Silva Ayçaguer, L. C.: Diseño razonado de muestras y captación de datos para la investigación sanitaria. Ediciones Días de Santos. S. A. Madrid. 2000.

[14] Popper, K.: The logic of Scientific Discovery. Routledge Classics. London. New York. 2002.

[15] Hill, A.B.: Principles of Medical Statistics. Oxford University Press. New York. 1971.

[16] Feinstein, A. R.: Clinical Biostatistics XX. The epidemiologic trohoc , the relative risk ratio and retrospective research . Clinical Pharmacology and Therapeutics. 1973; 14: 291-307.

[17] Rothman, K.J. : Modern Epidemiology. Little Brown . Boston. 1986.

[18] Ellenberg , J.H., Nelson , K.B. Sample selection and the natural history of disease – Studies of febrile seizures. Journal of American Medical Association. 1980; 243: 1337- 1340.

[19] Feinstein AR, Walter SD, Horwitz RI.: An analysis of Berkson's bias in case-control studies. J Chronic Dis. 1986; 39: 495 - 504.

[20] Snyder N, Atterbury CE, Pinto Correia J, Conn HO. : Increased concurrence of cirrhosis and bacterial endocarditis. A clinical and post-mortem study. Gastroenterology. 1977; 73: 1107 - 13.

[21] Conn HO, Snyder N, Atterbury CE. : The Berkson bias in action. Yale J Biol Med. 1979; 52: 141 - 147.

[22] Peritz E. : Berkson's bias revisited. J Chronic Dis. 1984 ; 37 : 909 - 916.

[23] Walter SD: Berkson's bias and its control in epidemiologic studies. J Chronic Dis. 1980 ; 33 : 721 - 725.

[24] Roberts RS, Spitzer WO, Delmore T, Sackett DL.: An empirical demonstration of Berkson's bias. J Chronic Dis. 1978; 31: 119 - 128.

[25] Delgado Rodríguez,M.: Discordancias entre los estudios de ámbitos hospitalario y comunitario cuando evalúan la misma pregunta de investigación. Gaceta. Sanitaria. 2002;16:344-353.

[26] Hill,A.B.: The environment and disease : association or causation? Proc. Royal Soc. Med. 1965;58:295-306.

[27] Hulley,S.B.,Cummings,S.R. editors. : Designing clinical research: an epidemiologic approach. Baltimore: William & Wilkins.1987.

[28] Rothman,K.: Causes. Am.J.Epidemiol. 1976;104:587-592.

[29] Thijs,C.,Knipschild,P.,Leffers,P.: Does alcohol protect against the formation of gallstones? A demonstration of protophatic biass. J. Clin. Epidemiol .1991;44:941-946.

[30] Roth.,H.D.,Levy,P.S.,Shi,L.,Post,E.: Alcoholic beverages and breast cancer: some observations on published case-control studies. J.Clin.Epidemiol. 1994 ;47 :207-216.

[31] Dosemici, M., Wacholder,S.,Lubin,J.H. : Does nondifferential missclasification of exposure always bias a true effect toward the null value? Am.J.Epidemiol 1990;132:746-748.

[32] Altman,D.G., Royston,P : What do we mean by validating a prognostic model? Statistics in Medicine. 2000; 19 : 453-473.

[33] Coste, J., Fermanian, J., Venor, A.: Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. Statistics in Medicine. 1995; 14:331-345.

[34] Ortega Calvo, M., Cayuela Domínguez, A.: Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. Revista Española de Salud Pública.2002; 76: 85-93.

[35] Brier, G.W.: Verification of weather forecasts expressed in terms of probability. Monthly Weather Review.1950;78:1-3.

[36] Braitman, L.E., Davidoff, F.: Predicting clinical states in individual patients . Annals of Internal Medicine. 1996;125:406-412.

[37] Miller, M.E., Hui, S.L., Tierney, W.M. : Validation techniques for logistic regression models. Statistics in Medicine. 1991;10:1215-1226.

[38] Van Houwelingen , J.C. Thorogood , J. : Construction , validation and updating of a prognostic model for kidney graft survival. Statistics in Medicine. 1995; 14:1999-2008.

[39] Cook, R.J., Sackett, D.L.: The number needed to treat: a clinically useful measure of treatment effect. British Medical Journal.1995;310:45

[40] Justice,A.C., Covinsky,K.E..,Berlin,J.A.: Assessing the generalizability of prognostic information. Ann. Intern. Med. 1999; 130: 515 - 524.

[41] Hanley,J.A.,McNell,B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29-36.

[42] Feinstein AR, Wells CK, Walter SD. : A comparison of multivariable mathematical methods for predicting survival -- I. Introduction, rationale, and general strategy. J Clin Epidemiol. 1990;43:339-347

[43] Charlson, M. E., Ales, R.J., Simon, R., MacKenzie, R.: Why predictive indexes perform less well in validation studies. Is it magic or methods? Arch. Intern. Med. 1987;147: 2155-2161.

[44] Harrel , F.E. , Lee , L.L. , Mark , D.B. : Multivariable prognostic models : issues in developing models , evaluating assumptions and adequacy , and measuring and reducing errors. Statistics in Medicine. 1996; 15: 361 – 387.

[45] Picard , R.R., Bark, K.N. : Data splitting . American Statician . 1990;44:140-147.

[46] Efron, B.: Estimating the error rate of a prediction rule : improvement on cross-validation. Journal of the American Statistical Association . 1983 ; 78 : 316-331.

[47] Efron,B. , Gong, B. : A leisure look at the bootstrap , the jackknife and cross-validation . American Statician. 1983;37:36-48.

[48] Sauerbrei , W. , Schumacher , M. : A bootstrap resampling procedure for model building : application to the Cox regression model. Statistics in Medicine. 1992 ; 11 : 2093-2109.

[49] Gifi , J. : Non linear multivariate analysis . Wiley,Chichester. 1990.

[50] McGee , D. Reed , Yano , K. : The results of logistic analysis when the variables are highly correlated : an empirical example using diet and CHD incidence. Journal of Chronical Diseases . 1984; 37 : 713-719.

[51] Chen , C.H., George , S.L. : The bootstrap and identification of prognostic factors via Cox`s proportional hazards regression model Statistics in Medicine. 1985; 4: 39-46.

[52] Altman , D.G. , Andersen , P.K: : Bootstrap investigation of the stability of a Cox regression model. Statistics in Medicine .1989:8:771-783.

[53] Efron, B., Tibshirani , R. : Bootstrap methods for standards errors , confidence intervals and other measures of statistical accuracy. Statistical Science. 1986 ; 1 : 54 - 77.

[54] http://www.worldwidewords.org./qa/qa-boo2.htm . Visitada el 12 de noviembre de 2004.

[55] http://en.wikipedia.org/wiki/Bootstrapping . Visitada el 12 de noviembre de 2004.

[56] Diaconis, P., Efron, B.: Computer-intensive methods in statistics. Scientific American.1983; May: 116-130.

[57] Efron, B.: Bootstrap methods: Another look at the jackknife. The Annals of Statistics.1979; 7: 1-26.

[58] Efron, B.: Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika.1981; 68: 589-599.

[59] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Society of Industrial and Applied Mathematics. CBMS - NSF. 1982. Monographs. n 38.

[60] Efron, B., Tibshirani, R. J.: An introduction to the bootstrap. New York: Chapman & Hall. 1993.

[61] Yu, Chong Ho : Resampling methods: concepts, applications, and justification. Practical Assessment, Research & Evaluation. 2003; 8. http://PAREonline.net/getvn.asp?v=8&n=19. Visitada el 14 de Noviembre de 2004.

[62] Fan, X., Wang, L. : Comparability of jackknife and bootstrap results: An investigation for a case of canonical correlation analysis. Journal of Experimental Education.1996; 64: 173-189.

[63] Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. New York: Springer-Verlag.2000.

[64] Edgington, E. S.: Randomization tests. New York: M. Dekker.1995.

[65] Lunneborg, C. E. : Data analysis by resampling: Concepts and Applications. Pacific Grove,CA.Duxbury.2000. 13.- Thompson, B., Snyder, P. A. : Statistical significance testing practices in the Journal of Experimental Education. Journal of Experimental Education. 1997 ; 66: 75-83.

[66] Ludbrook, J. & Dudley, H. : Why permutation tests are superior to t and F tests in biomedical research. American Statistician, 1998; 52: 127-132.

[67] Rodgers, J. L. : The bootstrap, the jackknife , and the randomization test: A sample taxonomy. Multivariate Behavioral Research. 1999 ; 34: 441- 456.

[68] Noreen, E. : Computer-intensive methods for testing hypothesis: An introduction. New York: John Wiley & Sons. 1989.

[69] Watson , J.D. : ADN . El secreto de la vida. Taurus-Pensamiento. Santillana Ediciones S.L. 2003. pp.:88-90.

[70] Rose, G.: Blackburn, H., Keys, A. Shipley, M.J. : Colon cancer and blood cholesterol. Lancet. 1974; 1: 181 - 183.

[71] Méndez, J.L., Ortega, M., Toapanta, G., Fabiani, F., Cantillana, J., Martínez Manzanares, C. : Perfil lipídico en una serie de 34 pacientes con carcinoma colorrectal esporádico: estudio transversal. Clínica e Investigación en Arteriosclerosis. 1997; 9 : 253 - 261.

[72] Law, M.R., Thompson, S.G.: Low serum cholesterol and the risk of cancer: an analysis of the published prospective studies. Cancer Causes Control. 1991; 2:253-261.

[73] Kritchevsky, S.B., Kritchevsky, D.: Serum cholesterol and cancer risk: An epidemiologic perspective. Annu. Rev. Nutr. 1992; 12: 391-416.

[74] Kono, S., Ikeda, H., Yanai, F., Yamamoto, M., Shigematsu, T.: Serum lipids and colorrectal adenoma among male self-defence officials in nothern Kyushu. Japan. International Journal of Epidemiology. 1990; 19. 274:278.

[75] Winawer, S. J., Flehinger, B. J., Buchalter, J., Herbert, E., Shike, M.: Declining serum cholesterol levels prior to diagnosis of colon cancer. A time - trend, case - control study. J.A.M.A. 1990; 263: 2083- 2085.

[76] Fernandez Bañares, F., Esteve, M., Navarro, E. Cabre, E., Boix, J., Abad Lacruz, A. et al.: Changes of the mucosal n3 and n6 fatty acid status occur early in the colorectal adenoma-carcinoma sequence. Gut. 1996; 38:254-259.

[77] Törnberg, S.A., Holm, L.E., Carstensen, J.M., Eklund, G.A.: Risk of cancer of

the colon and rectum in relation to serum cholesterol and betalipoprotein. N.Engl. J.Med. 1986;315:1629-1633.

[78] Yamada, K., Araki, S., Tamura , M., Sakai, I., Takahashi, Y., Kashihara, H., Kono, S.: Relation of serum total cholesterol, serum triglycerides and fasting plasma glucose to colorectal carcinoma in situ. International Journal of Epidemiology. 1998;27:794-798.

[79] Jacobs, D., Blackburn , H., Higgins, M., Redd,D., Iso, H. MacMillan, G. et al.: Report of the Conference on Low Blood Cholesterol :Mortality associations. Circulation. 1992; 86 : 1046-1060.

[80] Park, SK, Joo, JS, Kim, DH, Kim, YE, Kang, D, Yoo, KY: Association of serum lipids and glucose with the risk of colorectal adenomatous polyp in men: a case-control study in Korea. J. Korean Med. Sci. : 2000;15 : 690-695

[81] Eichholzer ,M. ,Stahelin, HB., Gutzwiller, F., Ludin, E., Bernasconi, F.: Association of low plasma cholesterol with mortality for cancer at various sites in men: 17-years follow-up of the prospective Basel study. Am J Clin Nutr.2000; 71 : 569 - 574.

[82] Sugarbaker, P.H..: Carcinoembryonic antigen (CEA) assays in obstructive colorectal cancer . Ann. Surg. 1976; 184: 752-757.

[83] Yuste,A.L., Aparicio,J., Segura,A., López-Tendero,P., Girones , R., Pérez Fidalgo,J.A., Díaz, R. , Calderero , V.: Analysis of clinical prognostic factors for survival a time to progression in patient with metastatic colorectal cancer treated with 5-fluoroucil-based chemotherapy. Clin. Colorrectal Cancer. 2003; 4: 231-234.

[84] Sanz Rubiales,A., García Alvarez, G., : Valor del antígeno cárcinoembrionario en el seguimiento del cáncer colorrectal. Medicina Clínica (Barcelona). 1998; 110 : 277 - 278.

[85] Guillermo Bannura C, Miguel A Cumsille G, Jaime Contreras P, Alejandro Barrera E, Carlos Melo L, Daniel Soto C. : Antígeno carcinoembrionario preoperatorio como factor pronóstico independiente en cáncer de colon y recto. Rev Méd Chile 2004; 132: 691-700.

[86] Filella,X., Molina,R., Piqué, J.M. , García-Valdecasas , J.C., Grau , J.J., Novell , F. et al.: Use of Ca 19.9 in the early detection of recurrences in colorectal cancer: comparison with CEA. Tumor Biology. 1994 ; 15 :1 - 6.

[87] Mendez Mora JL, Ortega Calvo M, Cayuela Dominguez A, Villadiego Sanchez JM, Barros Perez MM, Cantillana Martinez J. CA 19.9 and HDL-cholesterol behaviour in a sporadic colorectal carcinoma sample.
An Med Interna.(Madrid) 2004 May;21:227-230

[88] Dukes, C.E., Bussey, H.J.R. : The spread of rectal cancer and its effect on prognosis. Br. J. Cancer. 1958;12:308-320.

[89] Fabiani, F.: Métodos recomendados para la determinación de lípidos en suero. Manual de las Clínicas de Lípidos Españolas. Sociedad Española de Arteriosclerosis. Barcelona. 1992; pp .25 - 32.

[90] Friedwald, W.T., Levy, R.I., Frederickson, D.S.: Estimation of plasma low density lipoprotein cholesterol concentration with use of preparative ultracentrifugation. Clin. Chem. 1972;18 : 499-509.

[91] Vella , J.C.: Valoración del estado nutricional. Contribución del laboratorio. Rev. Diagn. Biol. 1998;47:1-9.

[92] Delgado - Rodríguez M, Medina - Cuadros M, Gómez - Ortega A, Martínez - Gallego G, Mariscal -Ortiz M, Martínez - González MA, Sillero - Arenas M.: Cholesterol and serum albumin levels as predictors of cross infection, death, and length of hospital stay. Arch Surg. 2002 ; 137 : 805 - 812.

[93] Sacks GS, Dearman K, Replogle WH, Cora VL, Meeks M, Canada T.: Use of subjective global assessment to identify nutrition-associated complications and death in geriatric long-term care facility residents. J Am Coll Nutr.2000;19 : 570 - 577. Principio del formularioFinal del formulario

[94] Rudman D, Mattson DE, Nagraj HS, Fe26.-Rudman D, Mattson DE, Nagraj HS, Feller AG, Jackson DL, Caindec N, Rudman IW: Prognostic significance of serum cholesterol in nursing home men. J Parenter Enteral Nutr 1988;12: 155 -158.

[95] Cowan LD, O'Connell DL, Criqui MH, Barrett-Connor E, Bush TL, Wallace RB.

Cancer mortality and lipid and lipoprotein levels. Lipid Research Clinics Program Mortality Follow-up Study. Am J Epidemiol. 1990;131 : 468 - 482.

[96] Lieberman DA, Prindiville S, Weiss DG, Willett W; VA Cooperative Study Group 380. Risk factors for advanced colonic neoplasia and hyperplastic polyps in asymptomatic individuals. JAMA. 2003; 290: 2959 - 2967.

[97] Moran,J.L., Solomon, P.: Worrying about normality . Critical Care and Resuscitation. 2002, 4: 316-319.

[98] Oliver D, Mahon SM. : Reading a research article part II: parametric and nonparametricstatistics. Clin J Oncol Nurs. 2005;9:238-40.

[99] Doménech-Massons,J.M., Granero,R. : Correlación y Regresión lineal. En : Métodos Estadísticos en Ciencias de la Salud . Ed. Signo Barcelona. 1999. Unidad Didáctica nº 13. p.: 24-26.

[100] Pavai S, Yap SF.: The clinical significance of elevated levels of serum CA 19-9.

[101] . Parra JL, Kaplan S, Barkin JS : Elevated CA 19-9 caused by Hashimoto's thyroiditis: review of the benign causes of increased CA 19-9 level. Dig Dis Sci. 2005;50:694-695

[102] Fuszek P, Lakatos P, Tabak A, Papp J, Nagy Z, Takacs I, Horvath HC, Lakatos PL, Speer G. : Relationship between serum calcium and CA 19-9 levels in colorectal cancer. World J Gastroenterol. 2004 ;10:1890-1892.

[103] Varol E, Ozaydin M, Dogan A, Kosar F.: Tumour marker levels in patients with chronic heart failure. Eur J Heart Fail. 2005;7:840-843.

[104] Carpelan-Holmstrom M, Louhimo J, Stenman UH, Alfthan H, Jarvinen H, Haglund C. : Estimating the probability of cancer with several tumor markers in patients with colorectal disease. Oncology. 2004;66:296-302.

[105] Mroczko B, Szmitkowski M, Wereszczynska-Siemiatkowska U, Okulczyk B. : Stem cell factor (SCF) and interleukin 3 (IL-3) in the sera of patients with colorectal cancer. Dig Dis Sci. 2005;50:1019-1024.

[106] Notarnicola M, Altomare DF, Correale M, Ruggieri E, D'Attoma B, Mastrosimini A, Guerra V, Caruso MG. : Serum lipid profile in colorectal cancer patients with and without synchronous distant metastases.Oncology. 2005;68:371-374.

[107] .- Caruso MG, Notarnicola M. : Biochemical changes of mevalonate pathway in human colorectal cancer. Anticancer Res. 2005;25:3393-3397.

[108] Shah BR, Laupacis A, Hux JE, Austin PC : Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005; 58 : 550 - 559.

[109] Song JH, Venkatesh SS, Conant EA, Arger PH, Sehgal CM. : Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. Acad Radiol. 2005;12:487-495.

[110] Austin PC, Tu JV : Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol. 2004 ;57:1138-1146.

[111] Greenland , S. : Introduction to Regression Models. In: Rothman, K.J. , Greenland , S. : Modern Epidemiology. 2nd.edition. Lippincot Williams & Wilkins. Philadelphia. 1998. pp.385-386.

[112] Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58:475-483.

[113] Cepeda MS, Boston R, Farrar JT, Strom BL. : Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am. J. Epidemiol. 2003; 158 : 280 - 287.

[114] Abraira, V.: El control de la confusión en los estudios observacionales: el índice de propensión. SEMERGEN. 2003; 29:529-531.

[115] Winkelmayer WC, Kurth T. : Propensity scores: help or hype? Nephrol Dial Transplant. 2004;19 :1671-1673.

[116] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. : Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf.. 2005; 14 : 227 - 238.

[117] Hsieh FY, Bloch DA, Larsen MD.: A simple method of sample size calculation for linear and logistic regression. Stat Med 1998 ; 17 :1623 -1634

UNIVERSIDAD DE SEVILLA

Reunido el tribunal en el día de la fecha, integrado por los abajo firmantes, para evaluar la tesis doctoral de D. José Mª Villadiego Sánchez titulada "Un modelo multivariante en el carcinoma de colon y recto. Su utilización como instrumento de ayuda del cirujano en los Centros de Atención Primaria" acordó otorgarle la calificación de

Sevilla, a 4 de Noviembre de 2006

Vocal,

Vocal,

Vocal,

Presidente,

Secretario,

Doctorando,