Universidad de Murcia
Departamento de Cirugía, Pediatría, Obstetricia y Ginecología
Facultad de Medicina

VALIDACIÓN Y ESTUDIO PROSPECTIVO COMPARATIVO DE LA APLICABILIDAD DE SEIS ÍNDICES PRONÓSTICO INTERNACIONALES DE MORBILIDAD Y MORTALIDAD EN PACIENTES INTERVENIDOS DE FORMA PROGRAMADA EN UN SERVICIO DE CIRUGÍA GENERAL Y DIGESTIVA

Tesis doctoral realizada por el licenciado en Medicina y Cirugía
Álvaro Campillo Soto, para optar al grado de Doctor en Medicina y Cirugía. Diciembre 2009

ÍNDICE

ÍNDICE

1. INTRODUCCIÓN, OBJETIVOS E HIPÓTESIS

INTRODUCCIÓN

Al comienzo del siglo XXI, España presenta uno de los mejores balances del mundo en materia de salud. Según las últimas clasificaciones de la Organización Mundial de la Salud (OMS), nuestro país ocupa el sexto lugar entre 191 países por sus resultados en materia de nivel de salud de su población, a pesar de que por su nivel de gasto sanitario, tanto total como público, en porcentaje de PIB, ocupa el lugar número 29. El principal resultado de un sistema sanitario eficiente se traduce en una esperanza de vida aumentada y libre de incapacidades en la medida de lo posible. Un elevado gasto sanitario no es condición suficiente para la buena salud de la población. De hecho puede suceder todo lo contrario, además, por encima de los 1000 dólares empleados por habitante en gasto sanitario no se observa un aumento significativo de la esperanza de vida ajustada asociado al aumento de dicho gasto[1].

Según la Organización para la Cooperación y el Desarrollo Económico (OCDE) el gasto sanitario en España se incrementará del 6,7 % del PIB de la actualidad hasta el 13 % del PIB en 2050. De los 4 factores de los que depende el crecimiento sanitario: a) Sanitarios demográficos (envejecimiento de la población); b) Sanitarios no demográficos (nuevas tecnologías); c) Demográficos asociados a dependencia (incremento de la relación anciano/joven) y d) No demográficos asociados a dependencia (costes de las enfermedades), solamente podemos influir sobre este último, reduciendo los costes de la enfermedad, lo que puede llegar a suponer hasta una reducción en el 3% del PIB destinado a asistencia sanitaria[1-4].

Se ha demostrado que las complicaciones quirúrgicas se asocian con un incremento en los costes de hospitalización, contribuyendo de forma significativa al aumento del coste por proceso. Por tanto, la reducción de dichas complicaciones es un objetivo deseable de la gestión clínica, que contribuye a reducir los costes de hospitalización, al mismo tiempo que mejora los resultados en la atención médica[5-7]. Se estima que el coste anual de las complicaciones quirúrgicas en un hospital medio es de unos 6 millones de dólares, y que, con las medidas adecuadas, se podrían reducir en un 30 – 40 % fácilmente[8].

Cuando hablamos de resultados de la atención médica nos referimos a aquellos cambios producidos en la salud, los hábitos o actitudes de los individuos, grupos o comunidades que pueden ser atribuidos a la atención médica recibida. Estos cambios no podrán ser atribuidos a los cuidados médicos mientras que otras posibles causas o factores no se descarten o controlen. Estos resultados pueden ser: *favorables* (se traducen en mejora) o *adversos* (producen un deterioro).

Las estrategias posibles para investigar problemas de calidad a partir de resultados adversos son básicamente dos:

1.- La identificación de aquellos casos individuales que reclaman una revisión del proceso asistencial en busca de problemas: consiste en identificar “sucesos centinela”, cuya característica es poseer una excelente validez a la hora de ser atribuidos a cuidados deficientes.

2.- La medición y análisis de tasas (ajustadas por riesgo previo), que se apoya en la medición de determinados sucesos que, sin justificar un estudio del proceso en cada caso individual al poderse producir aun cuando se reciben cuidados excelentes, se repiten de una forma sistemática.

Centrándonos en el caso de los resultados quirúrgicos, estos van a estar condicionados básicamente por: 1) Estado fisiológico previo del paciente; 2) Complejidad o severidad de la intervención; 3) Calidad y adecuación de la provisión de cuidados. Estos tres parámetros permiten establecer *tasas ajustadas por riesgo*, de tal manera que podemos realizar:

1. Auditorías de resultados mediante el ajuste de las tasas de mortalidad y morbilidad a la casuística de cada centro o cirujano.
2. La monitorización periódica de las razones observadas/esperadas (ratio O/E) proporcionan información acerca de la mejora o deterioro en la práctica clínica.
3. Un aumento progresivo de las ratio O/E debe servir de punto de partida para el análisis de las causas que están contribuyendo a empeorar la práctica clínica.

Al mismo tiempo que evitan los problemas derivados de las auditorías basadas en tasas brutas:

1. Hacer juicios, a veces temerarios, sobre resultados de unidades clínicas, lo que ha llevado al cierre de unidades y a la interrupción de programas de formación.

2. Realizar sesiones de morbi-mortalidad (SMM) de pacientes sin saber si el resultado obtenido era o no esperable.
3. La no valoración de los éxitos obtenidos en pacientes con alto riesgo de morbi-mortalidad[9-14].

En base a lo anterior, para medir el impacto de las complicaciones es necesario recurrir a índices pronósticos de morbi-mortalidad, que nos permitan valorar de forma objetiva el resultado en la atención médica en cuanto a la probabilidad de complicación en cada paciente determinado y así poder intentar detectar causas evitables que nos posibiliten mejorar nuestros resultados ante futuras situaciones “similares”. A pesar de la gran cantidad de índices pronósticos de morbi-mortalidad que se han ido desarrollando hasta la actualidad, no se han publicado estudios que las estudien de forma comparativa con el fin de validarlos en cada medio y definir sus características propias en cuanto a aplicabilidad, valor predictivo, precisión y capacidad de discriminación de cada una de ellas.

OBJETIVOS

El objetivo general del proyecto es validar y comparar la aplicabilidad de seis escalas de riesgo en cuanto a capacidad de predecir morbilidad y mortalidad en pacientes intervenidos quirúrgicamente en el Servicio de Cirugía General y Digestiva del HGU “JM Morales Meseguer” de Murcia, con el fin de facilitar el manejo adecuado de cada uno de ellos en nuestro medio y, por tanto, contribuir al conocimiento y a la eficiencia de nuestra práctica clínica.

Los objetivos específicos del proyecto son:

1. Definir y analizar las características propias de cada uno de los índices pronóstico a estudio.
2. Identificar la correcta utilización y selección de cada índice en función de sus propiedades, valores predictivos y ámbito donde vaya a ser aplicado.
3. Describir y facilitar la utilización adecuada de estos índices pronósticos en nuestro medio en función del tipo de paciente y tratamiento al que vaya a ser sometido.

HIPÓTESIS

1. La casuística de cada centro o servicio es diferente (case-mix) por tanto, la medición de tasas brutas de morbi-mortalidad (TB) no permiten la comparación entre centros, servicios ni cirujanos.

2. Las TB usadas habitualmente para expresar los resultados en atención médica no son buenos indicadores de la actividad sanitaria.

3. Las TB no permiten realizar auditorías de resultados de forma eficiente y real.

4. La monitorización periódica de las TB no proporciona información acerca de la mejora o deterioro de la práctica clínica.

5. Las Tasas ajustadas de morbi-mortalidad (TA) permiten realizar sesiones de morbi-mortalidad sabiendo si el resultado obtenido era o no esperable.

6. Las TA permiten valorar y estudiar los éxitos obtenidos en pacientes con alto riesgo de morbi-mortalidad.

7. Las TA permiten hacer juicios justos, reales y comparables sobre los resultados de unidades clínicas, lo que facilita la detección de deterioros en la práctica clínica, mejora de programas de formación y distribución equitativa de los recursos.
8. Cada Unidad Clínica debe seleccionar un sistema eficiente y adaptado a su actividad para monitorizar su morbi-mortalidad ajustada.

2. REVISIÓN Y ESTADO ACTUAL DEL TEMA

2.1. DEFINICIÓN Y CONCEPTO DE CALIDAD ASISTENCIAL

Introducción

Cualquier programa de mejora de la calidad debe comenzar con una definición conceptual de la calidad de forma general, junto con una referencia especial al contexto particular en el que se va a realizar la evaluación. No resulta fácil definir qué es la calidad y es por esto por lo que muy diversos autores han aportado diferentes definiciones, introduciendo nuevos y diversos matices[15].

La Real Academia Española define la calidad como “la propiedad o conjunto de propiedades inherentes a una cosa que permiten apreciarla como igual, mejor o peor que las restantes de su misma especie”. Esta definición nos permite identificar varias de sus características inherentes: 1) La calidad es un concepto cuantitativo y relativo (se puede medir y podemos tener mayor o menor calidad), por tanto, la calidad responde a una idea de mejora continua y siempre podemos buscar la forma de incrementarla; 2) Hablar de calidad implica la necesidad de comparar y para ello tenemos que medir previamente a la comparación y para medir es necesario saber qué es lo importante valorar y comparar. Por tanto, la calidad trabaja con hechos y datos, con criterios de qué es una buena práctica, con datos que nos permiten evaluarla (indicadores) y con una buena definición de qué consideramos hacerlo bien (estándares de calidad); 3) Es el cliente quien hace la comparación, entendido en sentido general como quien se beneficia de la labor profesional del otro, esta búsqueda de la satisfacción de las necesidades y las expectativas de los clientes, es uno de los objetivos importantes en todos los modelos de calidad. En el sector

sanitario, si sustituimos la palabra cliente por paciente, calidad implica que los profesionales sanitarios, en el ejercicio de su profesión, tengan en cuenta, además de las necesidades de los pacientes, sus preferencias sobre tratamiento, horarios, necesidades de información, etc[16].

Hablar de calidad en cuanto a asistencia sanitaria es algo más complejo, entre las definiciones propuestas, Lee y Jones[17] sugieren que: "La buena asistencia médica es aquella que se limita a la práctica de una medicina racional basada en las ciencias médicas. Caracterizándose la medicina racional por enfatizar en la prevención, la colaboración entre el usuario y los profesionales de la medicina, el tratamiento integral del sujeto, la coordinación con el resto de servicios médicos y de asistencia sociosanitaria y la aplicación de todos los servicios necesarios, según la medicina científica moderna, a las necesidades de la población". De esa definición podemos destacar tres aspectos más de la calidad en el ámbito sanitario: 1) Que la calidad se formula en términos de conducta normativa, como la adecuación de los procedimientos médicos al conocimiento científico; 2) Que la calidad se limita exclusivamente a la aplicación del conocimiento científico y tecnológico disponible (calidad ideal) y 3) Que la calidad está estrechamente relacionada con la práctica clínica real, ya que si no evaluamos lo que hacemos e intentamos mejorarlo no estaremos haciendo ni una buena praxis ni ofreciendo los mejores cuidados a nuestros pacientes. Desde este punto de vista, no se considera que la aplicación de la tecnología y la ciencia médica deban estar condicionadas por restricciones de tipo económico, cuando se espera algún beneficio marginal neto en el paciente[18].

La calidad ideal no es compatible con un modelo sanitario basado en la eficiencia y orientado a la atención de la comunidad. La definición de la calidad debe, por tanto, considerar los recursos disponibles (calidad relativa). El Institute of Medicine (IOM) define en este sentido que “el primer objetivo de un programa de garantía de calidad debe ser hacer los cuidados de salud más efectivos para mejorar el nivel de salud y satisfacción de la población, en relación a los recursos que la sociedad y los individuos han decidido invertir en ellos”[18]. Este enfoque presenta una serie de ventajas. Mientras que desde una perspectiva de calidad ideal, la mejora se basa fundamentalmente en la incorporación de nuevos recursos, el marco conceptual propuesto por el IOM (calidad relativa, en función de los recursos) favorece la utilización más racional de estos. Además, la calidad sólo será comparable entre distintas organizaciones cuando presenten recursos similares o los ajusten a su casuística.

Por tanto, la calidad debe orientarse hacia el óptimo social y la atención sanitaria, maximizando las preferencias y necesidades de todos los usuarios, a diferencia de la perspectiva propuesta por la calidad ideal, en la que se atienden las preferencias de cada usuario. La calidad se define en relación a la obtención de unos resultados determinados (mejora en salud y satisfacción), lo que supone un punto de ruptura con las definiciones anteriores que estaban basadas en la adecuación de buena práctica médica. El componente normativo de la calidad (lo que se debe hacer) se relaciona con su finalidad (conseguir los mejores resultados), así Brook distingue dos aspectos de la calidad de la atención: la selección de la actividad adecuada o combinación de actividades, y

la ejecución de estas actividades de modo que produzcan el mejor resultado[19]. Mientras que la especificación de las normas de la atención sanitaria se realiza exclusivamente por los propios profesionales, las definiciones de tipo finalista, como ésta y la de la IOM posibilitan que, además del proveedor de la asistencia (el médico), otros actores del proceso asistencial (administradores, usuarios) tengan un papel relevante en la formulación de los requisitos de calidad de la atención. Se introduce también un elemento de flexibilidad en la evaluación de la calidad, ya que las preferencias y la importancia de los resultados de la atención son relativos y muy diferentes según sean éstos formulados por la administración, los profesionales, o los usuarios[20].

La Joint Comisión on Accreditation of Health Care Organizations (JCAHO) propone que "la calidad es el grado en el que la atención sanitaria incrementa la probabilidad de obtener los resultados deseados por el paciente y reduce la probabilidad de resultados indeseables, ofrecidos por la situación actual del conocimiento"[21]. De este modo, las necesidades y deseos del paciente, junto con el tipo de servicio asistencial proporcionado y los resultados de ese servicio constituyen los elementos que deben ser considerados en la mejora de la calidad[22]. Al mismo tiempo se pone de manifiesto que, en ocasiones, la dificultad para mejorar el nivel de salud de una persona no es prueba de que la calidad pueda mejorarse, sino que sólo evidencia los límites de los actuales conocimientos. Por tanto, la evaluación de la calidad asistencial debe centrarse sobre los aspectos susceptibles de mejora, teniendo en cuenta las circunstancias y el estado actual de conocimientos[23].

El concepto de calidad relativa se enriquece con la aportación de Donabedian que define la calidad de la atención como “el tipo de atención que se espera que va a maximizar el bienestar del paciente, una vez tenido en cuenta el balance de ganancias y pérdidas que se relacionan con todas las partes del proceso de atención”[24]. Desde una óptica exclusivamente científica la calidad de la atención médica sería el grado en que se consiguiera restaurar la salud de un paciente teniendo en cuenta solamente la ciencia y la tecnología médicas. Cuando, desde una perspectiva individual, es el usuario el que define la calidad de la atención médica, intervienen sus expectativas y valoración sobre los costes y los beneficios y riesgos que comporta la asistencia, y obliga al paciente a implicarse en la toma de decisiones a partir de la información proporcionada por el profesional sanitario. Desde una óptica social, además de los factores enunciados en la perspectiva individual habría que considerar nuevos criterios: el beneficio o la utilidad netos de toda una población, el modo de distribución del beneficio a toda la comunidad y procurar producir, al menor coste social, los bienes y servicios más valorados por la sociedad[15].

Según la Oficina Europea de la OMS, “la calidad de la asistencia sanitaria es asegurar que cada paciente reciba el conjunto de servicios diagnósticos y terapéuticos más adecuados para conseguir una atención sanitaria óptima, teniendo en cuenta todos los factores y los conocimientos del paciente y del servicio médico, y lograr el mejor resultado con el mínimo riesgo de efectos iatrogénicos, y la máxima satisfacción del paciente con el proceso”. Esta definición destaca la adecuación de los medios empleados, la situación

del enfermo y sus conocimientos, la competencia profesional, los resultados y la seguridad y satisfacción del paciente[25].

Esselstyn propone que la calidad de la asistencia depende “del grado en el que ésta sea disponible, aceptable, extensa y documentada, así como el grado en que una terapia adecuada esté basada sobre un diagnóstico preciso y no sintomático”[26]. Incorpora dos componentes básicos: la accesibilidad, para que la asistencia esté disponible, y la aceptabilidad. El proceso asistencial y, en consecuencia, la aportación de calidad al producto salud no acaba en el profesional médico, sino que es el usuario, en función del grado de cumplimiento de las pautas recomendadas por el profesional (aceptabilidad), y la facilidad con la que obtiene la atención, el que determinará el resultado final de la atención médica.

Palmer y el Programa Ibérico (Programa de cooperación ibérica para la formación y puesta en marcha de actividades de garantía de calidad en atención primaria), incorporan al concepto de calidad asistencial la responsabilidad social del sistema. Este nuevo elemento es esencial en la calidad de cualquier sistema de salud que, como es el caso del español, tenga una orientación comunitaria. Para Palmer, la calidad de la atención es “la producción de mejora de la salud y satisfacción de una población con las limitaciones de la tecnología existente, los recursos y las circunstancias de los usuarios”[27].

En la definición adoptada por el Programa Ibérico, siguiendo la misma línea de Palmer, calidad de la atención es “la provisión de servicios accesibles y equitativos, con un nivel profesional óptimo, que tiene en cuenta los recursos disponibles y logra la adhesión y satisfacción del usuario con la atención recibida”[28]. Esta definición además de incluir componentes comunes a la mayoría de las definiciones como son el nivel profesional y la satisfacción del usuario, subraya la importancia de la equidad en la provisión de los servicios, de forma que se preste más atención a quien más lo necesite, y la limitación impuesta por los recursos existentes.

La gran cantidad y variedad de definiciones de calidad existentes, implica que no existe una definición universal y válida para todos sobre la calidad, por tanto, cada institución debe seleccionar, elaborar o adoptar la que considere más apropiada en función de la actividad que desarrolla y los servicios y resultados que obtiene[29].

2.1.1. Dimensiones de la calidad

Para que la definición de la calidad sea operativa, debe especificar los componentes que permitan su medición (dimensiones de la calidad):

1.- *La competencia profesional o calidad científico-técnica* es una de las dimensiones tradicionales de la calidad[20]. Se ha definido como la capacidad de los proveedores de utilizar el más avanzado nivel de conocimiento científico para producir salud y satisfacción en los usuarios. En la mayoría de las ocasiones se incluyen, los aspectos científico-técnicos y el trato interpersonal[29]. Es la dimensión que mejor se entiende y más frecuentemente medida como

representante de la calidad de los servicios de salud, significa atender de forma científica las necesidades sanitarias[15].

2.- *La eficacia y la efectividad* se refieren a la capacidad de la atención sanitaria de mejorar la salud y aumentar la satisfacción de la población. Aunque en ocasiones se han utilizado de forma indistinta, en la actualidad se acepta que la eficiencia es la "probabilidad de beneficio de una determinada tecnología en condiciones ideales de uso", mientras que al efectividad es la "probabilidad de beneficio de una determinada tecnología en condiciones de uso normales"[30]. Se puede medir de forma directa en función de los resultados de la asistencia, o bien de forma indirecta, mediante la evaluación de criterios de proceso que cuando se cumplen, existe una constancia previa de que produce una mejora de los resultados asistenciales[31]. Para medir la calidad científico-técnica se comprueba si en la resolución de un determinado problema de salud se han tomado o no decisiones de una comprobada efectividad y eficiencia, a la luz de los conocimientos previamente existentes sobre estas cuestiones. Mientras los estudios sobre eficiencia buscan responder a la pregunta ¿qué es lo que debe hacerse en casos como éstos?, al evaluar la competencia profesional se trata de responder a al pregunta ¿se hizo lo que había que hacer y se hizo de manera correcta?[32]. En aquellos casos en que no hay respuesta adecuada sobre eficacia de una determinada intervención no es posible contestar a la pregunta de la efectividad. La calidad se ha definido, a partir de este marco, como la reducción de las diferencias entre eficacia y efectividad atribuibles a la atención médica[32]. Para Copeland aunque estas dimensiones de la calidad, junto con la anterior, son de las que más se miden y valoran no siempre se

calculan de forma adecuada, y por tanto, su medición puede llevarnos a conclusiones irreales[34].

3.- *La eficiencia* se define en términos de la relación entre coste y producto, es decir, un máximo de efectividad o unidades de producto dado un determinado coste, o un mínimo coste dadas unas determinadas exigencias de efectividad o unidades de producto. Una intervención eficiente es aquella que maximiza los resultados para un determinado nivel de recursos[35].

4.- *La accesibilidad* es la facilidad con que la atención sanitaria puede obtenerse en relación con los aspectos (barreras) organizacionales, económicos, culturales y emocionales[15]. No tiene el mismo significado para todos los individuos porque su definición depende de la atención que se necesita, la que se requiere, la que se busca, la que se obtiene y la que se financia[20]. Para Donabedian no es un componente de la calidad, sino que se encuentra estrechamente relacionada con esta, pero cuando rebasa un nivel determinado puede provocar una atención redundante, perjudicial y costosa[24]. En todo caso, lo que subyace es la necesidad de cuantificar si la atención sanitaria llega o no a quien la necesita y cuando la necesita.

5.- *La equidad* supone que los cuidados científicamente óptimos sean accesibles, y además, que el sistema y sus profesionales estén en disposición de ofrecer mayor atención a quien más la necesita e igual atención a igual necesidad[36].

6.- L*a continuidad de los cuidados,* se refiere al grado en que los servicios médicos, que son necesarios para el usuario, se reciben como una secuencia de eventos ininterrumpidos y coordinados[37]. Se considera la dimensión de la calidad de la atención más compleja de conceptualizar y medir[15].

7.- *La satisfacción del usuario* es el grado en el que la atención sanitaria y el estado de salud resultante cumplen las expectativas del usuario[36].

8.- *La adecuación, apropiación o resultado asistencial* es otra de las dimensiones mencionadas con frecuencia. No existe una definición operativa uniforme para este término. Para la JCAHO la adecuación es la medida en que la atención médica se corresponde con las necesidades del paciente, es decir, adecuado como sinónimo de correcto, conveniente o necesario para la patología concreta que es atendida. Para otros autores, atención apropiada es, sobre todo, aquella que merece la pena hacer (beneficio de salud esperado mayor que las posibles consecuencias negativas, por un margen lo bastante grande como para deducir que merece la pena)[35]. Para Saturno la característica definitoria de esta dimensión hace referencia a si la elección del tipo y cantidad de atención es razonable a la luz del conocimiento científico existente[15]. Es un concepto muy difícil de separar a la hora de ser medido, del de calidad científico-técnica o competencia profesional. En definitiva, hace referencia al resultado de la atención médica recibida (adecuada o inadecuada). Esta dimensión constituye el eje central para medir la calidad asistencial en nuestro estudio y será tratada en profundidad en los siguientes apartados.

Una vez que hemos estudiado diferentes definiciones de calidad y sus dimensiones, podemos establecer que el marco conceptual que ha sido aplicado en nuestro trabajo es el que propuso el Programa Ibérico y ha sido adoptado por el Programa EMCA de la Región de Murcia y que define la calidad de la atención como “la provisión de servicios accesibles y equitativos, con un nivel profesional óptimo, que tiene en cuenta los recursos disponibles y logra la adhesión y satisfacción del usuario”[38].

2.2. LA VARIABILIDAD EN LA PRÁCTICA CLÍNICA

Introducción

La práctica clínica es el proceso de la actuación médica en relación con el cuidado del paciente. Sus componentes son el cuerpo de conocimientos clínicos disponibles, los datos clínicos del paciente, las percepciones, juicios, razonamientos y decisiones de los médicos, los procedimientos que estos utilizan y las intervenciones que aplican.

Mientras que la medicina se ha caracterizado por prestar mucha atención a la investigación de las causas y mecanismos biológicos de la enfermedad, se ha investigado muy poco sobre el proceso de la práctica clínica y se desconoce mucho de cómo los médicos obtienen y usan la información clínica, aplican los conocimientos diagnósticos y terapéuticos, predicen los desenlaces y evalúan los intereses y preferencias de los pacientes, es decir, acerca de los determinantes y consecuencias de las decisiones clínicas[39,40].

La variabilidad de la práctica clínica se refiere a la existencia de diferencias en la utilización de los recursos, los servicios y los procedimientos sanitarios, entre los proveedores de la atención, una vez que se han controlado los factores demográficos, sociales, de nivel de salud, tipo de casuística (case-mix) y económicos[33,41]. Como es lógico, esta variabilidad en la práctica clínica va a dar lugar a diferencias en cuanto a la efectividad de la asistencia prestada por diferentes proveedores de la atención (médicos, servicios, hospitales)[41].

Puede ser de dos tipos: aleatoria y asignable. La variabilidad aleatoria se caracteriza porque en su origen intervienen múltiples causas, se debe exclusivamente al azar y puede determinarse mediante una fórmula matemática[42], pero, dada su naturaleza azarosa, es difícil de controlar de forma eficiente. La variabilidad asignable no se puede predecir mediante fórmulas matemáticas y, se debe a un número reducido de causas que se pueden identificar y controlar de forma eficiente[42].

La existencia de variaciones importantes en relación a un punto de referencia en la práctica clínica, ya sea éste de tipo descriptivo (media, mediana, moda de la práctica profesional habitual) o de tipo normativo (práctica profesional que se acepta como adecuada) indica que una elevada proporción de procedimientos pueden ser innecesarios o inadecuados. Supone que se aplican diferentes criterios científico-técnicos en la atención de un mismo problema, y se traduce en falta de efectividad (no se consiguen los resultados esperados) y/o ineficiencia (se consiguen los resultados esperados, pero con costes más elevados que otros procedimientos normativos) como consecuencia de la utilización innecesaria o inadecuada de recursos. Por todo esto es importante que seamos capaces de controlar todos los factores de confusión, con el fin de poder medir de forma real qué se debe a variabilidad en la práctica clínica y qué se debe a factores de confusión que nos pueden llevar a obtener resultados “sesgados” y conclusiones erróneas[34]. Por otra parte, debemos recordar que la variabilidad clínica puede estar influenciada por el tipo de actividad asistencial (servicios quirúrgicos frente a servicios médicos), ya que tan sólo el 10-20% de las decisiones en cuanto a tratamiento en los

servicios quirúrgicos son en base a evidencias científicas sólidas, frente al 30-50%, de las que se toman en los servicios médicos[43].

Los estudios sobre variabilidad en la práctica clínica se pueden agrupar en dos tipos según sean estudios poblacionales o de base individual. Por una parte están aquellos trabajos de tipo ecológico, que relacionan el número de residentes en las áreas geográficas a estudio, que han recibido un determinado servicio sanitario en un periodo de tiempo definido, con la población total de tales áreas en dicho periodo. El denominador empleado en estos trabajos es población en riesgo durante un determinado periodo y, por tanto, el término epidemiológico preciso es incidencia acumulada. El objetivo es comparar las tasas de frecuentación hospitalaria[44-48] y de procedimientos médicos o quirúrgicos[49-59] y valorar si la variabilidad entre áreas implica una diferente utilización de los servicios estudiados. Los resultados obtenidos suelen interpretarse como evidencia indirecta de la existencia de componentes evitables de la atención sanitaria que, según la magnitud de las variaciones halladas, pueden tener implicaciones en los costes y en los resultados de la atención médica[60,13,41].

Por otra parte están los estudios sobre variabilidad de indicadores del proceso y/o resultado de procedimientos en pacientes con patologías específicas. La variabilidad en la práctica clínica se caracteriza por la existencia de diferencias en el proceso asistencial y/o en el resultado de la atención de un problema clínico concreto entre diversos proveedores (variabilidad interproveedores), o un mismo proveedor a lo largo de un periodo de tiempo (variabilidad intraproveedor), una vez que se han controlado los factores

demográficos, socioculturales y de nivel de salud[61,34]. En estos trabajos se analiza principalmente la forma en que se presta el servicio, como la diferente utilización de pruebas diagnósticas[62-65], prescripción de medicamentos[66-70], duración de la hospitalización[71] u otros, en pacientes en situaciones clínicas similares. Estos estudios a diferencia de los anteriores, se desarrollan sobre una base individual y sus objetivos son evaluar la efectividad o la eficiencia de los centros o profesionales sanitarios, o buscar determinantes de variabilidad en función de los pacientes (como sexo[72-76], grupo étnico[77-79] y nivel socioeconómico[80]), del médico (especialidad[71,81,82], sexo[83], formación, experiencia[84-86], y sistema de pago), del hospital (público o privado, rural o urbano, docencia, tamaño) o del sistema sanitario (financiación, organización, cobertura u otras[87-93]).

En España se han publicado en los últimos años diversos trabajos en esta línea relacionados con frecuentación hospitalaria[94,95], variabilidad en procedimientos médicos o quirúrgicos[96-102], en la atención según género del paciente[103-106], determinación de pruebas diagnósticas[107-109] o prescripción de medicamentos[110-112].

La variabilidad de la práctica clínica está adquiriendo cada vez mayor interés por motivos clínicos y económicos. Por un lado, cada vez más la comunidad científica y la población exigen la constatación de que las variaciones no entrañen diferencias en los resultados obtenidos en los pacientes. Por otro lado, planificadores, compradores, y gestores de servicios sanitarios identifican la presencia de variabilidad como un factor a controlar, por la repercusión que tiene en el consumo de recursos los patrones de gran utilización.

El aspecto clave de los estudios sobre variabilidad es averiguar si un alto consumo de servicios lleva consigo un mayor beneficio o riesgo para la población que accede a ellos, o si una baja utilización de determinados procedimientos condiciona un peor pronóstico. Lo importante es conocer cuál es la tasa óptima de realización de una técnica o de utilización de un servicio en una población de características controladas, que nos permitiese garantizar que se está prestando una asistencia de calidad óptima[113]. Estudios recientes realizados por la OMS[114] muestran claramente que un elevado gasto sanitario no es una condición suficiente para la buena salud de la población. De hecho, puede suceder todo lo contrario. En general, sin embargo, un mayor gasto sanitario por habitante (muy correlacionado con la renta per cápita), por ejemplo, está asociado con una mayor esperanza de vida ajustada, como se muestra en la figura 2.1. Puede apreciarse que a medida que aumenta el gasto por habitante la esperanza de vida ajustada aumenta rápidamente; pero a partir de los 200 dólares (en 1997), los aumentos de la esperanza de vida no son tan significativos. El ajuste de una función logarítmica a la nube de puntos muestra una serie de intervalos significativos. Para que la esperanza de vida ajustada por discapacidad aumente de 60 a 70 años, aproximadamente, hay que aumentar el gasto sanitario per cápita, de 200 a 1200 dólares, mientras que para que esta esperanza de vida aumente a partir de los 70 años, el aumento del gasto por habitante debería ser asintótico, es decir, desproporcionado. Por encima de los 1000 dólares por habitante, de hecho, no se observa un aumento significativo de la esperanza de vida ajustada asociado al aumento de dicho gasto.

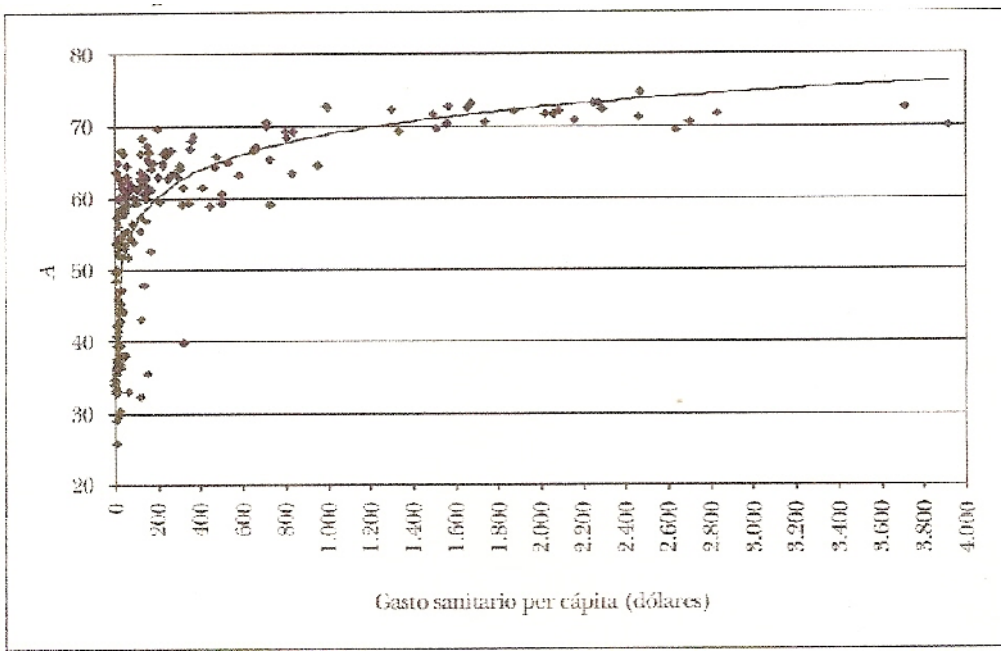


Figura 2.1. Gasto sanitario por habitante y esperanza de vida (OMS 1997)[1].

El primer estudio sobre variabilidad en la práctica clínica se debe a Glover[115] (1938), que demostró que en los distritos escolares británicos, la tasa de amigdalectomía podía variar hasta en 8 veces en función del distrito en el que se estuviera adscrito, sin que existiera ninguna explicación aparente. Posteriormente, en la década de los 60 y 70, tuvieron especial impacto el estudio de Wennberg y Gittelsohn[116] sobre las variaciones en las tasas de intervenciones de adenoidectomía, prostatectomía, histerectomía, hernia inguinal y colecistectomía, que sentaron las bases conceptuales para el análisis actual de las variaciones en la práctica, y en buena medida, para el desarrollo de los programas de investigación sobre efectividad de los tratamientos y de difusión de resultados. Durante su estudio, estos autores observaron que, en Vermont, las tasas de intervenciones según áreas hospitalarias vecinas variaban hasta en 6 veces para la adenoidectomía y en 4 veces para la

histerectomía y la prostatectomía. Estos autores estimaron la probabilidad de haber sido intervenido en función de la edad y área de residencia, observando que en el área con mayor incidencia de intervenciones, la probabilidad de haber sido amigdalectomizado antes de los 25 años era del 64 % frente a un 8 % en las de menor incidencia y un 25 % para el conjunto de las áreas. Para la probabilidad de haber sido histerectomizada antes de los 75 años las proporciones variaban del 15 % al 55 % con una media del 40 % y para la prostatectomía a los 85 años, las probabilidades oscilaban entre el 35 % y 60 % con un promedio del 50 %. Tras demostrar que no existían diferencias importantes en la salud de la población ni en su situación económica, y argumentar que las diferencias en el número de médicos, camas hospitalarias y nivel de cobertura de los planes de seguro sanitario no podían ser responsables de la totalidad de la variación, concluyeron que el factor más importante debía ser el estilo de práctica.

Desde la aparición de estos dos trabajos, han sido muchos los autores que han publicado tasas diferentes de intervenciones entre países, regiones, áreas hospitalarias o áreas pequeñas, poniendo de manifiesto la amplia variabilidad existente en la práctica clínica[117-122]. Diversos trabajos realizados en áreas pequeñas vecinas, con similares condiciones socio-económicas, demográficas y de estado de salud han puesto de manifiesto que las diferencias en cuanto a variabilidad de práctica no se deben a diferencias en cuanto a la morbilidad de la población[123-130].

2.2.1. Factores explicativos de la variabilidad.

Los médicos pueden tener diferentes opiniones diferentes sobre los méritos relativos de las diversas opiniones de tratamiento o las estrategias diagnósticas para una misma condición. Básicamente, el origen de estas diferencias de opinión[131] se halla en la presencia de la incertidumbre (no existe evidencia científica sobre los resultados de las posibles alternativas de tratamiento o sobre el valor de determinadas pruebas diagnósticas en situaciones concretas), o en la ignorancia (existe evidencia científica sobre el valor de las pruebas o tratamientos, pero el médico la desconoce o, aun conociéndola, emplea otras pautas)[132,133]. El análisis de la variabilidad apunta hacia la discrecionalidad de las decisiones clínicas en situaciones de incertidumbre, los estilos de práctica clínica, como la principal explicación del fenómeno de la variabilidad.

Los estilos de práctica serían determinantes de variabilidad en la utilización de los servicios a nivel poblacional, sólo para aquellas situaciones en las que existe incertidumbre, que serían las situaciones que mostrarían variaciones importantes. Dada la frecuencia de las situaciones de incertidumbre en la realidad clínica[33,134,135], y la ausencia de evidencia científica respecto a muchas de las prácticas médicas habituales, esta condición no resta importancia a los estilos de práctica como factor explicativo de variabilidad.

Las propuestas básicas de la hipótesis de la incertidumbre[136] pueden resumirse en:

1. Las diferencias en morbilidad y otras variables de la población no explican sustancialmente la variabilidad.
2. La variabilidad es escasa cuando existe acuerdo entre los clínicos sobre el valor de un procedimiento.
3. En aquellos casos en que existe incertidumbre sobre la utilidad de un procedimiento, los clínicos desarrollan estilos de práctica diferentes que son la principal fuente de variabilidad.
4. Los factores de la oferta, volumen, incentivos y otros, pueden ser relevantes en los procesos de alta incertidumbre, pero su influencia será escasa en aquellos casos en que exista consenso ante qué hacer en una situación dada.
5. Las variaciones son un indicador de utilización inadecuada en las áreas con mayores tasas debido al exceso de utilización por demanda inducida.

El principal problema en la hipótesis de la incertidumbre, es que dado que no existen mecanismos formales para categorizar los procedimientos por su grado de incertidumbre salvo la propia presencia de variabilidad, la teoría se ha desarrollado sobre un razonamiento circular[137]. Sin embargo, existen algunas evidencias indirectas que dan soporte a la proposición sobre el estilo de práctica como fuente clave de variabilidad. Así, los estudios han mostrado descensos importantes en las tasas de intervenciones de una población tras el cambio de los profesionales que las atendían, tras intervenciones de

retroinformación a los clínicos sobre la variabilidad en las tasas de intervenciones o tras campañas de información a la población[138], refuerzan la importancia del estilo de práctica como factor de variabilidad, al menos en intervenciones quirúrgicas como las amigdalectomías e histerectomías.

Por otra parte, no es obvio si las variaciones indican uso inadecuado por exceso en las áreas de alta utilización o por defecto (subprovisión de cuidados) en las áreas de baja utilización, existiendo evidencias contradictorias al respecto. Aunque los reducciones en las tasas de procedimientos quirúrgicos, tras algún tipo de intervención, sugieren la existencia de sobreutilización en las áreas con tasas de intervenciones elevadas, los escasos trabajos que han estudiado directamente la proporción de intervenciones inadecuadas en áreas de alto y bajo uso no hallaron diferencias entre ellas[117,122,139]. La mayor parte de estos trabajos se limitan a algunas intervenciones concretas, en poblaciones muy determinadas (habitualmente pacientes cubiertos por Medicare) y su planteamiento es metodológicamente discutible, ya que no estudian las tasas de intervenciones adecuadas e inadecuadas, sino la proporción de intervenciones inadecuadas entre las realizadas (aunque la proporción de intervenciones inadecuadas sea igual, por ejemplo del 30 %, en áreas de alta y baja utilización, no tiene el mismo impacto el 30 % de una tasa de 10/1000, que de una tasa 100/1000). En este sentido, y aunque las variaciones son consideradas en muchas ocasiones como un indicador de utilización inadecuada, conocer la tasa adecuada de una intervención en una población requiere la investigación de resultados y, por tanto, no es posible conocer si indican uso inadecuado por defecto o por exceso[140] a partir de estudios ecológicos de tasas de utilización.

2.2.2. Estrategias de actuación ante las variaciones en la práctica clínica

El estilo de práctica es el factor con más influencia en la realización de los procedimientos cuyos criterios de indicación son más controvertidos, es decir, aquellas técnicas cuyos beneficios y riesgos han sido menos estudiados y en los que, por tanto, no existe consenso general sobre cual es la práctica de más calidad. A partir de esta premisa, Wennberg se ha constituido en el líder de un grupo cuya principal propuesta para disminuir la variabilidad se basa en la realización y difusión de estudios de investigación sobre la efectividad de los procedimientos más frecuentes, para así disponer de suficiente evidencia científica que dé soporte a los clínicos en la toma de decisiones[113,141].

La general aceptación de que la variabilidad traduce problemas de calidad de las actuaciones médicas, debidos al uso inadecuado de los recursos, es el origen de la preocupación de la comunidad sanitaria por la variabilidad[142-144]. Además, planificadores, compradores y gestores están interesados en su estudio, por la suposición de que constituyen una oportunidad de reducir el gasto sanitario. La importancia de las variaciones estriba en que pueden ser reflejo de otros problemas (incertidumbre, ignorancia, problemas organizativos, gastos innecesarios o infrautilización), cuyo abordaje redundaría en una mejora de la calidad de la atención, aunque no necesariamente siempre en una disminución de los gastos.

La variabilidad de la práctica clínica también puede ser reflejo de fenómenos, en principio, no tan fácilmente abordables (distinta morbilidad o distintas preferencias informadas de la población). El caso de las preferencias de la población es cierto que es difícilmente abordable, pero en el de la distinta morbilidad sí existe una oportunidad de mejora importante, ya que si conseguimos ajustarla a riesgo, podríamos detectar diferencias reales entre áreas, hospitales, médicos, servicios, y de este modo estudiar el porqué de estas diferencias (factor médico o cirujano, factor cuidados perihospitalarios o perioperatorios, factor cuidados postquirúrgicos…)[13,34] y actuar en consecuencia.

Por tanto, ante un procedimiento donde existe un claro consenso sobre efectividad y sus indicaciones, cabe pensar que la variabilidad sólo puede ser explicada por diferencias en la demanda (morbilidad y preferencias de los pacientes)[138] o por ignorancia de los profesionales sanitarios[145,146] ya sea en el sentido de no utilizar un procedimiento de efectividad comprobada o de usar uno cuya efectividad no está demostrada[147].

La política sanitaria debe apoyar las estrategias clínicas de reducción de la variabilidad y mejora de la efectividad, adecuación eficiencia de los procesos médicos, estimulando la gestión clínica, los movimientos del tipo de medicina basada en la evidencia, la investigación de resultados y su difusión entre la comunidad sanitaria. Más aún si se tiene en cuenta la aceptabilidad de estas estrategias por su cercanía al pensamiento médico y su mayor legitimidad social, al intentar basar el control de la utilización en la reducción de los

servicios innecesarios. Desde una óptica en la que se considere la salud como un derecho de la población, la disminución de la variabilidad, con su consiguiente efecto sobre la efectividad clínica, es uno de los pasos fundamentales para que otro individuo pueda beneficiarse de la atención médica, con el consiguiente incremento de la eficiencia social del sistema sanitario[148].

2.3. MEDICINA BASADA EN EVIDENCIA Y BENCHMARKING

2.3.1. Por qué una medicina basada en evidencia (MBE)

El médico ejerce su profesión mediante el uso del conocimiento y habilidades clínicas aprendidos en sus años de formación. Mientras que los conceptos y conocimientos con los que el médico afronta su práctica clínica diaria se van reduciendo progresivamente, la investigación biosanitaria y las publicaciones se van incrementando de forma exponencial, por tanto, el resultado es que lo que aprendimos en nuestros años de formación se va quedando obsoleto.

Cada año aparecen en todo el mundo unos 2 millones de artículos, publicados en unas 25.000 revistas médicas especializadas. Pero, la investigación médica anual produce muchos más datos, que nunca serán publicados. Se estima que entre el 50 y el 70 % de todos los resultados de estudios no son publicados, estos datos “no publicados” suelen corresponder a resultados negativos o desfavorables. Por tanto, existe un sesgo de publicación hacia los resultados favorables (selección de datos) o que apoyen un tratamiento o procedimiento en lugar de otro[149]. Este fenómeno de selección de datos además de reducir la transparencia y objetividad del método científico puede dar lugar a recomendaciones terapéuticas erróneas. Whittington et al[150] compararon los datos publicados sobre los inhibidores de la recaptación de serotonina frente a los no publicados por la industria, y observaron que los ensayos publicados sugerían que estos psicofármacos podían ser útiles en el tratamiento de niños depresivos, mientras que los datos no publicados

demostraban que no sólo no reducían la dolencia en éstos sino que incluso podían aumentar el riesgo de suicidio.

Por un lado, se ha calculado que cada 45 años la mitad de los conocimientos quirúrgicos son sobrepasados y hace falta renovarlos, y se ha demostrado que un internista, para poder practicar un diagnóstico y tratamientos actualizados, debe leer 19 artículos por día[151,152]. Por otro lado, los médicos no dedican más de 30-60 minutos por semana a leer literatura científica. Este desequilibrio entre el enorme volumen de información y la escasa disponibilidad de tiempo dedicado al estudio, supone que la mayoría de los profesionales se informen de los avances médicos a través de los medios de difusión de la información médica secundarios (revistas de difusión científica, noticiarios, agentes técnicos sanitarios, etc), en lugar de leer directamente los trabajos originales en las revistas científicas[43].

La investigación en cirugía se encuentra en peor situación que la médica, ya que no sólo el 10-20% de las decisiones en cuanto a tratamiento en los servicios quirúrgicos son en base a evidencias científicas sólidas, frente al 30-50%, de las que se toman en los servicios médicos[43], sino que se estima que sólo el 3 % de los estudios clínicos realizados en cirugía satisfacen los estándares de calidad exigidos[153]. Esto nos lleva a pensar que la medicina del siglo XXI, a pesar de los avances científicos y tecnológicos, sigue guiándose por prácticas antiguas, con escasa evidencia frente a otras alternativas menos agresivas y con poca confianza en su utilidad por quienes las practican[154].

Domenighetti[154], comparó la frecuencia con se realizaban siete intervenciones muy frecuentes, aunque no urgentes, entre personas que presentaban un estado similar de salud, para ello dividió a los pacientes en dos grupos (médicos y no médicos), y demostró que con excepción de la operación de apendicectomía (similar en ambos grupos), el resto de intervenciones se aplicaban con una frecuencia de un 33 % más en los pacientes no médicos (Figura 2.2). En el análisis por subgrupos, se observó que el grupo de abogados presentaba las mismas tasas de intervenciones que el grupo médico. Domenighetti concluyó que: "los abogados son pacientes de riesgo" porque en el caso de que una operación escasamente indicada saliese mal, podrían defenderse mejor que los ciudadanos normales".

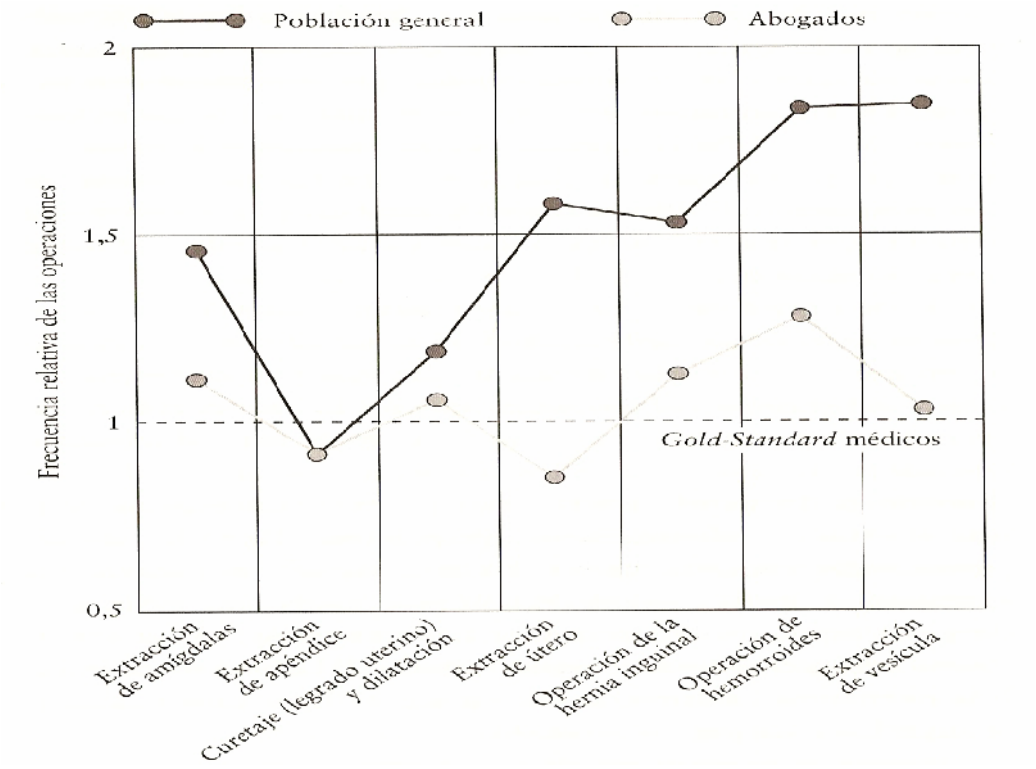


Figura 2.2 Diferencias en la utilización de procedimientos quirúrgicos entre médicos y la población general. (Adaptado de Domenighetti G et al. Revisiting the most informed consumer of surgical services – the physician patient. International Journal of Technology Assessment in Health Care 1993)[154].

2.3.2 Cáncer, pruebas de cribado y MBE

Aproximadamente un 0,1 % de las personas mayores enferman de cáncer de tiroides y presentan síntomas clínicamente reconocibles. Black y Welch[155] extirparon la glándula tiroides a personas que habían fallecido por otras causas e hicieron secciones de 2,5 mm que examinaron con técnicas de anatomía patológica y diagnosticaron un 36 % de carcinomas tiroideos en las piezas, concluyeron que: “debido a que muchos tumores miden únicamente 0,5 mms de diámetro, si hubieran hecho cortes más finos habrían diagnosticado muchísimos más tumores, incluso si las secciones hubieran sido lo suficientemente finas habrían hallado un carcinoma en casi todas las piezas”.

Investigaciones posteriores han confirmado estos resultados para otros tipos de tumores. Folkman y Kalluri[156] confirmaron estos hallazgos y concluyeron que: “Si se aplican las técnicas diagnósticas más avanzadas casi todas las personas tienen en algún lugar de su cuerpo pequeños tumores, a pesar de gozar de un perfecto estado de salud y de que estas células cancerígenas no les causarán ningún problema en el futuro”. La explicación de lo anterior parece estar en que para crecer estos minitumores necesitan, como toda célula eucariótica, oxígeno y nutrientes abundantes, y son precisamente estas dos cosas las que el cuerpo humano les “niega” como mecanismo defensivo antitumoral. Para que se desarrolle y extienda un cáncer en el organismo son necesarios dos pasos: 1) Mutaciones al azar en una célula normal que la transforman en una célula con características tumorales (ausencia del límite de Hayckflick, no inhibición por contacto…) y 2) Angiogénesis, por la cual la nueva célula mutada crea sus vasos para poder nutrirse y oxigenarse. Si en un organismo se da el primer paso, pero no el

segundo, el tumor no supondrá una amenaza para la vida del organismo, ya que al no poder crecer de forma desproporcionada, el cuerpo podrá mantenerlo controlado, pero si se produce también el segundo paso, entonces es cuando aparece el tumor maligno y de crecimiento rápido y fatal[156,157]. Sabemos que el sistema inmune del ser humano dispone de forma natural de mecanismos para inhibir la angiogénesis y que por tanto, la mayoría de estos tumores serán autolimitados y no tendrán consecuencias clínicas fatales, sino que simplemente se limitarán a ser hallazgos incidentales en autopsias y pruebas de imagen. Por tanto, si sabemos todo esto, ¿por qué la terapéutica del cáncer no sigue estas líneas de investigación?, aún a sabiendas de que, como veremos en el siguiente apartado (resultados en asistencia médica), los fármacos quimioterápicos no han producido un incremento neto de la supervivencia desde que aparecieron los primeros en los años 70 hasta la actualidad.

Todo lo anterior nos permite explicar por qué algunos programas de screening de cáncer son cuestionables desde el punto de vista de sus beneficios reales[158]. En el caso de cáncer de mama (tumor maligno más frecuente en la mujer) se estima que una de cada 10 mujeres contraerá cáncer de mama, los partidarios del cribado defienden que la detección temprana de los carcinomas de mama reduciría en un 20 - 30 % la tasa de mortalidad entre las mujeres sometidas a la prueba. Mühlhauser et al[159] demostraron que la única utilidad real de los programas de cribado de cáncer de mama radica en el hecho de que alargan el tiempo entre el diagnóstico y la muerte, y con ello el tiempo en que una mujer es paciente de cáncer de mama, pero sin que sirvan

para retrasar la progresión de la enfermedad, debido a que, por lo general, el screenning detecta variantes del cáncer que progresan de forma relativamente lenta, e incluso una detección más tardía no empeoraría el pronóstico. Además, la afirmación de que una de cada 10 mujeres contraerá cáncer de mama es cierta pero sólo si se aplica a mujeres mayores de 80 años, ya que en realidad, afirman estas autoras: “El cáncer de mama es la causa de muerte únicamente para 3 ó 4 mujeres de cada 100, las 96 ó 97 restantes mueren por otras causas”. Mühlhauser et al[159] elaboraron un estudio con un seguimiento de 10 años a 2000 mujeres de entre 50 y 60 años, divididas en dos grupos de 1000 pacientes cada uno (con y sin mamografía) y evaluaron los resultados pasado estos 10 años (figura 2.3), siendo la mortalidad global de ambos grupos semejante, y del 0,8 % por cáncer de mama en el grupo sin mamografía frente al 0,6 % para el grupo con mamografía, no existiendo diferencias significativas entre los resultados de ambos grupos. En el grupo sometido a mamografía un 6 % de las mujeres fueron sometidas a cirugía mamaria por resultados falsos positivos de la mamografía. Si analizamos estos resultados podemos ver que la reducción de la mortalidad del 20-30 % gracias al cribado no es cierta, ya que la diferencia de mortalidad entre ambos grupos es del 0,2% (sin significación estadística), además, hay que considerar los problemas derivados de los falsos positivos y las cirugías innecesarias, así como los problemas postquirúrgicos y radioterápicos, por otro lado, la mamografía presenta otros dos inconvenientes: 1) La tasa de falsos negativos (mujeres con cáncer no detectado en la mamografía y diagnosticados dentro de los 12 meses siguientes), en torno al 0,2 % y 2) Los tumores radioinducidos por la mamografía; estadísticamente si 20000 mujeres se sometieran de forma regular a la mamografía de screening

desde los 40 años, entre 3 y 9 de ellas desarrollaran un cáncer como consecuencia de la radiación a la que serían expuesta[158,159].

Sin mamografía	*Con mamografía*	
8	6	Muertas de cáncer de mama
72	74	Muertas por otras causas
920	920	No muertas
25	30	Diagnóstico: CM.
975	970	Diagnóstico: no CM.
	5000	Cifra total de los exámenes de mamografía
En cada uno de los dos grupos se observa durante un período de 10 años a 1.000 mujeres de entre 50 y 60 años. Las mujeres del grupo «con mamografía» han pasado cinco exámenes de mamografía.	200	Mujeres con al menos un resultado de mamografía sospechoso
	60	Mujeres a las que se extrae tejido del pecho para aclarar falsos resultados positivos

Figura 2.3. Utilidad de la mamografía. (Adaptado de www.nationales-netzwerk-frauengesundheit.de)[159].

Otro ejemplo de la incertidumbre y confusión que existe en torno a las pruebas de cribado lo tenemos en cuanto al cáncer de próstata. Concentraciones del antígeno específico de la próstata (PSA) mayor de 4 ng/ml halladas en los tests de cribado suelen ser sospechosas de malignidad, tras este hallazgo, el paso siguiente suele ser una biopsia prostática. Stamey[158,160] realizó un estudio sobre 1300 muestras de próstata extraídas en el departamento de urología de la universidad de Stanford entre 1983 y 2003, comparando los hallazgos anatomopatológicos con los valores correspondientes de PSA, observando que en los primeros años de realización del cribado sí existía una buena correlación entre los valores elevados de PSA y tumores malignos y de gran volumen, mientras que conforme se avanzó en los screening y hombres cada vez más jóvenes eran sometidos a estos tests, el valor de PSA no guardaba correlación con los hallazgos histológicos, incluso los resultados obtenidos en los últimos cinco años del estudio demostraban que los valores de este antígeno estaban relacionados con la hiperplasia benigna de la próstata, por tanto, concluyó que el PSA no es un test fiable para el screening del cáncer prostático. Estos hallazgos también fueron confirmados por investigadores del Memorial Sloane-Kettering Cancer Center de New York[161], ya que tras examinar 2000 próstatas que habían sido extraídas en base a los hallazgos de la biopsia, observaron que en el 30 % de ellas, o bien encontraron sólo carcinomas microscópicos o bien no hallaron ningún cáncer en absoluto.

Otras desventajas publicadas del uso del PSA son:

1. Elevada tasa de falsos positivos (PSA elevado (> 4ng/ml) en las pruebas de screening sin confirmación de cáncer en los estudios histológicos posteriores)[158,160,161]: Entre el 10 y el 15 % de los resultados considerados como patológicos se consideran errores diagnósticos.

2. Falsos negativos (PSA normal (< 4ng/ml) en las pruebas de screening y cáncer en los estudios histológicos posteriores)[162]: Hasta en un 15% de los pacientes con cáncer de próstata avanzado presentan cifras normales de PSA.

Por todo lo anterior, la Red Alemana para la Medicina Basada en la Evidencia[163] considera que: “No existe un criterio unánime que certifique que el uso del PSA en el screening del cáncer de próstata logre alargar la vida de los hombres afectados por este tumor”

Estos ejemplos ponen de manifiesto la necesidad de una medicina basada en evidencia y de la necesidad de la investigación en resultados en asistencia sanitaria, ya que, todo esto demuestra que a pesar del esfuerzo económico tan elevado para los programas de cribado su utilidad real es escasa.

2.3.3. Qué es la MBE

Con el concepto de MBE se apela a cuestiones relativas a la naturaleza de la ciencia médica o del conocimiento por una parte y, por otra, al uso de este mismo conocimiento en el ejercicio clínico diario. Además, la cuestión del conocimiento aparece unida a la de su organización, generación, comprensión, contraste y difusión como cuestiones intrínsecas relacionadas. Por otra parte, la MBE debe constituir una cierta manera de ejercer la medicina que puede implicar organizarse de forma diferente y con otra logística. Los aspectos filosóficos de la naturaleza del saber médico han sido objeto de estudio y debate desde los inicios de la medicina. Los aportes de la investigación clínica cuantitativa han constituido una herramienta adicional para comprobar la relativa idoneidad de los modelos conceptuales de enfermedad aumentando el campo de las observaciones[43].

Los orígenes de la MBE se remontan a la primera mitad del siglo XIX, cuando, por un lado, el médico francés Louis decidió aplicar su “método numérico” para valorar la eficiencia de la sangría en distintas patologías. Comparó los resultados obtenidos con los pacientes con las mismas enfermedades pero que no habían sido sometidos a esta terapia y no halló diferencia alguna. Louis creó en 1834 un movimiento que denominó “Medecine d’observation” y mediante diversos experimentos contribuyó a la erradicación de terapias inútiles como la sangría. Por otro, en 1840 el médico J Gavarret publicó “ Principes généraux de statistique médicale”, sentando las bases de la estadística médica moderna[164,165]. Las aportaciones de esto dos pioneros de la “ciencia médica” tuvieron una gran repercusión en Europa (Francia e Inglaterra) y en EE.UU. Aunque los orígenes filosóficos de la MBE son más antiguos, el

“status” legal fue establecido en el Acta de Farmacéuticos de 1815, la cual concedía licencias de boticarios con el fin de proteger a la población de un creciente número de herboristas y boticarios sin ninguna cualificación profesional. El Acta Médica de 1858 lleva a la creación de un registro de médicos, el cual contiene los nombres de todos los doctores con una preparación médica reconocida e implica que la medicina practicada por estos médicos se basaba en la evidencia mientras que la medicina alternativa se basaba en rumores, viejos cuentos o remedios ancestrales[43].

A pesar de todos estos antecedentes, no es hasta mediados del siglo XX cuando Bradford Hill sienta las bases de la metodología de los ensayos clínicos y los fundamentos estadísticos de la MBE. A partir de los años 80 Feinstein et al conceptualizan la epidemiología clínica como disciplina[166]. En la década de los 90, Sackett publicó las bases conceptuales de la MBE[167]: “la MBE integra dos componentes: por un lado la experiencia clínica individual y de otro la mejor evidencia clínica disponible derivada de la búsqueda sistemática. La pericia clínica individual se adquiere como resultado de la práctica clínica y esto significa que un clínico no va a seguir de forma servil las reglas dictadas por otros cuando trata a un paciente en particular. Por otro lado, los resultados de una búsqueda clínica excelente proporcionan un soporte científico válido para el cuidado del paciente”. Ambos componentes son necesarios, ya que la experiencia clínica sin la aplicación de los resultados de las nuevas investigaciones puede llevar a un estancamiento, ya que no se puede esperar mejorar sin la educación continua que proporcionan las buenas publicaciones clínicas.

Por tanto, la MBE se puede definir como: “La integración de la experiencia clínica personal con la mejor evidencia posible externa disponible procedente de la investigación sistemática”. También se ha definido como: “La utilización juiciosa de la mejor evidencia proveniente de la investigación clínica para la toma de decisiones en el cuidado de cada paciente en particular”. Nos debe quedar claro que las evidencias clínicas pueden conformar pero nunca sustituir a la pericia clínica, y es esta maestría la que decide si las evidencias externas se pueden aplicar por completo a un paciente en particular y, si así ocurre, cómo debe integrarse en una decisión. Una mala conceptualización de la MBE puede llevar a un mal uso de ésta y finalmente a problemas en la asistencia sanitaria, por tanto, debemos tener claro que la MBE no es y no debe ser utilizada para[167]:

- Recortar los gastos sanitarios en asistencia sanitaria: ya que el control presupuestario en los hospitales es organizado por los gestores, y la MBE debe ser manejada por los médicos clínicos, no es lógico ni correcto que sea utilizada para recortar o impedir la realización de determinadas terapéuticas en determinados pacientes o circunstancias. Es el clínico que practica la MBE el que aplicando e identificando las intervenciones más eficaces en cada caso concreto aumentará al máximo la calidad y la cantidad de vida de cada paciente.

- Tampoco se puede considerar que la MBE es solamente todo lo relacionado con los ensayos clínicos aleatorizados (ECA), es decir, no podemos basar nuestra práctica clínica únicamente en aquello que está

demostrado que es eficaz por medio de ECA. Existen muchos tipos de evidencia, y además, no toda la investigación clínica, desde el punto de vista ético, permite la realización de estos ECA. Como veremos en el apartado sobre resultados en atención sanitaria la investigación en resultados estandarizados, no utiliza la metodología de los ECA, pero es un arma más potente, incluso en las circunstancias de la práctica clínica diaria, que los ensayos clínicos, ya que en estos, en la mayoría de las ocasiones, las poblaciones a estudio no siempre son representativas de los pacientes tratados en la realidad diaria[168], mientras que las conclusiones que se obtienen con los estudios de resultados asistenciales, son de aplicación directa sobre nuestra práctica clínica y nos permiten ver el estado de la misma, así como aportarnos las oportunidades de mejora[34,41].

2.3.4. MBE, Niveles de evidencia y benchmarking.

La gran aportación de la MBE ha sido racionalizar y profesionalizar la toma de decisiones terapéuticas sobre los tratamientos a administrar en las diferentes enfermedades y en los distintos grupos de pacientes, al tener en cuenta en el proceso de toma de decisiones solamente aquellas evidencias que hayan sido obtenidas a través de diseños metodológicos válidos, precisos y creíbles. Además, ha logrado ordenar y graduar las evidencias que deben contemplarse en el proceso de toma de decisiones, de tal manera que siempre se tengan en cuenta las mejores evidencias existentes y las que sean más válidas y fiables[169] . La apreciación crítica de un problema clínico se puede dividir en dos componentes, el primero es la evaluación de los artículos

publicados. El fin de esto es clasificarlo dentro de los ensayos controlados y randomizados y conocer su valor en el contexto del problema clínico a tratar. El segundo, es la revisión sistemática, la cual va a evaluar solamente los ensayos controlados y randomizados y va a organizar los datos en un metanálisis o en otras formas de valoración, definiendo de esta forma los grados de evidencia. Estos conceptos han llevado a establecer categorías de los datos obtenidos y evaluarlos dentro de niveles de calidad de la evidencia que, a su vez, llevan a grados de recomendación[203]. Fletcher y Sackett[132, 204, 205], introdujeron estas ideas en 1979 relacionadas con las recomendaciones con respecto a los exámenes periódicos de salud; posteriormente, estos conceptos han sufrido diversas modificaciones (figura 2.4 a y b).

Nivel	*Fuente de evidencia*
I	Metanálisis de estudios controlados bien diseñados. Estudios randomizados con escasos falsos positivos o falsos negativos.
II	Al menos un estudio experimental bien diseñado. Estudios randomizados con alto número de falsos positivos o negativos o los dos.
III	Estudios cuasi-experimentales bien diseñados, tal como series no randomizadas, controladas, grupos pequeños, comparación preoperatoria-posoperatoria, cohortes, tiempo o casos-control emparejados.
IV	Estudios no experimentales bien diseñados.
V	Informe de casos y ejemplos clínicos.

Figura 2.4a. Niveles de evidencia.

Grado	Grado de recomendación
A	Evidencia de tipo I o hallazgos consistentes de estudios múltiples de tipo II, III o IV.
B	Evidencia de tipo II, III o IV y hallazgos generalmente consistentes.
C	Evidencia de tipo II, III o IV peor con hallazgos inconsistentes.
D	Evidencia empírica pequeña o no sistemática.

Figura 2.4b. Grados de recomendación.

Por otro lado, en la actualidad contamos con otra herramienta en relación con la MBE, el benchmarking (BMK), que es una técnica que sirve para identificar, comparar y aprender de los mejores productos, servicios y prácticas que existan, para configurar un programa para el cambio y promover una cultura de mejora continua dentro de una organización. Existen diversas definiciones de BMK en la literatura; una sencilla es la establecida por McKenon[206] que lo define como: "el proceso de medir las prácticas internas comparándolas con parámetros externos con el fin de mejorar los procesos existentes", otra definición, quizás más completa es la elaborada por la American Productivity and Quality Center (APQC): "BMK es un proceso de evaluación continuo y sistemático; un proceso mediante el cual se analizan y comparan permanentemente los procesos empresariales de una organización respecto de las organizaciones líderes de cualquier lugar del mundo, con el fin de obtener la información necesaria para ayudar a mejorar la actuación"[207].

Cuando se aplica el BMK a la atención sanitaria, hay que considerar cinco principios fundamentales[206-208]:

- Mejorar las prácticas de atención al paciente, ya que el propósito fundamental del BMK en las organizaciones sanitarias es mejorar la calidad de la atención.
- Centrarse en los procesos y servicios de gran impacto económico, de tal manera que el coste del proyecto de BMK no sea superior a los beneficios que se conseguirán.
- Adoptar la actitud de un aprendiz. La capacidad de aprender puede acelerar o enlentecer el proyecto.
- Adaptar la mejor práctica que se ajuste a la organización.
- Orientarse en producir comunidades más sanas.

Para que poder llegar a desarrollar estos principios son necesarias dos herramientas previas fundamentales:

- Herramienta logística: hay que saber elegir el momento más adecuado para ponerlo en marcha, contar con el apoyo de la dirección y compartir la información mediante el desarrollo de las infraestructuras necesarias.

- Herramienta metodológica: De nada sirve disponer de todas las estructuras adecuadas, si no somos capaces de evaluar y valorar el grado de evidencia, relevancia y seriedad científica de que disponemos en cuanto a la atención sanitaria, por tanto, para desarrollar este punto es fundamental poder medir los resultados de la atención sanitaria de forma fiable, precisa y en la realidad clínica.

Para llevar a cabo esto es necesario el desarrollo y conocimiento de la MBE y de la medicina basada en resultados en salud.

2.3.5. MBE versus Medicina Basada en Resultados en Salud (MBRS).

Como vimos en el apartado anterior la MBE preconiza la necesidad de una práctica clínica basada en pruebas científicas que demuestren de forma fehaciente que los tratamientos administrados para tratar las enfermedades son eficaces y seguros, con un elevado cociente beneficio/riesgo. La MBE considera que las evidencias científicas deben provenir de diseños metodológicos de alta validez y precisión, como son los ECA con orientación explicativa, donde es posible controlar los diferentes sesgos existentes (selección de pacientes, seguimiento, análisis estadístico, etc) y donde los diversos factores de confusión y factores pronósticos se van a distribuir de manera más o menos homogénea en los grupos en evaluación (siempre que el tamaño muestral sea adecuado)[169,170].

La MBE ha ayudado tremendamente al profesional sanitario a elegir la mejor opción terapéutica existente para tratar una enfermedad específica en un paciente concreto y, por lo tanto, ha incrementado la calidad asistencial de los sistemas sanitarios[171]. Sin embargo, también presenta algunas limitaciones y aspectos negativos, que al final limitan de manera su uso en la toma de decisiones terapéuticas acertadas en la práctica clínica diaria[172,173].

1- Sólo considera evidencias relevantes y libres de sesgos (y, por lo tanto, a tener en cuenta) los resultados provenientes de ensayos clínicos controlados (datos de eficacia y seguridad, básicamente), y

no tiene en cuenta las evidencias que vengan de diseños observacionales (o las considera evidencias de segundo orden, poco creíbles y de escasa relevancia), las cuales nos van a proporcionar datos de efectividad con una elevada validez externa[174,175].

2- No evalúa de manera sistemática otros aspectos y valores de los tratamientos aplicados distintos de la eficacia y seguridad, tales como la calidad de vida relacionada con la salud, el nivel de satisfacción de los pacientes con el tratamiento, el grado de cumplimiento terapéutico y persistencia con los tratamientos y la eficiencia (relación coste/efectividad)[176].

Estas deficiencias de la MBE pueden llevar a que en la práctica clínica diaria puedan tomarse decisiones equivocadas o no totalmente correctas, ya que no se han tenido en cuenta todos los datos e información necesarios para que la decisión sea la más adecuada desde todos los puntos de vista (clínico, paciente, financiadores, proveedores, etc)[177]. En la actual realidad médica de los centros sanitarios, es muy importante tener en cuenta todos estos factores, ya que el entorno sanitario está cambiando de manera rápida y han aparecido agentes y actores con una creciente información y una progresiva mayor capacidad de decisión (asociaciones de pacientes)[178,179]. Además, en estos momentos es cada vez mayor la limitación de los recursos existentes para financiar una atención sanitaria con una demandas incontrolables, lo que obliga a los agentes financiadores a implantar continuas medidas de control del gasto y a tomar decisiones en que la eficiencia y el valor terapéutico añadido tienen

un papel cada vez más predominante a la hora de decidir las opciones a utilizar en la práctica asistencial[180].

En este contexto, sería deseable que la MBE evolucionase e incorporase, además, de los datos de eficacia y seguridad provenientes de ECA, los datos del valor terapéutico añadido de los tratamientos existentes (evaluado a través de la investigación de resultados en salud) a la hora de tomar decisiones terapéuticas. Por lo tanto, la disciplina de la MBE debería evolucionar de una manera lógica y natural hacia una MBRS, uniendo los datos de eficacia y seguridad con otra información complementaria obtenida del comportamiento de los tratamientos una vez que empiezan a emplearse en condiciones de uso habitual[169,181].

¿Qué nos aporta la MBRS a la hora de tomar decisiones en la práctica habitual?[169,181]:

1.- Una mayor información y más elementos de juicio sobre los efectos beneficiosos de los tratamientos disponibles en condiciones de uso habitual.

2.- Permite dimensionar la verdadera utilidad terapéutica y social de las opciones terapéuticas existentes, ya que nos permite obtener información desde las diferentes perspectivas y visiones (clínico, paciente, financiadores, proveedores, etc).

3.- Cuantificar cuánto del efecto es atribuible a la intervención realizada.

4.- Qué resultados son los más relevantes a medir y cómo desarrollar herramientas de medición que sean válidas y fiables.

Para conseguir sus objetivos, la MBRS debe recurrir a los estudios de investigación de resultados en atención médica o investigación de resultados en salud (IRS), este tipo de estudios van a cuantificar, analizar e interpretar los resultados en salud que generan los tratamientos en condiciones de práctica médica habitual (resultados clínicos, económicos, humanísticos y de gestión sanitaria), lo que permitirá conocer su verdadero valor y utilidad a la hora de curar y/o controlar las enfermedades [41,182, 183].

2.4. INVESTIGACIÓN DE RESULTADOS EN ATENCIÓN MÉDICA.

2.4.1. Desarrollo de la investigación en resultados en atención médica.

Cuando hablamos de resultados en atención médica nos referimos a aquellos producidos en la salud, los hábitos o actitudes de los individuos, grupos o comunidades que pueden ser atribuidos a la atención médica recibida. Estos cambios no podrán atribuirse a los cuidados médicos mientras que otras posibles causas o factores no se descarten o controlen. Estos resultados pueden ser de dos tipos:

- Favorables: Cuando se traducen en una mejora del estado de salud, hábitos o actitudes.
- Adversos: Cuando lo que se produce es un deterioro.

Posiblemente la primera referencia que encontramos en la historia respecto a los resultados de la atención médica proviene de la antigua Babilonia, del Código de Hammurabi (1750 A.C):

“Si un cirujano efectuara una incisión profunda en el cuerpo de un hombre con un bisturí de bronce y salvara la vida de este hombre, o le hubiera abierto un absceso en el ojo y le hubiera salvado el ojo, recibirá 10 siclos de plata. Si el cirujano hubiera efectuado una incisión profunda en el cuerpo de un hombre con un bisturí de bronce pero no hubiera salvado la vida de este hombre, o le hubiera abierto un absceso en el ojo, se le cortará la mano”.

Tres resultados de la atención son considerados: curación, muerte y pérdida de un ojo. En esta redacción se asume que el tipo de tratamiento practicado tiene un efecto específico en el resultado y que la habilidad o incapacidad del cirujano también cuenta para el mismo.

Tras estos primeros planteamientos en relación con los resultados de la asistencia sanitaria, no hay prácticamente trazos históricos sobre ello hasta 3500 años después. Se produce una sorprendente discontinuidad histórica hasta que Florence Nightingale se interesó por los resultados obtenidos como parámetro para medir la calidad de los hospitales. Ella clasificó estos resultados como “curación”, “no curación” y “fallecimiento” y sugirió que las tasas de mortalidad hospitalaria podrían ser útiles para valorar la efectividad de la asistencia prestada.

A lo largo del siglo XX, Codman y otros autores sugirieron que la monitorización de los resultados de la asistencia resultaría útil para detectar resultados indeseables y tratar de identificar o buscar causas (prevenibles o evitables) por las que un tratamiento no obtiene los resultados adecuados y así modificarlos, en el futuro, ante otros pacientes similares[41].

A partir de los años 60, se realizaron diversos estudios que demostraron que el ajuste de resultados en relación con distintas características de los pacientes atendidos reduce, a veces muy sustancialmente, las diferencias observadas en las tasas crudas o brutas, pero que aun así, en determinados procedimientos seguían manteniéndose diferencias en mortalidad que

resultaban significativas. Estas variaciones son atribuidas realmente a diferencias en cuanto a la efectividad de la asistencia prestada[34,41,].

Las estrategias posibles para investigar problemas de calidad a partir de resultados adversos son básicamente dos:

1.- La identificación de aquellos casos individuales que reclaman una revisión del proceso asistencial en busca de problemas.

2.- La medición y análisis de las tasas de este tipo de sucesos obtenidos a partir de muchos casos.

La primera estrategia hace referencia a la identificación de “sucesos centinela” que deben reunir dos características fundamentales: Una muy baja probabilidad de ocurrencia y una alta probabilidad de ser debidos al tipo de cuidados prestados, o no dados.

En la segunda estrategia, el estudio de tasas (ajustadas a riesgo previo), se apoya en la medición de determinados sucesos que, sin justificar un estudio del proceso en cada caso individual al poderse producir, aun cuando se reciben cuidados excelentes, se repiten de una forma sistemática.

En el caso de los resultados quirúrgicos, van a estar condicionados básicamente por[13,34,41]:

- Estado fisiológico preoperatorio del paciente.
- Complejidad o severidad de la intervención.
- Calidad y adecuación de la provisión de cuidados.

A pesar de los intentos por establecer esta disciplina a lo largo de todo el siglo XX, la IRS no surgió de forma sistemática hasta finales de los 80, tras numerosas observaciones de la falta de correspondencia entre los resultados observados en las condiciones experimentales de los ECA (resultados en cuanto a eficacia) y los resultados atribuibles a la misma intervención en condiciones habituales de práctica clínica (resultados en cuanto a efectividad), y puede definir como: “El estudio de los desenlaces producidos por las intervenciones sanitarias en condiciones de práctica clínica habitual”[184].

Las diferencias entre la eficacia y efectividad (calculadas en torno al 30%) se han relacionado con: 1) La selección de la muestra en los ECA (paciente ideal versus paciente real); 2) Las condiciones de aplicación de la intervención, que en la vida real no puede garantizarse en las condiciones ideales (dosis y posología indicada); 3) Escasa cantidad de resultados incluidos en los ECA, muy centrados en los desenlaces intermedios (cambios de medidas fisiológicas, etc), que raramente incorporan resultados en los pacientes (tipo de mortalidad, morbilidad, etc)[185].

2.4.2. Neoplasias, quimioterapia e IRS.

Un claro ejemplo de la necesidad de introducir de forma sistemática la IRS en los sistemas sanitarios lo tenemos en cuanto al tratamiento con quimioterápicos, debido a que constantemente se van introduciendo sustancias nuevas en la práctica clínica diaria, con un coste para el sistema sanitario y el paciente cada vez más elevado, sin que se haya demostrado su efectividad en la reducción de la morbi-mortalidad de pacientes reales, ni frente al no tratamiento farmacológico, ni al quimioterápico utilizado previamente a la comercialización de este último.

Según los datos recogidos en el Registro del Cáncer de la Universidad de Múnich (el más amplio registro de Alemania y Europa) (figura 2.5), para carcinomas metastásicos de colon, mama, pulmón y próstata, no se ha producido ningún avance en cuanto a lograr una mayor supervivencia de estos pacientes desde la introducción de los primeros quimioterápicos a finales de los años 70, hasta la actualidad con el uso de los citostáticos modernos. Si revisamos las cifras de supervivencia para tumores sólidos del Instituto Nacional del Cáncer de Estados Unidos, obtenemos las mismas conclusiones expuestas anteriormente[186-190]. Garattini S et al[188], al comparar los datos clínicos de doce quimioterápicos legalizados entre 1995 y 2000 por la Agencia Europea de Medicamentos (Emea) con los citostáticos utilizados de forma estándar hasta entonces, no observaron ninguna mejoría en cuanto a supervivencia, calidad de vida ni seguridad clínica, a pesar de tener un coste mucho mayor.

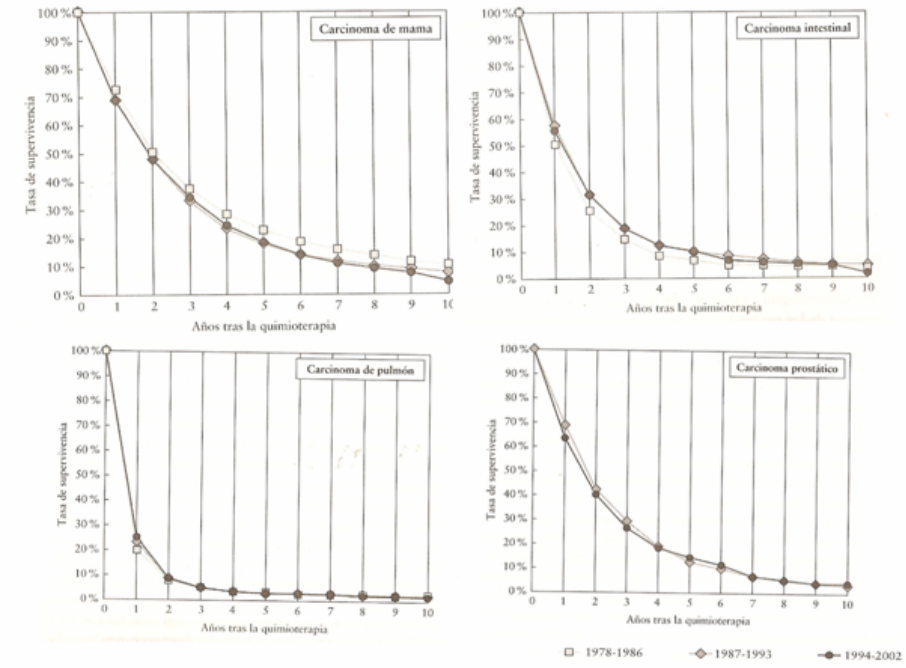


Figura 2.5. Tasas de supervivencia en cáncer. (Adaptado del registro de la Universidad de Múnich)[168]

Los problemas detectados en cuanto a la escasa efectividad de los tratamientos quimioterápicos se pueden resumir en tres apartados:

1) Los requisitos previos a la comercialización de un nuevo fármaco citostático por parte de la Administración de Medicamentos y Alimentos de Estados Unidos (FDA) no incluyen una demostración del aumento en las tasas de supervivencia en los pacientes estudiados en el ECA. El 75 % de estos fármacos aprobados (53 de 71) desde 1990 a 2002, no han demostrado mayor supervivencia, sino que para su aprobación y comercialización basta con que demuestren alguna diferencia estadísticamente significativa en algún parámetro del estudio (menor número de efectos secundarios, por ejemplo)[191].

2) Por consideraciones éticas no está permitido comparar en un ECA un nuevo fármaco frente a placebo, si ya está comercializado un fármaco previo con indicación terapéutica para la patología en estudio, por tanto, en los estudios de estos fármacos, no existe un grupo control con placebo, tal y como hemos dicho en el punto anterior, si solamente el 25 % de los quimioterápicos aprobados han demostrado mejorar la supervivencia en ECA (resultados en cuanto eficacia), ¿cómo vamos a comparar diferencias en supervivencia entre fármacos, si no disponemos de los resultados en el incremento de supervivencia del fármaco control en 3 de cada 4 fármacos?[186,192].

3) Debido a los avances en las pruebas diagnósticas y las amplias campañas de screenings para algunas neoplasias, en la actualidad el cáncer puede ser diagnosticado antes que en épocas pasadas, de tal manera que el intervalo desde su detección hasta la muerte es más largo, lo que se traduce en una mejoría en las tasas de supervivencia, sin que haya intervenido tratamiento alguno, este efecto se conoce como fenómeno Will-Rogers (una mejoría en la tasa de supervivencia "per se" no permite inferir ningún avance en el tratamiento)[189]. Esto ha de ser tenido en cuenta a la hora de estudiar las tasas de supervivencia en series prologadas de pacientes, ya que de lo contrario podemos caer en errores graves en cuanto a efectividad de los tratamientos[192].

2.4.3. Particularidades de los ensayos clínicos en cirugía.

Del total de estudios publicados en cirugía, solamente un 8% corresponden a ensayos clínicos, siendo el 3% de estos ECA. Esto es debido a que existen grandes controversias, en cuanto a la metodología a utilizar para la distribución aleatoria de los pacientes, intervenciones y especialmente de los cirujanos, derivadas de las particularidades propias de esta especialidad[193]:

1) Poca experiencia de los cirujanos en la realización de ensayos clínicos: Esto probablemente se deba a que la FDA norteamericana no impone la certificación de las variaciones técnicas de los distintos procedimientos quirúrgicos, como sí ocurre con los tratamientos médicos.

2) Escasa financiación de los ECA quirúrgicos: Debido al elevado coste y de infraestructura que se requiere para llevar a cabo EC son las empresas farmacéuticas las que promueven la mayoría de ellos, y éstas suelen estar más interesadas en los estudios de nuevos fármacols que en el de técnicas quirúrgicas, ya que presentan a medio-largo plazo una relación coste/rendimiento económico mejor.

3) Irreversibilidad del tratamiento quirúrgico: Los ECA con fármacos disponen de una cláusula de escape predeterminada si la respuesta al tratamiento es poco satisfactoria, debido a que en cirugía el tratamiento suele ser irreversible, la cláusula de escape no tiene sentido en este tipo de estudios quirúrgicos. Además, los ECA en cirugía privan al paciente de recibir, una vez finalizado el ensayo, el tratamiento que haya demostrado mayor eficacia, como se suele hacer en los ECA farmacológicos. Todo esto supone una reducción de las posibilidades de inclusión de pacientes en los estudios quirúrgicos en forma de ECA, tanto desde la perspectiva del paciente (falta de aceptación), del cirujano (dilemas éticos) y del promotor del ensayo (dilemas éticos, riesgo financiero y escaso beneficio económico).

4) Técnicas quirúrgicas y curvas de aprendizaje: Dado que en la cirugía la técnica manual es fundamental y la pericia para una técnica dada no es la misma entre cirujanos, ni en un mismo cirujano para técnicas diferentes de la misma intervención, esto supone un sesgo para la realización de ECA en cirugía.

Como hemos visto, la cirugía, debido a sus características propias, se presta poco a la realización de ECA, por tanto, según los conceptos clásicos de la MBE, la mayoría de actos realizados en esta disciplina contarían con escasa evidencia científica. La IRS, presentan unas características que la hacen idónea para el estudio de resultados quirúrgicos y médicos.

2.4.4.- Características de la IRS.

La IRS es una actividad en la que están implicadas varias disciplinas científicas debido a los distintos tipos de resultados que miden. La IRS utiliza métodos de investigación experimentales y observacionales para examinar de forma sistemática no solo las consecuencias de las intervenciones, sino también los determinantes de las diferencias entre eficacia y efectividad. De tal manera que presenta 5 características principales[194]:

1) Son estudios centrados en la práctica clínica habitual: Se estudian gran número de pacientes, no seleccionados y con observaciones prolongadas, que permiten medir la consecución de resultados finales tal como se producen en la realidad.

2) Analiza la efectividad de intervenciones sanitarias.

3) Pone énfasis en los beneficios del paciente: Presta atención a los resultados importantes para el paciente, tales como la calidad de vida, mortalidad, satisfacción, etc, sin tener en cuenta variables intermedias.

4) Utiliza metodología sistematizada: Los métodos que aplica en el diseño de sus estudios derivan de los establecidos en la epidemiología y la investigación biomédica. Haciendo especial hincapié en desarrollar los estudios en condiciones de práctica clínica habitual.

5) Enfocada a individuos o poblaciones: El análisis de desenlaces puede referirse a pacientes individuales (encuesta de satisfacción, por ejemplo) o a datos agregados (como las tasas de mortalidad).

2.4.5. Tipos de Resultados en IRS

La selección precisa del tipo de desenlace que se estudia es fundamental, ya que la IRS trata de relacionar los desenlaces de las intervenciones con los atributos y el proceso de cuidados prestados[195].

El tipo de resultado a medir dependerá en cada estudio de la intervención a evaluar y del ámbito al que se dirija la intervención, si es el paciente individual o a grupos de pacientes. Por su parte, los resultados que se miden condicionan el tipo de estudio y las características de precisión que debe tener la medición, para definir hasta qué punto el efecto es atribuible a la intervención sobre el estado de salud previo.

En la tabla 2.1 se aprecian las características de las variables que pueden medirse en los pacientes. Se puede observar que las de mayor interés para el paciente son resultados finalistas que tratan de medir el estatus de salud de la persona.

TIPOS:	Variables físicas y fisiológicas	Estatus sintomático	Estatus funcional	Percepción de salud	Calidad de vida global
EJEMPLO:	T. Arterial	Disnea	Papel Social	Sensación Cansancio	Satisfacción
CARACTERÍSTICAS:	Prueba Física	Percepción Subjetiva	Encuesta de Capacidad	Sensación	Encuesta Global
VARIABLE IRS:	Intermedia	Intermedia	Final/ intermedia	Final	Final
INTERÈS:	Profesional	Profesional	Paciente/ familia	Paciente	Paciente

Tabla 2.1. Atributos de los diversos tipos de resultados medibles.
(Tomado de Badía X. La investigación en salud. Barcelona 2000)[194].

En la tabla 2.2 se recoge una clasificación de los tipos de resultados que más habitualmente se miden:

- En cuanto a los resultados clínicos, el área de mayor interés radica en disponer de datos sobre los beneficios terapéuticos de las opciones existentes en la práctica médica real, esto es, conocer su grado de efectividad clínica. Otros datos de interés son la evolución de los síntomas de las enfermedades y los factores de riesgo existentes en la población, disponer de sus datos de morbimortalidad a medio-largo plazo, averiguar el porcentaje de pacientes que alcanzan objetivos terapéuticos, diseñar herramientas de cribado para diagnosticar rápidamente las enfermedades y conocer el nivel de cumplimiento terapéutico y el grado de persistencia de los pacientes con el tratamiento prescrito por el médico[196].

- Sobre los resultados económicos, estos estudios están muy orientados a conocer la eficiencia de las opciones existentes (relación entre los resultados clínicos obtenidos y los costes necesarios para su consecución) y las ventajas económicas (ahorro de recursos) derivadas de su utilización sistemática a través de análisis de evaluación económica, estudios de coste de enfermedad y análisis del impacto presupuestario[197].

- En lo referente a los resultados humanísticos, se intenta conocer cómo los tratamientos administrados afectan a la calidad de vida y el nivel de satisfacción de los pacientes, así como medir el estado de salud de éstos y conocer su grado de preferencia por las alternativas terapéuticas disponibles, junto con la valoración de la discapacidad y el estado funcional que produce la enfermedad y su tratamiento en el paciente[198].

- En relación con la gestión sanitaria, estos estudios se centran en evaluar la calidad asistencial de los servicios sanitarios, así como la búsqueda y conocimiento de indicadores sanitarios que puedan reflejar los resultados en salud existentes en la población, la evolución de los indicadores de calidad de la prestación farmacéutica y la realización de estudios de utilización de medicamentos y otros tratamientos.

Resultados	Tipos	Ejemplos
A nivel individual		
Clínicos		
	Signos y síntomas	Listado de síntomas
	Eventos clínicos	Fractura
	Medidas fisiológicas y metabólicas	Nivel de glucosa en sangre
	Muertes	Muerte por causas específicas o mortalidad general
Variables de interés para el paciente		
	Calidad de la vida	Cuestionario SF-36, EuroQol-5D
	Satisfacción con el tratamiento	Satisfacción con el tratamiento de la diabetes
	Adherencia al tratamiento	Cuestionario de acherencia de Morisky
Económicos		
	Costes directos	Ingresos hospitalarios, visitas médicas, fármacos
	Costes indirectos	Absentismo laboral, restricción actividad laboral
	Costes intangibles	Absentismo escolar, dolor, ansiedad
A nivel de agregado o poblacional		
Mortalidad		Tasa de mortalidad anual
Morbilidad		Comorbilidad
Incidencia y prevalencia		Prevalencia en población general
Productividad social y económica		Pérdidas en productividad

Tabla 2.2. Tipos de resultados medidos más frecuentemente.
(Tomado de Badía X. La investigación en salud. Barcelona 2000)[194].

La IRS se va a valer de diversas fuentes y diseños metodológicos a la hora de elaborar sus estudios y análisis, en unos casos serán fuentes primarias (Ensayos clínicos pragmáticos, diseños observacionales, estudios epidemiológicos, evaluaciones económicas, estudios de calidad de vida y satisfacción, etc.), mientras que en otros casos tendrá que recurrir a fuentes secundarias (metaanálisis, revisiones sistemáticas, revisiones de seguridad, estudios de coste de la enfermedad, etc.), lo que va a generar que la información final disponible sea de gran rigor científico, validez y relevancia para el clínico y el agente decisor[199].

Como el análisis de los resultados pretende aislar el efecto producido por la intervención, del efecto del resto de circunstancias, se hará una medición de todos los factores que influyen, para controlar y hacer un ”análisis ajustado” de los resultados. Las escalas o herramientas específicas utilizadas para los diferentes estudios a realizar deben estar validadas (deben realizar lo que dicen que realizan) en el medio en el que se van a emplear[200].

Para el médico asistencial la IRS es una herramienta clave, debido a que le aporta datos sobre la efectividad de múltiples intervenciones diagnósticas, preventivas o terapéuticas que toma a diario. Además los resultados de estas intervenciones los obtiene en términos de variables finales aplicadas a los pacientes: tales como supervivencia, años libre de enfermedad, calidad de vida ganada, etc. La IRS le permiten relacionar las variables fisiológicas que utiliza en la asistencia, con variables humanísticas.

Por un lado la práctica clínica diaria es la realidad donde se comprueban o refutan las hipótesis de la ciencia biomédica, por otro, genera todos los días una cantidad inmensa de datos, que son con frecuencia infrautilizados para contrastar dichas teorías científicas. Cuando el clínico experimentado contrasta las verdades aceptadas de la biomedicina, como el efecto de un fármaco, con la realidad de los hechos, constata unas diferencias que no hacen más que señalar que tales “verdades aceptadas” no son más que hipótesis que quedan parcialmente refutadas por la realidad.

Este tipo de investigación está en alza y su misión es validar o refutar las propuestas vigentes y suscita nuevas preguntas.

La importancia de este tipo de investigación para el Sistema Nacional de Salud es tal que la Ley de Cohesión y Calidad del Sistema Nacional de Salud[201], y la Ley de Garantías y uso Racional de los Medicamentos y Productos Sanitarios[202], ya recalca la necesidad de medir y evaluar los resultados en salud que se producen en nuestro Sistema Nacional de Salud, como una de las estrategias para incrementar el uso racional de los medicamentos y elevar la calidad de la atención sanitaria.

2.5. MEDICIÓN DE RESULTADOS QUIRÚRGICOS.

En los últimos años se ha desarrollado gran cantidad de escalas y clasificaciones de gravedad o severidad, basadas en la respuesta fisiológica ante la enfermedad. Al ajustar a riesgo previo del paciente este tipo de instrumentos permiten[9-14,34,41]:

1. IRS mediante el ajuste de las tasas de mortalidad y morbilidad a la casuística de cada centro o cirujano.

2. Monitorizar de forma periódica las razones observadas/esperadas (ratio O/E) con el fin de proporcionar información acerca de la mejora o deterioro en la práctica clínica.

3. Detectar el empeoramiento en la práctica clínica, mediante el aumento progresivo de las ratio O/E.

4. Evitar hacer juicios, a veces temerarios, sobre resultados de unidades clínicas no basados en ajuste de riesgo.

5. Realizar sesiones de morbi-mortalidad (SMM) valorando pacientes que a pesar de tener una escasa probabilidad de morbilidad o mortalidad, alguna de éstas a ocurrido.

6. Valoración de los éxitos obtenidos en pacientes con alto riesgo de morbi-mortalidad.

A continuación vamos a describir las 6 escalas de medición del riesgo quirúrgico más utilizadas.

2.5.1. POSSUM (Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity)

El sistema POSSUM fue desarrollado por Copeland et al[209] en 1991 con el fin de poder predecir riesgo ajustado de morbilidad y mortalidad en pacientes diferentes (case-mix), es decir, es un sistema que permite demostrar si las diferencias o no en cuanto al resultado para con pacientes diferentes son debidas a los cuidados prestados, eliminando de la comparación factores de confusión como pueden ser la edad, comorbilidades, etc. Además, este sistema permite conocer el estado de la práctica clínica dentro de un hospital, departamento quirúrgico e incluso por cirujano y compararlo con otros hospitales, servicios y profesionales[13,209].

El sistema consta de 2 tipos de variables:

- Variables fisiológicas: son 12, e incluyen signos y síntomas cardiopulmonares, determinaciones de hemograma y bioquímica, y alteraciones electrocardiográficas. Si alguna de las variables no puede ser recogida se le asigna un valor de 1. Se obtienen antes de la intervención quirúrgica y la suma de puntos varía entre 12 y 88.

- Variables quirúrgicas: son 6, divididas en 4 puntuaciones que crecen exponencialmente (1, 2, 4 y 8). La puntuación quirúrgica se obtiene tras la intervención quirúrgica.

Una vez que se obtienen las puntuaciones, se calcula el riesgo predicho de mortalidad y morbilidad, usando las siguientes ecuaciones desarrolladas por Copeland et al[209] (Siendo R_1 el riesgo de mortalidad y R_2, el riesgo de morbilidad):

- $L_n\ R_1 / 1 - R_1 = -7{,}04 + (0{,}13 \times$ puntuación fisiológica) + (0,16 x puntuación de gravedad operatoria).

- $L_n\ R_2 / 1 - R_2 = -5{,}91 + (0{,}16 \times$ puntuación fisiológica) + (0,19 x puntuación de gravedad operatoria).

El sistema POSSUM además del riesgo esperado de morbi-mortalidad, permite calcular las razones de mortalidad y morbilidad observada (O) y esperada (E) (ratio O:E) tanto de forma individual (por cirujano) como de forma global (por servicio, hospital, etc), de tal manera que; una ratio de 1 indica una correlación perfecta entre lo esperado y lo observado; si es < 1 expresa que los resultados obtenidos son mejores que los esperados; y si es > 1, los resultados obtenidos son peores que los esperados.

Esta escala fue desarrollada y validada por Copeland para gran variedad de cirugías, exceptuando la cirugía cardíaca[209], posteriormente ha sido aplicada a gran cantidad de procedimientos quirúrgicos, especialidades y subespecialidades: traumatología[210-212], neurocirugía[213], cirugía pancreática[214-215], cirugía abdominal urgente y programada[216-221], cirugía colorrectal maligna y

diverticular[222-229], cirugía gástrica neoplásica[230], cirugía hepátobiliar[231], cirugía vascular[232-235], Cirugía bariátrica[236], cirugía torácica[237-239], cirugía esofágica[240-242], cirugía en pacientes de elevado riesgo quirúrgico[243] y comparación entre cirujanos[13,228,235,244].

2.5.1.A. POSSUM y cirugía traumatológica.

Existen varios estudios que demuestran la posible aplicación del sistema POSSUM para calcular el riesgo de morbi-mortalidad tras la cirugía de la fractura de cadera[210-212]. Estos estudios concluyen que este sistema de cálculo de riesgo quirúrgico, aunque sobreestima la mortalidad, sobre todo en pacientes de bajo riesgo, puede ser adecuado para la realización de auditorías de resultados y estudios de comparativos entre distintos centros de trabajo.

2.5.1.B. POSSUM y neurocirugía.

En 2008 Ramesh et al[213] han publicado sus resultados en cuanto a la aplicación del sistema POSSUM a un total de 285 pacientes neuroquirúrgicos sometidos a craneotomía electiva, demostrando que dicho sistema no es válido para calcular mortalidad postoperatoria, debido a que la sobreestima de forma muy elevada, no pudiendo extraerse conclusiones válidas tras su aplicación. Aunque la muestra de pacientes estudiados por los autores de este trabajo es amplia, serán necesarios más estudios para obtener conclusiones de forma definitiva.

2.5.1.C. POSSUM y cirugía pancreática.

Por un lado, Pratt et al[214] en su artículo publicado recientemente (enero 2008) estudian de forma prospectiva la aplicación del POSSUM en 326 pacientes sometidos a resecciones pancreáticas mayores (227 duodenopancreatectomías cefálicas, 7 pancreatectomías centrales, 87 pancreatectomías distales y 5 pancreatectomías totales), concluyendo que el POSSUM, además de un sistema adecuado de predicción de morbi-mortalidad postoperatoria en pacientes sometidos a cirugía pancreatica, permite predecir de forma eficaz los pacientes que requerirán estancias en la Unidad de Cuidados Intensivos, rehabilitación al alta y mayores costes de hospitalización.

Por otro lado, Khan et al[215] en 2003 comunicaron una serie de 50 enfermos sometidos a duodenopancreatectomía parcial a los que de forma retrospectiva se les aplicó la escala POSSUM, observándose que sobreestima la morbi-mortalidad de forma muy desproporcionada, por tanto, concluyen que no es un buen sistema para ser utilizado en cirugía pancreática.

2.5.1.D. POSSUM y cirugía abdominal urgente y programada.

La mayoría de trabajos publicados sobre la aplicación del POSSUM a la cirugía abdominal, tanto urgente como programada, demuestran su utilidad de forma general[216-218,220,221]. Algunos autores han observado una leve sobrepredicción de mortalidad tanto en pacientes mayores de 80 años[217] como en los grupos de bajo riesgo, por tanto, recomiendan el uso de la escala P-POSSUM (Portmouth-POSSUM), al menos en estos grupos de pacientes[219,220].

2.5.1.E. POSSUM y cirugía colorrectal.

En cuanto al cáncer colorrectal la literatura demuestra una leve sobrepredicción de la mortalidad con el sistema POSSUM[224], siendo el cálculo de morbilidad predicha semejante al real. Ramkumar et al[229] demostraron sobre una serie de 347 pacientes con cáncer colorrectal sometidos a cirugía mayor, que no existían diferencias en cuanto a las predicciones de morbi-mortalidad entre la escala POSSUM, la P-POSSUM y el sistema Colorectal-POSSUM, desarrollado específicamente para estimaciones en este tipo de cirugías. Resultados similares fueron publicados por Senagore el al[224].

Sagar et al[228] demostraron que el sistema POSSUM contribuía a hacer más reales y prácticas las auditorías de resultados en cirugía del cáncer colorrectal entre hospitales.

En cuanto a la enfermedad diverticular, existen resultados contradictorios. Por un lado, Constantinides et al[222] observaron una falta de calibración de la escala POSSUM y P-POSSUM pacientes con diverticulitis aguda complicada, mientras que Oomen et al[223] la consideran útil para predecir morbi-mortalidad postoperatoria en este tipo de pacientes.

Wang et al[226] comunicaron en 1998 una serie de 4 pacientes con perforación de colon secundaria a enemas de bario con buena correlación entre la mortalidad esperada por la escala POSSUM y la observada en realidad.

2.5.1.F. POSSUM y cirugía gástrica neoplásica.

La aplicación del sistema POSSUM a la cirugía del cáncer gástrico ha sido escasa hasta la actualidad. Bollschweiler et al[230] la aplicaron a 137 pacientes sometidos a gastrectomía por laparotomía, observando una buena correlación entre los resultados del POSSUM y lo observado en la realidad, por lo que recomiendan su utilización en este tipo de pacientes, con el fin de poder comparar y evaluar de forma objetiva los resultados (auditorías externas e internas).

2.5.1.G. POSSUM y cirugía hepatobiliar.

En un estudio retrospectivo sobre 259 pacientes intervenidos de resección hepática mayor Lam et al[231] observaron una sobreestimación de la mortalidad por parte del POSSUM frente a un cálculo correcto de la misma por parte del sistema P-POSSUM.

2.5.1.H. POSSUM y cirugía vascular.

Los artículos publicados hasta la actualidad sobre el uso del sistema POSSUM en cirugía vascular, demuestran que es útil para el cálculo de morbi-mortalidad postoperatoria, tanto en la cirugía vascular de urgencias (aneurismas rotos), como en la cirugía programada[13,232-235].

2.5.1.I. POSSUM y cirugía bariátrica.

Sólo existe un artículo reportado sobre este tema y en él únicamente se estudian 20 pacientes obesos mórbidos intervenidos mediante gastroplastia vertical anillada. Los autores[236] concluyen que el sistema

POSSUM permite pronosticar las complicaciones y la morbilidad de forma eficaz, además, recomiendan su inclusión en los protocolos de tratamiento de la obesidad mórbida, ya que permitiría las auditorías de resultados entre distintos grupos de trabajo.

2.5.1.J. POSSUM y cirugía torácica.

El POSSUM permite calcular morbi-mortalidad de forma general en los pacientes sometidos a resecciones mayores pulmonares[238,239], pero no permite predecir complicaciones específicas de este tipo de cirugía[237].

2.5.1.K. POSSUM y cirugía esofágica.

Los estudios comunicados son escasos y su metodología es retrospectiva, pero demuestran la falta de precisión del sistema POSSUM para medir complicaciones y mortalidad en cáncer de esófago[240-242].

2.5.1.L. POSSUM y cirugía en pacientes de alto riesgo quirúrgico.

Tanto el sistema POSSUM como otros métodos (P-POSSUM, Surgical Risk Store) son útiles y no muestran diferencias de predicción en cuanto a los pacientes de alto riesgo quirúrgico[243].

2.5.1.M. POSSUM y cirugía comparación entre cirujanos.

Se ha demostrado que esta escala de riesgo es adecuada para comparar la actividad y los resultados de cirujanos tanto dentro de una misma unidad, como de distintas especialidades o de hospitales distintos e incluso con case-mix(casuística) muy distinta[13,227, 228,235,244]. Diversos autores[13,216,228]

recomiendan su utilización de forma sistemática debido a su fácil manejo, escaso tiempo de aplicación y sobre todo porque permite objetivar los resultados quirúrgicos de forma real.

2.5.2. P-POSSUM (Portmouth-Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity)

Posteriormente al desarrollo del sistema POSSUM otro grupo de investigadores[220], tras utilizarlo en sus propios pacientes llegaron a la conclusión de que con las ecuaciones descritas por Coopeland se sobreestimaba la mortalidad en el grupo de riesgo más bajo (<10%), lo que suponía el grupo en el que se incluyen la mayoría de los pacientes quirúrgicos por tanto es muy importante ser capaces de predecir mortalidad de la forma más correcta en estos pacientes, con el fin de poder realizar auditorías de forma fiable. Con el fin de evitar estos problemas derivados del uso del POSSUM, definieron una nueva fórmula a aplicar para el cálculo de la mortalidad:

- $L_n\ R_1 / 1 - R_1 = -9{,}07 + (0{,}17 \times$ puntuación fisiológica) + (0,16 x puntuación de gravedad operatoria).

A continuación vamos a hacer una revisión concisa pero exhaustiva de las aplicaciones del sistema P-POSSUM publicadas hasta el momento actual.

2.5.2.A. P-POSSUM y cirugía traumatológica.

Varios estudios demuestran no sólo la utilidad del sistema P-POSSUM para calcular el riesgo de morbi-mortalidad tras la cirugía de la fractura de cadera[210,245] , sino que en pacientes de bajo riesgo (<10%) ha demostrado ser mejor que la escala POSSUM, por tanto, es un sistema útil de forma global para monitorizar la actividad traumatológica de un servicio o departamento y de esta manera facilitar las auditorías de resultados y estudios de comparativos entre distintos centros de trabajo.

2.5.2.B. P-POSSUM y neurocirugía.

En 2008 Ramesh et al[213] han comunicado sus resultados en cuanto a la aplicación del sistema P-POSSUM a un total de 285 pacientes neuroquirúrgicos sometidos a craneotomía electiva (hubo 9 muertes, lo que supone el 3,16% de los pacientes), demostrando que el sistema P-POSSUM es mucho más preciso que el POSSUM para predecir mortalidad en pacientes programados de neurocirugía (craneotomías); siendo la mortalidad predicha por P-POSSUM del 3,16% frente al 31% por POSSUM. Por tanto, recomiendan el uso de P-POSSUM en los servicios de neurocirugía.

2.5.2.C. P-POSSUM y cirugía pancreática.

Khan et al[215] en 2003 publicaron una serie de 50 enfermos sometidos a duodenopancreatectomía parcial a los que de forma retrospectiva se les aplicó la escala POSSUM y P-POSSUM, observándose que POSSUM sobreestima la morbi-mortalidad de forma muy desproporcionada, por tanto, concluyen que no es un buen sistema para ser utilizado en cirugía pancreática, mientras que P-POSSUM sí era un sistema eficaz para predecir mortalidad postquirúrgica en este tipo de enfermos. Debido a que no existen más estudios comparativos en esta superespecialidad quirúrgica, que el número de enfermos estudiado es pequeño y a la metodología retrospectiva del trabajo, serán necesarios más estudios para obtener conclusiones sólidas a este respecto.

2.5.2.D. P-POSSUM y cirugía abdominal urgente y programada.

Por un lado, Mohil et al[218] (2004) y Hobson et al[221] (2007) han aplicado de forma prospectiva el sistema P-POSSUM a casi 300 pacientes sometidos a cirugía urgente, demostrando su utilidad en la predicción de morbi-mortalidad en este grupo de pacientes, si bien, parece que tiene tendencia a infraestimar ligeramente la mortalidad en pacientes con muy alto riesgo quirúrgico (>90%). Por otro lado, otros autores[219,220] han demostrado su utilidad en la cirugía programada, recomendando su uso, ya que evita la sobrepredicción de mortalidad que ocurre con el POSSUM en pacientes mayores de 80 años[217] y en los grupos de bajo riesgo.

2.5.2.E. P-POSSUM y cirugía colorrectal.

En cuanto a la cirugía colorrectal, por un lado, Tekkis et al[246] en un estudio prospectivo y multicéntrico desarrollado entre 1993 y 2001 en el Reino Unido demostraron que el sistema P-POSSUM no presentaba una buena calibración en cuanto a predicción de mortalidad en cirugía colorrectal tanto maligna o benigna como urgente o programada. Ramkumar et al[229] (2006) obtuvieron hallazgos similares en una serie de 347 pacientes con cáncer colorrectal sometidos a cirugía mayor. Resultados semejantes fueron publicados por otros autores[222,224,247]. Por otro lado, Vather R et al[248] (2006) en un estudio prospectivo con 308 pacientes de Nueva Zelanda han concluido que el sistema P-POSSUM es útil para la predicción de mortalidad en pacientes sometidos a cirugía colorrectal mayor, otro estudio del grupo de Poon et al[249] demostró sobre una muestra de 160 pacientes sometidos a cirugía urgente por obstrucción de cáncer colorrectal la utilidad del sistema P-POSSUM en las predicciones de mortalidad postquirúrgica.

2.5.2.F. P-POSSUM y cirugía hepatobiliar.

En un estudio retrospectivo sobre 259 pacientes intervenidos de resección hepática mayor Lam et al[231] observaron una sobreestimación de la mortalidad por parte del POSSUM frente a un cálculo correcto de la misma por parte del sistema P-POSSUM.

2.5.2.G. P-POSSUM y cirugía vascular.

Los artículos publicados hasta la actualidad sobre el uso del sistema P-POSSUM en cirugía vascular, demuestran que es útil para el cálculo de morbi-mortalidad postoperatoria, tanto en la cirugía vascular de urgencias (aneurismas rotos), como en la cirugía programada[232-235,250,251].

2.5.2.H. P-POSSUM y cirugía esofágica.

Tekkis et al[252] reportaron en 2004 un estudio retrospectivo con 1042 pacientes sometidos a cirugía esofágica y/o gástrica tanto urgente como programada donde demostraban una sobrepredicción de la mortalidad por parte del sistema P-POSSUM. Posteriormente, Lai et al[240] (2007) sobre 313 pacientes sometidos a cirugía esofágica programada demuestran la utilidad del P-POSSUM en este tipo de cirugía. Nagabhushan et al[241] (2007), confirman los resultados publicados por el grupo de Lai. Dada la metodología retrospectiva de estos trabajos, será necesario seguir investigando sobre su utilidad en este tipo de cirugía.

2.5.2.I. P-POSSUM y cirugía en pacientes de alto riesgo quirúrgico.

El sistema P-POSSUM ha demostrado su utilidad en predicciones de mortalidad en los pacientes de alto riesgo quirúrgico[243]. Aunque existen trabajos que refieren que esta escala podría infraestimar la mortalidad en pacientes de muy alto riesgo quirúrgico[218,221].

2.5.2.J. P-POSSUM y cirugía hepática.

Lam et al[231] (2004) en un estudio retrospectivo con 250 pacientes demostraron la utilidad de P-POSSUM en predicciones de mortalidad en cirugía hepática mayor, frente a POSSUM que la sobreestimaba. Un año después de que este trabajo apareciera publicado, Markus et al[253] publicaron en la misma revista (Br J Surg) una serie prospectiva de 190 hepatectomías mayores en la que demostraban una sobreestimación de la mortalidad por parte del sistema P-POSSUM. En la actualidad, en base a los datos publicados en la literatura no podemos considerar ni su uso sistemático en este tipo de cirugía ni su rechazo hasta que no dispongamos de una mejor evidencia al respecto.

2.5.2.K. P-POSSUM y cirugía ginecológica oncológica.

Das et al[254] evaluaron la precisión del sistema P-POSSUM como modelo predictivo de mortalidad operatoria en cirugía de tumores ginecológicos, para ello lo aplicaron de forma prospectiva durante un año a 482 pacientes con cáncer ginecológico: ovario (63,5%); cuerpo del útero (19%); cérvix (9%); otros (5,7%) y demostraron que sobreestimaba la mortalidad de forma global, sólo ajustándose a la mortalidad real en el estrato de muy bajo riesgo (<4%). Los autores concluyen que el sistema POSSUM podría ser una

herramienta útil en la predicción de mortalidad en este tipo de operaciones si se modificaran algunas de sus variables, incluyendo algunas otras muy importantes dentro de la cirugía neoplásica ginecológica, como son el índice de masa corporal y la determinación de albúmina sérica.

2.5.2.L. P-POSSUM y cirugía comparación entre cirujanos.

Al igual que POSSUM, P-POSSUM ha demostrado ser adecuada para comparar la actividad y los resultados de cirujanos tanto dentro de una misma unidad, como de distintas especialidades o de hospitales distintos e incluso con case-mix (casuística) muy distinta[13,227, 228,235,244]. Diversos autores[13,216,228] recomiendan su utilización de forma sistemática debido a su fácil manejo, escaso tiempo de aplicación y sobre todo porque permite objetivar los resultados quirúrgicos de forma real.

2.5.2.M. P-POSSUM y auditorías por países.

En 2002, Yii et al[255] publicaron en Br J Surg un estudio de auditoría quirúrgica en Malasia que, tal y como los autores lo definen, se trata de un país en vías de desarrollo, la importancia que dan al trabajo es la de auditar y valorar la eficacia de este sistema para monitorizar y, dado el caso, detectar los puntos a mejorar de los cuidados quirúrgicos en un país con limitados recursos sanitarios. Con un total de 605 pacientes estudiados prospectivamente mediante el sistema POSSUM para el cálculo de la morbilidad y P-POSSUM para el de la mortalidad, los autores concluyen que son sistemas eficaces para las auditorías quirúrgicas y proponen que se extienda su uso con el fin de facilitar las comparaciones entre regiones y países y contribuir todos a mejorar de forma conjunta.

Posteriormente, en 2003, Bennett-Guerrero et al[256] publicaron un estudio comparativo de resultados ajustados a riesgo mediante el sistema P-POSSUM entre pacientes quirúrgicos en USA y Reino Unido (RU). Compararon de forma prospectiva 1056 pacientes sometidos a cirugía mayor no cardíaca en USA con 1539 pacientes similares en RU, el estudio concluyó que el sistema P-POSSUM sobrepredijo de forma muy elevada la mortalidad en USA y se ajustó bastante a la mortalidad real en RU, pero lo más interesante de este estudio no es esto, sino que la mortalidad observada en RU fue más de 4 veces superior a la de USA. Entre las críticas que se le hacen al artículo están: 1) Sólo se comparan dos hospitales en RU con 1 de USA por tanto no se pueden extrapolar los resultados al resto de la sanidad de estos países; 2) Dado que USA dedica mayor PIB a sanidad que RU esto ha de ser tenido en cuenta a la hora de realizar estudios de ajuste de riesgos; 3) En USA hay mayor número de camas de UCI y más enfermería por cama que en RU, por tanto, esto también podría ser un factor de confusión a la hora de hacer comparaciones entre países; 4) En USA las intervenciones quirúrgicas realizadas por los médicos residentes y los cirujanos con poca experiencia suelen estar supervisadas por cirujanos expertos en mayor proporción que en RU. Al margen de todas estas críticas, estudios de este tipo contribuyen a detectar puntos de mejora asistencial y motivan a mejorar nuestras prácticas clínicas diarias “adoptando” los protocolos empleados por los mejores (benchmarking).

2.5.3. APACHE II (Acute Physiology and Chronic Health Evaluation)

El APACHE II se obtuvo a partir del APACHE I (que consta de 34 variables) en 1985[257], reduciéndose el número de variables fisiológicas a 12, más la edad y el estado de salud previo. Se divide en dos componentes; el primero, llamado APS o Acute Physiology Score que califica las variables fisiológicas. Para la determinación de los parámetros fisiológicos se determinan: temperatura, tensión arterial media, frecuencia cardiaca, frecuencia respiratoria, PaO2, pH arterial, sodio, potasio y creatinina sérica, hematócrito, leucocitos, y la puntuación de la escala de coma de Glasgow (se puede usar el HCO3 en caso de no contar con el PaO2 arterial). A cada variable se le asigna un valor que va del 0 al 4. La suma de las puntuaciones de estas variables proporcionan el primer componente del APACHE II (el APS), que se considera una medida de la gravedad de la enfermedad aguda del paciente. El segundo componente, denominado Chronic Health Evaluation, califica la edad y el estado de salud previo. Si el paciente está inmunocomprometido, tiene insuficiencia hepática, cardíaca, renal o respiratoria y es sometido a un procedimiento quirúrgico programado deberán sumarse 2 puntos al total, pero si es sometido a un procedimiento de urgencias, deberán sumarse 5 puntos. La suma de ambas escalas constituye la puntuación Acute Physiology And Chronic Health Evaluation II o APACHE II. La puntuación máxima posible del sistema APACHE II es 71, aunque apenas existe supervivencia sobrepasando los 55 puntos.

El APACHE II ha sido validado en gran cantidad de países tanto en Unidades de cuidados intensivos generales[257-260], como en quirúrgicas[261-262], obteniéndose los mejores resultados tras su aplicación en los pacientes quirúrgicos urgentes frente a los programados y no quirúrgicos[258,263]. Se han publicado estudios que demuestran tanto infraestimación de la mortalidad global[263] o en pacientes de alto riesgo[264,265], como sobreestimación de la misma en pacientes de alto y bajo riesgo[264,266].

2.5.3.A.- APACHE II y cáncer oral y orofaríngea.

De Cássia et al[267] comunicaron en 2003 un estudio prospectivo con 530 pacientes diagnosticados de cáncer cervical y oral y a los que les aplicaron las escalas APACHE II, POSSUM y ASA con el fin de valorar su utilidad en la predicción de morbilidad en este tipo de pacientes, y obtuvieron una muy buena correlación entre lo esperado y lo observado, en cuanto a morbilidad para la escala APACHE II y POSSUM, sin que existieran diferencias significativas entre ellas, mientras que el sistema ASA demostró muy malos resultados, por tanto, los autores concluyen que el uso de alguna de estas 2 escalas puede permitir al cirujano detectar a los pacientes potencialmente complicados y de este modo garantizarles un seguimiento más estrecho.

2.5.3.B.- APACHE II y cirugía abdominal urgente.

Existen varios trabajos que demuestran la utilidad de la escala APACHE II en la predicción de mortalidad en la cirugía abdominal de urgencias[268-270]. Aunque es un sistema útil para predecir mortalidad, siempre que se ha comparado con el sistema POSSUM ha resultado menos potente

que éste[13,209,270]. Moshe Schein et al[269] publicaron un estudio prospectivo en 154 pacientes con úlcera péptica y sangrado masivo que requirieron cirugía urgente, demostrando que los índices APACHE II elevados (>11) se correlacionaban con una mayor mortalidad operatoria si se realizaban grandes cirugías (gastrectomía), mientras que si se tomaban medidas quirúrgicas menos agresivas (sutura simple de úlcera y piloroplastia) se podía reducir el riesgo de mortalidad de estos pacientes, y al contrario, los pacientes con APACHE II bajo (<10), tolerarán sin riesgo apenas una cirugía definitiva, por tanto, este sistema de cálculo de riesgo puede servir para estratificar a los pacientes críticos en el preoperatorio y realizarles un tipo de cirugía menos agresiva (control de daños).

2.5.3.C.- APACHE II y cirugía esofagogástrica.

Aunque retrospectivos, existen trabajos que revelan la utilidad del cálculo seriado del APACHE II en el postoperatorio de la cirugía esofágica con anastomosis intratorácica, de tal manera que si el APACHE II se eleva durante el postoperatorio es muy indicativo de complicación quirúrgica[271,272].

2.5.3.D.- APACHE II y neurocirugía.

El sistema APACHE II no ha demostrado utilidad en la evaluación pronóstica de muerte temprana tras traumatismo craneoencefálico, siendo los resultados obtenidos con la aplicación seriada de la escala de Glasgow superiores en este tipo de patología. Para lo que sí es útil y tiene uno resultados similares al APACHE III es para la estimación de mortalidad tardía en estos enfermos[273,274].

2.5.3.E.- APACHE II y trasplantes.

Existen trabajos que demuestran que el sistema APACHE II es útil y eficaz para predecir mortalidad en pacientes sometidos a trasplante hepático[275,276], de tal manera que estos estudios concluyen que no son necesarios sistemas específicos de valoración de este tipo de pacientes si se dispone y realiza de forma seriada el APACHE II. Por otro lado, el ensayo clínico de Wei-Huang S et al[277] con 56 pacientes divididos en 2 grupos: Grupo A (trasplante hepático + sepsis) y grupo B (no trasplante + sepsis) concluyó que los factores pronósticos más importantes en los enfermos trasplantados hepáticos son la presencia de inmunosupresión y el lactato elevado de forma persistente. No se demostró que el sistema APACHE II tuviera una eficacia elevada en cuanto a capacidad pronóstica con estos enfermos. El número de enfermos en los estudios llevados a cabo es escaso y algunos de los estudios realizados presentan metodología retrospectiva, por tanto, serán necesarios más estudio para poder establecer conclusiones sólidas.

No existen evidencias de la utilidad del sistema APACHE II en los pacientes trasplantados de riñón.

2.5.3.F.- APACHE II y cirugía vascular.

Existen estudios prospectivos no randomizados que estudian la posible utilidad del sistema APACHE II en la predicción de morbilidad y mortalidad en enfermos con problemas vasculares periféricos, pero en ninguno de ellos se ha demostrado que tenga capacidad de predecir ni las complicaciones de la cirugía programada en el bypass aortobifemoral (oclusión vascular aguda), ni la supervivencia tras la cirugía urgente de la ruptura de aneurismas aórticos infrarrenales[278,279].

2.5.4. SAPS II (Simplified Acute Physiology Score)

Es una versión simplificada del apartado de afectación fisiológica aguda de APACHE, desarrollado en 8 UCIs polivalentes de Francia[280], que permite mediante la valoración de datos clínicos sencillos y habituales en la clínica rutinaria de la UCI establecer un índice de gravedad y una estimación pronóstica.

Al igual que APACHE II, SAPS requiere datos de las primeras 24 horas del ingreso, siendo estas variables las siguientes:

- Edad
- Frecuencia cardiaca
- Presión arterial sistólica
- Temperatura corporal
- Débito urinario
- Hematocrito
- Recuento leucocitario
- Glucosa plasmática
- Urea plasmática
- Potasio plasmático
- Sodio plasmático
- Bicarbonato sérico
- Glasgow Coma Score

Sumando la puntuación de estas variables puede obtenerse una estimación de mortalidad para cada paciente.

Al igual que sucedió con APACHE, SAPS también sufrió modificaciones para mejorar su rendimiento, incrementando los parámetros valorados hasta 15 variables y con puntuaciones mas ajustadas a su peso estadístico, y se validó mediante un gran estudio internacional en el que se incluyeron 13.152 pacientes de 137 UCIs europeas y norteamericanas[281-282]. A las variables valoradas en SAPS se añaden parámetros de disfunción hepática, renal y respiratoria, tipo de paciente (médico, quirúrgico programado o quirúrgico urgente), presencia de SIDA, neoplasias hematológicas o tumoraciones metastásicas. El sistema SAPS II ha sido estudiado en las siguientes subespecialidades y circunstancias:

2.5.4.A. SAPS II y cirugía pancreática.

Padalino P et al[287] (2005) demostraron en un estudio con 21 pacientes con pancreatitis aguda grave necrotizante la utilidad del sistema APACHE II, SAPS II y SOFA tanto en la predicción de indicación de cirugía debido a las complicaciones del proceso agudo, como en el pronóstico de muerte, destacando una ligera superioridad del sistema SOFA con respecto a los otros.

2.5.4.B. SAPS II y cirugía abdominal urgente.

Ertan et al[283] (2008) han comunicado en una serie prospectiva de 102 pacientes con cáncer colorrectal complicado, sometidos a colectomía urgente sus resultados en cuanto a validación y aplicación del sistema SAPS II, obteniendo unos resultados muy buenos en cuanto a validación (curva ROC = 0,83) y aplicación en este tipo de pacientes (17 muertes observadas frente a 15 esperadas (p=0,982), concluyendo que es una herramienta adecuada para pacientes con neoplasia colorrectal complicada.

2.5.4.C. SAPS II y cirugía colorrectal.

En un trabajo prospectivo publicado este año por Can MF et al[284] sobre cirugía colorrectal programada comparan APACHE II, SAPS II, POSSUM y P-POSSUM en cuanto a su capacidad predictiva para mortalidad a los 30 días postoperatorios (mortalidad observada del 3,6%), concluyendo que SAPS II (mortalidad esperada del 3,7%; ROC=0,854) y P-POSSUM (mortalidad esperada del 5,2%; ROC=0,831) son sistemas útiles para este tipo de cirugías programadas, aunque concluyen que serán necesarios más estudios para aceptar su uso de forma rutinaria. APACHE II (mortalidad esperada del 9,1%; ROC=0,786) y POSSUM (mortalidad esperada del 13,4%; ROC=0,793) no demostraron una buena correlación con la realidad observada).

2.5.4.D. SAPS II y cirugía vascular.

En un estudios prospectivo no randomizado donde se estudió la eficacia del sistema APACHE II y SAPS II en la predicción de morbilidad y mortalidad en 107 enfermos con problemas de obstrucción vascular periférica sometidos a by-pass aorto-bifemoral programado, no se demostró la utilidad de ninguno de ellos ni para predecir mortalidad ni morbilidad postoperatoria[278].

2.5.4.E. SAPS II y cirugía en pacientes de alto riesgo quirúrgico.

En un estudio prospectivo sobre 24 pacientes ingresados en UCI por sepsis grave secundaria a colecistitis aguda litiásica[286], se observó un descenso del riesgo de muerte esperado con buena correlación con la mortalidad observada, tras ser colecistectomizados mediante abordaje laparotómico, con los sistemas APACHE II, SOFA y SAPS II.

2.5.4.F. SAPS II y cirugía cardíaca.

Kern H et al[289] demostraron en un estudio prospectivo con 680 pacientes sometidos a cirugía cardíaca la utilidad del sistema SAPS II en la predicción de complicaciones y necesidad de ventilación mecánica con una curva ROC de 0,93.

2.5.3.G.- SAPS II y trasplante hepático.

Bein T et al[275] publicaron un estudio retrospectivo con 123 pacientes sometidos a trasplante hepático, en el que demostraban unas buenas sensibilidades de los sistemas SAPS II (S=72%), APACHE II (S=71%) y MPM (S=84%) en el pronóstico de mortalidad de estos pacientes. El escaso número de pacientes y el diseño retrospectivo no permiten obtener conclusiones muy sólidas respecto a este tipo de pacientes.

2.5.4.H. SAPS II y traumatismo craneoencefálico.

En un trabajo sobre 401 pacientes con traumatismo craneoencefálico ingresado en UCI, Álvarez et al[290] aplicaron de forma prospectiva los sistemas Glasgow, SAPS II, APACHE II y MPM, demostrando la superioridad en predicción de mortalidad del MPM II frente a los otros 3.

2.5.4.I. SAPS II y pacientes oncológicos no operados.

Por un lado, González-Pérez et al[285] comunicaron en 2007 un estudio prospectivo sobre 250 pacientes oncológicos no subsidiarios de tratamiento quirúrgico e ingresados en UCI para tratar de establecer su riesgo de mortalidad, para ello compara los resultados esperados para mortalidad por SAPS II, MPM, APACHE II y III, SOFA y MODS con los observados en

realidad, concluyendo que aunque ninguno de estos sistemas fue eficaz en el cálculo pronóstico de mortalidad, recomiendan el uso de SAPS II debido a que en su cálculo incluye variables oncológicas y su cálculo es muy sencillo. Por otro, Berghmans T et al[288], en un estudio prospectivo sobre 247 pacientes oncológicos ingresados por complicaciones médicas en una UCI específica para pacientes neoplásicos (mortalidad observada del 34%) demostraron la utilidad del sistema SAPS II (mortalidad esperada del 24%), APACHE II (mortalidad esperada del 32%) y ICM (ICU Cancer Mortality Model) (mortalidad esperada del 28%), demostrando unos resultados ligeramente superiores para ICM con los mejores valores en la comparación de las curvas ROC (0,79).

2.5.5. MPM (Mortality Prediction Model)

Se desarrolló en un único hospital, con una muestra de 755 pacientes, asignando valores a los parámetros con capacidad predictiva de mortalidad hospitalaria mediante regresión logística[291]. Su modificación posterior (MPM II) se basó en un estudio internacional con 12.610 pacientes y fue validado en otra muestra posterior de 6.514 pacientes[292].

MPM utiliza variables clínicas simples, obtenidas en el momento del ingreso (MPM_0) y a las 24 horas del mismo (MPM_{24}), además de la edad y el estado de salud previa.

Las variables utilizadas en MPM_0 son:

- **Edad**
- **Alteración fisiológica aguda**:
    - Coma o estupor
    - Frecuencia cardiaca ≥ 150 ppm.
    - Tensión arterial sistólica ≤ 90 mmHg
    - Ventilación mecánica
    - Fracaso Renal Agudo
    - Arritmias cardiacas graves
    - Accidente cerebrovascular
    - Sangrado gastrointestinal
    - Efecto masa craneal
    - Reanimación cardiopulmonar previa al ingreso

- **Estado crónico de salud**:
    - Insuficiencia renal crónica
    - Cirrosis
    - Neoplasia metastásica
- **Tipo de paciente**:
    - Paciente médico o quirúrgico urgente

MPM_{24} utiliza algunos de los parámetros de ingreso y valora los cambios evolutivos en las primeras 24 horas. Las variables utilizadas son:

- **Edad**
- **Parámetros evaluados al ingreso**:
    - Cirrosis
    - Neoplasia metastásica
    - Efecto masa craneal
    - Paciente médico o quirúrgico urgente
- **Parámetros evaluados a las 24 horas de tratamiento**:
    - Coma o estupor profundo a las 24 horas
    - Creatinina > 2 mg/dl
    - Infección confirmada
    - Ventilación mecánica a las 24 horas del ingreso
    - PO2 < 60 mmHg
    - Tiempo de Protrombina

El MPM valora la presencia o no de las variables descritas previamente y les asigna una puntuación en función de su peso estadístico, lo que permite

una estimación de la probabilidad de supervivencia de forma directa. La realización de MPM seriada de forma diaria (MPM $_{24}$, MPM $_{48}$, MPM $_{72}$) permite discernir si la evolución del paciente es adecuada a las medidas terapéuticas aplicadas; así, un paciente que pese al tratamiento intensivo mantenga coeficientes de MPM estables incrementa su probabilidad de muerte de forma significativa[293].

Entre los estudios en los que MPM ha sido empleado destaca su utilidad en predicción de mortalidad en los pacientes con traumatismo craneoencefálico[290], trasplante hepático[275], pancreatitis grave de origen litiásico[294] y en úlcera péptica perforada[295].

2.5.6. MODS (Multiple Organ Dysfunction Score)

Es otro sistema de valoración de la gravedad de pacientes críticos basada en la afectación orgánica derivada de la agresión[296]. Esta escala permite una visión evolutiva de la enfermedad y su repercusión sobre la fisiología en función de la respuesta al tratamiento y la capacidad de recuperación funcional del paciente.

Los órganos y sistemas valorados por MODS son la función respiratoria, función renal, función hepática, sistema cardiovascular, hemostasia y estado neurológico, a los que se les asigna una puntuación entre 0 y 4 en función de la desviación respecto a la normalidad. Con el sumatorio de los puntos obtenidos puede calcularse la probabilidad de muerte del paciente.

En un estudio observacional prospectivo con 949 pacientes Peres Bota D et al[297] demostraron la capacidad de MODS en la predicción de mortalidad en pacientes ingresados en UCIs generales, sin que hubiera diferencias significativas entre éste y los sistemas APACHE II y SOFA (Sepsis-related Organ Failure Assessment). En la calibración por sistemas específicos, fue mejor SOFA que MODS en la disfunción cardiovascular, tanto al ingreso como en la evolución posterior. Por tanto, los pacientes que presentan shock son mejor calibrados por el sistema SOFA.

MODS ha demostrado ser mejor que APACHE II en el cálculo del pronóstico de los pacientes con sepsis tras trasplante hepático[277]. En cuanto a su uso en cirugía cardíaca, se han obtenido valores similares a APACHE II[298] siendo útil su uso en la predicción de mortalidad postoperatoria de estos pacientes.

3. MATERIAL Y MÉTODOS

3.1. ÁMBITO

El estudio se llevó a cabo en el Servicio de Cirugía General y Digestiva (SCGD) del Hospital General Universitario "JM Morales Meseguer" (HGUMM) de la Región de Murcia perteneciente primero al Instituto Nacional de la Salud (INSALUD) y luego al Servicio Murciano de Salud.

El SCGD del HGUMM constituye una unidad clínica que se ocupa de la *actividad asistencial,* (consulta, hospitalización, intervención y atención de urgencias) de la especialidad correspondiente, en el Área de Salud VI de Murcia, a la vez que considera imprescindibles e inseparables de la anterior la realización de *actividades de docencia e investigación clínica.*

El HGUMM se trata de un hospital de Área (nivel II) de 418 camas (de las que 48 camas están destinadas al SCGD) que trata cualquier patología aguda de adultos, con exclusión de las especialidades de tercer nivel (excepto Oncohematología, con la que sí cuenta), Pediatría y Obstetricia y Ginecología. Se estima que el total de ciudadanos que atiende nuestro centro es de 300.000.

El organigrama del SCGD del GUM está compuesto: 1 Jefe de Servicio; 2 Jefes de Sección, 15 médicos Facultativos Especialistas de Área (FEA) y 4 Médicos Internos residentes (MIR). A su vez el servicio está compuesto por las siguientes Unidades Funcionales:

- Unidad de Coloproctología.
- Unidad de Cirugía Endocrina.
- Unidad de Mama.
- Unidad de Pared Abdominal.
- Unidad de Pie Diabético.
- Unidad de Dispositivos de Acceso Venoso y Varices.
- Unidad de Estómago y Cirugía Bariátrica.

El número de ingresos anuales por parte del SCGD está en torno a los 3000 ingresos/año. Así como el número de intervenciones programadas en nuestro servicio está en torno a las 2300 operaciones/año. Y el número de ingresos por urgencias es de unos 1500 ingresos por año por esta vía, interviniéndose de urgencia, unos 900 casos/año.

3.2. UNIDADES DE ESTUDIO

Las unidades de estudio son las complicaciones y la mortalidad postoperatoria en pacientes intervenidos de cirugía abdominal con ingreso de forma programada en el SCGD del HGUMM y seguidos en consultas externas de cirugía. Se considera morbi-mortalidad postoperatoria a la ocurrida dentro de los 30 primeros días tras la cirugía. Quedan excluidos los pacientes compartidos o derivados a otros servicios quirúrgicos, y a los que no se ha podido realizar el seguimiento, en consultas externas, tras el alta hospitalaria.

El periodo de observación del estudio fue desde el 1 de enero de 2007 hasta 31 de diciembre de 2008, momento en el que llegamos al “end-point” de nuestro proyecto.

El periodo de evaluación comprende desde el día de la intervención quirúrgica hasta el 30 día postoperatorio. Si el paciente no se encuentra ingresado en el hospital durante todo este periodo, el seguimiento de morbi-mortalidad se realizó en consultas externas del SCGD al mes de la operación, recogiéndose de forma prospectiva.

La evaluación se ha realizado sobre los pacientes intervenidos incluidos en el estudio, mientras que las herramientas evaluadas son 6 escalas de riesgo quirúrgico con el objetivo de valorar su capacidad predictiva en cuanto a morbi-mortalidad en este tipo de pacientes.

3.3. CRITERIOS DE EVALUACIÓN Y HERRAMIENTAS UTILIZADAS

Para este estudio hemos utilizado 6 de los índices pronósticos más utilizados a nivel internacional para valorar complicaciones en pacientes quirúrgicos ingresados[13,209,218,257,280,292,296].

- *APACHE* II (Acute Physiology and Chronic Health Evaluation).
- *POSSUM* (Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity).
- *P-POSSUM* (Portmouth-Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity).

- *SAPS* II (Simplified Acute Physiology Score).
- *MPM* (Mortality Prediction Model).
- *MODS* (Multiple Organ Dysfunction Score).

Para evaluar la capacidad predictiva de los 6 índices pronósticos estudiados, realizamos una primera fase de validación de las escalas, para ello seguimos la metodología clásica de validación:

1.- Tamaño muestral: 10 sujetos por cada uno de los ítems a evaluar. (POSSUM y P-POSSUM se componen de 18 ítems; APACHE de 15; SAPS de 15; MPM de 14; MODS de 6) por tanto, el número mínimo de pacientes necesarios para esta fase del estudio es de 180 para POSSUM y P-POSSUM; 150 para APACHE y SAPS; 140 para MPM; y 60 para MODS).

2.- Traducción de la escala seleccionada: Para el estudio utilizaremos la traducción inglés-español que ya realizamos, validamos y publicamos (junio de 2006)[299] en nuestro servicio de Cirugía de la escala POSSUM y P-POSSUM. Para el resto de escalas se procederá a la traducción independiente (a partir de los artículos originales en inglés donde se describe cada escala) por parte de dos observadores y su posterior comparación y correlación.

3.- Prueba de valoración de comprensión de los distintos ítems de la escala: 2 evaluadores previamente entrenados (recibieron un curso de medición de resultados en salud) aplicaran, de forma independiente y separada, la escala a los pacientes seleccionados para cada índice pronóstico.

4.- Confiabilidad interobservador: Se analizará la concordancia interobservador para aquellos ítems que podían estar sometidos a diferentes interpretaciones según el evaluador que aplicara la escala. Se evaluará la consistencia interna de cada escala, aplicando el *test alfa de Cronbach*, como medida de la homogeneidad.

Para la comparación de las distintas escalas emplearemos los siguientes métodos:

5.- Bondad de la prueba: Para evaluarla se procederá al cálculo del área bajo la curva (ABC), mediante las *curvas ROC* de cada uno de los índices pronóstico, junto con sus respectivos intervalos de confianza al 95% ($IC_{95\%}$).

6.- Precisión de la prueba: Se evaluará con el *índice de Shannon*; para calcularla en cada paciente mediante la comparación entre el valor predicho por cada escala y el resultado observado, de acuerdo con la siguiente fórmula:

$$S = \{ [(1+o) \ln (1+e) + 2 o] \ln (2-e) - \ln 2 \} / \ln 2$$

(Donde *ln;* es el logaritmo neperiano; *o;* es la presencia (o=1) o la ausencia (o=0) del evento de muerte y *e;* es el valor esperado obtenido con cada escala para cada paciente. A partir de cada índice se calculará el promedio más una desviación estándar (DE) para el total de los pacientes (índice de Shannon global), y el promedio más una DE de los fallecidos solamente (índice de Shannon para fallecidos); de esta forma se determinará la precisión de cada

modelo en una y otra situación. Este índice de precisión puede oscilar entre 0 (precisión nula) y 1 (precisión perfecta).

7.- Métodos gráficos: De forma complementaria se usarán métodos gráficos para comparar el rendimiento de los distintos índices pronóstico. Por un lado se comparará en un gráfico la mortalidad observada frente a la esperada por intervalos de riesgo y para cada uno de los índices pronóstico, además, se representará la ratio observada:esperada para cada una de las escalas por intervalos de riesgo. De igual forma se procederá con la morbilidad en caso de las escala POSSUM y P-POSSUM.

8.- Análisis estadístico: Los datos se recogerán como promedios, desviación estándar (DE) o error estándar (EE). El uso de intervalos de confianza del 95% se indicará oportunamente en cada caso. La comparación de múltiples medias, se realizará con ANOVA de medidas repetidas. Para determinar si existen diferencias estadísticamente significativas entre los resultados obtenidos y los esperados según las diferentes escalas de riesgo se aplicará el test X^2 de Pearson.

3.3.1. Descripción de las escalas pronóstico.

APACHE II (Acute Physiology and Chronic Health Evaluation) (Tabla 3.1)

Correlaciona la gravedad de la enfermedad actual y la evaluación del estado de salud previa del paciente. La escala permite mediante cálculos matemáticos establecer una probabilidad de mortalidad hospitalaria.

Los parámetros medidos en la escala de afectación fisiológica aguda son: temperatura rectal, presión arterial media, frecuencia cardíaca, frecuencia respiratoria, oxigenación, pH arterial, sodio sérico, potasio sérico, creatinina sérica, hematocrito, recuento leucocitario, escala de coma de Glasgow y la edad según los intervalos recogidos en la tabla 3.2. Cada variable puntúa de 0-4 en función de la desviación de la normalidad.

Variable	+ 4	+3	+2	+1	0	+1	+2	+3	+4
Temperatura	>41	39-409	—	38.5-38.	36-38.4	34-35.9	32-33.9	30-31.9	< 29.9
TAM	>160	130-159	110-129	—	70-109	—	50-69	—	< 49
FC	>180	130-159	110-129	—	70-109	—	50-69	40-54	< 39
FR	>50	35-49	—	25-34	12-24	10-11	6-9	—	< 5
A a PO2	> 500	350-499	200-349	—	<200	—	—	—	—
Po2	—	—	—	—	> 70	61-70	—	55-60	< 55
PH Arterial	> 7.7	7.6-7.69	—	7.5-7.59	7.33-7.49	—	7.25-7.32	7.15-7.24	< 7.15
HCO3	> 52	41-51.9	—	32-40.9	23-31.9	—	18-21.9	15-17.9	< 15
Sodio	> 180	160-179	155-159	150-154	130-149	—	120-129	111-119	< 110
Potasio	> 7	6-6.9	—	5.5-5.9	3.5-5.4	3-3.4	2.5-2.9	—	< 2.5
Creatinina	> 3.5	2-3.4	1.5-1.9	—	0.6-1.4	—	< 0.6	—	—
Hto	> 60	—	50-59.9	46-49.9	30-45.9	—	20-29.9	—	< 20
Recuento Leucocitos	> 40	—	20-39.9	15-19.9	3-14.9	—	1-2.9	—	< 1
Glasgow	—	—	—	—	—	—	—	—	—
Puntaje fisiológico agudo	—	—	—	—	—	—	—	—	—

Tabla 3.1. Escala APACHE II

≤ 44 años	0 puntos
5-54 años	2 puntos
55-64 años	3 puntos
65-74 años	5 puntos
> 75 años	6 puntos

Tabla 3.2. Edad y puntuación en APACHE II

El estado de salud previa diferencia tres tipos de pacientes: pacientes no quirúrgicos, pacientes quirúrgicos urgentes y pacientes quirúrgicos programados, valorando en cada uno de ellos la presencia de fracasos orgánicos crónicos graves o estado de inmunodepresión (tabla 3.3).

Tipo de paciente	**Estado de salud previo**	**Puntos**
Paciente no quirúrgico	Fracaso orgánico crónico o inmunodepresión	5
	Sin fracaso e inmunocompetente	0
Cirugía urgente	Fracaso orgánico crónico o inmunodepresión	5
	Sin fracaso e inmunocompetente	0
Cirugía programada	Fracaso orgánico crónico o inmunodepresión	2
	Sin fracaso e inmunocompetente	0

Tabla 3.3. Estado salud previo en APACHE II

La probabilidad de fallecer calculada en el APACHE II viene dada por la fórmula:

$$\text{Probabilidad} = [e^x / (1 + e^x)] \times 100$$

Siendo el cálculo de x para el sistema APACHE II:

X= -3,517 + puntuación APACHE II x 0,146 + coeficiente categoría diagnóstica

SAPS II (Simplified Acute Physiology Score) (Tabla 3.4)

Las variables que valora son: Edad, frecuencia cardiaca, presión arterial sistólica, temperatura corporal, débito urinario, hematocrito, recuento leucocitario, glucosa plasmática, urea plasmática, potasio plasmático, sodio plasmático, bicarbonato sérico, escala de coma de Glasgow, disfunción hepática, renal y respiratoria, tipo de paciente (médico, quirúrgico programado o quirúrgico urgente), presencia de SIDA, neoplasias hematológicas o tumoraciones metastásicas.

1.- **Edad** en años :

0	7	12	15	16	18
<40	40-59	60-69	70-74	75-79	>80

2.- **Frecuencia cardiaca** en latidos minuto:

11	2	0	120	7
<40	40-69	70-119	120-159	>160

3.- **Presión arterial sistólica** en mmHg:

11	5	0	2
<70	70-99	100-199	>200

4.- **Temperatura** en grados centígrados C°:

0	2
<39	>39

5.- **PO2/FiO2**

11	9	6
<100	100-199	>200

6.- **Diuresis** en litros por 24 h

11	4	0
< 0,5	0,5-0,999	>1,000

7.- **Urea** en mgr %:

0	6	10
<28	28-83	>84

8.- **Leucocitos** en 10°/ litro:

12	0	3
<1,0	1,0-19,9	>20,0

9.- **Potasio** en Mmol/litro:

3	0	3
<3,0	3,0-4,9	>5,0

10.- **Sodio** en Mmol/litro:

5	0	1
< 125	125 -144	>145

11.- **Bicarbonato** en mEq / litro:

6	3	0
<15	15-19	>20

12.- **Bilirrubina** mg/dl:

0	4	9
<4,0	4,0-5,9	>6,0

13.- **Glasgow Coma Score** en puntos:

26	13	7	5	0
<6	6-8	9-10	11-13	14-15

14.- **Enfermedad crónica**:

9	10	17
Cáncer metastásico	Neoplasia hematológica	SIDA

15.- **Tipo de admisión**:

0	6	8
Cirugía programada	Causa médica	Cirugía urgente

Tabla 3.4. Escala SAPS II.

La probabilidad de fallecer calculada en el SAPS II viene dada por la misma fórmula que en el APACHE II:

$$\text{Probabilidad} = [e^{logit} / (1 + e^{logit})] \times 100$$

Siendo el cálculo del logit para el sistema SAPS II:

$$\text{Logit} = -7{,}7631 + \text{puntuación SAPS II} \times 0{,}0737 + 0{,}997\ [\ln(\text{puntuación SAPS II} + 1)]$$

En la tabla 3.5 puede comprobarse como la previsión de mortalidad hospitalaria por SAPS II (%) sigue una curva de incremento significativo de la mortalidad asociada al incremento en la puntuación de SAPS II con una morfología sigmoidea, siendo únicamente los valores bajos y los extremadamente altos poco expresivos de cambios significativos en el % de mortalidad prevista.

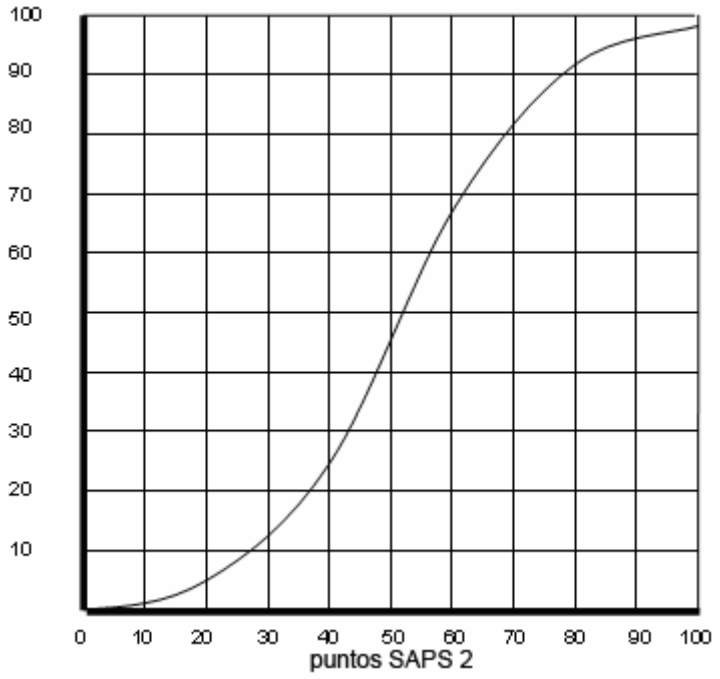


Tabla 3.5. Probabilidad de muerte según el SAPS II.

MPM (Mortality Probability Model) (Tabla 3.6)

El sistema MPM utiliza variables clínicas simples, obtenidas en el momento del ingreso (MPM_0) y a las 24 horas del mismo (MPM_{24}), además de la edad y el estado de salud previa.

Las variables utilizadas en MPM_0 son: Edad, alteración fisiológica aguda (coma o estupor), frecuencia cardiaca ≥ 150 ppm, tensión arterial sistólica ≤ 90 mmHg, ventilación mecánica, fracaso renal agudo, arritmias cardiacas graves, accidente cerebrovascular, sangrado gastrointestinal, efecto masa craneal, reanimación cardiopulmonar previa al ingreso, estado crónico de salud (insuficiencia renal crónica, cirrosis, neoplasia metastásica), tipo de paciente (paciente médico o quirúrgico urgente).

	β	X
Término constante β_0	**- 5,46836**	**Ausente = 0 o presente = 1**
Diagnósticos fisiológicos		
a. Coma o estupor profundo	1,48592	
b. Frecuencia cardíaca ≥ 150/min	0,45603	
c. Tensión arterial sistólica ≤ 90 mmHg	1,06127	
Diagnósticos crónicos		
a. Insuficiencia renal crónica	0,91906	
b. Cirrosis	1,13681	
c. Carcinoma metastásico	1,19979	
Diagnósticos agudos		
a. Fracaso renal agudo	1,48210	
b. Arrímia cardíaca	0,28095	
c. Accidente cerebrovascular	0,21338	
d. Hemorragia gastrointestinal	0,39653	
e. Efecto masa intracraneal	0,86533	
Otros		
a. Edad en años	0,03057	
b. Resucitación cardiopulmonar previa	0,56995	
c. Ventilación mecánica	0,79105	
d. Causa médica o cirugía no programada	1,19098	

Tabla 3.6. Cálculo de probabilidades de mortalidad del MPM II_0

MPM_{24} utiliza algunos de los parámetros de ingreso y valora los cambios evolutivos en las primeras 24 horas de tratamiento en UCI. Las variables utilizadas son: Edad, Parámetros evaluados al ingreso (cirrosis, neoplasia metastásica, efecto masa craneal, paciente médico o quirúrgico urgente), parámetros evaluados a las 24 horas de tratamiento (coma o estupor profundo a las 24 horas, creatinina > 2 mg/dl, infección confirmada, ventilación mecánica a las 24 horas del ingreso, PO2 < 60 mmHg, tiempo de protrombina).

Para el cálculo de la probabilidad individual de morir según cualquier modelo de MPM II, cada uno de las variables (X), queda expresada dicotómicamente (presente=1 ó ausente=0), en su valor absoluto. Este valor se multiplica por el coeficiente de ponderación β, obtenido mediante regresión logística múltiple del estudio original.

El polímero de cálculo (logit) es:

$$\text{Logit} = \sum \beta_i X_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots \beta_{15} X_{15}$$

Este logit se sustituye en la fórmula general del cálculo de la probabilidad de fallecer:

$$\text{Probabilidad} = [e^{logit} / (1 + e^{logit})] \times 100$$

MODS (Multiple Organ Dysfunction Score) (Tabla 3.7)

Es otra escala de valoración de la gravedad de los pacientes críticos basada en la afectación orgánica derivada de la agresión. Esta escala permite una visión evolutiva de la enfermedad y su repercusión sobre la fisiología en función de la respuesta al tratamiento y la capacidad de recuperación funcional del paciente.

Los órganos y sistemas valorados por MODS son la función respiratoria, función renal, función hepática, sistema cardiovascular, hemostasia y estado neurológico, a los que se les asigna una puntuación entre 0 y 4 en función de la desviación respecto a la normalidad.

Puntuación MODS					
Parámetro	Puntos				
	0	1	2	3	4
PAO_2/FiO_2	> 300	226-300	151-225	76-150	≤ 75
Creatinina sérica	≤ 100	101-200	201-350	351-500	≥ 500
Bilirrubina sérica	≤ 20	21-60	61-120	121-240	> 240
Frecuencia cardiaca ajustada al pulso	≤ 10	10.1-15	15.1-20	20.1-30	> 30
Plaquetas	> 120	81-120	51-80	21-50	≤ 20
GCS	15	13-14	10-12	7-9	≤ 6

Tabla 3.7. Escala MODS.

Si no se dispone de algún parámetro se considera la normalidad del mismo. Con el sumatorio de los puntos obtenidos puede calcularse la probabilidad de muerte del paciente (tabla 3.8).

MODS: Disfunción orgánica múltiple y mortalidad	
Puntuación	Mortalidad (%)
0	0
1-4	1
5-8	3
9-12	25
13-16	50
17-20	75
>20	100

Tabla 3.8. Puntuación MODS y riesgo mortalidad.

POSSUM y P- POSSUM (Tabla 3.9)

Son sistemas de evaluación específicos de riesgo quirúrgico, que constan de 2 tipos de variables:

- Variables fisiológicas: son 12, e incluyen signos y síntomas cardiopulmonares, determinaciones de hemograma y bioquímica, y alteraciones electrocardiográficas. Si alguna de las variables no puede ser recogida se le asigna un valor de 1. Se obtienen antes de la intervención quirúrgica y la suma de puntos varía entre 12 y 88.
- Variables quirúrgicas: son 6, divididas en 4 puntuaciones que crecen exponencialmente (1, 2, 4 y 8). La puntuación quirúrgica se obtiene tras la intervención quirúrgica. Los principales ejemplos del grado de intervención en cirugía general se recogen en la tabla 3.10.

Puntuación	1	2	4	8
Variables fisiológicas				
Edad	< 60	61-70	> 70	–
Sistema cardíaco	No	Fármacos	Edema, cardiopatía	Cardiomegalia
Sistema respiratorio	-	EPOC	EPM	Grave
PAS	110-129	130/170 o 10/9	> 170 o 90-99	< 90
Pulso	50-80	81-100 o 40-49	101-120	> 120 o < 40
Glasgow	15	12-14	9-11	< 9
Urea (mmol/l)	< 7,5	7,5-10	10,1-15	> 15
Sodio	> 136	131-135	126-130	< 126
Potasio	3,5-5	3,1-3,4/5,1-5,3	2,9-3,1/5,4-5,9	< 2,9 o > 5,9
Hemoglobina (g/l)	13-16	11,5-12,9/16,1-17	10-11,4/17,1-18	< 10 o > 18
Leucocitos	4-10.000	10,1-20/3,1-3,9	> 20 o < 3,1	–
ECG	Normal	–	F.A. Contr.	Otro
Variables quirúrgicas				
Grav. quir.	Menor	Moderada	Mayor	Mayor +
N.° interv. quir.	1	2	> 2	–
Transf. (µl)	< 100	101-500	501-1.000	> 1.000
Exudado peritoneal	No	Seroso	Pus local	Peritonitis difusa
Malignidad	No	Tumor localizado	Adenopatías	Metástasis
Tipo de cirugía	Programada	–	Urgente resuc. posible	Urgencia inmediata

ECG: electrocardiograma; EPM: enfermedad pulmonar moderada; EPOC: enfermedad pulmonar obstructiva crónica; Grav. quir.: gravedad de la cirugía; N.° interv. quir.: número de intervenciones quirúrgicas; PAS: presión arterial sistólica; POSSUM: Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity; resuc.: reanimación previa a cirugía; Transf.: transfusión sanguínea.

Tabla 3.9. Escalas POSSUM y P-POSSUM.

Menor

- Hernias
- Tumoraciones subcutáneas extensas
- Biopsias de piel y partes blandas
- Cirugía perianal

Intermedia

- Colecistectomías laparotómica / laparoscópica
- Apendicectomía
- Amputaciones menores
- Hemitiroidectomía

Mayor

- Resección intestinal
- Colectomías
- Amputaciones mayores
- Cirugía vía biliar principal
- Tiroidectomía total
- Gastrectomías parciales

Mayor +

- Resección abdominoperineal de Milles
- Gastrectomía totales
- Duodenopancreatectomía cefálica (whipple)
- Hepatectomías totales

Tabla 3.10. Ejemplos de grados de intervención quirúrgica para POSSUM y P-POSSUM.

Una vez que se obtienen las puntuaciones, se calcula el riesgo predicho de mortalidad y morbilidad, usando las siguientes ecuaciones desarrolladas por Copeland et al[209] (Siendo R_1 el riesgo de mortalidad y R_2, el riesgo de morbilidad):

- $L_n\ R_1 / 1 - R_1 = -7{,}04 + (0{,}13 \times$ puntuación fisiológica) + (0,16 x puntuación de gravedad operatoria).

- $L_n\, R_2 / 1 - R_2 = -5{,}91 + (0{,}16 \times$ puntuación fisiológica) + (0,19 x puntuación de gravedad operatoria).

La única diferencia técnica entre el POSSUM y el P-POSSUM está en el tipo de ecuación que utilizan para el cálculo del riesgo de mortalidad, siendo la ecuación utilizada por P-POSSUM la siguiente:

- $L_n\ R_1 / 1 - R_1 = -9{,}07 + (0{,}17 \times$ puntuación fisiológica) + (0,16 x puntuación de gravedad operatoria).

3.4. DETERMINACIÓN DEL TAMAÑO MUESTRAL

El tamaño muestral se ha calculado considerando una precisión del 95% y una probabilidad esperada del 50% y una confianza del 95%, mediante la fórmula:

$$n = Z^2 \cdot p \cdot (1\text{-}p) / i^2$$

Lo que supone un total de 384 casos de pacientes intervenidos de cirugía abdominal programada con ingreso en el SCGD del HGUMM. Este número se incrementó en un 10% para controlar la pérdida de pacientes durante el seguimiento y aquellos casos en que fuera imposible acceder a la historia clínica del caso, quedando definido el tamaño muestral definitivo en 422 pacientes.

A partir del tamaño muestral previsto la selección se realizó mediante muestreo estratificado. La muestra se obtuvo de forma prospectiva de los pacientes del Servicio de Cirugía General. Dado que la duración del estudio fue calculada para 2 años, establecimos la revisión de unos 4 pacientes por semana.

3.5. PROCESO DE REVISIÓN DE LOS CASOS A ESTUDIO

La identificación de los casos a incluir en el estudio se realizó a partir del parte quirúrgico semanal del SCGD del HGUMM. A partir de éste se efectuó un muestreo sistemático seleccionando un paciente cada cierto intervalo fijo de casos. Esta fracción muestral se calculó, dividiendo el número total de casos del marco muestral (definido por el número total de intervenciones quirúrgicas programadas con ingreso esa semana) por el tamaño de la muestra de esa semana.

La revisión de casos incluyó el seguimiento de cada paciente desde la intervención quirúrgica hasta el 30 días postoperatorio, si el paciente seguía ingresado en el hospital, dicho día acababa el seguimiento del mismo, y si ya había sido dado de alta del hospital se citaba en consultas externas de cirugía al mes de la intervención. Se recogieron todas las morbilidades descritas por Coopeland et al[209] y en el caso de haberla, la mortalidad y su causa.

Todos los documentos de la historia clínica necesarios (hoja operatoria, hoja de preanestesia, electrocardiograma, analítica preoperatorio, evolución clínica y seguimiento en consultas externas de cirugía, informe de alta, gráficas y notas de enfermería, informes de interconsulta y resultados de pruebas complementarias) fueron utilizados para la revisión y cumplimentación de los protocolos de cálculo de las distintas escalas de riesgo valoradas. Los datos obtenidos se trasladaron a una base de datos y se compararon los resultados predichos por las escalas de riesgo con los resultados observados en el seguimiento clínico real de los pacientes.

3.6. CONCORDANCIA ENTRE OBSERVADORES Y CONSISTENCIA INTERNA DE LAS ESCALAS

Los médicos revisores recibieron formación y entrenamiento específico en el manejo de las 6 escalas de riesgo evaluadas y se realizaron talleres de manejo de dichas escalas previamente a la realización del proceso de validación y recogida de datos. Además, se identificaron las dificultades en el uso del cuestionario, las discrepancias en la revisión y se consensuaron los aspectos que plantearon alguna duda al aplicar las escalas. Los resultados en esta fase sirvieron para definir los términos con el fin de asegurar que la recogida de cada uno de los ítems no daba lugar a interpretaciones ambiguas.

La fiabilidad de las escalas de riesgo evaluadas, definida como el grado de reproductibilidad de los resultados cuando el instrumento es utilizado por observadores diferentes, se evaluó mediante el *índice Kappa* (*k)*. El índice Kappa mide la concordancia total que existe si excluimos la debida al azar, o el acuerdo real más allá del azar. Su cálculo responde a la fórmula general:

$$K = P_o – P_e \ / \ 1 - P_e$$

Donde P_o es la concordancia observada, y P_e es la concordancia esperada debida exclusivamente al azar[300,301]. El estadístico Kappa de Cohen se utiliza para corregir el acuerdo debido al azar, permitiendo estimar la significación estadística y los correspondientes intervalos de confianza, entre la diferencia en el grado de acuerdo que sería esperable simplemente por el azar (valor 0) y el grado de acuerdo observado (el acuerdo perfecto no debido al azar recibe el valor de 1). Los cálculos de P_0 y P_e, así como el cálculo de la desviación estándar se realizaron según las fórmulas indicadas por Fleiss tanto

para parejas de evaluadores como para el caso de evaluadores múltiples[300]. Existen publicadas en la literatura varias propuestas para su interpretación: Según Fleiss, se considera que un Kappa traduce buena fiabilidad si es mayor o igual a 0,6[300]. Según Landis y Koch[301], Kappa negativo ó 0 indica acuerdo bajo o inexistente; hasta 0,2 acuerdo ligero; por encima de 0,4 es acuerdo moderado; a partir de 0,6 acuerdo notable; a partir de 0,8 acuerdo alto; y 1 acuerdo absoluto. En general, la mayoría de los autores coinciden en dar por aceptables kappas mayores de 0,4[302,303].

Para medir la consistencia interna de las escalas hemos empleado el *test Alfa de Cronbach*, cuya fórmula general responde a la fórmula:

$$\frac{np}{1+p(n-1)}$$

Siendo;
n el número de ítems,
p el promedio de las correlaciones lineales entre cada uno de los ítems.

El test Alfa de Cronbach se trata de un índice de consistencia interna que toma valores entre 0 y 1 y que sirve para comprobar si el instrumento que se está evaluando recopila información defectuosa y por tanto nos llevaría a conclusiones equivocadas o si se trata de un instrumento fiable que mide lo que dice que mide. Alfa es por tanto un coeficiente de correlación al cuadrado que, a grandes rasgos, mide la homogeneidad de las preguntas promediando todas las correlaciones entre todos los ítems para ver que, efectivamente, se parecen. Su interpretación será que, cuanto más se acerque el índice al extremo 1, mejor es la fiabilidad, considerando una fiabilidad respetable a partir de 0,80.

3.7. ANÁLISIS DE LOS DATOS

El análisis de los datos se realizó mediante los programas informáticos SPSS 11.0® versión para Windows (SPSS Inc., Chicago. IL. USA) y Microsoft Excel® (Microsoft Corporation, Redmond, Washington, USA). Se calcularon las razones de mortalidad y morbilidad observada (O) y esperada (E) (*ratio O:E*) para el sistema POSSUM y P-POSSUM, así como las razones de mortalidad observadas (O) y esperadas (E) para el resto de escalas; una ratio de 1 indica una correlación perfecta entre lo esperado y lo observado; si es < 1 expresa que los resultados obtenidos son mejores que los esperados; y si es > 1, los resultados obtenidos son peores que los esperados.

Para evaluar la bondad de las escalas, procedimos al cálculo del área bajo la curva (ABC) tanto para la mortalidad (en las 6 escalas) como para la morbilidad (en el caso de POSSUM y P-POSSUM), con sus intervalos de confianza del 95% ($IC_{95\%}$), mediante las *curvas ROC* (Receiver Operating Characteristic). Estas curvas pueden tomar valores entre 1 (prueba perfecta) y 0,5 (prueba inútil), de modo que una prueba tendrá mayor capacidad de discriminación cuanto más se aproxime al 1 su ABC.

Para determinar si existían diferencias estadísticamente significativas entre los resultados obtenidos y los esperados según las diferentes escalas de riesgo se aplicó el test X^2 de Pearson.

En todos los casos se consideró una diferencia entre lo observado y lo esperado o entre dos escalas de riesgo como significativa cuando el nivel de significación resultante era menor de 0,05 ($p< 0,05$).

4. RESULTADOS

4.1. VALIDACIÓN DE LAS ESCALAS DE RIESGO (Tabla 4.1).

4.1.1. Validación de la escala POSSUM y P-POSSUM.

Para el proceso de validación de estas 2 escalas se utilizaron 180 historias clínicas, seleccionadas por muestreo aleatorio estratificado, obteniendo los siguientes resultados:

1.- Traducción de la escala[299]: En cuanto a las variables numéricas, no se observaron problemas ni diferencias significativas en la comparación de las traducciones llevadas a cabo. Para las variables cualitativas y descriptivas se observó una muy buena correlación interobservador (Kappa = 0,9).

2.- Prueba de valoración de comprensión de los distintos ítems de la escala: Para las variables cuantitativas no se encontraron diferencias en la interpretación y aplicación de sus categorías. Para las variables intraoperatorias, tampoco hubo diferencias. En cuanto a las variables cualitativas, hubo algunas diferencias en cuanto a distinción entre normalidad y patología leve, llegándose a un acuerdo entre los evaluadores, y formulando e incluyendo en el protocolo estandarizado de recogida de datos una descripción sobre lo que incluye y a lo que se refiere cada una de ellas.

3.- Confiabilidad interobservador: No hubo diferencias significativas ni entre las variables fisiológicas ni entre las quirúrgicas para distintos observadores. Existiendo una concordancia *K* mayor del 0,8 para todas las variables del score fisiológico, y un grado de concordancia entre el 0,7 y el 0,9 para las quirúrgicas.

En cuanto a la evaluación de la homogeneidad se observó una alfa de Cronbach de 0,8 para el score fisiológico y de 0, 72 para el quirúrgico.

4.- Tiempo medio: Fue de 8,5 minutos (rango: 4-15 minutos).

4.1.2. Validación de la escala APACHE II.

Para su validación fue necesario el uso de 150 historias clínicas, seleccionadas por muestreo aleatorio estratificado, obteniendo los siguientes resultados:

1.- Traducción de la escala: No hubo diferencias significativas en ninguno de los ítems valorados, siendo el índice Kappa calculado para esta escala de 0,85

2.- Prueba de valoración de comprensión de los distintos ítems de la escala: No hubo diferencias en la interpretación de las variables cuantitativas. Hubo algunas diferencias en cuanto a la determinación de la escala de coma de Glasgow, por lo que se llegó a un acuerdo entre los evaluadores para su determinación, incluyéndose en el protocolo de recogida de datos la misma con sus definiciones por categorías.

3.- Confiabilidad interobservador: No hubo diferencias significativas para los distintos observadores. Existiendo una concordancia *K* mayor del 0,8.

En cuanto a la evaluación de la homogeneidad se observó una alfa de Cronbach de 0,75.

4.- Tiempo medio: Fue de 11 minutos (rango: 7-20 minutos).

4.1.3. Validación de la escala SAPS II.

Para su validación fue necesario el uso de 150 historias clínicas, seleccionadas por muestreo aleatorio estratificado, obteniendo los siguientes resultados:

1.- Traducción de la escala: No hubo diferencias significativas en ninguno de los ítems valorados, siendo el índice Kappa calculado para esta escala de 0,85.

2.- Prueba de valoración de comprensión de los distintos ítems de la escala: No hubo diferencias en la interpretación de las variables cuantitativas. Al igual que para la escala APACHE II hubo algunas diferencias en cuanto a la determinación de la escala de coma de Glasgow, por lo que se llegó a un acuerdo entre los evaluadores para su determinación, incluyéndose en el protocolo de recogida de datos la misma con sus definiciones por categorías.

3.- Confiabilidad interobservador: No hubo diferencias significativas para los distintos observadores. Existiendo una concordancia *K* mayor del 0,85.

En cuanto a la evaluación de la homogeneidad se observó una alfa de Cronbach de 0,8.

4.- Tiempo medio: Fue de 10,3 minutos (rango: 8-19 minutos).

4.1.4. Validación de la escala MPM.

Para su validación fue necesario el uso de 140 historias clínicas, seleccionadas por muestreo aleatorio estratificado, obteniendo los siguientes resultados:

1.- Traducción de la escala: No hubo diferencias significativas en ninguno de los ítems valorados, siendo el índice Kappa calculado para esta escala de 0,9

2.- Prueba de valoración de comprensión de los distintos ítems de la escala: No hubo diferencias significativas en cuanto a la valoración de las variables cuantitativas. Sí hubo desacuerdos significativos en cuanto a la valoración de los diferentes ítems dicotómicos: 1) Presencia o no de cirrosis; 2) Presencia o no de efecto masa intracraneal; 3) Neoplasia metastásica o no; 4) Presencia de coma o estupor profundo a las 24 horas del ingreso; 5) Infección confirmada o no; 6) Tratamiento con drogas vasoactivas durante más de una hora. No fue posible establecer un acuerdo entre observadores para consensuar la valoración de dichos ítems.

3.- Confiabilidad interobservador: Hubo diferencias significativas para los distintos observadores. Existiendo una concordancia *K* del 0,55.

En cuanto a la evaluación de la homogeneidad se observó una alfa de Cronbach de 0,65.

4.- Tiempo medio: Fue de 10 minutos (rango: 5-13 minutos).

4.1.5. Validación de la escala MODS.

Para su validación fue necesario el uso de 60 historias clínicas, seleccionadas por muestreo aleatorio estratificado, obteniendo los siguientes resultados:

1.- Traducción de la escala: No hubo diferencias significativas en ninguno de los ítems valorados, siendo el índice Kappa calculado para esta escala de 0,95

2.- Prueba de valoración de comprensión de los distintos ítems de la escala: No hubo diferencias en la interpretación de las variables cuantitativas. Al igual que para la escala APACHE II y SAPS II hubo algunas diferencias en cuanto a la determinación de la escala de coma de Glasgow, llegando al mismo acuerdo que para las dos anteriores.

3.- Confiabilidad interobservador: No hubo diferencias significativas para los distintos observadores. Existiendo una concordancia *K* mayor del 0,9.

En cuanto a la evaluación de la homogeneidad se observó una alfa de Cronbach de 0,8.

4.- Tiempo medio: Fue de 6 minutos (rango: 3-11 minutos).

Escalas	Trad. Escalas (*K*)	Comprensión ítems (p<0,05)	Conf. Interobs (*K*)	α-Cronbach	tº (min)
(P)- POSSUM	0,9	N.S	0,85	0,75	8,5
APACHE II	0,85	N.S	0,8	0,75	11
SAPS II	0,85	N.S	0,85	0,8	10,3
MPM	0,9	SI	0,55	0,65	10
MODS	0,95	NS	0,9	0,8	6

Tabla 4.1. Validación de escalas de riesgo. (Abreviaturas de la tabla: Trad. Escalas: traducción escalas; Conf. Interobs: confiabilidad interobservador; tº: tiempo medio de aplicación; N.S: diferencias no significativas).

4.2. COMPARACIÓN PROSPECTIVA DE LAS ESCALAS DE RIESGO EN CUANTO A MORTALIDAD.

4.2.1. Resultados generales de mortalidad observada y esperada.

384 pacientes fueron incluidos en el estudio, los diagnósticos de los pacientes se recogen en la tabla 4.2. La edad media fue de 67,7 años (rango: 17- 94 años), con una distribución por sexo de 223 hombres (58%) y 161 mujeres (42%). Se han excluido del estudio aquellos pacientes compartidos o derivados a otros servicios, y a los que no se ha podido realizar el seguimiento dentro de los 30 días postoperatorios.

Diagnósticos	Número pacientes
Carcinomatosis peritoneal	7
Neoplasia de páncreas	6
Colitis ulcerosa	4
Neoplasia de recto	63
Eventración	15
Neoplasia de colon	48
Enfermedad de Cronh	3
Obesidad mórbida	75
Neoplasia gástrica	9
Cirugía bilio-pancreática	154

Tabla 4.2. Diagnósticos indicativos de laparotomía/laparoscopia.

Fallecieron 34 pacientes (6,25%) dentro de los 30 días tras la cirugía; y 148 pacientes tuvieron algún tipo de morbilidad (38,5%). Los resultados en cuanto a morbilidad se recogen en la tabla 4.3 y la tabla 4.4 muestra las causas de mortalidad.

Tipo de comorbilidad	Número de casos
Infección herida	22
Peritonitis	17
Hemorragia digestiva	5
Dehiscencia pared	8
Hemoperitoneo	3
Infección urinaria	9
Fuga anastomótica	21
SDRA	4
Sepsis/FMO	29
Insuficiencia renal	3
IAM	2
ACV	1
EAP	12
TEP	2
Neumonía	2
Absceso intraabdominal	8

Tabla 4.3. Tipo y frecuencia de las comorbilidades. (Abreviaturas de la tabla: FMO: fracaso multiorgánico; IAM: Infarto agudo de miocardio; EAP: Edema agudo de pulmón; SDRA: Síndrome de distress respiratorio del adulto; ACV: Accidente cerebro-vascular; TEP: Tromboembolismo pulmonar).

Causa de mortalidad	Número de casos
Shock séptico	14
Fallo multiorgánico	6
TEP	1
Peritonitis aguda	4
Fallo cardio-respiratorio	5
EAP	4

Tabla 4.4. Causas y frecuencia de mortalidad. (Abreviaturas de la tabla: EAP: Edema agudo de pulmón; TEP: Tromboembolismo pulmonar).

Hubo un 6,25% de mortalidad en el grupo de pacientes estudiados, siendo la mortalidad estimada por el sistema POSSUM y P-POSSUM del 12,5% (48 pacientes) y 12% (46 pacientes) respectivamente sin observarse diferencias significativas entre lo estimados por estos sistemas y lo observado en la realidad. La escala APACHE II estimó una mortalidad del 5,7% (22 pacientes) existiendo diferencias significativas con la observado (p=0,04). La mortalidad obtenida por SAPS II y MODS fue del 3,6 % (14 pacientes) con una p=0,001 (tabla 4.5).

Escala	Mortalidad observada	%	Mortalidad esperada	%	Significación estadística
POSSUM	34	6,25	48	12,5	N.S
PPOSSUM	34	6,25	46	12	N.S
APACHE II	34	6,25	22	5,7	SI (p=0,04)
SAPS II	34	6,25	14	3,6	SI (p=0,001)
MODS	34	6,25	14	3,6	SI (p=0,001)

Tabla 4.5. Mortalidad observada y predicha por cada escala de riesgo. (Abreviaturas de la tabla: N.S: No significativo considerando la significación estadística en p<0,05; %: porcentaje de mortalidad).

4.2.2.- Mortalidad por intervalos de riesgo para cada una de las escalas.

En cuanto al sistema POSSUM (tabla 4.6), observamos que su capacidad de predicción en cuanto a mortalidad es buena tanto en todos los intervalos de riesgo como de forma global, sin observarse diferencias significativas en ninguno de ellos, además, las ratios O:E son buenas en todos los intervalos, ya que son menores de 1 en general y de forma global (ratio O:E global 0,7), solamente en el grupo con riesgo del 80-100% es igual a 1. Los resultados obtenidos para el sistema P-POSSUM son muy similares a los de la escala POSSUM (tabla 4.7).

Riesgo de Muerte	Nº Intervenciones	Muertes Esperadas	Muertes Reales	Ratio O/E	Significación Estadística
<20%	247	10	5	0,5	N.S
20-39%	110	22	17	0,7	N.S
40-59%	5	2	2	1	N.S
60-79%	20	12	8	0,6	N.S
80-100%	2	2	2	1	N.S
Total	384	48	34	0,7	N.S

Tabla 4.6. Mortalidad observada y esperada por el Sistema POSSUM. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema POSSUM; N.S: No significativo considerando la significación estadística en p<0,05).

Riesgo de Muerte	Nº Intervenciones	Muertes Esperadas	Muertes Reales	Ratio O/E	Significación Estadística
<20%	229	6	5	0,8	N.S
20-39%	127	25	17	0.7	N.S
40-59%	4	2	2	1	N.S
60-79%	22	13	8	0,6	N.S
80-100%	2	2	2	1	N.S
Total	384	46	34	0,7	N.S

Tabla 4.7. Mortalidad observada y esperada por el Sistema P-POSSUM. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema P-POSSUM; N.S: No significativo considerando la significación estadística en p<0,05).

Los resultados para el APACHE II (tabla 4.8) muestran una agrupación de todos los pacientes en los 3 primeros intervalos de riesgo: < 20%, 20-39% y 40-59% existiendo diferencias significativas entre el riesgo estimado de mortalidad por esta escala para los pacientes del grupo entre 20-39% (p=0,04) y de forma global (p=0,04).

Riesgo de Muerte	Nº Intervenciones	Muertes Esperadas	Muertes Reales	Ratio O/E	Significación Estadística
<20%	320	9	11	0,8	N.S
20-39%	61	12	20	1,6	SI (p=0,04)
40-59%	3	1	3	3	N.S
60-79%	0	-	-	-	-
80-100%	0	-	-	-	-
Total	384	22	34	0,7	SI (p=0,04)

Tabla 4.8. Mortalidad observada y esperada por el Sistema APACHE II. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema APACHE II; N.S: No significativo considerando la significación estadística en p<0,05).

La mortalidad calculada por la escala SAPS II (tabla 4.9) muestra diferencias significativas tanto por estratos como por riesgo global con respecto a la mortalidad observada en la realidad. Además, agrupa a todos los pacientes estudiados en los dos primeros grupos de riesgo (riesgo menor del 20% y riesgo entre 20-39%).

Riesgo de Muerte	Nº Intervenciones	Muertes Esperadas	Muertes Reales	Ratio O/E	Significación Estadística
<20%	369	11	22	2	SI (p=0,002)
20-39%	15	3	12	4	SI (p=0,001)
40-59%	0	-	-	-	-
60-79%	0	-	-	-	-
80-100%	0	-	-	-	-
Total	384	14	34	2,4	SI (p=0,001)

Tabla 4.9. Mortalidad observada y esperada por el Sistema SAPS II. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema SAPS II; N.S: No significativo considerando la significación estadística en p<0,05).

El sistema MODS (tabla 4.10) presenta resultados muy similares a los de la escala SAPS II, tanto globalmente como por intervalos de riesgo, siendo significativas las diferencias observadas entre lo calculado por la escala y lo ocurrido en realidad. Al igual que la escala SAPS II y APACHE II, agrupa todos los pacientes en los estratos de riesgo más bajos (< 20%, 20-39% y 40-59%).

Riesgo de Muerte	Nº Intervenciones	Muertes Esperadas	Muertes Reales	Ratio O/E	Significación Estadística
<20%	370	11	23	2	SI (p=0,01)
20-39%	12	2	9	4,5	SI (p=0,01)
40-59%	2	1	2	2	N.S
60-79%	0	-	-	-	-
80-100%	0	-	-	-	-
Total	384	14	34	2,4	SI (p=0,001)

Tabla 4.10. Mortalidad observada y esperada por el Sistema MODS. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema MODS; N.S: No significativo considerando la significación estadística en p<0,05).

4.2.3.- Cálculo de las curvas ROC para cada una de las escalas estudiadas.

En la tabla 4.11 se recogen los resultados del cálculo de las curvas ROC para cada una de las escalas de riesgo, observándose un valor muy cercano al 1 (prueba perfecta) tanto para la escala POSSUM (área ROC = 0,78) como para la P-POSSUM (área ROC = 0,79) sin que existan diferencias significativas entre los dos sistemas.

Los sistemas APACHE II, SAPS II y MODS presentan valores ROC cercanos a 0,5 (prueba inútil) existiendo diferencias significativas entre ellas y las escalas POSSUM (p<0,001) y P-POSSUM (p<0,001).

Entre APACHE II y SAPS II no existen diferencias significativas, pero sí entre éstas y la escala MODS.

Escalas	ROC mortalidad	$IC_{95\%}$
POSSUM	0,78	0,68-0,88
P-POSSUM	0,79	0,71-0,87
APACHE II	0,62	0,56-0,68
SAPS II	0,58	0,52-0,64
MODS	0,56	0,48-0,64

Tabla 4.11. Comparación de las curvas ROC para mortalidad de las escalas de riesgo. (Abreviaturas de la tabla: $IC_{95\%}$: Intervalo confianza del 95%).

4.2.4.- Cálculo del índice de Shannon para cada una de las escalas de riesgo.

La precisión de las escalas estudiadas se evaluó con el índice de Shannon (IS), para ello se ha procedido al cálculo del IS para fallecidos (su cálculo sólo incluye los casos de mortalidad) (tabla 4.12) y el IS global (para su cálculo es necesario computar los pacientes vivos y muertos) (tabla 4.13).

Escalas	Índice de Shannon fallecidos	DE
POSSUM	0,92	± 0,24
P-POSSUM	0,91	± 0,29
APACHE II	0,64	± 0,49
SAPS II	0,44	± 0,50
MODS	0,41	± 0,49

Tabla 4.12. Índices de Shannon para fallecidos calculado para cada una de las escalas de riesgo. (Abreviaturas de la tabla: DE: Desviación estándar).

Escalas	Índice de Shannon global	DE
POSSUM	0,103	± 0,29
P-POSSUM	0,102	± 0,30
APACHE II	0,057	± 0,23
SAPS II	0,036	± 0,18
MODS	0,036	± 0,18

Tabla 4.13. Índices de Shannon global calculado para cada una de las escalas de riesgo. (Abreviaturas de la tabla: DE: Desviación estándar).

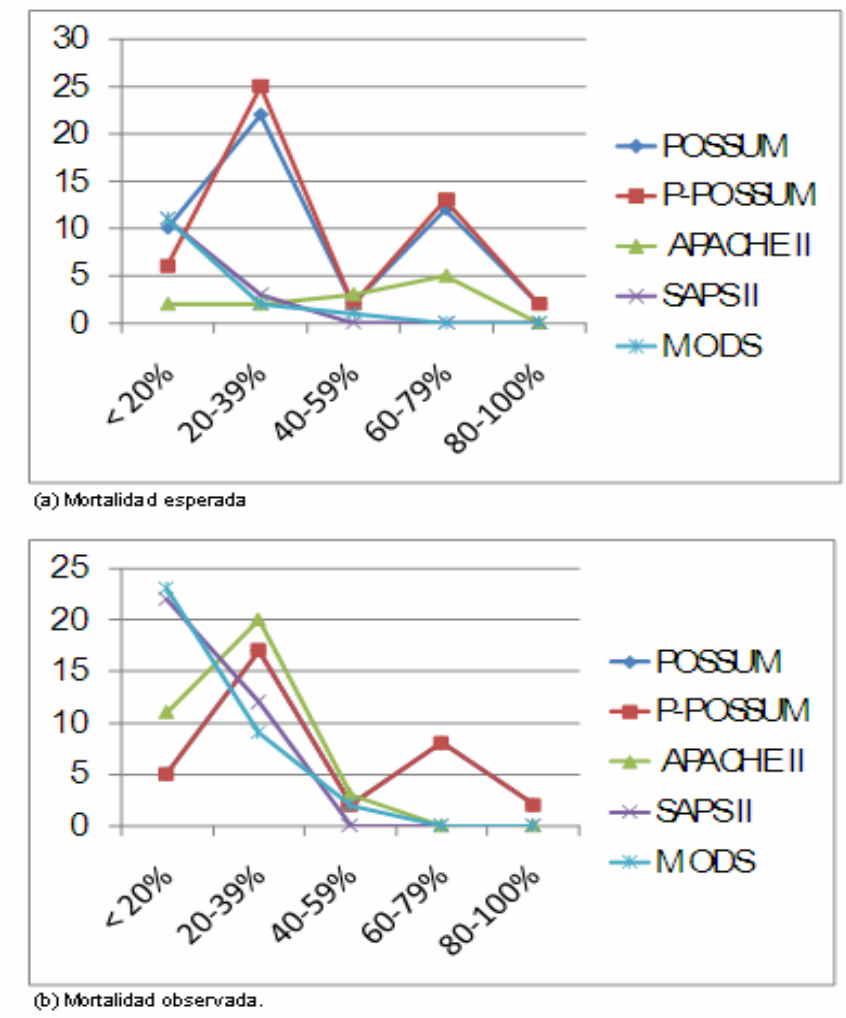
Dentro del cálculo del IS para fallecidos se obtuvieron los mejores resultados para POSSUM (0,92) y P-POSSUM (0,91) sin diferencias significativas entre ellos. Siendo el IS para fallecidos de 0,64 para APACHE II, 0,44 para SAPS II y de 0,41 para MODS. Hubo diferencias significativas entre el valor obtenido para POSSUM y los valores para APACHE II (p=0,002), SAPS II (p=0,000003) y MODS (p=0,0000009), así como para los resultados de P-POSSUM y APACHE II (p=0,004), SAPS II (p=0,000008) y MODS (p=0,000002). No hubo diferencias significativas entre estas tres últimas escalas. Estos resultados indican que POSSUM y P-POSSUM les otorgó valores de riesgo elevado a los pacientes que murieron, mientras que los otros tres índices no realizaron la labor de discriminación de los fallecidos de forma eficiente.

De forma similar, en el cálculo del IS global se obtuvieron los valores más altos para POSSUM (0,103) y P-POSSUM (0,102), y los valores más bajos para MODS (0,036), SAPS II (0,036) y APACHE II (0,057). Hubo diferencias significativas entre los resultados obtenidos para POSSUM y APACHE II (p=0,01), SAPS II y MODS (p=0,000005). También hubo diferencias significativas entre P-POSSUM y APACHE II (p=0,009), MODS y SAPS II (p=0,0001). No hubo diferencias significativas entre los valores de estos tres últimos índices pronóstico.

4.2.5.- Métodos gráficos y resultados en cuanto a mortalidad.

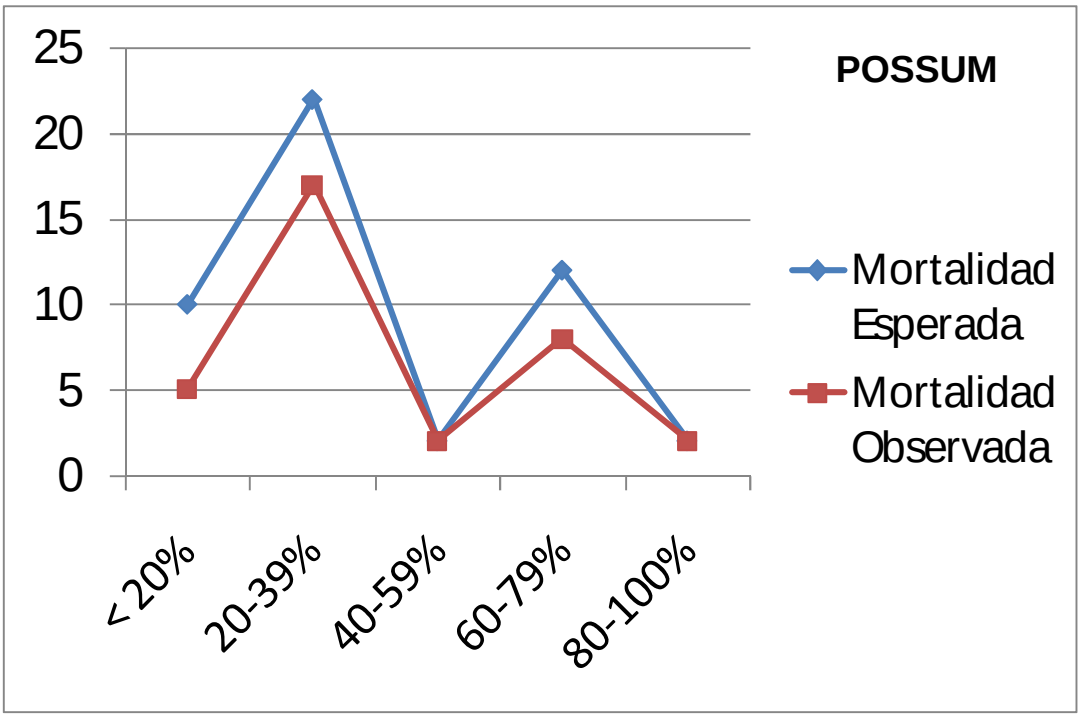
4.2.5.A.- Representaciones de mortalidad esperada y observada.

En este apartado se ha representado gráficamente la mortalidad esperada para cada escala de riesgo, comparándola con la observada para cada una de ellas (gráfica 4.1). Además, en representaciones gráficas sucesivas, se ha desglosado por escala pronóstica los resultados de mortalidad esperada *por intervalos de riesgo* y se han "superpuesto" a la gráfica de mortalidad observada para cada una de ellas. Brevemente comentamos los resultados de cada una de las escalas:



Gráfica 4.1. Comparación gráfica entre mortalidad esperada (a) y observada (b) para cada escala de riesgo.
(Ejes: Vertical: número de muertes; Horizontal: intervalos de riesgo de mortalidad).

El sistema POSSUM (gráfica 4.2) muestra unas gráficas de mortalidad esperada: observada muy similares, prácticamente superponibles sin diferencias significativas.

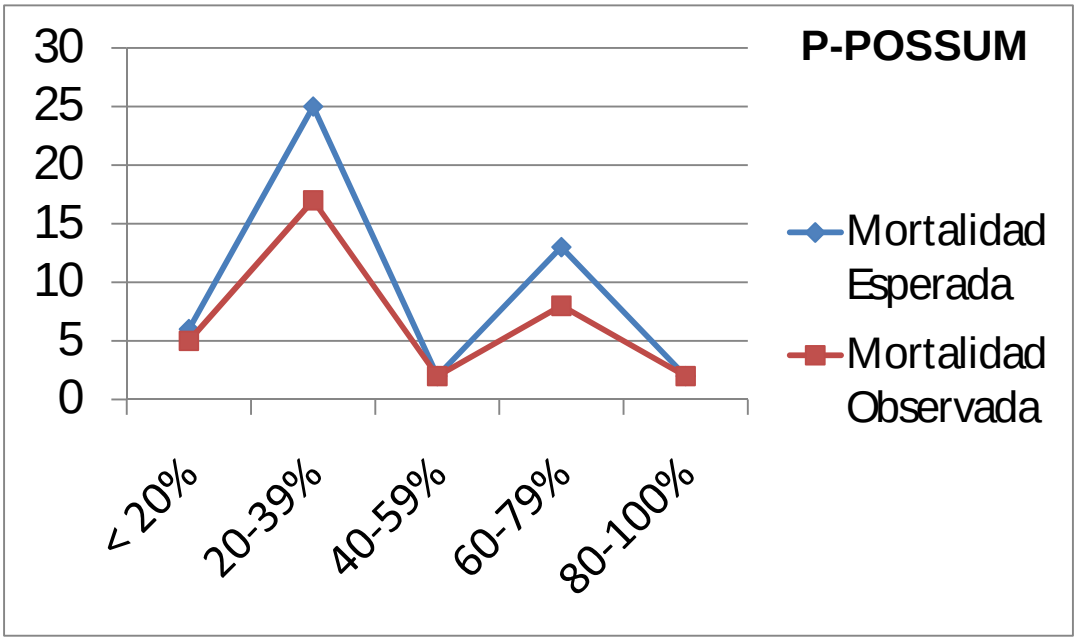


Gráfica 4.2. Comparación gráfica entre mortalidad esperada (azul) y observada (roja) para la escala POSSUM. (Ejes: Vertical: número de muertes; Horizontal: intervalos de riesgo de mortalidad).

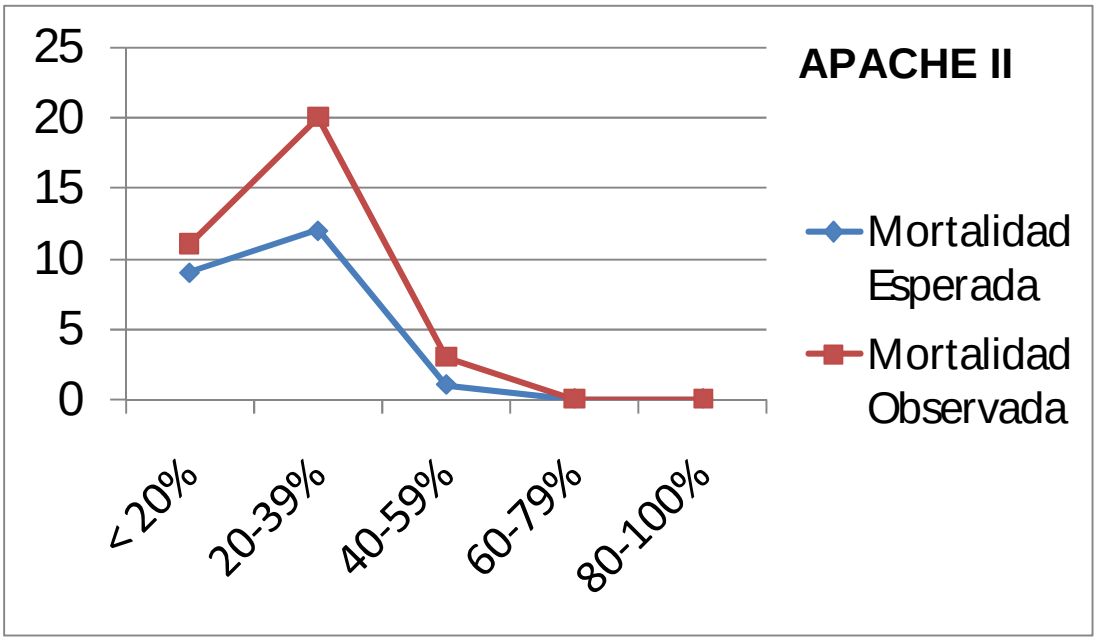
La escala P-POSSUM (gráfica 4.3) muestra resultados semejantes a los de la escala POSSUM, sin que existan diferencias significativas entre ellas.

El índice APACHE II (gráfica 4.4) sobreestima la mortalidad de forma significativa en el intervalo de riesgo de 20-39%, estrato en el que mayor número de pacientes han fallecido.

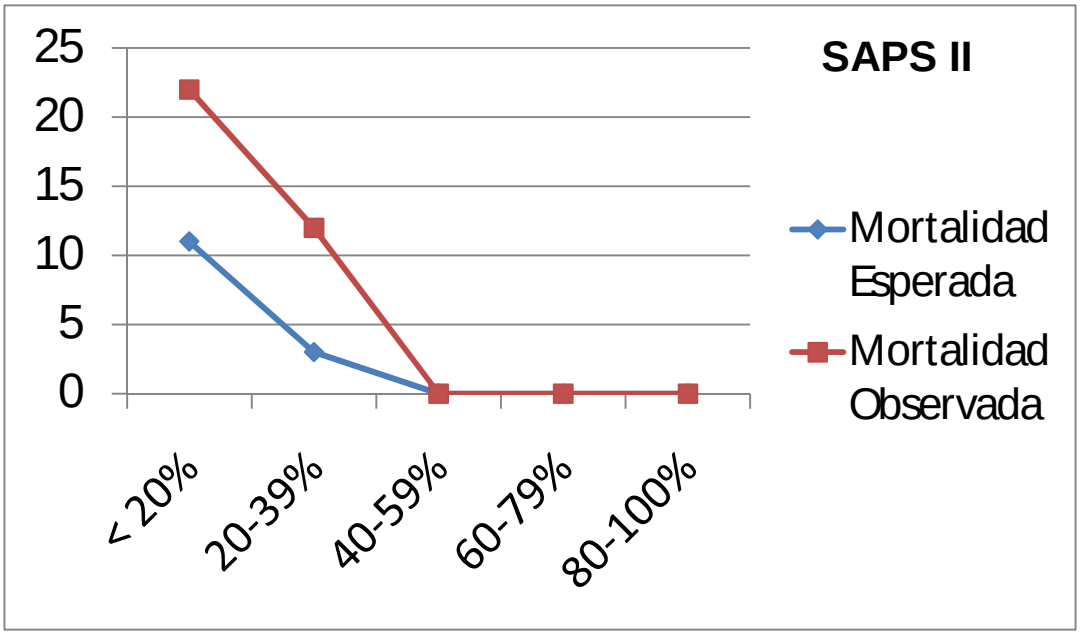
El estudio gráfico del sistema SAPS II (gráfica 4.5) indica una sobreestimación de la mortalidad estadísticamente significativa en los dos intervalos de riesgo en los que ha clasificado los pacientes.



Gráfica 4.3. Comparación gráfica entre mortalidad esperada (azul) y observada (roja) para la escala P-POSSUM. (Ejes: Vertical: número de muertes; Horizontal: intervalos de riesgo de mortalidad).

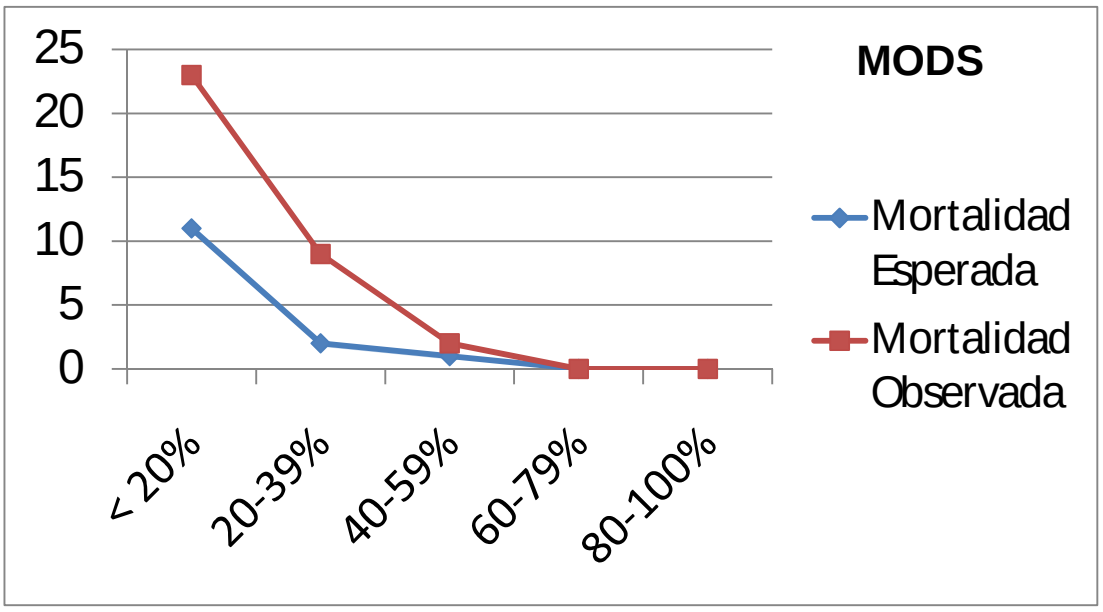


Gráfica 4.4. Comparación gráfica entre mortalidad esperada (azul) y observada (roja) para l la escala APACHE II. (Ejes: Vertical: número de muertes; Horizontal: intervalos de riesgo de mortalidad).



Gráfica 4.5. Comparación gráfica entre mortalidad esperada (azul) y observada (roja) para la escala SAPS II. (Ejes: Vertical: número de muertes; Horizontal: intervalos de riesgo de mortalidad).

La gráfica 4.6 demuestra que el sistema MODS sobreestima de forma significativa la mortalidad esperada de forma global.

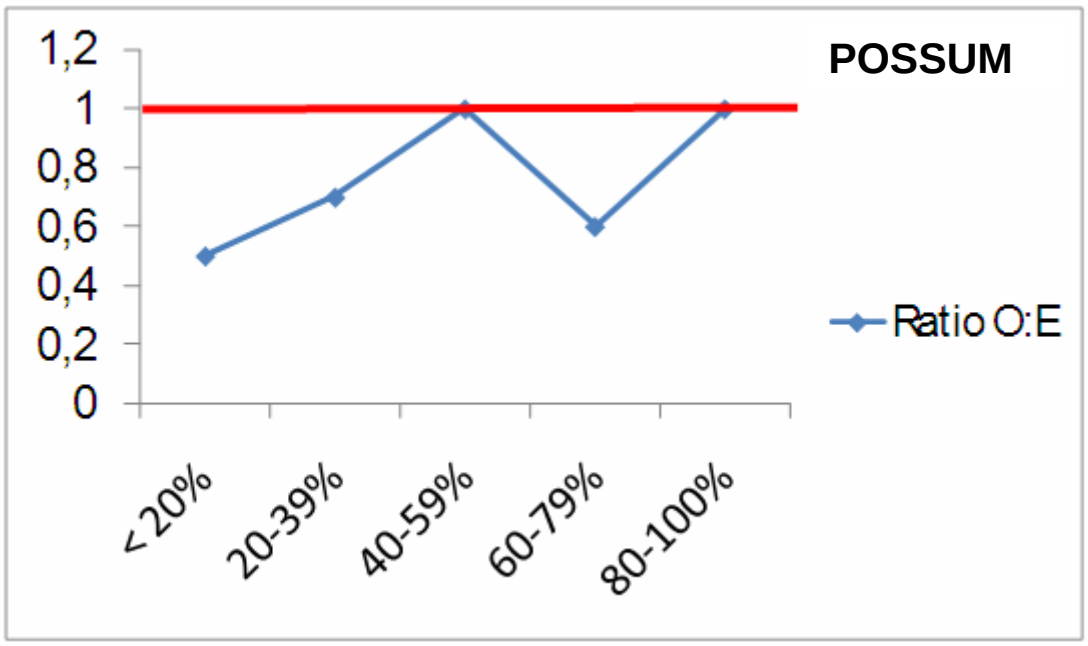


Gráfica 4.6. Comparación gráfica entre mortalidad esperada (azul) y observada (roja) para la escala MODS. (Ejes: Vertical: número de muertes; Horizontal: intervalos de riesgo de mortalidad).

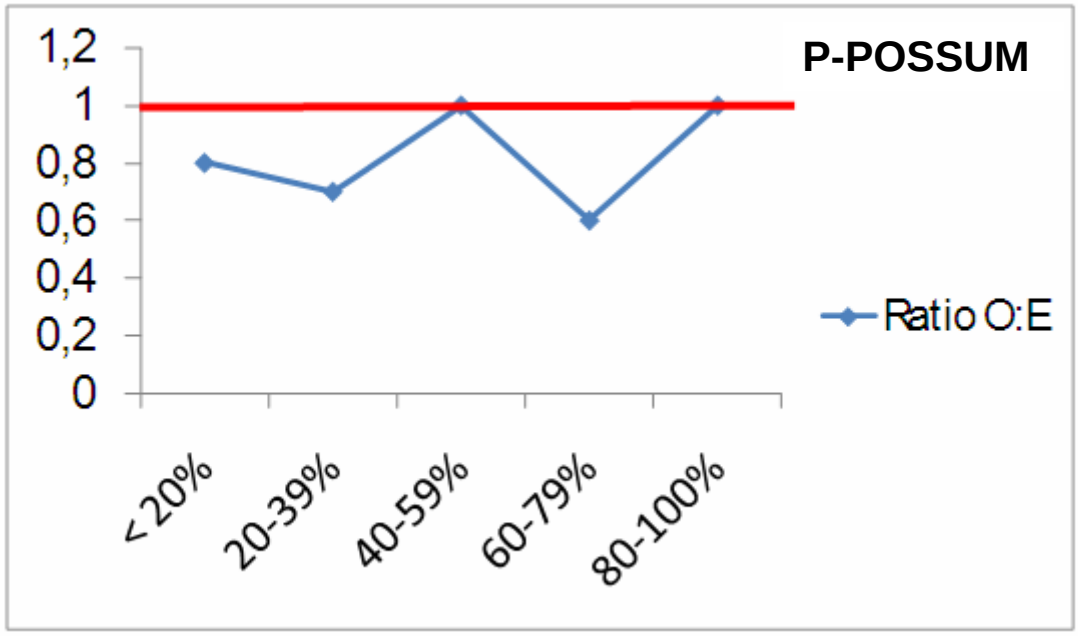
4.2.5.B.- Representaciones del ratio observado:esperado de mortalidad.

El estudio gráfico se ha completado con la representación gráfica de las razones de mortalidad observadas:esperadas *por intervalo de riesgo*, representando con una línea horizontal el punto de corte en 1; valor a partir del cual, los resultados que aparezcan por encima de él indican resultados observados peores de los esperados y por debajo de él lo contrario (resultados obtenidos mejores de los esperado), siendo dicho punto 1 el equilibrio (resultados esperados iguales a los observados).

En la gráficas 4.7 y 4.8 se observa que las ratios O:E en cuanto a mortalidad para POSSUM y P-POSSUM respectivamente se ajustan a una buena predicción (valores iguales o menores de 1).

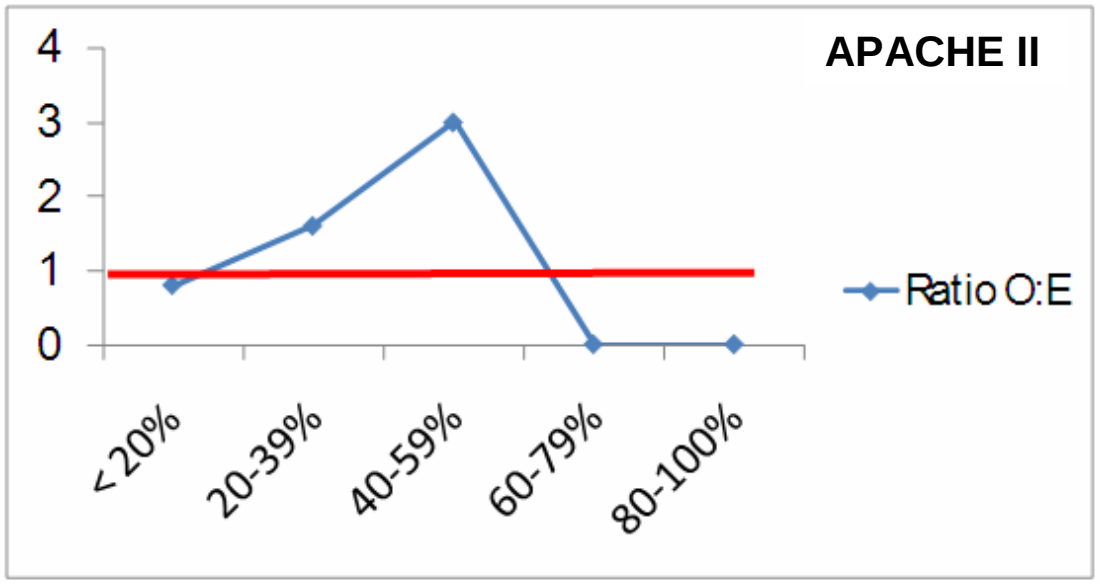


Gráfica 4.7. Ratio de mortalidad Observada:Esperada por intervalos de riesgo calculado para la escala POSSUM.



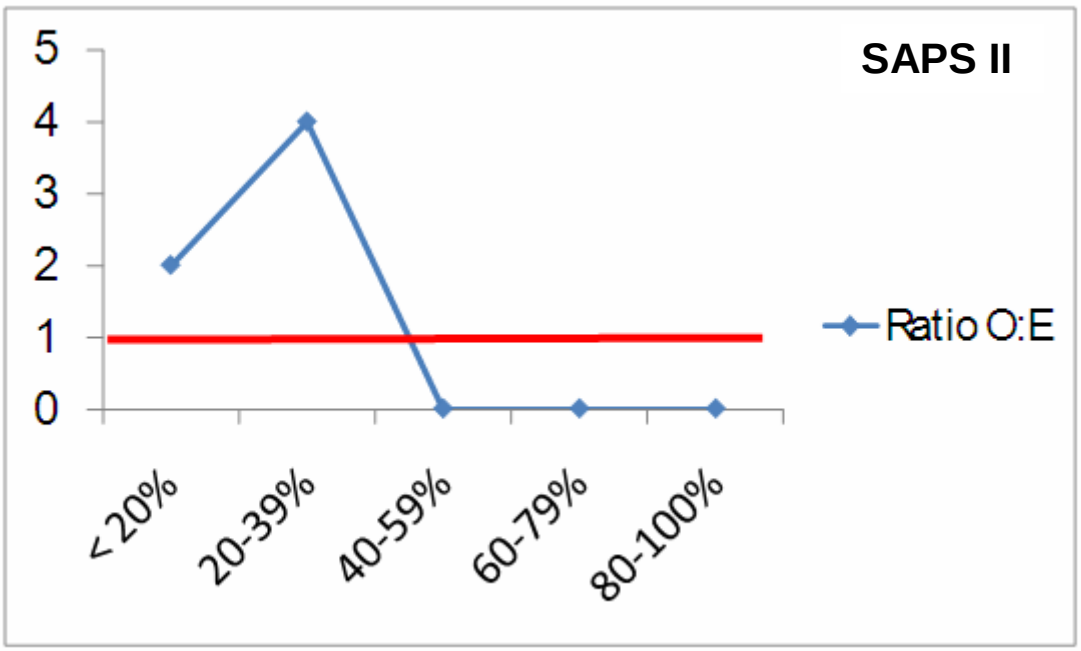
Gráfica 4.8. Ratio de mortalidad Observada:Esperada por intervalos de riesgo calculado escala P-POSSUM.

La gráfica 4.9 indica que las predicciones del sistema APACHE II son peores que las obtenidas en todos los intervalos de riesgo, excepto para los pacientes de riesgo más bajo (<20%) que se ajusta a lo obtenido en la realidad.

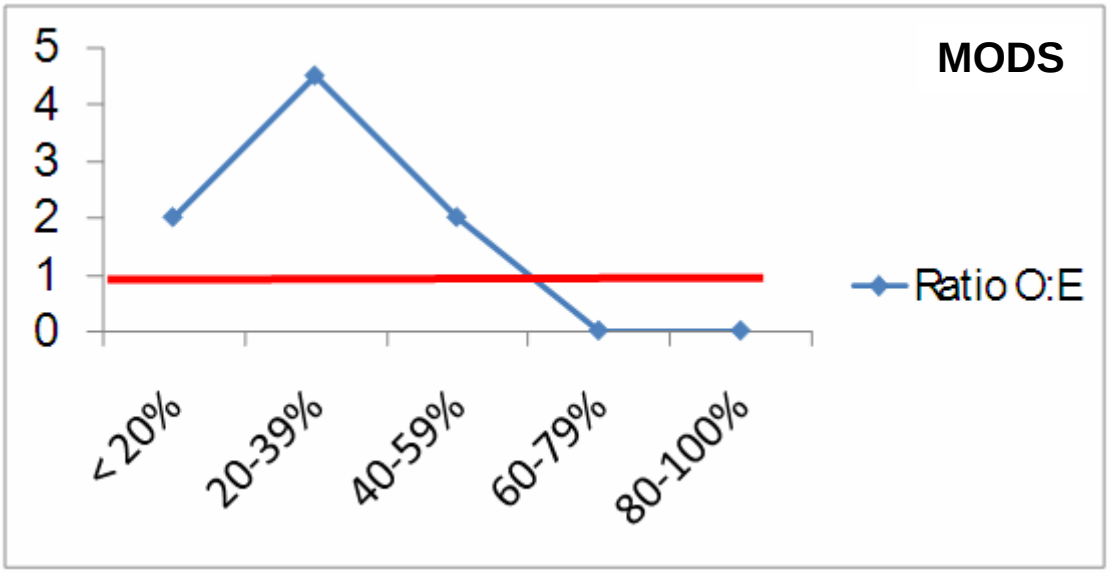


Gráfica 4.9. Ratio de mortalidad Observada:Esperada por intervalos de riesgo calculado para la escala APACHE II.

Las gráficas 4.10 y 4.11 son muy similares y representan ratios O:E muy por encima de lo obtenido en la realidad para las escalas SAPS II y MODS respectivamente, siendo las diferencias estadísticamente significativas.



Gráfica 4.10. Ratio de mortalidad Observada:Esperada por intervalos de riesgo calculado para la escala SAPS II.



Gráfica 4.11. Ratio de mortalidad Observada:Esperada por intervalos de riesgo calculado para la escala MODS.

4.3. ESTUDIO DE LA CAPACIDAD PREDICTIVA EN CUANTO A MORBILIDAD DE LAS ESCALAS DE RIESGO POSSUM Y P-POSSUM.

De las 6 escalas estudiadas, las únicas que permiten medir y predecir morbilidad, además de mortalidad, son la escala POSSUM y la P-POSSUM, por tanto, el estudio en cuanto a morbilidad lo realizaremos únicamente con estas dos escalas. Se estudiará la capacidad de ambas escalas de predecir morbilidad en comparación con las complicaciones reales observadas y recogidas en los pacientes a estudio.

En la tabla 4.14 se representan los resultados globales en cuanto a morbilidad calculados por el sistema POSSUM y P-POSSUM y los observados en la realidad. Como se observa en la tabla no hubo diferencias significativas en cuanto a la predicción de morbilidad de estas escalas y la observada en la realidad. Siendo el riesgo de morbilidad esperado (33,6%) similar al observado (38,5%).

Riesgo	Eventos Esperados	Eventos Reales	Ratio O/E	Riesgo Esperado	Riesgo Real	Significación Estadística
Morbilidad	129	148	1,14 (0,92-1,40)	33,6%	38,5%	N.S

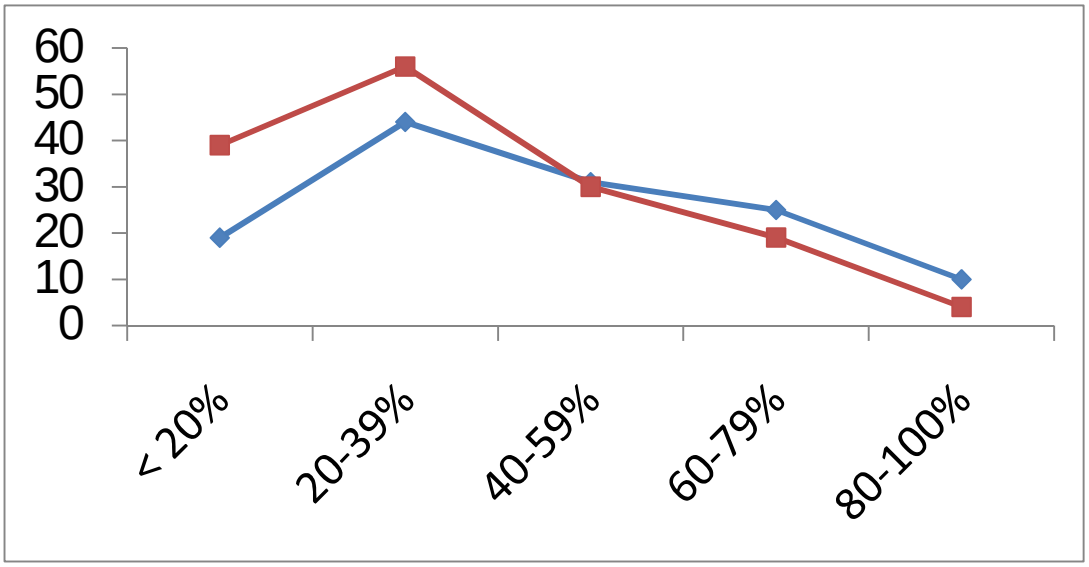
Tabla 4.14. Resultados obtenidos en cuanto a morbilidad con POSSUM y P-POSSUM. (Abreviaturas de la tabla: Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema POSSUM y P-POSSUM; N.S: No significativo considerando la significación estadística en p<0,05).

En la tabla 4.15 se representan los valores de morbilidad esperada y la ratio O:E por intervalos porcentuales de riesgo. Al analizar la morbilidad por grupos de riesgo atribuible por el sistema POSSUM y P-POSSUM, no se observaron diferencias significativas en ninguno de ellos.

Riesgo Complicaciones	Nº Intervenciones	Complicaciones Esperadas	Complicaciones Reales	Ratio O/E	Significación Estadística
<20%	115	19	39	2	N.S
20-40%	156	44	56	1,27	N.S
40-60%	64	31	30	0,96	N.S
60-80%	36	25	19	0,76	N.S
80-100%	13	10	4	0,4	N.S
Total	384	129	48	1,14	N.S

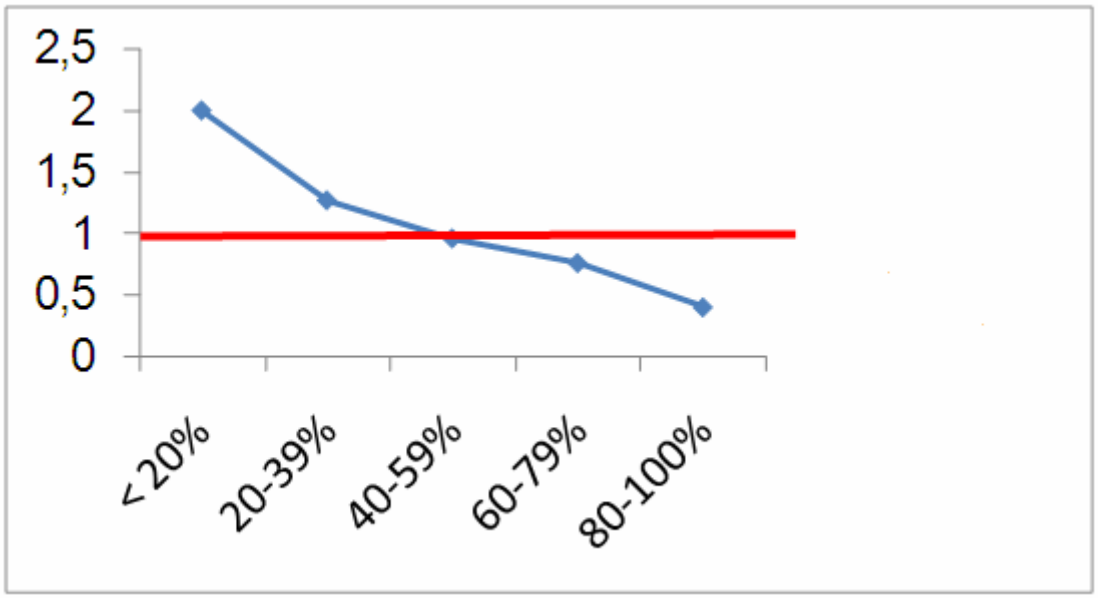
Tabla 4.15. Resultados obtenidos en cuanto a morbilidad con POSSUM y P-POSSUM desglosado por intervalos porcentuales de riesgo. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema POSSUM y P-POSSUM; N.S: No significativo considerando la significación estadística en p<0,05).

La gráfica 4.12 demuestra gráficamente que el sistema POSSUM y P-POSSUM permite predecir de forma muy exacta el riesgo de complicaciones tanto de forma global como por intervalos de riesgo.



Gráfica 4.12. Comparación gráfica entre morbilidad esperada (azul) y observada (roja) para las escalas POSSUM y P-POSSUM.

El estudio gráfico de la ratio observada:esperada para morbilidad con los sistemas POSSUM y P-POSSUM indica que estas escalas infraestiman la morbilidad, de forma no significativa, en los grupos de más bajo riesgo (< 20% y 20-39%), siendo adecuada la estimación para el resto de estratos (gráfica 4.13).



Gráfica 4.13. Ratio morbilidad Observada:Esperada para las escalas POSSUM y P-POSSUM.

En la tabla 4.16 se recoge la morbilidad observada y esperada por cirujano. Podemos observar que las tasas de morbilidad según el sistema POSSUM y P-POSSUM están dentro de lo considerado como óptimo en 11 de ellos (73,3%), mientras que en 3 (20%), se ha observado un ligero aumento de la morbilidad por encima de lo esperado, estadísticamente significativo. El cirujano 3 no ha podido ser estudiado de forma individual al no tener muestra suficiente de intervenciones y morbilidades observadas y esperadas.

Riesgo Complicaciones	Nº Intervenciones	Complicaciones Esperadas	Complicaciones Reales	Ratio O/E	Significación Estadística
Cirujano 1	19	3,6	3	0,8	N.S
Cirujano 2	63	21,3	19	0,9	N.S
Cirujano 3	6	1,7	0	0	-
Cirujano 4	20	8,2	17	2	P<0,05
Cirujano 5	24	5	5	1	N.S
Cirujano 6	52	22,3	40	1,8	P<0,01
Cirujano 7	14	3,5	6	1,7	N.S
Cirujano 8	50	12,7	9	0,7	N.S
Cirujano 9	30	11,1	10	0,9	N.S
Cirujano 10	28	8,1	6	0,7	N.S
Cirujano 11	29	14,6	13	0,9	N.S
Cirujano 12	14	3,5	8	2,3	P<0,01
Cirujano 13	14	4,8	5	1	N.S
Cirujano 14	11	3,5	3	0,8	N.S
Cirujano 15	10	5,1	4	0,8	N.S
Total	384	129	148	1,15	

Tabla 4.16. Morbilidad esperada y observada desglosada por cirujano según POSSUM y P-POSSUM. (Abreviaturas de la tabla: Nº: número; Ratio O/E: Relación entre los eventos Observados en la realidad y los Esperados por el sistema POSSUM y P-POSSUM; N.S: No significativo considerando la significación estadística en p<0,05).

5. DISCUSIÓN

5.1.SOBRE LA METODOLOGÍA

5.1.1. Tasas brutas, tasas ajustadas y sistemas de medición estándar.

Por un lado, las tasas brutas o crudas de mortalidad y morbilidad siguen siendo muy utilizadas como indicadores tanto en auditorías quirúrgicas como en la presentación de resultados en estudios publicados. Sin embargo, en ellas aparecen mezclados el comportamiento del fenómeno con otras variables que influyen de manera decisiva en el mismo y que pueden justificar la existencia de diferencias razonables. Así, debido a que no tienen en cuenta el ajuste según el riesgo por paciente, las tasas brutas no permiten la correcta valoración de resultados de diferentes modalidades o técnicas terapéuticas, ni la comparación entre cirujanos, unidades u hospitales, y su uso puede llevar a conclusiones erróneas o no precisas, para el fin que se busca, como recomendar tratamientos proponer o aconsejar el cierre de unidades o la interrupción de programas de formación[13]. Por otro, se está intentando extender el uso de tasas ajustadas al riesgo fisiológico y quirúrgico del paciente, además de a la suma de cuidados recibidos durante su proceso.

En la actualidad, no sólo se nos plantea la difícil tarea de “desterrar” el uso de las tasas brutas, sino que debemos tratar de “concienciar” a los profesionales sanitarios de la necesidad de utilizar sistemas estandarizados de medición de resultados en salud, y es en el concepto de estandarización donde surge el problema, ya que la existencia de una determinada herramienta publicada no es indicativo de que sea útil en todos los tipos de hospitales y pacientes. Beck DH[304] demostró que las diferencias en la casuística (case-mix), y una mayor prevalencia de eventos importantes (mayor tasa de mortalidad

observada que esperada) influían negativamente en la posibilidad de medir con exactitud los resultados asistenciales y realizar comparaciones entre instituciones, aunque se emplearan escalas ajustadas a riesgo. Por tanto, es fundamental la validación y re-calibración “in situ” de los modelos estandarizados antes de que puedan ser aplicados con confianza y precisión en nuevas poblaciones. Realizar este tipo de “personalización” de los modelos existentes es una estrategia muy importante para obtener información auténtica sobre la gravedad de las enfermedades, el plan de cuidados que estamos realizando y el estado de nuestra práctica clínica, siendo un requisito previo para las comparaciones fiables en cuanto a calidad de la asistencia y costes sanitarios.

Tener esto en cuenta es fundamental para entender, al menos en parte, la disparidad de resultados publicados hasta la actualidad en cuanto a la aplicación de escalas de riesgo, ya que hay mucha bibliografía sobre empleo de las mismas[210-215,217-219,236,248-251,258-267,284-290], pero muy poca que incluya un proceso de validación previo de la herramienta a utilizar[209,216,228,238,261,299,305].

Por todo lo anterior, la primera parte de nuestro estudio ha sido la validación en nuestro medio de las 6 escalas de riesgo a utilizar.

5.1.2. Proceso de validación de escalas.

Dicho proceso consiste en determinar si el instrumento que estamos utilizando “mide realmente lo que tiene que medir”. Dado que el proceso de validación sigue unos pasos muy concretos y establecidos, su aplicación debería ser obligada en todos los estudio de este tipo. El primer paso es la traducción de la escala, si ésta se encuentra descrita originariamente en otro idioma distinto al nuestro. Este apartado, no busca realizar una traducción literal del texto, sino definir el sentido conceptual que cada ítem encierra. Además, es importante que el proceso sea realizado, al menos, por dos revisores que dispongan de: 1) Conocimientos del idioma a traducir; 2) Desarrollen su actividad profesional en relación con el campo de estudio y en el ambiente donde se va a aplicar la escala; 3) Tengan conocimientos básicos sobre tests de medición en salud y escalas de riesgo. En nuestro estudio, a los dos revisores seleccionados para el mismo, se les impartió un curso sobre medición de resultados en salud y escalas de riesgo, previamente a la validación y aplicación de las escalas. Una vez traducida debe ser aplicada a un número suficiente de pacientes y de características similares a los que posteriormente se vaya a utilizar, para ello nosotros consideramos el uso de 10 pacientes por cada ítem de las escalas a utilizar[306].

El siguiente aspecto importante es evaluar si la escala será igualmente reproductible si es utilizada por observadores diferentes. En nuestro estudio empleamos el índice de Kappa, debido a que es un parámetro muy sencillo de calcular, ampliamente usado en la literatura con este fin[300-304] y que expresa de una manera muy clara el concepto que queremos medir ($k > 4$, implica

resultados aceptable o buenos en cuanto a reproductibilidad). Si los ítems que conforman la escala, tienen unos adecuados niveles de correlación entre ellos, es decir, si conforman una estructura "sólida" que le confiere cierta estabilidad al instrumento (*consistencia Interna* u *homogeneidad)* es otro parámetro muy importante a considerar, ya que, un instrumento puede dar mediciones semejantes en distintas situaciones (*k* alto), pero si está mal calibrado (consistencia interna baja) no nos será de ninguna utilidad; para evaluar este parámetro hemos empleado el test alfa de Cronbach. Al evaluar los resultados de estos coeficientes debe tenerse en cuenta que sus valores se afectan por el número de ítem en la escala; según esto, al aumentar el número de ítems del instrumento el valor del coeficiente alfa se incrementa artificialmente. Los valores que se recomiendan para estos índices son entre 0.7 y 0.9 (70% a 90%). Valores bajos sugieren que la escala es poco homogénea, que puede estar evaluando diferentes fenómenos y que no muestra consistencia ante diferentes condiciones de aplicación; valores mayores de 0.9 sugieren una estructura demasiado homogénea, en la cual probablemente existan ítems redundantes[307].

Aunque no es un parámetro obligatorio en los estudios de validación, nosotros recomendamos calcular de forma sistemática el tiempo medio de aplicación de la escala o instrumento a estudio. Siendo realistas, una herramienta que presente una validación muy buena en un medio concreto pero que para ser cumplimentada sea necesario mucho tiempo, nunca llegará a ser implantada de forma efectiva en el sistema y por tanto, aunque cumpla los estándares de calidad de "medir lo que realmente queremos medir" no será práctica en el trabajo diario y dejará de ser utilizada, por tanto, a igualdad de

calibración será mejor seleccionar aquella que se cumplimente de forma más rápida y sencilla. Incluso, ante dos escalas en las que una de ellas es mucho más larga de completar que la otra pero tiene una calibración ligeramente superior, probablemente sería mejor seleccionar la más simple, ya que la finalidad de todos estos procedimientos es que sea utilizados y permitan monitorizar la práctica clínica en el tiempo y no que sólo sirvan para realizar estudios de investigación tipo corte transversal y ya nunca más vuelvan a emplearse.

Una de las limitaciones más importantes a tener en cuenta a la hora de hacer la validación de las escalas es el tipo de diseño del estudio (retrospectivo o concurrente). Los estudios retrospectivos tienen como desventaja fundamental el partir de registros ya disponibles pero que no se crearon para el fin con el que los estamos utilizando y, por tanto, es posible que no dispongamos de datos necesarios para el estudio, siendo sus ventajas el ahorro de tiempo y dinero en su realización. Los estudios prospectivos evitan estas situaciones, pero tienen el inconveniente de su alto coste y mayor tiempo de realización. En nuestro caso, aunque el estudio de comparación ha sido prospectivo, el de validación de las distintas escalas fue retrospectivo por razones de tiempo y dinero, ya que de haber sido prospectivo la fase de validación habría durado más de un año (frente a un mes de duración con el diseño actual), con un incremento de costes asociados.

El cumplimiento de todos estos puntos debería ser incluido en todos los trabajos que tratan sobre validación de escalas, ya que son los que garantizan el correcto empleo de las mismas y la comparación de resultados entre estudios diferentes.

5.1.3. Precisión de las pruebas. Índice de Shannon (IS) y curvas ROC.

El IS está basado en la teoría de la información, de acuerdo con el valor de entropía (cantidad de información aportada por el modelo de ajuste de riesgo). El grado de precisión se calcula en cada paciente mediante la comparación entre el valor predicho por cada sistema de riesgo y el resultado observado de acuerdo con la fórmula comentada previamente; dándolo el valor 1 a la presencia del evento muerte y 0 a su ausencia. Una escala que otorga valores de riesgo muy bajos para los que sobreviven y probabilidades de muerte muy altas a los que fallecen tendrá índices de Shannon globales y para fallecidos con muy buena precisión y viceversa. Este índice presenta unas características especiales debidas a que el método adjudica el valor 0 al paciente que sobrevive y 1 al paciente que fallece, mientras que en forma ideal, todos las escalas de riesgo adjudican 0 ó un riesgo muy bajo (cercano a 0) al que sobrevive y un porcentaje que suele superar el 0,5 para el paciente con mayor riesgo de fallecer en el postoperatorio. Por tanto, una escala que adjudique un valor cercano a 1 a los pacientes fallecidos tendrá un *índice de Shannon para fallecidos* cercano a 1 (buena precisión para fallecidos), mientras que si da un valor cercano a 0 a los pacientes vivos y la proporción de fallecidos es baja (menos de un 10%), el índice de Shannon global estará próximo a 0. En este sentido el índice de Shannon es el mejor método para determinar la precisión de un sistema de riesgo para predecir mortalidad a nivel del paciente individual, mientras que el resto de modelos de análisis enfocan la predicción sobre poblaciones[308-310].

La toma de decisiones clínicas es un proceso extremadamente complejo en el que finalmente debe valorarse la utilidad de una determinada herramienta diagnóstica. En este contexto es imprescindible conocer detalladamente la exactitud de las pruebas diagnósticas, es decir, su capacidad para clasificar correctamente a los pacientes en categorías o estados en relación con la enfermedad, en nuestro caso, conocer si las escalas de riesgo evaluadas clasifican adecuadamente a los pacientes en cuanto a su probabilidad de desarrollar morbilidad y/o mortalidad. Generalmente la exactitud diagnóstica se expresa como sensibilidad (S) y especificidad (E) diagnósticas. Cuando se usa una prueba dicotómica (sus resultados se pueden interpretar directamente como positivos o negativos), la S es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba (fracción de verdaderos positivos). La E es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo (fracción de verdaderos negativos). Pero este enfoque sólo es válido para el caso de pruebas dicotómicas (2 opciones de resultado), por tanto, todas las pruebas diagnósticas cuyos resultados se miden en una escala continua o discreta nominal (como es el caso de las escalas de riesgo), quedan excluidas. Para solucionar esta situación se realiza la elección de distintos *niveles de decisión* o *valores de corte* que permitan una clasificación dicotómica de los valores de la prueba según sean superiores o inferiores al valor elegido. La diferencia esencial con el caso más simple es que ahora contamos no con un único par de valores de S y E que definan la exactitud de la prueba, sino más bien con un conjunto de pares correspondientes cada uno a un distinto nivel de decisión[311]. Este procedimiento constituye la esencia del

análisis ROC, una metodología que fue desarrollada en base a la Teoría de la Decisión en los años 50 y cuya aplicación fue motivada por problemas prácticos en la detección de señales por radar (aunque el detalle parezca anecdótico, la correspondencia entre el operador que interpreta los picos en la pantalla del radar para decidir sobre la presencia de un misil y el médico que emplea el resultado de una prueba diagnóstica para decidir sobre la condición clínica del paciente es muy buena)[312]. Su uso en medicina fue introducido por Swets y Pickett[313] para la interpretación de los resultados de pruebas radiológicas.

Para obtener la curva ROC de una determinada prueba diagnóstica, se calcula la S y E para cada uno de los diferentes valores observados en nuestros datos y se representan en una gráfica con la S en el eje de las ordenadas (eje Y), y [1-Especificidad] en el eje de abscisas (eje X). Una prueba que discrimine perfectamente tiene una curva que pasa a través de la esquina superior izquierda (punto más alto del eje y) donde la fracción de verdaderos positivos es de 1 ó 100% (S perfecta); y la fracción de falsos positivos es 0 (E perfecta), mientras que una prueba que no discrimine nada en absoluto, corresponde con la línea diagonal (a 45º) de los dos ejes. Entre las ventajas de la representación gráfica de la curva ROC destacan: 1) Es fácil de dibujar y de comprender la precisión de una prueba desde el punto de vista gráfico; 2) No requiere seleccionar un umbral de decisión particular porque es incluido todo el rango de posibles umbrales; 3) Es independiente de la prevalencia; 4) Proporciona una comparación visual directa entre las pruebas sobre una escala común. Sus desventajas son: 1) No se muestran los umbrales de decisión reales; 2) No se muestra el número de sujetos, y a medida que el tamaño de la

muestra decrece, la representación gráfica tiende a volverse progresivamente desigual; 3) Siempre requieren para su cálculo programas informáticos.

El cálculo del área bajo la curva (ABC) ROC soluciona todos estos problemas y es el mejor indicador global de la precisión de una prueba diagnóstica, ya que hace factible expresar el valor de una prueba diagnóstica mediante un número simple. Esta área es siempre mayor o igual a 0,5. El rango de valores se mueve entre 1 (discriminación perfecta) y 0,5 (no hay diferencias en la distribución de los valores de la prueba entre los 2 grupos). Viendo un ejemplo, comprenderemos su interpretación real: un área de 0,8 significa que un individuo seleccionado aleatoriamente del grupo de enfermos tiene un valor de la prueba mayor que uno seleccionado aleatoriamente del grupo de sanos en el 80% de las veces. Por todas estas razones, hemos considerado fundamental el cálculo del ABC en nuestro estudio con el fin de comparar de forma clara y sencilla la exactitud diagnóstica de cada una de las escalas de riesgo estudiadas[314,315].

5.2. SOBRE LOS RESULTADOS OBTENIDOS

5.2.1. Resultados de la validación

Todas las escalas (APACHE II, POSSUM, P-POSSUM; SAPS II y MODS), excepto MPM II, han obtenido valores muy altos tanto de consistencia interna como de confiabilidad interobservadores (con valores *K* por encima de 0,8), por tanto, estas 5 escalas en la versión que hemos validado miden lo que tienen que medir y además de forma reproductible con independencia del observador que la aplique. Por otro lado, tanto su traducción desde los originales en inglés, como la comprensión de los ítems por los distintos observadores no han planteado ningún problema como se recoge en los resultados. El tiempo medio de aplicación de cada una de ellas es muy similar y corto, lo que no debería plantear problemas con ninguna a la hora de ser llevadas a la práctica diaria (si los tiempos de implementación fueran elevados se correría el riesgo de no hacer práctica su utilidad real en la clínica diaria). Globalmente no hemos observado diferencias significativas en cuanto a la validación de estos 5 escores.

En cuanto a MPM II, no hemos podido obtener una validación óptima. Entre las razones para ello, consideramos que el hecho de que esta parte del estudio tenga un carácter retrospectivo a contribuido a ello, ya que la escala MPM II consta de algunos parámetros tanto de escasa recolección en informes retrospectivos como de difícil valoración (coma o estupor profundo, efecto masa intracraneal, resucitación cardiopulmonar previa) siendo el peso específico de éstos en el resultado final de su aplicación elevado. Además, el uso de coeficientes decimales para cada uno de los ítems a valorar en función

de su peso específico y de la aparición o no del evento puede dificultar el proceso de consenso y concordancia de los evaluadores. Mientras que en el resto de las escalas valoradas, la mayoría de parámetros suelen estar recogidos en las pruebas preoperatorias y /o en la valoración preanestésica y la puntuación de los diferentes ítems viene dada en números enteros. Debido a que no hemos podido conseguir una buena validación de esta última escala, no se ha incluido en el estudio comparativo y prospectivo de estimación de morbi-mortalidad.

Los estudios de validación de escalas son muy escasos[209,228,238,261,317], pero al compararlos con nuestro estudio podemos observar que, exceptuando el trabajo de Álvarez del Castillo M[317] en el que consigue una muy buena validación y aplicación del sistema MPM II en pacientes con traumatismo craneoencefálico, nuestros resultados son muy buenos, habiendo obtenido puntuaciones en cuanto a consistencia interna y confiabilidad interobservadores más elevadas que el total de trabajos. Entre las causas que consideramos para haber obtenido estos resultados consideramos el hecho de haber realizado un entrenamiento previo y formación de los revisores, además de redefinir, consensuar e incluir definiciones y tablas de ayuda en los protocolos utilizados para los ítems más problemáticos. Estas medidas las consideramos muy efectivas y contribuyen a reducir la variabilidad.

5.2.2. Resultados de la aplicación prospectiva en mortalidad.

Las escalas de riesgo han sido aplicadas en nuestro servicio de cirugía general a suficiente número de pacientes, desde el punto de vista estadístico, como para poder obtener conclusiones. Sólo se han incluido pacientes sometidos a cirugía programada y con indicación de cirugía abdominal, a través de laparotomía o laparoscopia. La variedad de pacientes en cuanto a las etiologías de los procesos indicativos de cirugía creemos que ha sido lo suficientemente amplia, tanto en número de casos, como en clases de procesos, como para que nuestro estudio haya podido demostrar la utilidad, o no, de las escalas de riesgo en un servicio de cirugía general en el que se interviene a todo tipo de pacientes y no sólo a pacientes con diagnósticos muy específicos, es decir, ¿es útil algún tipo de sistema de estratificación de riesgos para pacientes con procesos muy distintos en etiología y gravedad?.

Globalmente las escalas POSSUM y P-POSSUM no obtuvieron diferencias significativas en cuanto a estimar la mortalidad del estudio, mientras que APACHE II, MODS y SAPS II infraestimaron la mortalidad global de forma significativa. Este hecho creemos que se debe a 2 razones: 1) Estas tres últimas escalas sólo evalúan parámetros fisiológicos de forma específica y no variables intraoperatorias; 2) El estudio sólo incluye pacientes sometidos a cirugía programada. Estudios publicados demuestran que APACHE II[13,209,268-270], SAPS II[281-283,287,286] y MODS[277,297] son buenos estimadores del riesgo de mortalidad en pacientes críticos o sometidos a cirugía urgente, pero no existen trabajos que los evalúen solamente en pacientes programados. La explicación a este diferente comportamiento en pacientes urgentes o graves (buena

estimación de riesgo) frente a programados (mala estimación de riesgo) puede ser debida a que los pacientes que van a ser sometidos a intervenciones programadas mantienen sus parámetros fisiológicos (con fármacos o sin ellos) lo más cercanos a la normalidad, Por tanto, mientras que en el otro grupo de pacientes su fisiología está tan afectada que pueden llegar, incluso, a alcanzar valores extremos, por tanto, el hecho de no valorar hallazgos intraoperatorios es decisivo para los pacientes programados, ya que va a ser el factor diferencial principal entre distintos sujetos. Por otro lado, la circunstancia de que los sistemas POSSUM y P-POSSUM también consideren un incremento en el riesgo si el paciente recibe fármacos para las enfermedades cardiovasculares y respiratorias y los otros sistemas no, es otro factor que aumenta la sensibilidad de estas escalas en cuanto la predicción de riesgo se refiere.

5.2.3. Resultados por intervalos de riesgo de mortalidad para POSSUM.

Al igual que, globalmente, en el estudio por intervalos de riesgo no se han observado diferencias significativas entre lo calculado por POSSUM y lo observado en la realidad, estos hallazgos son compatibles con los publicados por Copeland[13,209] y, en general, por casi todos los grupos que han publicado resultados de POSSUM en cirugía general[216-221]. En el grupo de menor riesgo (<20%) no hemos observado una sobreestimación de la mortalidad al contrario de lo que observaron Prytherch et al[220], razón por la cual propusieron otra ecuación para el cálculo de la mortalidad (regresión lineal frente a la regresión logística propuesta por Copeland). Copeland[13] considera que, dentro de los rangos de riesgo normales en los que se suelen encontrar la mayoría de los

pacientes, si los métodos estadísticos son los adecuados, cualquiera de los dos tipos de regresión son igualmente útiles.

Todas las ratios O:E son iguales o están por debajo de 1 en nuestro estudio, sin diferencias significativas, lo que indica que la práctica clínica para con estos pacientes ha sido correcta. Incluso globalmente se ha observado una menor mortalidad de la esperada para este grupo de pacientes.

5.2.4. Resultados por intervalos de riesgo de mortalidad para P-POSSUM.

Los resultados obtenidos para P-POSSUM son similares a los de POSSUM, por tanto, todo lo que hemos expuesto en el apartado anterior es válido para éste. Los hallazgos comprobados en nuestro estudio apoyan las conclusiones sostenidas por Copeland[13,209].

5.2.5. Resultados por intervalos de riesgo de mortalidad para APACHE II.

Como ya comentamos, el APACHE II ha infraestimado la mortalidad de forma global, pero, en el estudio por intervalos de riesgo observamos que ha clasificado a más del 80% de los pacientes en el intervalo de bajo riesgo de mortalidad (<20%), sin que existan diferencias significativas en la mortalidad estimada y la observada en este intervalo. Para el resto de pacientes, APACHE II ha infraestimado la mortalidad en el grupo de riesgo intermedio de mortalidad (20-39%) y ha sido óptimo el cálculo en el de mayor riesgo de moratlidad (40-59%). Dado que de forma global APACHE II infraestima la mortalidad en este tipo de pacientes, el hecho de que en el grupo de bajo riesgo no lo haya hecho, va a favor de nuestro razonamiento, ya que, esto nos indica que este sistema ha incluido en este grupo a pacientes (riesgo de mortalidad <20%) tanto con

riesgo extremadamente bajo como pacientes con mayor riesgo real, y por tanto, la mortalidad de este grupo ha quedado equilibrada. Otro dato a favor de esta afirmación es el hecho de que en la mortalidad para el grupo de mayor riesgo (hasta 60%) tampoco se hayan observado diferencias significativas, ya que, siguiendo con el razonamiento, estos pacientes, fisiológicamente tienen una alteración tal que la gravedad de la operación no debe aportarles “mucho mayor riesgo calculado”. Evidentemente todas estas consideraciones necesitarán ser confirmados por estudios posteriores en condiciones similares.

5.2.6. Resultados por intervalos de riesgo de mortalidad para SAPS II.

Los resultados obtenidos para SAPS II en los diferentes intervalos de riesgo son similares al resultado del global de pacientes, así que no se pueden sacar nuevas conclusiones que no hayamos comentado previamente. Además, al no existir bibliografía en este tipo de enfermos (quirúrgicos programados), serán necesarios estudios para concluir si la escala SAPS II es o no válida para este tipo de pacientes.

5.2.7. Resultados por intervalos de riesgo de mortalidad para MODS.

Los mismos comentarios realizados para SAPS II son válidos para MODS, ya que, aunque en el grupo de riesgo de 40-59% no hay diferencias significativas entre lo observado y lo esperado, la muestra de pacientes es muy escasa (n=2) como para establecer algún tipo de conclusión.

5.2.8. Resultados en cuanto a las curvas ROC y áreas bajo la curva.

POSSUM y P-POSSUM han mostrado unas áreas bajo la curva muy elevadas (casi 0,8) sin que existan diferencias significativas entre ellas. Estos datos indican que su exactitud diagnóstica para predicción de morbi-mortalidad es muy alta y son compatibles a los observados por otros autores[13,209,220]. Por tanto, el uso indistinto de una u otra de ellas es muy recomendable en cualquier departamento quirúrgico previa validación de las mismas.

El resto de escalas han obtenido áreas bajo la curva muy bajas (próximas al valor inútil) con diferencias significativas con respecto a las dos primeras. Por tanto, no se recomienda su uso en nuestro tipo de pacientes, ya que, no nos permiten discriminar de forma más exacta que el simple azar.

5.2.9. Resultados para los índices de Shannon (IS).

Los resultados conseguidos con este índice son concordantes con los hallazgos previos de nuestro estudio, ya que, como POSSUM y P-POSSUM han estimado de forma muy fiable la mortalidad, sus respectivos índices de Shannon para fallecidos son muy próximos a 1 (muy buena capacidad de predecir fallecidos), no existiendo diferencias significativas entre ellos. Y como los otros 3 índices pronósticos no han predicho de forma aceptable los fallecidos y han infraestimado la mortalidad, sus índices de Shannon para fallecidos son bajos (no predicen bien la mortalidad), existiendo diferencias significativas entre ellos y POSSUM y P-POSSUM. No existe ningún estudio en la literatura que haya establecido el IS para estas escalas, así que no podemos establecer comparaciones ni obtener conclusiones desde el punto de vista global.

Los IS globales son muy bajos para todas las escalas, existiendo diferencias significativas entre los dos sistemas POSSUM (ligeramente superiores al resto de escalas) y los otros tres. Estos resultados tan bajos indicarían que ninguna de las escalas tiene buena capacidad de predecir supervivencia (no muerte). Lo que es cierto es que este índice es tan pequeño, fundamentalmente porque computa con un 1 a los eventos muerte y con un 0 a los eventos "no muerte", que, dado que el porcentaje global de muertes observadas es inferior al 10%, el cálculo de este índice será siempre bajo. Tampoco hay estudios publicados con estas escalas que calculen este índice.

5.2.10. Resultados de la aplicación prospectiva en morbilidad.

Para el estudio de la morbilidad solamente hemos considerado las escalas POSSUM y P-POSSUM, porque; 1) La fórmula para el cálculo de morbilidad es la misma y 2) Aunque en la actualidad existe algún trabajo que tratan de demostrar la utilidad del APACHE II en la predicción de morbilidad[267] esta escala no fue validada originariamente para predecir morbilidad[257]. El resto de escalas estudiadas tampoco fueron desarrolladas para este propósito.

Los resultados de morbilidad esperados son similares a los observados en la realidad sin que existan diferencias significativas. Estos datos son compatibles con los publicados en la literatura[206-221]. Tampoco existen diferencias significativas en cuanto a morbilidad por intervalos de riesgo.

En cuanto a las ratios O:E podemos ver que globalmente están dentro de lo que se considera buena práctica (en torno a 1 sin diferencias significativas con lo observado). Por estratos de riesgo se observa un aumento de las complicaciones observadas frente a las esperadas en dos grupos de más bajo riesgo (<20% y <40%), lo que indicaría que el estado de la práctica clínica en estos dos estratos está por debajo de los estándares de calidad recomendados. Por tanto, consideramos que este tipo de estudios permiten: 1) Conocer el estado de la práctica clínica; 2) Si se detecta un deterioro de la práctica clínica, pueden establecerse sistemas para detectar los puntos donde ha fallado y por tanto cómo resolverlo.

En cuanto a los resultados desglosados por cirujano podemos observar que globalmente el estado de la práctica clínica es adecuado a los estándares, y que más del 80% de los cirujanos del servicio en el momento del desarrollo del estudio presentaba una práctica clínica adecuada con el estándar, incluso por encima del mismo. En 3 de ellos se observó un estándar bajo en su práctica clínica, que habrá que seguir y estudiar para valorar sus posibles causas.

El hecho de poder predecir morbilidad es una de las ventajas del sistema POSSUM frente al APACHE II, además de su simplicidad de recolección de ítems, ya que sólo requiere datos clínicos y de laboratorio simples. Aunque POSSUM y P-POSSUM se elaboraron para auditar poblaciones y no para predecir evoluciones individuales[13], ya existe algún estudio que demuestra esta capacidad predictiva individual[318]. Nosotros no hemos analizado la morbi-mortalidad de forma individual, pero consideramos que podría ser objeto de nuevos estudios por nuestra parte.

5.3. SOBRE LAS CARACTERÍSTICAS DE LA ESCALA DE RIESGO IDEAL.

Entre los factores fundamentales que debe cumplir una escala de riesgo quirúrgico ideal podemos considerar los siguientes:

1.- Capaz de medir tanto morbilidad como mortalidad. La mayoría de escalas existentes se desarrollaron para medir una u otra, pero no las dos a la vez. POSSUM y P-POSSUM son de las pocas escalas que sí cumplen esta premisa.

2.- Aplicable a una gran variedad de intervenciones. Este punto es muy importante por dos razones: 1) En los países que no disponen de hospitales dedicados exclusivamente a un tipo de patología, como ocurre en Europa, de no cumplirse este propósito sería necesario el uso de diferentes sistemas de medición dentro de un mismo departamento, lo que dificultaría el manejo y cumplimentación de las mismas (diferentes escalas requieren información de diferentes ítems) y 2) El uso de diferentes escalas no permitiría comparar la actividad clínica entre diferentes departamentos, incluso entre los miembros de un mismo servicio (por ejemplo, no se podrían comparar, de forma directa, los resultados de la unidad de coloproctología con los de obesidad mórbida o hepatobiliar, ya que usarían sistemas diferentes). De las 6 escalas valoradas en este estudio, solamente POSSUM y P-POSSUM fueron concebidas desde su creación con este propósito[209,220].

3.- Fácil de cumplimentar en tiempo y en ítems. Que las variables necesarias para su uso sean de determinación común y que la cumplimentación no requiera mucho tiempo es decisivo para que sea elementos útiles y aplicables a la realidad clínica. Todas las escalas valoradas

en nuestro trabajo cumplen estos requisitos, según hemos podido demostrar. En la actualidad, en un intento de mejorar la cumplimentación de datos sin reducir la precisión diagnóstica de las escalas, Prytherch et al[319] (el grupo que definió el P-POSSUM), por un lado, y Sutton et al[320], por otro, han propuesto modelos basados en el menor número posible de variables relevantes, obteniendo resultados buenos y comparables a los de POSSUM y P-POSSUM[321]. Serán necesarios más trabajos para valorar realmente la utilidad de estas nuevas escalas.

4.- Aplicable a pacientes médicos y quirúrgicos. El disponer de una herramienta que fuera aplicable a todos los pacientes ingresados en un hospital de forma precisa sería fundamental, ya que podría estar incluida en los servicios informáticos centrales y ser calculada de forma casi automática para todos los pacientes, con la consiguiente ventaja de disponer en cualquier momento de los datos sobre el estado de la práctica clínica global y desglosada por tipo de especialidad. Lamentablemente esto todavía no es una realidad, ya que es difícil calibrar una herramienta de forma precisa para un espectro tan amplio y variado de pacientes, aunque ya se ha publicado algún trabajo relacionado con esto[319].

5.- Aplicable a pacientes individuales. La mayoría de sistemas que estratifican riesgo fueron elaborados para medir poblaciones y no el riesgo individual de los pacientes. Es cierto que se puede extrapolar a pacientes individuales[318] pero sería deseable que el diseño de las escalas lo tuviera en cuenta.

6.- Que estimara el riesgo antes de actuar sobre el paciente. Este punto está muy relacionado con el anterior, ya que, por un lado, no han sido diseñadas para estimar riesgo de forma individual y, por otro, en el caso de las escalas que recogen variables intraoperatorias, como POSSUM y P-POSSUM, no disponemos de la estimación precisa hasta que hemos actuado sobre el paciente. Probablemente, esta sea la crítica más extendida en cuanto a las escalas de riesgo se refiere, ya que ninguna de ellas es capaz de permitir descartar realizar maniobras sobre los pacientes que realmente no se van a beneficiar de ellas, sino que lo único que les va a suponer es un "encarnizamiento terapéutico".

7.- Que sea validada en cada medio previamente a su uso. Si no se realiza una validación y recalibración de las escalas antes de ser utilizadas en un sistema determinado no nos servirán para mejorar nuestra práctica clínica, sino que nos aportarán información sesgada y poco realista, igual que lo hacen las tasas brutas pero con más trabajo futil realizado por nuestra parte.

6. CONCLUSIONES

CONCLUSIONES

1. Las tasas brutas de mortalidad y morbilidad no son indicadores fiables ni válidos para monitorizar y evaluar la actividad asistencial de un servicio, departamento u hospital.

2. Las escalas*:* Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM); Portmouth-Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (P-POSSUM); Acute Physiology and Chronic Health Evaluation (APACHE II); Simplified Acute Physiology Score (SAPS II) y Multiple Organ Dysfunction Score (MODS) han sido validadas, para pacientes sometidos a cirugía programada mediante laparotomía o laparoscopia, de forma óptima en el Hospital General Universitario “JM Morales Meseguer” de Murcia.

3. La escala Mortality Prediction Model (MPM II) no ha podido ser validada aceptablemente en los pacientes quirúrgicos programados (laparotomías y laparoscopias) del HGU “JM Morales Meseguer”.

4. Los escores APACHE II, SAPS II y MODS no presentan buena fiabilidad en su aplicación para predecir mortalidad en los pacientes quirúrgicos programados.

5. Los sistemas POSSUM y P-POSSUM tienen una elevada fiabilidad en su aplicación para medir riesgo de mortalidad y morbilidad en los pacientes estudiados.

6. El POSSUM y P-POSSUM han demostrado alta reproductibilidad en su uso en este tipo de pacientes.

7. No se han observado diferencias significativas ni en cuanto a mortalidad ni morbilidad observada en la realidad y la esperada por las escalas POSSUM y P-POSSUM.

8. La mortalidad observada en los pacientes estudiados está por encima de los estándares de calidad exigidos, tanto globalmente como por intervalos de riesgo.

9. La morbilidad observada está ligeramente por debajo de los estándares de calidad aceptados.

10. En contra de lo previsto, la morbilidad de los pacientes con menor riesgo quirúrgico (0-39%) está por encima de lo aceptable. Mientras que la morbilidad de los pacientes con mayor riesgo quirúrgico (40-100%) está por debajo de lo exigido por los estándares de calidad.

11. La morbilidad por cirujano está dentro de los estándares de calidad en más del 80%.

12. El uso del sistema POSSUM o P-POSSUM es altamente recomendable en los servicios de cirugía para monitorizar y detectar errores en la práctica clínica.

7. BIBLIOGRAFÍA

BIBLIOGRAFÍA

1.- Namkee A, Meseguer JA. Gasto sanitario y envejecimiento de la población en España. Herce San Miguel. Bilbao. Fundación BBVA, D.L. 2003.

2.- Jencks SF, Williams DK, Kay TL. Assesing hospital-associated deaths from discharge data: the role of length of stay and comorbidities. JAMA 1988; 260:2240-2246.

3.- Lezzoni LI, Shwartz M, Ash AS. Using severity-adjusted stroke mortality rates to judge hospitals: how severity is measured may affect perceptions of hospital performance. Int J Qual Health Care 1995;7:81-94.

4.- Shwartz M, Lezzoni LI, Moskowitz MA. The importance of comorbidities in explaining differences in patient costs. Med Care 1996;34:767-782.

5.- Dimick JB, Chen SL, Taheri PA. Hospital costs associated with surgical complications: a report from the private-sector Nacional Surgical Quality Improvement Program. J Am Coll Surg 2004;4:531-537.

6.-- Kalish RL, Daley J, Duncan CC. Costs of potential complications of care for major surgery patients. Am J Med Qual 1995;10:48-54.

7.- Dimick JB, Pronovost PJ, Cowan JA. Complications and costs after high-risk surgery: where should we focus quality improvement initiatives? J Am Coll Surg 2003;1996:671-678.

8.- Davenport DL, Henderson WG, Khuri SF, Mentzer RM. Preoperative risk factor and surgical complexity are more predictive of costs than postoperative complications. Ann Surg 2005;242:463-471.

9.- Ahicart C. Técnicas de medición del case-mix hospitalario. Los procesos productivos en el hospital y la medición del producto sanitario. Hospital 2000 1998;Supl. 1:4-22.

10.- Ahicart C. Técnicas de medición del case-mix hospitalario II. Diagnostic Groups y Grupos Relacionados con el Diagnóstico. Hospital 2000 1998;Supl. 1:3-22.

11.- Ahicart C. Técnicas de medición del case-mix hospitalario III. AS-Score, Patient Severityof Iones, APACHE, Staging Disease, Patient Management Categories. Hospital 2000 1998;Supl. 1:3-22.

12.- Bonfill X, Gispert R. La mortalidad evitable: la eterna esperanza blanca para estudiar y comparar la efectividad hospitalaria. Gac Sanit 1995;9-14.

13.- Copeland GP. The POSSUM system of surgical audit. Arch Surg 2002:137:15-9.

14.- Shuhaiber JH. Augmented reality in surgery. Arch Surg 2004: 139:170-4.

15.- Saturno PJ, Quintana O, Varo J. ¿Qué es calidad? En: Saturno PJ, Gascón JJ, Parra P (Eds). Tratado de Calidad Asistencial en Atención Primaria. Barcelona. Du Pont Pharma, 1996;19-45.

16.- Lorenzo Martínez S. Sistemas de gestión de la calidad. En: Ruiz P, Alcalde J; Landa J (Eds). Gestión Clínica en Cirugía. Madrid, Arán Ediciones SL, 2005; 115-139.

17.- Lee RI, Jones LW. The Fundamentals of good medical care. Chicago: University of Chicago Press, 1933.

18.- Institute of Medicine. Advancing the quality of health care: A policy statement. Washington DC: National Academy of Sciences, 1974.

19.- Brook RH, Kosecoff JB. Commentary. Competition and quality. Health aff; 2000:7:150-161.

20.- Institute of Medicine. A strategy for Quality Assurance; Sources and Methods. Vol. II. Washington DC: National Academy Press, 1990.

21.- Join Commission on Accreditation of Health Care Organizations. Quality assurance in ambulatory care. Chicago. 2 ed. JCAHO, 1990.

22.- Join Commission on Accreditation of Health Care Organizations. Guide to Quality Assurance. Chicago. 2 ed. JCAHO, 1988.

23.- Palmer RH. Evaluación de la asistencia ambulatoria: Principios y práctica. Madrid: Ministerio de Sanidad y Consumo, 1990.

24.- Donabedian A. Explorations in Quality Asessment and Monitoring: The definition of Quality and Approaches to its Asessment Vol. 1 Ann Arbor (Michigan): Health Administration Press, 1980.

25.- World Health Organization. Regional Office for Europe. The principles of quality assurance, report on a WHO meeting. Euro Reports and Studies Series: nº 94. Copenhagen: WHO, 1985.

26.- Esselstyn CB. Principles of physician remuneration. En: Papers and proceedings of the national conference on Labor Health Services. Washington DC: American Labor Health Association, 1958.

27.- Palmer RH. Considerations in defining Quality of Health care. En Palmer RH, Donabedian A, Povar GJ. Striving for quality in health care: An inquiry into policy practice. Ann Abor (Michigan), Health Administration Press.1991:pags:1-58.

28.- Saturno P, Imperatori E, Corbella A. Características del Programa Ibérico de formación. Estrategia para la introducción de actividades de garantía de calidad. En: Evaluación de la calidad asistencial en atención primaria. Experiencias en el marco de la cooperación ibérica. Madrid: Ministerio de Sanidad y Consumo, 1990.

29.- Saturno P. La definición de la calidad de la atención. En: Marquet R (Eds). Monografías clínicas en atención primaria: Garantía de Calidad en APS. Barcelona: Doyma, 1993: 7-18.

30.- Office of Technology Assessment. Assessing the efficacy and safety of medical Technologies. Washington DC: Government Printing Office, 1978.

31.- Nutting P, Burkhalter BR, Carney LP, Gallagher KM. Métodos de evaluación de la calidad en Atención Primaria: Guía para clínicos. Barcelona:S.G. (Eds). 1991.

32.- Wyszewianski L. Quality of care: Past achievements and future challenges. Inquiry 1988; 25:13-22.

33.- Brook RH, Lohr KN. Efficacy, effectiveness, variations, and quality. Boundary-crossing research. Med Care 1985; 7:10-22.

34.- Copeland GP. Comparative audit: fact versus fantasy. Br J Surg 1993:80:1424-1425.

35.- Hopkins A. Measuring the quality of medical care. London: Royal College of Physicians of London, 1990.

36.- Saturno P, Imperatori E, Corbella A. Introducción al concepto y dimensiones de calidad asistencial. Como empezar. En: Evaluación de la calidad asistencial en atención primaria. Experiencias en el marco de la cooperación ibérica. Madrid: Ministerio de Sanidad y Consumo, 1990.

37.- Shortell SM. Continuity of medical care: conceptualization and measurement. Med Care, 1976; 14:377-91.

38.- Saturno PJ, Parra P. Programa de Evaluación y Mejora de la Calidad Asistencial (EMCA) en la Región de Murcia: Estrategias y actividades. En: Libro de Ponencias y Comunicaciones de la I Conferencia Iberoamericana sobre Docencia y Calidad en los Servicios de Salud. Murcia, 1995: 133-140.

39.- Pozo F, Ricoy JR, Lázaro P. Una estrategia de investigación en el Sistema Nacional de Salud: I. La epidemiología clínica. Med Clin 1994; 102:664-669.

40.- Dombal FT. Toma de decisiones: análisis formal versus intuición clínica. En: Dombal FT (Ed). El proceso que conduce a la toma de decisiones en cirugía. Barcelona, Ediciones Científicas y Técnicas, SA. 1994, pags 45-57.

41.- Alberquilla S, González C. Sistemas de información y medida del producto sanitario. En: Ruiz P, Alcalde J; Landa J (Eds). Gestión Clínica en Cirugía. Madrid, Arán Ediciones SL, 2005, pags 43-89.

42.- Berwick DM. Controlling variation in health care: a consultation from Walter Shewhart. Med Care 1991;29:1212-1225.

43.- Villeta Plaza R, Landa García JI. Cirugía basada en la evidencia. En: Ruiz P, Alcalde J; Landa J (Eds). Gestión Clínica en Cirugía. Madrid, Arán Ediciones SL, 2005, pags 503-543.

44.- Black N, Glickman ME, Ding J, Flood AB. International variation in intervention rates. What are the implications for patients selection? Int J Technol Assess Health Care 1995; 11:719-732.

45.- Westerling R. Can regional variation in avoidable mortality be explained by deaths outside hospital? A study from Sweden, 1987-90. J Epidemiol Community Health 1996;50:326-333.

46.- Lewis CE. Variations in the incidence of surgery. N Engl J Med 1969; 281:880-884.

47.- Ruiz I, Hernández-Aguado I, Garrido P. Variation in surgical rates. A population study. Med Care 1998;36:1315-1323.

48.- Wolfe RA, Griffith JR, Mcmahon LF, Tedeschi PJ. Patterns of surgical and nonsurgical hospital use in Michigan communities from 1980 through 1984. Health Serv Res 1989;24:66-82.

49.-Kuh D, Stirling S. Socioeconomic variation in admission for disease of female genital system and breast in a national cohort aged 15-43. BMJ 1995;311:840-843.

50.- Legorreta AP, Silber JH, Costantino GN, Kobilinski RW, Zatz SL. Increased cholecystectomy rate after the introduction of laparoscopic cholecstectomy. JAMA 1993;270:1429-1432.

51.- Wennberg JE, Mulley AG, Hanley D, Timothy RP, Fowler FJ, Roos NP, Barri M. An assessment of prostatectomy for benign urinary tract obstruction. Geographic variations and the evaluation of medical care outcomes. JAMA 1988;259:3027-3030.

52.- Payne N, Saul C. Variations in use of cardiology services in a health authority: comparison of coronary artery revascularization rates with prevalence of angina and coronary mortality. BMJ 1997;314:257-261.

53.- Selby JV, Fireman BH, Lundstrom UJ, Swain BE, Truman AF, Wong CC, Froelicher ES, Barron HV, Hlatky MA. Variation among hospital in coronary-angiography practices and outcomes after myocardial infarction in a large health maintenance organization. N Engl J Med 1996;335:1888-1896.

54.- Vader JP, Burnand B, Froehlich F, Dupriez K, Larequi-Lauber T, Pache I. Appropriateness of upper gastrointestinal endoscopy:comparison of American and Swiss criteria. Int J Qual Health Care 1997;9:87-92.

55.- Cowper PA, DeLong ER, Peterson ED, Lipscomb J, Muhlbaier LH, Jollis JG. Geographic variation in resource use for coronary artery bypass surgery. Med Care 1997;35:320-333.

56.- Bengtson A, Herlitz J, Karlsson T, Brandrup-Wognsen G, Hjalmarson A. The appropriateness of performing coronary angiography and coronary artery revascularization in a Swedish population. Jama 1994;271:1260-1265.

57.- McGlynn EA, Naylor CD, Anderson GM, Leape LL. Comparison of the appropriateness of coronay angiography and coronary artery bypass graft surgery between Canada and New York State. JAMA 1994;272:934-940.

58.- Pipel D, Fraser GM, Kosecoff J, Weitzman S, Brook RH. Regional differences in appropriateness of cholecystectomy in a prepaid health insurance system. Public Health Rev 1993;20:61-74.

59.- Kahn KL, Kosecoff J, Chassin MR, Solomon DH, Brook RH. The use and misuse of upper gastrointestinal endoscopy. Ann Intern Med 1988;109:664-670.

60.- Marión J, Peiró S, Márquez S, Meneu R. Variaciones en la práctica clínica: importancia, causas e implicaciones. Med Clin 1998;110:382-390.

61.- Gascón JJ, Marión J, Peiró S. La variabilidad en la práctica clínica. En: Saturno PJ, Gascón JJ, Parra P (Eds). Tratado de calidad asistencial en Atención Primaria. Barcelona. Du Pont Pharma 1996;117-148.

62.- Walter D, Williams P, Tawn J. Audit of requests for preoperative chest radiography. BMJ 1994;309:772-773.

63.- Adam PA, Van Der Wouden JC, Van Der Does E. Influencing behavior of physicians ordering laboratory tests: a literatura study. Med Care 1993;31:784-794.

64.- Gortmaker SL, Bickford AF, Mathewson HO, Dumbaugh D, Tirrell PC. A successful experiment to reduce unnecessary laboratory use in community hospital. Med Care 1998;26:631-642.

65.- Winkens RAG, Pop P, Grol RPTM, Baugter-Maesen AMA, Kester ADM. Effects of routine individual feedback over nine years on general practitioners´ requests for tests. BMJ 1996;312:490.

66.- Kerr MP. Antidepressant prescribing: a comparison between general practitioners and psychiatrists. Br J Gen Pract 1994;44:275-276.

67.- Wilson RPH, Hatcher J, Barton S, Walley T. Influences of practice characteristics on prescribing in fundholding and non-fundholding general practices: and observational study. BMJ 1996;313:595-599.

68.- Gonzáles R, Steiner JF, Sande MA. Antibiotic prescribing for adults with colds, upper respiratory tract infections, and bronchitis by ambulatory care physicians. JAMA 1997;278:901-904.

69.- Buetow SA, Sibbald B, Cantril JA, Halliwell S. Prevalence of potentially inappropriate long term prescribing in general practice in the United Kingdom, 1980-95; systematic literature review. BMJ 1996;313:1371-1374.

70.- Cocburn J, Pit S. Prescribing behaviour in clinical practice: patients' expectations and doctors' perceptions of patients' expectations-a questionnaire study. BMJ 1997;315:520-523.

71.- Campillo-Soto A, Soria-Aledo V, Flores-Pastor B, Aguayo-Albasini. Ventajas del pase de visita sistemático los fines de semana. Med clínc (Barc) 2006; 127.555-7.

72.- Bertakis K, Jay Helms L, Callahan E, Azari R, Robbins J. The influence of gener on physician practice style. Med Care 1995;33:407-16.

73.- Delgado A, López-Fernández LA, de Dios J. Influence of the doctor's gender in the satisfaction of the users. Med Care 1993; 31:795-800.

74.- Sonke GS, Beaglehole R, Stewart AW, Jackson R, Stewart FM. Sex differences in case fatality before and after admission to hospital after acute cardiac events: analysis of community based coronary heart disease register. BMJ 1996;313:853-855.

75.- Pearson TA, Myerson M. Treatment of hypercholesterolemia in women. Equality, effectiveness, and extrapolation of evidence. JAMA 1997;277:1320-1321.

76.- Mustard CA, Kaufert P, Kozyrskyj A, Mayer T. Sex differences in the use of health care services. N Engl J Med 1998;338:1678-1683.

77.- Blustein J, Witzman BC. Access to hospitals with high-technology cardiac services: how is race important? Am J Public Health 1995;85:345-351.

78.- Gurwitz JH, Goldberg RJ. Coronary thrombolysis for the elderly. Is clinical practice really lagging behind evidence of benefit? JAMA 1997;277:1723-1724.

79.- Krumholz HM, Murillo JE, Chen J, Vaccarino V, Radford MJ, Ellerbeck EF, Wang Y. Thrombolytic therapy for eligible elderly patients with acute myocardial infarctation. JAMA 1997;1683-1688.

80.- Weingarten JP, Clay JC, Herckert A. Impact of socioeconomic status on health care utilization: factors influencing length of stay. Hosp Health Serv Adm 1997:385-409.

81.- Kahan JP, Park RE, Leape LL, Berstein SJ, Hilborne LH, Parker L. Variations by specialty in physician ratings of the appropriateness and necessity of indications for procedures. Med Care 1996;34:403-415.

82.- Rosenblatt RA, Dobie SA, Hart LG. Scheeweiss R, Gould D, Raine TR. Interspecialty differences in the obstetric care of low-risk women. Am J Public Health 1997;87:334-351.

83.- Britt H, Bhasale A, Miles DA, Meza A, Sayer GP, Angelis M. The sex of general practitioner. A comparison of characteristics, patients and medical condition managed. Med Care 1996;34:403-415.

84.- Kitzinger J. Variations in medical attitudes to postoperative recovery period. BMJ 1995;311:296.

85.- Sullivan F, Mitchell E. Has general practitioner computing made a difference to patient care? A systematic review of published reports. BMJ 1995;311:848-852.

86.- McLeod PJ, Tamblyn RM, Gayton D, Grad R, Snell L, Berkson L. Use of standardized patients to assess between-physician variations in resource utilization. JAMA 1997;278:1164-1168.

87.- Rafferty T, Wilson-Davis K, McGavoch H. How has fundholding in Northern Ireland affected prescribing patterns? A longitudinal study. BMJ 1997;315:166-170.

88.- Retchin SM, Brown RS, Yeh SCH, Chu D, Moreno L. Outcomes of stroke patients in medicare fee for service and managed care. Jama 1997;278:119-124.

89.- Webster JR, Feinglass J. Stroke patients, managed care, and distributive justice. JAMA 1997;278:161-162.

90.- Chan L, Koepsell TD, Deyo RA, Esselman PC, Haselkorn JK, Stolov WC. The effect of medicare's payment system for rehabilitation hospitals on length of stay, charges, and total payments. N Engl J Med 1997;337:978-985.

91.- Thomas DR, Davis KM. Physician awareness of cost under prospective reimbursement systems. Med Care 1987;25:181-184.

92.- Manning WG, Leibowitz A, Goldberg GA, Rogers WH, Newhouse JP. A controlled trial of the effect a prepaid group practice on use of services. N Engl J Med 1984;310:1505-1510.

93.- Newacheck PW, Stoddard JJ, Hughes DC, Pearl M. Health insurance and access to primary care for children. N Engl J Med 1998;338:513-519.

94.- Jané Camacho E, Barba Albós G, Salvador Vilala X, Salas Ibáñez T, Sánchez Ruiz E, Bustins Poblet M. Variaciones de la tasa de hospitalización por procedimientos quirúrgicos seleccionados. Aplicación del análisis de áreas pequeñas. Gac Sanit 1996;10:211-219.

95.- Marqués JA, Peiró S, Medrano J, Librero J, Pérez-Vázquez MT, Aranaz J, Ordeiñana R. Variabilidad en las tasas de intervenciones de cirugía general por áreas de salud. Cir Esp 1998;63:445-453.

96.- Sarria A, Sendra JM. Evolución de la tasa de cesareas en España: 1984-8. Gac Sanit 1994;8:209-214.

97.- Latour-Pérez J, Gutiérrez Vicent T, López-Camps V, Bonastre-Mora J, Giner-Boix JS, Rodríguez-Serra M, Rosado-Bretón L. Diferencias de esfuerzo terapéutico en razón del nivel socioeconómico en pacientes con infarto agudo de miocardio. Gac Sanit 1995;9:5-10.

98.- Granados A, Escarrabill J, Borrás JM, Sánchez V, Jovell AJ. Utilización apropiada y efectividad: la oxigenoterapia crónica domiciliaria en Cataluña. Med Clin 1996;106:251-253.

99.- Jiménez RE, Gutiérrez AR, Fariñas H, Suárez N, Fuentes E. Variaciones del tiempo de estancia postoperatoria según las características de los pacientes en un servicio de cirugía general. Gac Sanit 1994;8:180-188.

100.- Simó J, Gaztambide M, Morote MV, Palazón G, Gálvez J, Salto ML. Utilización de la mamografía de cribado y sus determinantes demográficos y de riesgo entre mujeres de 25 a 65 años. Med Clin 1997;108:767-771.

101.- Cimas JE, Arce MC, González ME, López A. Atención Especializada y Atención Primaria en el tratamiento del asma: ¿Existen diferencias? Aten Primaria 1997;19:477-481.

102.- Gutiérrez A, Núñez E, Sanz JC, Martínez M. Adecuación de transfusiones en urgencias. Med Clin 1997;109:396.

103.- Ruiz MT, Ronda E. Atención Sanitaria según el sexo de los pacientes. Med Clin 1996;103:537-538.

104.- Ceniceros I, Gastaldo R, Cadabés A, Cebrián J. El sexo femenino es un factor pronóstico independientes de mortalidad en la fase aguda del infarto de miocardio. Med Clin 1997;109:171-174.

105.- Clúa JL, Piñol JL, Pipió JM, Queralt ML. ¿Influye el género del paciente en la calidad de las historias clínicas de Atención Primaria? Aten Primaria 1997;20:75-81.

106.- Segura A, Rajmil L. Hospitalización infantil y género. Aten Primaria 1997;20:108-110.

107.- Rivera J, García- Monforte A. Variación en el número de análisis y pruebas diagnósticas en una consulta externa. Gac Sanit 1994;8:310-316.

108.- Sanfélic Genovés J, Pereiró Berenguer I, Oterino de la Fuente D, Altarriba Cano ML, Monzó MJ, Pérez de los Cobos J. Pertinencia de las peticiones analíticas en Atención Primaria. Aten Primaria 1996;18:87-89.

109.- Vázquez I, Pérez A, Alcantarilla G, Arjona I. Influencia del formulario de petición sobre la demanda de analítica. Todo Hospital 1997;139:13-17.

110.- SánchezJL, Larrabe J, Óscar J, Tsakiridu DO, Ruiz R, Bilbao J, Sologuren A. Prescripción de antiinflamatorios no esteroides y gastroprotectores. Adecuación a criterios de calidad en atención primaria. Aten Primaria 1997;20:127-132.

111.- Vilaseca J, Buxeda C, Cámara C, Flor F, Pérez R, Sánchez M. ¿Tienes riesgo coronario los pacientes que tratamos con fármacos hipolipemiantes? Aten Primaria 1997;20:49-53.

112.- Kanterewicz E, Iruela A; Pladevall M, Serrarols M, Pañella D, Brugués J, Diez A. Estudio de las prescripciones de calcitonina: estimación del gasto pro prescripción inadecuada. Med Clin 1998;110:411-415.

113.- Delgado Villa R. La variabilidad d la práctica clínica. Rev Calidad Asistencial 1996;11:177-183.

114.- OMS [ORGANIZACION MUNDIAL DE LA SALUD] (2000):Health systems: improving performance, The World Health Report 2000.

115.- Glover JA. The incidence of tonsillectomy in school children. Proc Royal Society Med 1938;31:1219-1236.

116.- Wennberg J, Gittelsohn A. Variations in medical care among small areas. Sci Am 1982;264:100-111.

117.- Chassin MR, Lisecoff J, Park RE, Winslow CM, Kahn KL, Merrick NJ. Does inappropriate use explain geographic variations in the use on health care services? A study of three procedures. JAMA 1987;258:2533-2537.

118.- McPherson K, Strong PM, Epstein A, Jones L. Regional variations in the use of common surgical procedures: within and between England and Wales, Canada and the United States of America. N Engl J Med 1969;273-288.

119.- McPherson K, Wennberg JE, Hovind OB, Clifford P. Small-area variations in the use of common surgical procedures: an international comparison of New England, England and Norway. N Engl J Med 1982;307:1310-1314.

120.- Sarría Santamera A, Sendra Gutierrez JM. Diferencias regionales en la utilización hospitalaria. Gac Sanit 1993;7:63-69.

121.- Leape LL, Park RE, Solomon DH, Chassin MR, Kosecoff J, Brook RH. Does inappropriate use explain small area in the use of health care services? JAMA 1990;263:669-672.

122.- Roos NP, Roos LL, Henteleff PD. Elective surgical rates – do high rates mean lower standars?tonsillectomy and adenoidectomy in Manitoba. N Engl J Med 1977;297:360-365.

123.- Shwartz M, Ash AS, Anderson J, Iezzoni LI, Payne SMC, Restuccia JD. Small area variations in hospitalization rates: how you see depends on how you look. Med Care 1994;32:189-201.

124.- Lluch JA, Peiró S. Flujo de pacientes entre distritos hospitalarios para atención al parto: análisis descriptivo e implicaciones. Todo Hospital 1996;125:15-20.

125.- Wennberg JE. Future directions for small area variations. Med Care 1993;31(5 Suppl)YS75-YS80.

126.- Berstein SJ, Hilborne LH, Leape LL, Fiske ME, Park RE, Kamber CJ, Brook RH. The appropriateness of use of coronary angiography in New York State. JAMA 1993;269:766-769.

127.- Berstein SJ, McGlynn EA, Siu AL, Roth CP, Sherwood MJ, Keesey JV. The appropriateness of hysterectomy a comparison of care in seven health plans. JAMA 1993;269:2398-2402.

128.- Kleinman LC, Kosecoff J, Dubois RW. The medical appropriateness of tympanostomy tubes proposed for children younger than 16 years in the United States. JAMA 1994;271:1250-1255.

129.- Hilborne LH, Leape LL, Bernstein SJ, Park RE, Fiske ME, Kamberg CJ. The appropriatneness of use of percutaneous transluminal coronary angioplasty in New York State. JAMA 1993;269:761-765.

130.- Stano M. Further issues in small area variations analysis. J Health Polit Policy Law 1991;16:573-588.

131.- McPherson K. The best and the enemy of good: randomised controlled trials, uncertainty, and assessing the role patient choice in medical decision making. J Epidemiol Community Health 1994;48:6-15.

132.- Sackett DL, Rosemberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't? Br Med J 1996;312:71-72.

133.- Goold SD, Hofer T, Zimmerman M, Hayward RA. Measuring physician attitudes toward cost, uncertainty, malpractice, and utilization review. J Gen Intern Med 1994;9:544-549.

134.- Brook RH, Park RE, Chassin MR, Solomon DH, Keesey J, Kosecoff J. Predicting appropriate use of carotid endarterectomy, upper gastrointestinal endoscopy, and coronary angiography. N Engl J Med 1990;323:1173-1177.

135.- McKinlay JB, Burns RB, Feldman HA, Freund KM, Irish JT, Kasten LE. Physician variability and uncertainty in the management of breast cancer. Med Care 1998;36:385-396.

136.- Wennberg JE. Unwanted variations in the rules of practice. JAMA 1991;265:1306-1307.

137.- Folland S, Stano M. Small area variations a critical review of propositions methods, and evidence. Med Care 1990;47:419-465.

138.- Domenighetti G, Luraschi P, Casablanca A, Gutzwiller F, Spinelli A, Pedrinis E, Repetto F. Effect of information campaign by the mass on hysterectomy rates. Lancet 1988;24:1470-1473.

139.- Payne SM, Donahue C, Rappo P, McNamara JJ, Bass J, First L. Variations in pediatric pneumonia and bronchitis/ashma ademission rates. Is appropriateness a factor? Arch Pediatr Adolesc Med 1995;149:162-169.

140.- Wennberg JE. Dealing with medical practice variations: a proposal for action. Health Aff 1984;3:6-32.

141.- Logan RL, Scott PJ. Uncertainty in clinical practice: implications for quality and costs health care. Lancet 1996;347:595-598.

142.- Editorial. Calidad y variabilidad ¿Consensos europeos en Atención Primaria? Aten Primaria 1993;11:62.

143.- Kosecoff J. Variaciones en la práctica profesional. V Congreso de la Sociedad Española e Salud Pública y Administración Sanitaria. Granada, 27-30 octubre, 1993.

144.- Cain KC, Diehr P. The relationship between small-area variations in the use of health care services an inappropriate use: a commentary. Health Serv Res 1993;28:411-418.

145.- European Secondary Prevention Study Group. Transferencia de los ensayos clínicos a la práctica médica: estudio europeo a nivel poblacional sobre el uso de trombolisis en el infarto agudo de miocardio. Lancet 1996;29:149-153.

146.- Raine R, Streetly A, Davis AM. Variation in local policies and guidelines for cholesterol management: national survey. BMJ 1996;313:1368-1369.

147.- Chassin MR. Explaining geographic variations. The enthusiasm hypothesis. Med Care 1993; 31:YS37-YS44.

148.- Peiró S, Meneu R, Marqués JA, Librero J, Ordiñana R. La variabilidad en la práctica médica: relevancia, estrategias de abordaje y política sanitaria. Papeles de economía española 1998;76: 165-175.

149.- Hensley, Scout, Abboud. Medical research has “black hole”. Wall Street Journal 2004;135-138.

150.- Whittington, Craig. Selective serotonin reuptake inhibitors in childhood depression:systematic review of published versus unpublished data. Lancet 2004; 363:1341-1345.

151.- McColl A, Smith H, White P, Field J. General practitioners. Perception of the route to evidence based medicine: a questionnaire study. BMJ 1998; 316;361-365.

152.- Meakins JL. Innovation in surgery. The American Journal of Surgery 2002;183:399-405.

153.- Blech J. La apariencia de la cirugía. En: Blech J (Eds). Medicina Enferma. Barcelona, Ediciones Destino SA, 2007, pags 170-186.

154.- Domenighetti G. Revisiting the most informed consumer of surgical services – the physician patient. International Journal of Technology Assessment in Health Care 1993;9:505-513.

155.- Black W, Welch G. Advances in diagnostic imaging and overestimation of disease prevalence and the benefits of therapy. N Engl J Med 1993;328:1237-1243.

156.- Folkman J, Kalluri R. Cancer without disease. Nature 2004;427:787-791.

157.- Geaves M. Cómo se las arreglan las células cancerosas para ganar el juego. En: Geaves M (Ed). Cáncer el legado evolutivo. Barcelona, Editorial Crítica SA, 2002, pags 69-86.

158.- Blech J. La detección precoz no significa curación. En: Blech J (Eds). Medicina Enferma. Barcelona, Ediciones Destino SA, 2007, pags 101-112.

159.- www.nationales-netwerk-frauengesundheit.de (consultada 12-04-2007).

160.- Stamey T. The era of serum prostate specific antigen as a marker for biopsy of the prostate and detecting prostate cancer is now over in USA. BJU Int 2004;94(7):963-4.

161.- Dockser Marcus A. At 32, a decision: Is cancer small enough to ignore? Prev Med 2004;38(6):799-803.

162.- Thompson I. Prevalence of prostate cancer among men with a Prostate-Specific Antigen Level 4.0 ng per mililiter. N Engl J Med 2004;350:2239-2246.

163.- www.ebm-netzwerk.de (consultada 12-04-2007).

164.- Laín Entralgo P. La mentalidad anatomoclínica y la anatomía patológica. En: Laín Entralgo P (Ed). Historia de la medicina. Barcelona, Masson-Salvat Ediciones SA, 1994, pags 465-476.

165.- Laín Entralgo P. El tratamiento y la prevención de la enfermedad. En: Laín Entralgo P (Ed). Historia de la medicina. Barcelona, Masson-Salvat Ediciones SA, 1994, pags 519-537.

166.- Feinstein AR, Horwitz RI. Problems in the 'evidence' of 'evidence-based medicine'. Am J Med 1997;103:529-535.

167.- Sackett DL, Rosemberg WMC, Gray J, Haynes RB, Richarson WS. Evidence based medicine:What it is and what it isn´t. BMJ 1996; 312:71-72.

168.- Blech J. Dudas sobre la quimioterapia. En: Blech J (Eds). Medicina Enferma. Barcelona, Ediciones Destino SA, 2007, pags 113-125.

169.- Soto J. Medicina basada en resultados en salud: la evolución lógica y deseable de la medicina basada en la evidencia. Med Clin (Barc) 2007;128:254-5.

170.- Abel U, Koch A. The role of randomization in clinical studies. Myths and beliefs. J Clin Epidemiol 1999;52:487-497.

171.- Pozo F. La medicina basada en la evidencia: una perspectiva desde la clínica. Med Clin (Barc) 1999;112 Supl 4:36-9.

172.- Caplan LR. Evidence-based medicine: concerns of a clinical neurologist. J Neurol Neurosurg Psychiatry 2001;71:569-74.

173.- Strauss SE, McAlister FA. Evidence-based medicine: a commentary on common criticisms. CMAJ 2000;163:837-41.

174.- Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence-based medicine". Am J Med 1997;103:539-41.

175.- Permayer-Miralde G, Ferreira-González I. ¿Hacia la perversión de la medicina basada en la evidencia? Med Clin (Barc) 2006;126:497-9.

176.- Soto J. Medicina basada en la evidencia, pero ¿en qué evidencia? Med Clin (Barc) 1998;111:539-41.

177.- García FM. Limitaciones y subterfugios de la medicina basada en la evidencia (carta). Med Clin (Barc) 2003;120:197-8.

178.- Mykhalovskiy E, Weir L. The problem of evidence-based medicine:directions for social science. Soc Sci Med 2004,120:197-8.

179.- Maynard A. Evidence-based medicine: an incomplete method for informing treatment choices. Lancet 1997;349:126-8.

180.- Soto J. Valor terapéutico añadido de los medicamentos: ¿qué es, cómo se evalúa y cuál debería ser su papel en política farmacéutica? An Med Interna (Madrid) 2005;22:39-42.

181.- Soto J. Obtención de datos de efectividad previos a la comercialización de los medicamentos: ¿utopía o realidad? Med Clin (Barc) 2006;127:736-7.

182.- Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science 1998;282:245-6.

183.- Rapier CM. An introduction to outcomes research. Brookwood: Brookwood Medical Publication;1996.

184.- American Medical Association. Principles of outcomes research. Outcomes research resource guide, 1996-97. American Medical Association. New York 1997.

185.- Freund D, Lave J, Clancy C. Patients outcomes research teams: contribution to outcomes and effectiveness research. Annu Rev Public Health 1999;20:337-359.

186.- Blech J. Dudas sobre la quimioterapia. En: Blech J (Eds). Medicina Enferma. Barcelona, Ediciones Destino SA, 2007, pags 113-124.

187.- Etzioni R. The case of early detection. Nature reviews of cancer 2003;3:1-10.

188.- Garattini S, Bertelé V. Efficacy, safety, and cost of new anticancer drugs. BMJ 2002; 325:269-71.

189.- Feinstein A. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. N Engl J Med 1985;312:1604-1608.

190.- Giordano S. Is breast cancer survival improving? Cancer 2004;100:44-52.

191.- Johnson J. End points and United States Food and Drug Administration Approval of Oncology Drugs. Journal of Clinical Oncology 2003;21:1404-11.

192.- André F. Breast cancer with synchronous metastases: Trends in survival during a 14-years period. Journal of Clinical Oncology 2004;22:3302-05.

193.- Flores B, Andrés B, Campillo A, Soria V, Candel MF, Miquel J, Aliaga F, Aguayo JL. Análisis de la fiabilidad de los informes de alta en un servicio de cirugía general. Rev Calidad Asistencial 2004; 19 (7): 443-5.

194.- Badía X. del Llano J. La investigación de resultados en salud. Med Clin (Barc) 2000;114 supl 3:1-7.

195.- Guyatt GH, Juniper EF, Walter SD. Interpreting treatment effects in randomized trials. BMJ 1998;316:690-693.

196.- Soto J. Investigación de resultados en salud: el conocimiento del valor terapéutico añadido de los medicamentos. Pharmacoeconomics Spanish Res Art 2005;2:111-115.

197.- Lancry PJ, Oconnor R, Stempel D, Raz M. Using health outcomes data to inform decision-making healthcare payer perspective. Pharmacoeconomics 2001;19 Suppl 2:39-47.

198.- Rodríguez Artalejo F. Investigación de resultados (outcomes research) en el área de envejecimiento. Rev Esp Geriatr Gerontol 2001;36:20-3.

199.- Epstein RS, Sherwood LM. From outcomes research to disease management: a guide for the perplexed. Ann Intern Med 1996;124:832-7.

200.- Hulley S, Cummings S. Designing clinical research an epidemiologic approach Baltimore:1998.

201.- Ley 16/2003, de 28 de mayo, de cohesión y calidad del sistema nacional de salud. BOE de 29 de mayo 2003.

202.- Ley 29/2006, de 26 de julio, de garantías y uso racional de los medicamentos y productos sanitarios. BOE de 27 de julio de 2006.

203.- Villeta Plaza R, Landa García I. Cirugía basada en la evidencia. En: Ruiz P, Alcalde J; Landa J (Eds). Gestión Clínica en Cirugía. Madrid, Arán Ediciones SL, 2005, pags 503-543.

204.- Haynes RB, Sánchez RG, Jadad AR. Herramientas para la práctica de la medicina basada en la evidencia (I). Actualizaciones en recursos de información basados en la evidencia para la práctica clínica. Med Clin 2000;115;258-260.

205.- Fletcher SW, Fletcher RH. Development of clinical guidelines. Lancet 1998;352: 1876.

206.- McKeon T. Benchmarks and performance indicators: two tools for evaluating organizational results and continuous quality improvement efforts. Nursing Care Quality 1996;10:12-7.

207.- Ruiz López P, Lorenzo Martínez S. Benchmarking: cómo aprender de los mejores. En: Ruiz P, Alcalde J; Landa J (Eds). Gestión Clínica en Cirugía. Madrid, Arán Ediciones SL, 2005, pags 581-90.

208.- Mosel D, Grift B. Collaborative benchmarking in health care. Joint Commission Journal on Quality Improvement 1994;20:239-49.

209.- Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. Br J Surg 1991;78:355-360.

210.- Wang TJ, Zhang BH, Gu GS. Evaluation of POSSUM scoring system in the treatment of osteoporotic fracture of the hip in elder patients. Chin J Traumatol 2008;11(2):89-93.

211.- Young W, Seigne R, Bright S, Gardner M. Audit of morbidity and mortality following neck of femur fracture using the POSSUM scoring system. N Z Med J 2006 May 19;119(1234) :U1986.

212.- Mohamed K, Copeland GP, Boot DA, Casserley HC, Shackleford IM, Sherry PG, Stewart GJ. An assessment of the POSSUM system in orthopaedic surgery. J Bone Joint Surg Br 2002;84(5):735-9.

213.- Ramesh VJ, Rao GS, Guha A, Thennarasu K. Evaluation of POSSUM and P-POSSUM scoring systems for predicting the mortality in elective neurosurgical patients. Br J Neurosurg 2008:22(2):275-8.

214.- Pratt W, Joseph S, Callery MP, Vollmer CM Jr. POSSUM accurately predicts morbidity for pancreatic resection. Surgery 2008;143(1):8-19.

215.- Khan AW, Shah SR, Agarwal AK, Davidson BR. Evaluation of the POSSUM scoring system for comparative audit in pancreatic surgery. Dig Surg 2003;20(6):539-45.

216.- Campillo-Soto A, Flores-Pastor B, Soria-Aledo V, Candel-Arenas M, Andrés-García B, Martín-Lorenzo JG, Aguayo-Albasini JL. The POSSUM scoring system: an instrument for measuring quality in surgical patients. Cir Esp 2006;80(6):395-9.

217.- Tambyraja AL, Kumar S, Nixon SJ. POSSUM scoring for laparoscopic cholecystectomy in the elderly. ANZ J Surg 2005;75(7):550-2.

218.- Mohil RS, Bhatnagar D, Bahadur L, Rajneesh, Dev DK, Magan M. POSSUM and P-POSSUM for risk-adjusted audit of patients undergoing emergency laparotomy. Br J Surg 2004;91(4):500-3.

219.- Tekkis PP, Kocher HM, Bentley AJ, Cullen PT, South LM, Trotter GA, Ellul JP. Operative mortality rates among surgeons: comparison of POSSUM and p-POSSUM scoring systems in gastrointestinal surgery. Dis Colon Rectum 2000;43(11):1528-32, discusssion 1532-4.

220.- Prytherch DR, Whiteley MS, Higgins B, Weaver PC, Prout WG, Powell SJ. POSSUM and Portsmouth POSSUM for predicting mortality. Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. Br J Surg 1998;85(9):1217-20.

221.- Hobson SA, Sutton CD, Garcea G, Thomas WM. Prospective comparison of POSSUM and P-POSSUM with clinical assessment of mortality following emergency surgery. Acta Anaesthesiol Scand 2007;51(1):94-100.

222.- Constantinides VA, Tekkis PP, Senapati A; Association of Coloproctology of Great Britain and Ireland. Comparison of POSSUM scoring systems and the surgical risk scale in patients undergoing surgery for complicated diverticular disease. Dis Colon Rectum 2006;49(9):1322-31.

223.- Oomen JL, Engel AF, Cuesta MA. Mortality after acute surgery for complications of diverticular disease of the sigmoid colon is almost exclusively due to patient related factors.. Colorectal Dis 2006;8(5):453.

224.- Senagore AJ, Warmuth AJ, Delaney CP, Tekkis PP, Fazio VW. POSSUM, p-POSSUM, and Cr-POSSUM: implementation issues in a United States health care system for prediction of outcome for colon cancer resection. Dis Colon Rectum 2004;47(9):1435-41.

225.- Isbister WH, Al-Sanea N. POSSUM: a re-evaluation in patients undergoing surgery for rectal cancer. The Physiological and Operative Severity Score for Enumeration of Mortality and Morbidity. ANZ J Surg 2002;72(6):421-5.

226.- Wang TK, Tu HH. Colorectal perforation with barium enema in the elderly: case analysis with the POSSUM scoring system. J Gastroenterol 1998;33(2):201-5.

227.- Whiteley MS, Prytherch D, Higgins B, Weaver PC, Prout WG. Comparative audit of colorectal resection with the POSSUM scoring system. Br J Surg 1995 Mar;82(3):425-6.

228.- Sagar PM, Hartley MN, Mancey-Jones B, Sedman PC, May J, Macfie J. Comparative audit of colorectal resection with the POSSUM scoring system. Br J Surg 1994;81(10):1492-4.

229.- Ramkumar T, Ng V, Fowler L, Farouk R. A comparison of POSSUM, P-POSSUM and colorectal POSSUM for the prediction of postoperative mortality in patients undergoing colorectal resection. Dis Colon Rectum 2006;49(3):330-5.

230.- Bollschweiler E, Lubke T, Monig SP, Holscher AH. Evaluation of POSSUM scoring system in patients with gastric cancer undergoing D2-gastrectomy. BMC Surg 2005;15;5:8.

231.- Lam CM, Fan ST, Yuen AW, Law WL, Poon K. Validation of POSSUM scoring systems for audit of major hepatectomy. Br J Surg 2004;91(4):450-4.

232.- Shuhaiber JH, Hankins M, Robless P, Whitehead SM. Comparison of POSSUM with P-POSSUM for prediction of mortality in infrarenal abdominal aortic aneurysm repair. Ann Vasc Surg 2002;16(6):736-41.

233.- Midwinter MJ, Tytherleigh M, Ashley S. Estimation of mortality and morbidity risk in vascular surgery using POSSUM and the Portsmouth predictor equation. Br J Surg 1999;86(4):471-4.

234.- Myers NA. Comparative vascular audit using the POSSUM scoring system. Ann R Coll Surg Engl 1993;75(6):449.

235.- Copeland GP, Jones D, Wilcox A, Harris PL. Comparative vascular audit using the POSSUM scoring system. Ann R Coll Surg Engl 1993;75(3):175-7.

236.- Cagigas JC, Escalante CF, Ingelmo A, Hernandez-Estefania R, Hernanz F, Castillo J, Fleitas MG. Application of the POSSUM system in bariatric surgery. Obes Surg 1999;9(3):279-81.

237.- Ferguson MK, Durkin AE. A comparison of three scoring systems for predicting complications after major lung resection. Eur J Cardiothorac Surg 2003;23(1):35-42.

238.- Brunelli A, Fianchini A, Gesuita R, Carle F. POSSUM scoring system as an instrument of audit in lung resection surgery. Physiological and operative severity score for the enumeration of mortality and morbidity. Ann Thorac Surg 1999;67(2):329-31.

239.- Brunelli A, Fianchini A, Xiume F, Gesuita R, Mattei A, Carle F. Evaluation of the POSSUM scoring system in lung surgery. Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity. Thorac Cardiovasc Surg 1998;46(3):141-6.

240.- Lai F, Kwan TL, Yuen WC, Wai A, Siu YC, Shung E. Evaluation of various POSSUM models for predicting mortality in patients undergoing elective oesophagectomy for carcinoma. Br J Surg 2007;94(9):1172-8.

241.- Nagabhushan JS, Srinath S, Weir F, Angerson WJ, Sugden BA, Morran CG. Comparison of P-POSSUM and O-POSSUM in predicting mortality after oesophagogastric resections. Postgrad Med J 2007;83(979):355-8.

242.- Zafirellis KD, Fountoulakis A, Dolan K, Dexter SP, Martin IG, Sue-Ling HM. Evaluation of POSSUM in patients with oesophageal cancer undergoing resection. Br J Surg 2002;89(9):1150-5.

243.- Brooks MJ, Sutton R, Sarin S. Comparison of Surgical Risk Score, POSSUM and p-POSSUM in higher-risk surgical patients. Br J Surg 2005;92(10):1288-92.

244.- Sagar PM, Hartley MN, MacFie J, Taylor BA, Copeland GP. Comparison of individual surgeon's performance. Risk-adjusted analysis with POSSUM scoring system. Dis Colon Rectum 1996;39(6):654-8.

245.- Gu GS, Zhang DB, Zhang BH, Sun NK. Evaluation of P-POSSUM scoring system in predicting mortality in patients with hip joint arthroplasty. Chin J Traumatol 2006 Feb;9(1):50-5.

246.- Tekkis PP, Prytherch DR, Kocher HM, Senapati A, Poloniecki JD, Stamatakis JD, Windsor AC. Development of a dedicated risk-adjustment scoring system for colorectal surgery (colorectal POSSUM). Br J Surg 2004 Sep;91(9):1174-82.

247.- Law WL, Lam CM, Lee YM. Evaluation of outcome of laparoscopic colorectal resection with POSSUM, Portsmouth POSSUM and colorectal POSSUM. Br J Surg 2006 Jan;93(1):94-9.

248.- Vather R, Zargar-Shoshtari K, Adegbola S, Hill AG. Comparison of the possum, P-POSSUM and Cr-POSSUM scoring systems as predictors of postoperative mortality in patients undergoing major colorectal surgery. ANZ J Surg 2006 Sep;76(9):812-6.

249.- Poon JT, Chan B, Law WL. Evaluation of P-POSSUM in surgery for obstructing colorectal cancer and correlation of the predicted mortality with different surgical options. Dis Colon Rectum 2005 Mar;48(3):493-8.

250.- Prytherch DR, Sutton GL, Boyle JR. Portsmouth POSSUM models for abdominal aortic aneurysm surgery. Br J Surg 2001 Jul;88(7):958-63.

251.- Wijesinghe LD, Mahmood T, Scott DJ, Berridge DC, Kent PJ, Kester RC. Wijesinghe LD, Mahmood T, Scott DJ, Berridge DC, Kent PJ, Kester RC. Comparison of POSSUM and the Portsmouth predictor equation for predicting death following vascular surgery. Br J Surg 1998 Feb;85(2):209-12.

252.- Tekkis PP, McCulloch P, Poloniecki JD, Prytherch DR, Kessaris N, Steger AC. Risk-adjusted prediction of operative mortality in oesophagogastric surgery with O-POSSUM. Br J Surg 2004 Mar;91(3):288-95.

253.- Markus PM, Martell J, Leister I, Horstmann O, Brinker J, Becker H. Predicting postoperative morbidity by clinical assessment. Br J Surg 2005 Jan;92(1):101-6.

254.- Das N, Talaat AS, Naik R, Lopes AD, Godfrey KA, Hatem MH, Edmondson RJ. Risk adjusted surgical audit in gynaecological oncology: P-POSSUM does not predict outcome. Eur J Surg Oncol 2006 Dec;32(10):1135-8.

255.- Yii MK, Ng KJ. Risk-adjusted surgical audit with the POSSUM scoring system in a developing country. Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. Br J Surg 2002 Jan;89(1):110-3.

256.- Bennett-Guerrero E, Hyam JA, Shaefi S, Prytherch DR, Sutton GL, Weaver PC, Mythen MG, Grocott MP, Parides MK. Comparison of P-POSSUM risk-adjusted mortality rates after surgery between patients in the USA and the UK. Br J Surg 2003 Dec;90(12):1593-8.

257- Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification System. Crit Care Med 1985; 13: 818-829.

258.- Porath A, Eldar N, Harman-Bohem I, Gurman G. Evaluation of the APACHE II scoring system in an Israeli intensive care unit. Isr J Med Sci 1994 Jul;30(7):514-20.

259.- Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. Crit Care Med 1994 Sep;22(9):1392-401.

260.- Wong DT, Crofts SL, Gomez M, McGuire GP, Byrick RJ. Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. Crit Care Med 1995 Jul;23(7):1177-83.

261.- Giangiuliani G, Mancini A, Gui D. Validation of a severity of illness score (APACHE II) in a surgical intensive care unit. Intensive Care Med 1989;15(8):519-22.

262.- Bohnen JM, Mustard RA, Oxholm SE, Schouten BD. APACHE II score and abdominal sepsis. A prospective study. Arch Surg 1988 Feb;123(2):225-9.

263.- Berger MM, Marazzi A, Freeman J, Chioléro R. Evaluation of the consistency of Acute Physiology and Chronic Health Evaluation (APACHE II) scoring in a surgical intensive care unit. Crit Care Med 1992 Dec;20(12):1681-7.

264.- Vassar MJ, Wilkerson CL, Duran PJ, Perry CA, Holcroft JW. Comparison of APACHE II, TRISS, and a proposed 24-hour ICU point system for prediction of outcome in ICU trauma patients. J Trauma 1992 Apr;32(4):490-9.

265.- Ratanarat R, Thanakittiwirun M, Vilaichone W, Thongyoo S, Permpikul C. Prediction of mortality by using the standard scoring systems in a medical intensive care unit in Thailand. J Med Assoc Thai 2005 Jul;88(7):949-55.

266.- Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 1997 Feb;23(2):177-86.

267.- de Cássia Braga Ribeiro K, Kowalski LP. APACHE II, POSSUM, and ASA scores and the risk of perioperative complications in patients with oral or oropharyngeal cancer. Arch Otolaryngol Head Neck Surg 2003 Jul;129(7):739-45.

268.- Wang BW, Mok KT, Chang HT, Liu SI, Chou NH, Tsai CC, Chen IS. APACHE II score: a useful tool for risk assessment and an aid to decision-making in emergency operation for bleeding gastric ulcer. J Am Coll Surg 1998 Sep;187(3):287-94.

269.- Schein M, Gecelter G. APACHE II score in massive upper gastrointestinal haemorrhage from peptic ulcer: prognostic value and potential clinical applications. Br J Surg 1989 Jul;76(7):733-6.

270.- Jones DR, Copeland GP, de Cossart L. Comparison of POSSUM with APACHE II for prediction of outcome from a surgical high-dependency unit. Br J Surg 1992 Dec;79(12):1293-6.

271.- Fahn HJ, Wang LS, Huang MS, Huang BS, Hsu WH, Huang MH. Leakage of intrathoracic oesophagovisceral anastomoses in adenocarcinoma of the gastric cardia: changes in serial APACHE II scores and their prognostic significance. Eur J Surg 1997 May;163(5):345-50.

272.- Crestanello JA, Deschamps C, Cassivi SD, Nichols FC, Allen MS, Schleck C, Pairolero PC. Selective management of intrathoracic anastomotic leak after esophagectomy. J Thorac Cardiovasc Surg 2005 Feb;129(2):254-60.

273.- Cho DY, Wang YC. Comparison of the APACHE III, APACHE II and Glasgow Coma Scale in acute head injury for prediction of mortality and functional outcome. Intensive Care Med 1997 Jan;23(1):77-84.

274.- Cho DY, Wang YC, Lee MJ. Comparison of APACHE III, II and the Glasgow Coma Scale for prediction of mortality in a neurosurgical intensive care unit. Clin Intensive Care 1995;6(1):9-14.

275.- Bein T, Fröhlich D, Pömsl J, Forst H, Pratschke E. The predictive value of four scoring systems in liver transplant recipients. Intensive Care Med 1995 Jan;21(1):32-7.

276.- Sawyer RG, Durbin CG, Rosenlof LK, Pruett TL. Comparison of APACHE II scoring in liver and kidney transplant recipients versus trauma and general surgical patients in a single intensive-care unit. Clin Transplant 1995 Oct;9(5):401-5.

277.- Huang SW, Guan XD, He XS, Chen J, Ouyang B. The scoring system for patients with severe sepsis after orthotopic liver transplantation. Hepatobiliary Pancreat Dis Int 2006 Aug;5(3):364-7.

278.- Wolters U, Mannheim S, Wassmer G, Brunkwall J. What is the value of available risk-scores in predicting postoperative complications after aorto-iliac surgery? A prospective non randomized study. J Cardiovasc Surg (Torino) 2006 Apr;47(2):177-85.

279.- Lazarides MK, Arvanitis DP, Drista H, Staramos DN, Dayantas JN. POSSUM and APACHE II scores do not predict the outcome of ruptured infrarenal aortic aneurysms. Ann Vasc Surg 1997 Mar;11(2):155-8.

280.- Le Gall JR, Loira P. A simplified acute physiology score for ICU patients. Crit Care Med 1984; 12: 975-977.

281.- Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 1993; 270: 2957-2963.

282.- Lemeshow S, Le Gall JR. Modeling the severity of illness in ICU patients: A systems update. JAMA 1994; 272: 1049-1055.

283.- Ertan T, Yoldas O, Kilic YA, Kilic M, Göcmen E, Koc M, Tez M. External validation of prognostic models among cancer patients undergoing emergency colorectal surgery. Am J Surg 2008 Apr;195(4):439-41.

284.- Can MF, Yagci G, Tufan T, Ozturk E, Zeybek N, Cetiner S. Can SAPS II Predict Operative Mortality More Accurately Than POSSUM and P-POSSUM in Patients with Colorectal Carcinoma Undergoing Resection? World J Surg 2008 Apr;32(4):589-95.

285.- González-Pérez L, Monedero P, de Irala J, Kadri C, Lushchenkov D. Prognostic factors for cancer patients in the postanesthetic recovery unit. Rev Esp Anestesiol Reanim 2007 Aug-Sep;54(7):405-13.

286.- Laurila J, Laurila PA, Saarnio J, Koivukangas V, Syrjälä H, Ala-Kokko TI. Organ system dysfunction following open cholecystectomy for acute acalculous cholecystitis in critically ill patients. Acta Anaesthesiol Scand 2006 Feb;50(2):173-9.

287.- Padalino P, Chiara O, Ravizzini C, Gattinoni MP, Canini T, Montagnolo G, Marini AM. Role of the severity score and of the multiple organ dysfunctions in the treatment of severe acute pancreatitis and its infective complications. Ann Ital Chir 2005 May-Jun;76(3):239-45.

288.- Berghmans T, Paesmans M, Sculier JP. Is a specific oncological scoring system better at predicting the prognosis of cancer patients admitted for an acute medical complication in an intensive care unit than general gravity scores? Support Care Cancer 2004 Apr;12(4):234-9.

289.- Kern H, Redlich U, Hotz H, von Heymann C, Grosse J, Konertz W, Kox WJ. Risk factors for prolonged ventilation after cardiac surgery using APACHE II, SAPS II, and TISS: comparison of three different models. Intensive Care Med 2001 Feb;27(2):407-15.

290.- Alvarez M, Nava JM, Rué M, Quintana S. Mortality prediction in head trauma patients: performance of Glasgow Coma Score and general severity systems. Crit Care Med 1998 Jan;26(1):142-8.

291.- Lemeshow S, Teres D, Pastides H. A method for predicting survival and mortality of UCI patients using objetively derived weights. Crit Care Med 1985; 13: 519-525.

292.- Lemeshow S, Teres D, Klar J. Mortality Probability Model (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993; 270: 2478-2486.

293.- Lemeshow S, Klar J, Teres D. Mortality probability models for patients in the intensive care unit for 48 or 72 hours: A prospective, multicenter study. Crit Care Med 1994; 22: 1351-1358.

294.- Göçmen E, Klc YA, Yoldaş O, Ertan T, Karaköse N, Koç M, Tez M. Comparison and validation of scoring systems in a cohort of patients treated for biliary acute pancreatitis. Pancreas 2007 Jan;34(1):66-9.

295.- Koç M, Yoldaş O, Kılıç YA, Göçmen E, Ertan T, Dizen H, Tez M. Comparison and validation of scoring systems in a cohort of patients treated for perforated peptic ulcer. Langenbecks Arch Surg 2007 Sep;392(5):581-585.

296.- Marsall JC, Coock DJ. Multiple Organ Dysfunction Score : A reliable descriptor of a complex clinical outcome. Crit Care Med 1995; 23: 1638-1652.

297.- Peres Bota D, Melot C, Lopes Ferreira F, Nguyen Ba V, Vincent JL. The Multiple Organ Dysfunction Score (MODS) versus the Sequential Organ Failure Assessment (SOFA) score in outcome prediction. Intensive Care Med 2002; 28: 1619-1624.

298.- Hekmat K, Kroener A, Stuetzer H, Schwinger RH, Kampe S, Bennink GB, Mehlhorn U. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. Ann Thorac Surg 2005 May;79(5):1555-62.

299.- Campillo-Soto A, Flores-Pastor B, Soria-Aledo V, Aguayo-Albasini. Sistema POSSUM: Implantación de una escala de riesgo para la gestión de la calidad asistencial en un servicio de Cirugía General. En Fundación Signo (ed). 4ª edición Premios Barea. Madrid 2006.

300.- Fleiss JL. The measurement of interrater agreement, Capítulo 13. En: Statistical methods for rates and proportions. 2ª ed. New York. John Wiley,1981. 21-236.

301.- Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 1997;33:159-74.

302.- Rossner B. The Kappa statistic. En: Fundamentals of biostatistics. 4ª ed, 2002. Duxbury Pres, Belmont.

303.- Thompson WD, Walter S: A reappraisal of Kappa coefficient. J Clin epidemiol 1994;47:1315-17.

304.- Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 2003;29(2):249-56.

305.- Campillo-Soto A, Flores-Pastor B, Del Pozo P, Soria-Aledo V, Aguayo-Albasini. Importancia de las escalas de riesgo en cirugía. Nuestra experiencia con la escala POSSUM. Cir Andal 2007;18:338.343.

306.- Sánchez Pedraza R, Gómez Restrepo C. Conceptos básicos sobre validación de escalas. Rev Col Psiquiatría 1998;27:121-130.

307.- Cronbach LJ.Coefficient alpha and internal structure of tests. Psychometrika 1951; 16:297-334.

308.- Diamond GA, Hirsch M, Forrester JS. Application of information theory to clinical diagnostic testing. The electrocardiographic stress test. Circulation 1981;63:915-21.

309.- Hlatky M, Botvinick E, Brundage B. Diagnostic accuracy of cariologísts compared with probability calculations using Bayes´ rule. Am J Cardiol 1982;49:1927-31.

310.- Pinna-Pintor P, Bobbio M, Colangelo S, Veglia F. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients, Eur J Cardiothorac Surg 2002;21:199-204.

311.- Lopez de Ullibarri I, Pita Fernández S. Curvas ROC. Cad Aten Primaria 1998;5(4):229-235.

312.- Robertson EA. Zweig MH. Use of receiver operating characteristic curves to evaluate the clinical performance of analytical systems. Clin Chem 1981:27:1569-1574.

313.- Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York:Academic Press;1982.

314.- Zweig MH, Campbell G. Receiver-Operating Characteristic (ROC) plots: A fundamenta evaluation tool in clinical medicine. Clin Chem 1993;39 (4):561-77.

315.- Boyd JC. Mathematical tools for demonstrating the clinical usefulness of biochemical markers. Scand J Clin Lab Invest Suppl 1997;227:46-63.

316.- Molina V, Garrido L, Manterola C. Evaluación de resultados quirúrgicos. Validación inicial del escore POSSUM. Rev Chilena de Cirugía. 2005;57(4):291-296.

317.- Alvárez del Castillo M. Utilidad del sistema general de medición de gravedad, Mortality Predict Model (MPM II), como predictor de mortalidad hospitalaria en pacientes adultos con traumatismo craneoencefálico, ingresados en cuidados intensivos. Universidad Autónoma de Barcelona. 2003. (tesis doctoral).

318.-Machado F, Barberousse C, Santiago P, Barrios E, Carriquiry L. Comparison of surgical outcomes in two populations with risk adjustment using the POSSUM system. Cir Esp 2007 Jan;81(1):31-7.

319.- Prytherch DR, Sirl JS, Weaver PC, Scmidt P, Higgins B, Sutton GL. Towards a national clinical minimum data set for general surgery. Br J Surg 2003;90:1300-1305.

320.- Sutton R, Bann S, Brooks M, Sarin S. The Surgical Risk Scale as an improved tool for risk-adjusted analysis in comparative surgical audit. Br J Surg 2002;89:763-768.

321.- Neary WD, Prytherch D, Foy C, Heather BP, Earnshaw. Comparison of different methods of risk stratification in urgent and emergency surgery. Br J Surg 2007;94:1300-1305.

1. INTRODUCTION, OBJETIVES AND HYPOTHESIS

1.1. INTRODUCTION

At the beginning of the twenty-first century, Spain has one of the best balance of world health. According to the latest rankings of the World Health Organization (WHO), our country occupies the sixth place among 191 countries by their performance in terms of health of its population, despite the fact that by their level of health expenditure, it is the number 29. The main result of an efficient health system results in an increased life expectancy and disability-free as possible. A high health spending is not a sufficient condition for good health. In fact the opposite can happen also in excess of $ 1000 per employee in health spending is not seen a significant increase in the adjusted life expectancy associated with it[1].

According to the Organization for Economic Cooperation and Development (OECD) health expenditure in Spain will rise from 6.7% of GDP now to 13% of GDP in 2050. Of the 4 factors that determines the growth of health: a) Health and demographic (aging population), b) non-demographic and health (new technologies); c) Demographic associated with dependence (increase in the ratio old / young) and d) not associated with demographic dependency ratio (cost of illness), we can only influence the latter, reducing the costs of the disease, which could lead to a reduction in 3% of GDP for health care[1-4].

It has been shown that surgical complications are associated with an increase in hospitalization costs, contributing significantly to the increase in the cost per illness. Therefore, the reduction of these complications is a desirable goal of clinical management, which helps to reduce hospitalization costs, while improving the outcomes of health care[5-7]. It is estimated that the annual cost of surgical complications in a hospital an average of about $ 6 million and, with appropriate measures, could be reduced by 30 - 40% easily[8].

When we speak about outcomes of care refers to those changes in health habits or attitudes of individuals, groups or communities can be attributed to received medical care. These changes can not be attributed to medical care while other possible causes or factors are not ruled out. These results can be: positive (resulting in improved) or negative (causing a deterioration).

Potential strategies for investigating quality problems from adverse outcomes are basically two:

1.- The identification of those individual cases that require a review of health care for problems is to identify "sentinel events" which have an excellent feature is the validity time to be attributed to poor care.

2.- The measurement and analysis of rates (adjusted for risk prior), which relies on the measurement of certain events, without a warrant study of the process in each individual case to be produced even when great care is repeated a systematic way.

Focusing on the case of surgical results, these will be influenced primarily by: 1) prior physiological state of the patient, 2) complexity or severity of the intervention, 3) Quality and adequacy of the provision of care. These three parameters allow risk-adjusted rates, so that we can obtain:

1. Performance audits through the adjustment of mortality and morbidity in the case of each centre or surgeon.

2. Regular monitoring of the observed / expected (ratio O / E) providing information about the improvement or deterioration in clinical practice.

3. A progressive increase in the ratio O / E providing a starting point for analyzing the causes that are contributing to worsening the clinical practice.

At the same time avoiding the problems arising from the audit based on gross rates:

1. Make judgments, sometimes reckless, the results of clinical units, which has led to the closure of units and the interruption of training programs.
2. Conduct meetings of morbidity and mortality (SMM) of patients without knowing if the result was or was not expected.
3. No assessment of the successes in patients at high risk of morbidity mortality[9-14].

Based on the previous idea, to measure the impact of these complications is necessary to forecast rates of morbidity and mortality, allowing us to objectively assess the outcome in health care as to the likelihood of complications in each patient and try to identify preventable causes that allow us to improve our performance in future "alike" situations. Despite the large number of prognostic indexes of morbidity and mortality that have been developed until now, no studies have been reported that a comparative study to validate each others and set their own characteristics in terms of applicability, predictive value, accuracy and discrimination capability of each of them.

1.2. OBJECTIVES

The project's overall objective is to validate and compare the applicability of six scoring systems in terms of accuracy to predict morbidity and mortality in patients undergoing surgery at the Surgery Department of the University General Hospital "JM Morales Meseguer" of Murcia, in order facilitate the proper management of each of them in our environment and therefore contribute to knowledge and efficiency in our clinical practice.

The specific objectives are:

1. Define and analyze the characteristics of each of the prognostic indexes to study.
2. Identify the correct use and selection of each index in terms of their properties, predictive values and level at which it will be implemented.
3. Describe and facilitate the appropriate use of these indexes in our forecasts by type of treatment to the patient and to be subjected.

1.3. HYPOTHESIS

1. The survey of each site or service is different (case-mix) therefore, measurement of gross rates of morbidity and mortality (GRs) do not allow comparison between sites, services or surgeons.

2. GRs usually used to express the results in medical care are not good indicators of the health activity.

3. GRs do not allow the audit results in an efficient and effective way.

4. Regular monitoring of GRs does not provide information about the improvement or deterioration in clinical practice.

5. Adjusted rates of morbidity and mortality (ARs) sessions allow morbimortality knowing if the result was or was not expected.

6. ARs allow assessing and studying the successes achieved in patients at high risk of morbidity and mortality.

7. ARs allow fair trials, actual results and comparable data on clinical units, which facilitates the detection of deterioration in clinical practice, improve training programs, and equitable distribution of resources.

8. Each unit must select a Clinical efficient and adapted to their activity to monitor their morbidity and mortality adjusted.

2. MATERIAL AND METHODS

2. MATERIAL AND METHODS.

2.1. SCOPE

The study was developed in the Department of General and Digestive Surgery of the University General Hospital "JM Morales Meseguer" of Murcia. Our hospital is a level II hospital of 418 beds (of which 48 beds are dedicated to our department) dealing with any acute disease of adults, excluding the specialties of third level (except oncohaematology, which is included within the hospital), Paediatrics and Obstetrics and Gynecology. It is estimated that the total number of people attending our center is 300,000.

2.2. UNITS OF STUDY

Units of study are the complications and postoperative mortality in patients undergoing abdominal surgery on a scheduled entry in our department and followed in outpatient visit. It is considered to postoperative morbidity and mortality occurred within the first 30 days after surgery. Excluded patients shared or diverted to other surgical services, and those who failed the follow-up, outpatient visits after discharge. The observation period of the study was from January 1, 2007 to December 31, 2008, when they arrived at the "end-point" of our project. The evaluation period includes the day of surgery until 30 days after surgery.

The evaluation was conducted on patients in the study, while the tools are evaluated 6 scales surgical risk in order to assess its predictive ability in terms of morbidity and mortality in these patients.

2.3. EVALUATION CRITERIA AND TOOLS USED.

For this study we used 6 of the prognostic systems used internationally to evaluate complications in surgical patients[13,209,218,257,280,292,296].

- APACHE II (Acute Physiology and Chronic Health Evaluation).
- POSSUM (Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity).
- P-POSSUM (Portmouth-Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity).
- SAPS II (Simplified Acute Physiology Score).
- MPM (Mortality Prediction Model).
- MODS (multiple organ Dysfunction Score).

To evaluate the predictive ability of the 6 prognostic indexes studied, we conducted an initial validation phase of the scales, so it followed the classic validation methodology:

1.- Sample size: 10 subjects from each of the items to be evaluated. (POSSUM and P-POSSUM are composed of 18 items, APACHE: 15, SAPS: 15; MPM: 14; MODS: 6) therefore, the minimum number of patients needed for this phase of the study is 180 for POSSUM and P - POSSUM; 150 for APACHE and SAPS, MPM: 140, and 60 for MODS).

2.- Translation of the selected scale: For the study will use the English-Spanish translation already performed, validated and published (June 2006)[299] in our department of Surgery of the P- POSSUM and POSSUM. For other scales will be independent translation (from the original articles in English that describes each of

them) by two observers and their subsequent comparison and correlation.

3.- Test to assess the understanding of items of the scale: 2 previously trained evaluators apply, independently and separately, the range of patients selected for each index forecast.

4.- Interobserver reliability: We analyze the interobserver agreement for those items that could be subject to different interpretations depending on the assessor to apply the scale. Assess the internal consistency of each scale using Cronbach's alpha test as a measure of uniformity.

For comparison of different scales using the following methods:

5.- Goodness-of-fit test: It will be to evaluate the calculation of the area under the curve (AUC) by AUC curves for each of the prognostic index, together with their respective confidence intervals at 95% ($CI_{95\%}$).

6.- Accuracy of the test: It was evaluated with the Shannon index, to calculate each patient by comparing the predicted value for each scale, the result observed, according to the following formula:

$$S = ([(1 + o) \ln (1 + e) + 2] \ln (2\text{-}e)\text{-}LN2) / LN2$$

(Where "ln" is the neperian logarithm, "o" is the presence (o = 1) or absence (o = 0) in the event of death, and "e" is the expected value obtained for each scale for each patient. For each index the average over a standard deviation (SD) for the total patients (Shannon index global) and the average plus one of the deceased only (index of Shannon deceased) is calculated, this will determine the accuracy of each

model in either situation. This index of precision can range from 0 (no accuracy) and 1 (perfect accuracy).

7.- Graphical Methods: Graphical methods were used to compare performance of different prognostic systems. On one side is compared in a graph versus the observed mortality by the expected risk and intervals for each of the prognostic indexes, moreover, the ratio observed: expected for each of the scales at intervals of risk is represented. The same thing shall apply with the disease in case of P-POSSUM and POSSUM.

8.- Statistical analysis: Data are collected as averages, standard deviation (SD) or standard error (SE). The use of confidence intervals of 95% will be indicated in each case. The comparison of multiple means was performed with ANOVA for repeated measures. To determine whether there were statistically significant differences between results and expected according to different scales of risk applies Pearson X2 test.

2.3.1. DESCRIPTION OF SCORING SYSTEMS.

APACHE II (Acute Physiology and Chronic Health Evaluation) (Table 3.1)

Correlate the severity of the disease and evaluating the current state of health prior to the patient. The scale is set using a mathematical probability of hospital mortality.

The parameters measured in the scale of acute physiological impairment are: rectal temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation, arterial pH, serum sodium, serum potassium, serum creatinine, hematocrit, leukocyte, Glasgow coma scale and the age intervals listed in Table 3.2.

SAPS II (Simplified Acute Physiology Score) (Table 3.4)

The variables assessed are: Age, heart rate, systolic blood pressure, body temperature, urinary output, hematocrit, leukocyte count, plasma glucose, plasma urea, plasma potassium, plasma sodium, serum bicarbonate, Glasgow coma scale, hepatic dysfunction, renal and respiratory rate of patient (medical, scheduled surgical or surgical emergency), presence of AIDS, hematologic malignancies or metastatic tumor.

The probability of dying calculated SAPS II is given by the same formula as in the APACHE II:

$$\text{Probability} = [\text{elogit} / (1 + \text{elogit})] \times 100$$

And the calculation of the logit for the SAPS II:

$$\text{Logit} = -7.7631 + \text{SAPS II score } 0997 + 0.0737 \times [\ln (\text{SAPS II score} + 1)]$$

MPM (Mortality Probability Model) (Table 3.6)

The MPM system using simple clinical variables obtained on admission (MPM0) and 24 hours of it (MPM24), in addition to age and previous health status. The variables used in MPM0 are: Age, acute physiological disturbance (coma or stupor), heart rate ≥ 150 bpm, systolic blood pressure ≤ 90 mmHg, mechanical ventilation, acute renal failure, serious cardiac arrhythmias, stroke, gastrointestinal bleeding, mass effect skull, cardiopulmonary resuscitation prior to admission, chronic health condition (chronic renal failure, cirrhosis, metastatic malignancy), type of patient (patient medical or surgical emergency).

MPM24 uses some of the admission parameters and valued evolutionary changes within the first 24 hours of treatment in ICU. The variables used are: Age, parameters assessed on admission (cirrhosis, metastatic neoplasm, cranial mass effect, and patient medical or surgical emergency), parameters to 24 hours of treatment (deep stupor or coma at 24 hours, creatinine> 2 mg / dl, confirmed infection, mechanical ventilation at 24 hours of admission, PO2 <60 mmHg, prothrombin activity).

To calculate the individual probability of dying by any model of MPM II, each of the variables (X), is expressed dichotomous (present = 1 or absent = 0), in its absolute value. This value is multiplied by the weighting coefficient β, obtained by multiple logistic regression of the original study.

The equation (logit) is:

Logit βiXi = Σ β0 + β1X1 = β2X2 + + + β4X4 + β3X3 ... β15X15

The logit is replaced in the general formula for calculating the probability of dying:

Probability = [elogit / (1 + elogit)] x 100

MODS (Multiple Organ Dysfunction Score) (Table 3.7)

It is another scale for evaluating the severity of critically ill patients based on functional impairment resulting from the aggression. This scale allows an evolutionary view of disease and its impact on the physiological function in response to treatment and the capacity of functional recovery of the patient.

The organs and systems are assessed by MODS respiratory function, renal function, liver function, cardiovascular system, haemostasis, and neurological status, which are assigned a score between 0 and 4 depending on the deviation from normality. If you do not have any parameters it is considered normal.

POSSUM and P-POSSUM (Table 3.9)

Systems are specific to surgical risk assessment, consisting of 2 types of variables:

- Physiological variables: they are 12, including cardiopulmonary signs and symptoms, and determinations of blood biochemistry, and electrocardiographic abnormalities. If any of the variables can not be collected a value of 1 is assigned. Before surgery punctuation varies between 12 and 88.
- Surgical Variables: They are 6, they are divided into 4 ratings grow exponentially (1, 2, 4 and 8). The score is obtained after the surgical operation. The main examples of the degree of intervention in general surgery are listed in Table 3.10.

Once you get the scores, calculate the predicted risk of mortality and morbidity, using the following equations developed by Copeland et al[209] (Whereupon, the R1 and R2 mortality risk, the risk of disease):

- Ln R1 / 1 - R1 = -7.04 + (0.13 x physiological score) + (0.16 x operative severity score).

- Ln R2 / 1 - R2 = -5.91 + (0.16 x physiological score) + (0.19 x operative severity score).

The only technical difference between POSSUM and P-POSSUM is the type of equation used to calculate the mortality risk by the P-POSSUM following:

- Ln R1 / 1 - R1 = -9.07 + (0.17 x physiological score) + (0.16 x operative severity score).

2.4. SAMPLE SIZE DETERMINATION

The sample size was calculated considering an accuracy of 95% and an expected probability of 50% and 95% confidence, using the formula:

$$Z2 = n \bullet p \bullet (1-p) / i2$$

Making a total of 384 patients undergoing abdominal surgery scheduled entry in our hospital. This number increased by 10% for control patients lost to follow up and where it is impossible to access the medical history of the case, being defined the final sample size in 422 patients.

From the sample size provided the selection is made by stratified sampling. The sample was obtained prospectively from patients in the Department of Surgery. As the duration of the study was calculated for 2 years, we review four patients a week.

2.5. PROCESS FOR REVIEW OF CASES

The review of cases included the monitoring of each patient from the surgery until 30 days after surgery. We collected mortality and its cause and all morbidities described by Coopeland et al[209]. The data was transferred to a database and compared the results predicted by the scales of risk with the results observed in actual clinical monitoring of patients.

2.6. CONCORDANCE BETWEEN OBSERVER AND INTERNAL CONSISTENCY OF SCALES

Reviewers were trained doctors and specific training in the handling of the 6 scales of assessment and risk management workshops were conducted on the rating scale prior to the completion of the validation process and data collection. In addition, we identified the difficulties in using the questionnaire, the discrepancies in the review and agreed on the aspects that raised any doubts in applying the scales. The results at this stage to define the terms used to ensure that the collection of individual items did not give rise to ambiguous interpretations.

The reliability of the scales of risk assessment, defined as the degree of reproducibility of results when the instrument is used by different observers was evaluated using the Kappa index (k). The Kappa index measures the total correlation that exists if we exclude due to at random, or the actual agreement beyond chance. Its calculation of the general formula:

$$K = Po - Pe / 1 - Pe$$

Where Po is the observed correlation, and Pe is the agreement expected due solely at random[300, 301]. Cohen's Kappa statistic was used to correct the agreement due to chance, allowing estimation of statistical significance and the corresponding confidence intervals, the difference between the degree of agreement that would be expected simply at random (value 0) and the degree of observed agreement (the agreement is not perfect due to at random given the value 1). Calculations of P0 and Pe, as well as calculating the standard deviation were performed according to the formulas set by Fleiss both pairs of assessors to the case of several evaluators.

Published in the literature there are several proposals for interpretation: According to Fleiss, if Kappa is equal or greater than 0.6 it can be considered good reliability. According to Landis and Koch, negative or 0 Kappa indicates agreement, or

no, up to 0.2: slight agreement; more than 0.4: agreement is moderate, more than 0.6: remarkable agreement, and 0.8 or more: high agreement; and 1: absolute agreement. In general, most authors agree to make acceptable kappa over.

To measure the internal consistency of the scales we used Cronbach alpha test, which responds to the general formula:

Being;

n, the number of items,

p, the average of the linear correlations between each of the items.

Cronbach alpha test is an index of internal consistency that takes values between 0 and 1 and is used to check whether the instrument being evaluated collects information defective and therefore lead to wrong conclusions or whether it is an instrument reliably measuring what it says it measures. Alfa is thus a correlation coefficient squared which, broadly speaking, measures the homogeneity of the questions by averaging all correlations between all items to see what it would really seem. The index is better when closer of 1. The reliability is better, considering a good reliability from 0.80.

2.7. DATA ANALYSIS

Data analysis was performed using SPSS software for Windows® version 11.0 (SPSS Inc., Chicago. IL. USA) and Microsoft Excel® (Microsoft Corporation, Redmond, Washington, USA). We calculated the ratios of mortality and morbidity observed (O) and expected (E) (ratio O: E) for the P-POSSUM and POSSUM systems and the observed mortality ratios (O) and expected (E) for the other scales, a ratio of 1

indicates perfect correlation between the observed and expected, if it is <1 says that the results are better than expected, and if> 1, the results are worse than expected.

To determine whether there were statistically significant differences between the results obtained and those expected according to different levels of risk was applied to the Pearson X2 test.

In all cases a difference was found between the observed and expected or between two scales of risk as significant when the resulting level of significance was less than 0.05 ($p < 0.05$).

3. RESULTS

3. RESULTS

3.1. VALIDATION OF THE RISK SCALE (TABLE 4.1).

3.1.1. Validation of the POSSUM and P-POSSUM scales.

For the validation process of these 2 scales were used 180 patient records, selected by stratified random sampling, obtaining the following results:

1.- Translation of the index: For numeric variables, there were no problems or differences in the comparison of translations carried out. For qualitative and descriptive variables showed a good interobserver correlation (Kappa = 0.9).

2.- Test to asses the understanding of items of the scale: For the quantitative variables we found no differences in the interpretation and application of its categories. For intraoperative variables, there were no differences. As for the qualitative variables, there were some differences as to distinguish between normal and mild disease, leading to an agreement between the evaluators, and developing and including in the protocol for collecting data on a description that includes what it is relates each of them.

3.- Interobserver reliability: There were no significant differences between the physiological variables between the surgery and for different observers. K was more than 0.8 for all variables in physiological score and it had a degree of correlation between 0.7 and 0.9 for surgery variables.

On the assessment of homogeneity showed a Cronbach alpha score of 0.8 for the physiological and 0, 72 for the surgery.

4 - Time: It was 8.5 minutes (range: 4-15 minutes).

3.1.2. Validation of APACHE II scale.

It was necessary to validate the use of 150 medical records, selected by stratified random sampling, obtaining the following results:

1.- Translation of the scale: There were no significant differences in any of the items rated, and the Kappa index calculated for this scale of 0.85

2.- Test to asses the understanding of items of the scale: There were no differences in the interpretation of quantitative variables. There were some differences in the determination of the Glasgow coma scale, so an agreement was reached between the assessors for their determination, including the protocol for data collection with the same definitions for their categories.

3.- Interobserver reliability: No significant differences were found for different observers. K index was greater than 0.8.

On the assessment of homogeneity showed a Cronbach alpha of 0.75.

4.- Time: It was 11 minutes (range: 7-20 minutes).

3.1.3. Validation of the SAPS II scale.

It was necessary to validate the use of 150 medical records, selected by stratified random sampling, obtaining the following results:

1.- Translation of the scale: There were no significant differences in any of the items rated, and the Kappa index calculated for this scale of 0.85.

2.- Test to assess the understanding of items of the scale: There were no differences in the interpretation of quantitative variables. There were some differences in the determination of the Glasgow coma scale, so an agreement was reached between the assessors for their determination, including the protocol for data collection with the same their definitions by category.

3.- Interobserver reliability: No significant differences were found for different observers. Concordance exists K greater than 0.85.

On the assessment of homogeneity showed a Cronbach alpha of 0.8.

4.- Time: It was 10.3 minutes (range: 8-19 minutes).

3.1.4. Validation of the MPM scale.

It was necessary to validate the use of 140 medical records, selected by stratified random sampling, obtaining the following results:

1.- Translation of the scale: There were no significant differences in any of the items rated, and the Kappa index calculated for this scale of 0.9

2.- Test to assess the understanding of items of the scale: There were no significant differences in the assessment of the quantitative variables. There were significant disagreements in the valuation of several dichotomous items: 1) presence or absence of cirrhosis, 2) presence or absence of intracranial mass effect, 3) or metastatic neoplasm; 4) presence of deep stupor or coma at 24 hours of admission; 5) confirmed infection or not, 6) treatment with vasoactive drugs for more than an hour. It was not possible to establish a consensus among observers for the valuation of these items.

3.- Interobserver reliability: There were significant differences for the different observers. K = 0.55.

On the assessment of homogeneity showed a Cronbach alpha of 0.65.

4 .- Time: It was 10 minutes (range: 5-13 minutes).

3.1.5. Validation of the MODS scale.

It was necessary to validate the use of 60 medical records, selected by stratified random sampling, obtaining the following results:

1. - Translation of the scale: There were no significant differences in any of the items rated, and the Kappa index calculated for this scale was 0.95.

2.- Test to assess the understanding of items of the scale: There were no differences in the interpretation of quantitative variables. There were some differences in the determination of the Glasgow coma scale, reaching the same agreement as for APACHE II and SAPS II.

3.- Interobserver reliability: No significant differences were found for different observers. K was more than 0.9.

On the assessment of homogeneity showed a Cronbach alpha of 0.8.

4.- Time: It was 6 minutes (range: 3-11 minutes).

3.2. PROSPECTIVE COMPARISON OF SCALES IN RELATION TO RISK OF DEATH.

3.2.1. Overall results of observed and expected mortality.

384 patients were included in the study; diagnoses of patients are listed in Table 4.2. The average age was 67.7 years (range: 17 - 94 years) with a sex distribution of 223 men (58%) and 161 women (42%). It has been excluded from the study patients shared or diverted to other services, and not able to follow up within the postoperative 30 days.

34 patients died (6.25%) within 30 days after surgery and 148 patients had some type of morbidity (38.5%). The findings regarding morbidity are listed in Table 4.3 and Table 4.4 shows the causes of mortality.

The actually mortality was of 6.25% in the group of patients studied. Estimated mortality for the POSSUM and P-POSSUM was of 12.5% (48 patients) and 12% (46 patients) respectively without significant differences between estimated by these systems and observed mortality. The APACHE II estimated a mortality of 5.7% (22 patients) with significant differences observed ($p = 0.04$). The mortality produced by SAPS II and MODS was 3.6% (14 patients) with $p = 0001$ (Table 4.5).

3.2.2. Mortality risk intervals for each of the scales.

For the POSSUM system (Table 4.6), we note that its predictive power with regard to mortality is good in all ranges of risk as a whole, no significant differences in any of them, moreover, the ratio O: E are good in all ranges, which are less than 1 in general and comprehensive manner (ratio O:E global 0.7), only in the group with 80-100% of the risk is equal to 1. The results obtained for the P-POSSUM are very similar to the POSSUM scale (Table 4.7).

Results for the APACHE II (Table 4.8) shows a grouping of all patients in the first 3 ranges of risk: <20% 20-39% 40-59% and significant difference between the estimated risk of mortality from this scale for patients in the group between 20-39% ($p = 0.04$) and overall ($p = 0.04$).

Mortality calculated by the SAPS II scale (Table 4.9) shows significant differences both within groups of risk and overall mortality.

The system MODS (table 4.10) shows very similar results to the scale of the SAPS II, both globally and at intervals of risk, with significant differences observed between the calculated by the scale and what actually happened. As the scale SAPS II and APACHE II includes all patients in the lowest strata of risk (<20%, 20-39% and 40-59%).

3.2.3. Calculation of the ROC curves for each of the scales studied.

Table 4.11 reflects the results of the calculation of the ROC curves for each of the risk scale, with a value close to 1 (perfect test) for the scale POSSUM (ROC area = 0.78) for the P-POSSUM (ROC area = 0.79) with no significant differences between two systems.

The APACHE II, SAPS II and MODS ROC values close to 0.5 (useless test) significant difference between them and the POSSUM scale ($p <0001$) and P-POSSUM ($p <0001$).

Between APACHE II and SAPS II there are no significant differences, but between them and the scale MODS.

3.2.4. Calculation of Shannon index for each of the scales of risk.

The accuracy of the scales studied were evaluated using the Shannon index (SI), this has been the calculation of SI for dead (his calculation only includes deaths) (table 4.12) and the global IS (for calculation necessary to compute the living and dead patients) (table 4.13).

Within the calculation of the mortality-SI the best results were for POSSUM (0.92) and P-POSSUM (0.91) without significant differences between them. Mortality-SI were for APACHE II (0,64), SAPS II (0.44) and for MODS (0,41). There were significant differences between the value obtained for POSSUM and the values for APACHE II (p = 0002), SAPS II (p = 0.000003) and MODS (p = 0.0000009) as well as results and P-POSSUM and APACHE II (p = 0004), SAPS II (p = 0.000008) and MODS (p = 0.000002). There was no significant difference between these latter three scales.

In a similar way, the calculation of the overall SI values were higher for POSSUM (0103) and P-POSSUM (0102), and the lowest values for MODS (0036), SAPS II (0036) and APACHE II (0057). There were significant differences between the results obtained for POSSUM and APACHE II (p = 0.01), SAPS II and MODS (p = 0.000005). There were also significant differences between P-POSSUM and APACHE II (p = 0009), MODS, and SAPS II (p = 0.0001). There was no significant difference between the values of these last three prognostic indexes.

3.2.5. Graphics methods and results in mortality.

3.2.5a. Representations of expected and observed mortality.

This section graphically represents the expected mortality for each risk scale, compared with those observed for each of them (Figure 4.1). In addition, graphical representations in successive scale is broken down by prognostic outcomes expected at intervals of mortality risk has been "superimposed" to the graph of observed mortality for each one. Brief discussion of the results of each of the scales:

POSSUM system (Figure 4.2) displays graphs of expected mortality: observed very similar, almost super imposable without significant differences.

P-POSSUM scale (Figure 4.3) shows results similar to those of the scale POSSUM, without significant differences between them.

APACHE II index (Figure 4.4) overestimated mortality significantly in the range of 20-39% risk, strata in which more patients died.

Graphic study of the SAPS II system (Figure 4.5) indicates a statistically significant overestimation of the mortality in the two intervals in which risk has classified patients.

4.6 graph shows that the system of MODS significantly overestimated the mortality expected manner.

3.2.5b. Representations of the ratio of observed: expected mortality.

The study has been completed with graphic representation of mortality ratios observed: expected range of risk, with a horizontal line representing the cut-off point where results that appear above it shows findings worst than expected and below him otherwise (results better than expected), being the balance 1 point (equal to the results observed).

In the graphs 4.7 and 4.8 shows that the ratio O: E in mortality for POSSUM and P-POSSUM respectively conform to a good prediction (values equal to or less than 1).

The graph 4.9 shows that the APACHE II predictions are worse than those obtained in all ranges of risk, except for lower risk patients (<20%) that is obtained in line with reality.

Graphs 4.10 and 4.11 are very similar and represent ratios O: E far above what obtained in the realities of SAPS II and MODS scales respectively, the differences being statistically significant.

3.3. STUDY OF THE PREDICTIVE CAPACITY OF MORBIDITY OF POSSUM & P-POSSUM SCALES.

Of the 6 scales studied, the only ones that can measure and predict morbidity, as well as mortality, are the POSSUM and P-POSSUM scales, therefore, the study in terms of morbidity only make it with these two scales. It will explore the ability of both scales to predict morbidity compared with actual outcomes and complications in patients under study.

In table 4.14 the results are represented in terms of overall morbidity calculated by the POSSUM and P-POSSUM systems and those observed in reality. As shown in the table there were no significant differences in the prediction of morbidity in these scales and the observed reality. As expected the risk of morbidity (33.6%) similar to that observed (38.5%).

Table 4.15 represents the values and expected morbidity ratio O: E for percentage of risk. When analyzing morbidity within risk groups by POSSUM and P-POSSUM, there are no significant differences in any of them.

The chart 4.12 shows that POSSUM and P-POSSUM scoring system can predict very accurately the risk of complications in both globally and at intervals of risk.

4. CONCLUSIONS

4. CONCLUSIONS

1. Gross rates of mortality and morbidity are not reliable or valid indicators to monitor and evaluate the activity of a service department or hospital.

2. POSSUM, P-POSSUM, APACHE II, SAPS II and MODS scoring systems have been validated for patients undergoing surgery by laparoscopy or laparotomy, in an optimal way in the UGH "JM Morales Meseguer of Murcia.

3. The Mortality Prediction Model scale (MPM II) could not be acceptably validated in surgical patients scheduled (laparotomy and laparoscopy) of UGH "JM Morales Meseguer.

4. APACHE II, SAPS II and MODS scores do not have good reliability in their application for predicting mortality in surgical patients.

5. POSSUM and P-POSSUM systems have a high reliability in application for measuring risk of mortality and morbidity in patients studied.

6. POSSUM and P-POSSUM have shown high reproducibility in use in these patients.

7. No significant differences were observed either in terms of mortality or morbidity observed in reality and expected for POSSUM and P-POSSUM.

8. The mortality observed in the patients studied is above the required quality standards, both overall and at intervals of risk.

9. Morbidity is observed slightly below the acceptable quality standards.

10. Contrary to expectations, the morbidity of patients with low surgical risk (0-39%) is above what is acceptable. While the morbidity of patients with increased surgical risk (40-100%) is below what is required by the standards of quality.

11. Morbidity surgeon is within quality standards in more than 80%.

12. The use of the POSSUM and P-POSSUM systems is highly recommended in the surgery services to monitor and detect errors in clinical practice.