# 7.7 Métodos de medición de casuística y ajuste de severidad


El ajuste de riesgos es una metodología para desarrollar comparaciones de resultados o costes entre centros sanitarios que atienden pacientes de diferente gravedad (casuística, severidad).

***Autor: Salvador Peiró***

*Área de Investigación en Servicios de Salud, Centro Superior de Investigación en Salud Pública (CSISP), Valencia*



Se recomienda imprimir 2 páginas por hoja

***Citación recomendada:***

Peiró S. Métodos de medición de casuística y ajuste de severidad [Internet]. Madrid: Escuela Nacional de Sanidad; 2013 [consultado día mes año]. Tema 7.7 Disponible en: direccion url del pdf.

## Resumen:

Comparar centros sanitarios es difícil porque sus resultados clínicos o de costes dependen, además de la calidad o eficiencia de los centros, de las características de los pacientes que han atendido. Para realizar estas comparaciones se ***ajustan*** los resultados obtenidos por cada centro según las características relevantes de los pacientes que ha atendido y que definen su ***riesgo previo*** (***severidad***, ***gravedad***) de tener el resultado concreto que se valora. Ajustar, en este contexto, es ponderar cada paciente por el riesgo que deriva de sus características

clínicas, demográficas u otras.

Ésta es la base para la construcción de ***sistemas de clasificación de pacientes*** o ***sistemas de ajuste de riesgos*** que, simplificando, no son más que ***sistemas que cuantifican la probabilidad que tienen los pacientes de obtener un determinado resultado clínico o de costes***. Algunos de estos sistemas, como los Grupos de Diagnósticos Relacionados (GDR), son ampliamente utilizados para valorar la eficiencia de los hospitales. Otros intentan valorar su calidad o establecer *ranking* de hospitales en función de diversas dimensiones.

*Introducción*

*1. Marco conceptual*
*1.1. Manejo*
*1.2. Características*

*2. Aspectos metodológicos*
*2.1 Definición del objetivo del sistema*
*2.2. Selección de los factores de riesgo*
*2.3. Construcción y desarrollo*
*2.4 Validación*
*2.5 Rendimiento*

*3. Uso práctico*

*4. Perspectivas*

*Referencias bibliográficas*

## Introducción

Comparar organizaciones sanitarias ofrece importantes oportunidades para identificar posibilidades de mejora. Aunque existen otras posibilidades en la práctica suelen compararse:

- ***Resultados clínicos***: mortalidad, complicaciones, reingresos, reintervenciones, etc.
- ***Procesos***: tasa de cesáreas, porcentaje de pacientes que han recibido determinados tratamientos, etc.
- ***Medidas de productividad***, tanto en forma de procesos por unidad de recurso (por ejemplo, el índice de rotación, la estancia media, el porcentaje de ocupación), como de costes por proceso.

La comparación de organizaciones sanitarias es compleja porque, más allá de su propia calidad o eficiencia, también depende de las características de los pacientes que ha atendido en cada organización: un hospital que tenga un 10% de ingresos para

intervención de cataratas y un 1% de ingresos oncológicos tendrá menos muertes y menores costes que otro que tenga un 10% de oncológicos y un 1% de cataratas. Incluso dentro de un mismo tipo de pacientes, las diferencias en gravedad pueden marcar importantes diferencias en tasas de mortalidad, costes por proceso o cualquier otro indicador.

La literatura sobre servicios sanitarios contiene numerosos ejemplos de comparaciones mecanicistas entre organizaciones (**cuadro 1**) que conducen a juicios temerarios y ponen en entredicho la propia utilidad de estas comparaciones para la toma de decisiones en atención sanitaria.

Las características -demográficas, clínicas y otras- de los pacientes definen su ***riesgo previo*** *(antes de entrar en el centro sanitario)* ***de obtener el resultado concreto que se está midiendo*** más allá de la calidad de la atención prestada. A este riesgo previo se le llama **gravedad** o severidad[1].

*Cuadro 1. Comparaciones de mortalidad entre hospitales en Estados Unidos (1986)*

*Entre 1986 y 1993 la Administración Federal de EE.UU. publicó una estadística de mortalidad hospitalaria con objeto de ofrecer a los usuarios una orientación sobre la "calidad" de cada hospital para que pudieran elegir con mayor información.*

*En la estadística publicada en 1986 (la primera de ellas, que incluía más de 5000 hospitales, y que ya utilizaba un sistema de ajuste de riesgos), en el hospital con peores resultados la mortalidad fue del 87,6%, mientras la mortalidad esperada conforme al promedio ajustado por el tipo de pacientes que había atendido era del 22%.*

*El dato es tan llamativo que implicaría un grave problema de calidad … de no ser porque se trataba de un hospital para pacientes terminales. Todos los pacientes fallecían y este dato no implicaba ningún problema de calidad.*

*El error obvio es que los pacientes ingresados en hospitales terminales no tienen la misma probabilidad de muerte que los pacientes ingresados en un hospital de agudos. Aunque fueran del mismo sexo, edad, diagnóstico y tuvieran la misma comorbilidad, existían otros factores que los diferenciaban (unos era terminales y otros no) que no se tuvieron en cuenta en ésta comparación.*

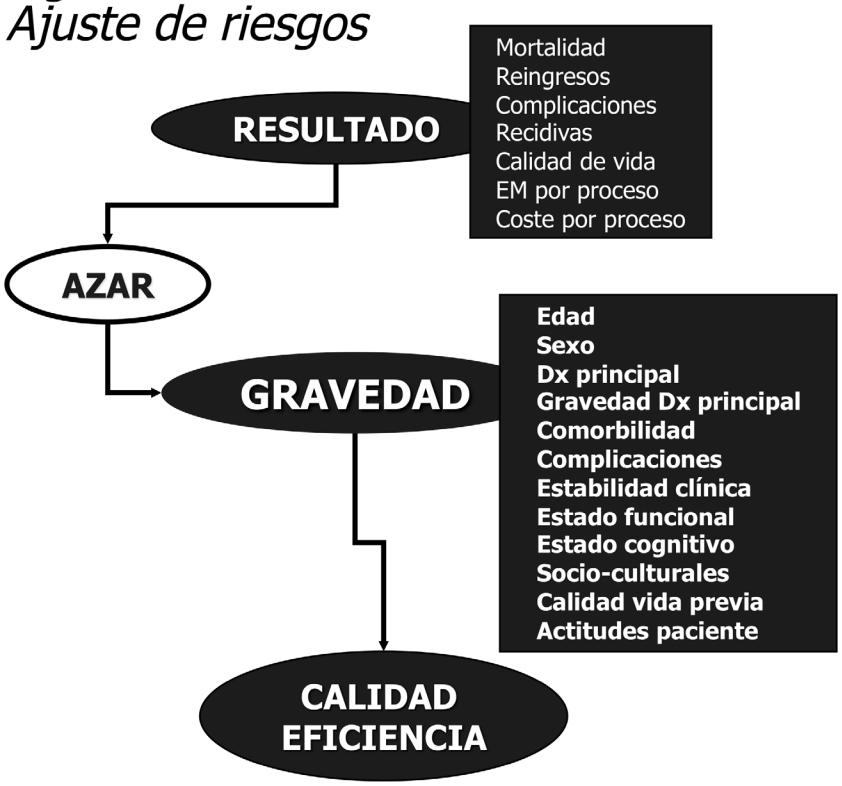
*La probabilidad previa (derivada de las características del paciente y no de la atención recibida) de obtener el resultado que se está midiendo*

La gravedad, en este contexto, no es (únicamente) equivalente a riesgo de muerte, sino la ***probabilidad previa (derivada de las características del paciente y no de la atención recibida) de obtener el resultado que se está midiendo***, que puede ser

*1 En España la traducción (incorrecta) de severity por "severidad" (en lugar de "gravedad") es tan popular que en este tema se usan como sinónimos.*

mortalidad o cualquier otro suceso adverso, pero también calidad de vida, consumo de recursos, etc. Este marco puede representarse mediante un esquema (**figura 1**) en el que el riesgo derivado de la características previas del paciente (gravedad), más la efectividad y eficiencia del proveedor, teniendo en cuenta el papel del azar, conducirán a diferentes resultados clínicos o de costes.

*Figura 1.*
*Ajuste de riesgos*




El abordaje de las diferencias en gravedad es crítico para comparar la calidad de las organizaciones sanitarias, porque el objetivo de la evaluación es comparar la calidad o eficiencia de estas organizaciones y, por tanto, es esencial que las diferencias en resultados (incluyendo los costes por proceso) se deban a diferencias en su desempeño, y no a diferencias en los pacientes que atienden.

Para ello, la metodología usada es *ajustar* los resultados obtenidos por cada centro, por las *características relevantes* de los pacientes que ha atendido y que definen su riesgo previo respecto al resultado concreto que se quiere medir. Ajustar, en este contexto, es ponderar cada paciente por el riesgo que deriva de sus características clínicas, demográficas u otras.

El supuesto que subyace bajo esta aproximación es que si los pacientes tratados fueran iguales (paciente hipotéticos con valores promedio en todos sus factores de riesgo), los resultados de la atención también deberían ser iguales y las posibles diferencias en resultados –de existir- serían atribuibles a problemas de calidad o de eficiencia. De este modo, si construimos grupos de pacientes con riesgos similares respecto al resultado que se desea medir, estos resultados serán comparables para tales grupos y, a su vez y mediante su ponderación, permitirán la comparación entre centros sanitarios.

Ésta es la base para la construcción de **sistemas de clasificación de pacientes** (SCP) que, simplificando, no son más que ***sistemas que cuantifican la probabilidad que tienen los pacientes de obtener un determinado resultado***. La denominación de SCP es relativamente impropia para los métodos de estandarización de riesgos que no construyen grupos o clasificaciones y en la literatura técnica se viene sustituyendo por la denominación de **sistemas de ajuste de riesgos** (SAR), que ha reunido bajo un mismo techo otros conceptos peor definidos, como casuística (*case mix*), gravedad (*severity*), enfermedad (*sickness*), intensidad (*intensity*), complejidad (*complexity*), comorbilidad (*comorbidity*) o carga de enfermedad (*burden of disease*).

Desde que en 1983 la *Health Care Financing Administration* (HCFA)[2], decidió utilizar uno de estos sistemas, los **Grupos de Diagnósticos Relacionados** (GDR), como base de un sistema de pago prospectivo para el reembolso de las hospitalizaciones (excluyendo los honorarios médicos y algunos otros aspectos como las prótesis), los SAR se han multiplicado y la literatura sobre ellos ha crecido extraordinariamente, no siendo difícil hallar excelentes revisiones sobre el tema, incluyendo algunos trabajos españoles.

## 1. Marco conceptual para el ajuste de riesgos

El marco conceptual para el ajuste de riesgos incluye varios elementos: la medida de resultados que se quiere emplear, el tipo de centros a comparar, las variables que definirán la gravedad, las fuentes de datos y la modelización estadística.

*2 La HCFA era el organismo encargado del reembolso a las organizaciones sanitarias y a los profesionales que prestan servicios a Medicare, el programa público de aseguramiento sanitario para mayores de 65 años de EEUU.*

## 1.1. Manejo de los sistemas de ajuste de riesgos

En la **tabla 1** se muestran las variables utilizadas por el **Mortality Probability Model II (MPM-II)**, un sistema de ajuste de riesgos utilizado en unidades de cuidados críticos (UCI). Los autores del MPM analizaron los datos de varias decenas de miles de pacientes ingresados en este tipo de unidades y estimaron que, por ejemplo, la presencia de cirrosis tenía el efecto independiente de incrementar hasta 3 veces (un 200%) el riesgo de muerte en los 30 días desde el ingreso, o que los ingresos médicos o quirúrgicos urgentes tenían un 130% más riesgo de muerte que los quirúrgicos programados (véase en la **tabla 1** la columna de las odds ratio, OR). Mediante un análisis estadístico los autores del MPM-II estimaron los coeficientes de regresión (coeficientes beta) asociados a cada variable de modo que sumados a la constante y despejando el logit permiten estimar la probabilidad de muerte de cada paciente.

*Tabla 1. VARIABLES, COEFICIENTES Y ODDS DEL MPM $II_{24}$*

| | | Beta | OR |
|---|---|---|---|
| Edad (Odds Ratio cada por 10 años) | 10 años | 0,03268 | 1,4 |
| Cirrosis | si/no | 1,08745 | 3,0 |
| Efecto masa intracraneal | si/no | 0,91314 | 2,5 |
| Neoplasia maligna metastásica | si/no | 1,16109 | 3,2 |
| Ingreso quirúrgico urgente ó médico | si/no | 0,83404 | 2,3 |
| Coma o estupor profundo (a las 24 h. del ingreso) | si/no | 1,68790 | 5,4 |
| Creatinina > 2 mg/dl (a las 24 h. del ingreso) | si/no | 0,72283 | 2,1 |
| Infección confirmada (a las 24 h. del ingreso) | si/no | 0,49742 | 1,6 |
| Ventilación mecánica (a las 24 h. del ingreso) | si/no | 0,80845 | 2,2 |
| PPO2 < 60 mmHg (a las 24 h. del ingreso) | si/no | 0,46677 | 1,6 |
| Tiempo protrombina > 3 "s/est. (a las 24 h. del ingreso) | si/no | 0,55352 | 1,7 |
| Orina < 150 ml en 8 h. (a las 24 h. del ingreso) | si/no | 0,82286 | 2,3 |
| Fármacos vasoactivos > 1 hora i.v. (a las 24 h. del ingreso) | si/no | 0,71628 | 2,0 |
| CONSTANTE | | -5,64592 | |

*Multiplicar cada coeficiente por el valor 1 si la variable es afirmativa y 0 si es negativa, o por grupos de 10 años para la edad. Los productos anteriores se suman para obtener el logit y se obtiene la probabilidad de muerte a través de la transformación $e^{logit}/1+e^{logit}$ .OR: Odds ratio; Basado en: Rue Monne M et al, 1996.*

Supongamos (**tabla 2**) un paciente de 70 años, no quirúrgico, que tras 24h del ingreso en UCI presenta una infección confirmada y esta ventilado, sin otras alteraciones en variables del MPM-II24.

Para calcular su riesgo de muerte tendríamos que sumar a la constante (-5,64592), los coeficientes de la edad (7*0,03268=2,18960, recordemos que este coeficiente es por cada 10 años de edad y al tener el paciente 70 años hay que multiplicar el coeficiente por 7), del tipo de ingreso (0,83404, al ser un ingreso médico), la infección (0,49742) y la ventilación mecánica (0,80845). En la **tabla 2** se muestran estos datos, y el valor del logit en este paciente.

*Tabla 2. COEFICIENTES DEL MPM $II_{24}$ EN UN PACIENTE HIPOTÉTICO*

| | |
|---|---|
| 70 años (7 · 0,03268) | 2,18960 |
| Ingreso quirúrgico urgente ó médico | 0,83404 |
| Infección confirmada (a las 24 h. del ingreso) | 0,49742 |
| Ventilación mecánica (a las 24 h. del ingreso) | 0,80845 |
| Constante | -5,64592 |
| Logit (sumatorio de coeficientes y constante) | -1,31640 |

La probabilidad de muerte de este paciente concreto puede despejarse mediante la formula [$\mathbf{p = e^{logit} / 1+e^{logit}}$] que, en este caso, $p(\text{muerte}) = e^{-1,3164} / 1+e^{-1,3164} = 0,21$. Esto es, el paciente tiene una probabilidad de muerte del 21% o, mejor expresado, de cada 100 pacientes como este, se espera que 21 mueran y 79 sobrevivan.

A los intensivistas les es útil hacerse una idea del pronóstico de sus pacientes. Pero adicionalmente, gracias a este tipo de sistemas podemos comparar las unidades de cuidados críticos. Por ejemplo, si dos UCI, que han tratado 100 pacientes cada una muestran la misma mortalidad, pongamos que del 37%, la comparación bruta sugiere una similar calidad de la atención. Pero si tenían la distribución de pacientes según probabilidad de muerte medida por el MPM que muestra la **tabla 3**, en una esperaríamos 36 muertes (vs. las 37 observadas) mientras que en la otra, dada la menor gravedad de sus pacientes, esperábamos sólo 26 (11 menos que las realmente se han producido), aspecto que sugiere la posible existencia de algún problema de calidad que convendría identificar.

**Tabla 3. COMPARACIÓN DE LA MORTALIDAD EN 2 UCI HIPOTÉTICAS**

| | Unidad de Cuidados Intensivos 1 | | | | Unidad de Cuidados Intensivos 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Ingresos UCI 1 | Muertes observadas | pMPM | Muertes Esperadas | Ingresos UCI 2 | Muertes observadas | pMPM | Muertes Esperadas |
| | 10 | 2 | 0,15 | 1,5 | 10 | 1 | 0,11 | 1,1 |
| | 20 | 5 | 0,24 | 4,8 | 20 | 4 | 0,18 | 3,6 |
| | 30 | 10 | 0,33 | 9,9 | 30 | 5 | 0,22 | 6,6 |
| | 30 | 14 | 0,45 | 13,5 | 30 | 17 | 0,34 | 10,2 |
| | 10 | 6 | 0,60 | 6,0 | 10 | 10 | 0,48 | 4,8 |
| Total | 100 | 37 | ---- -- | 35,7 | 100 | 37 | ---- -- | 26,3 |

*pMPM: probabilidad de muerte predicha por el Mortality Probability Model*

Al igual que hemos utilizado el MPM para comparar la mortalidad en dos UCI, podemos utilizar cualquier otro sistema de ajuste de riesgos para comparar otro tipo de unidades respecto a un resultado de interés. Por ejemplo, los GDR para comparar la estancia media entre dos centros. Nótese que este tipo de sistemas no es diferente de la tradición clínica de ajustar por algún instrumento para comparar o predecir resultados: la escala de coma de Glasgow para comparar la mortalidad post traumatismo-craneoencefálico entre dos hospitales, o el Apgar para comparar la mortalidad perinatal entre unidades de neonatología o de obstetricia, o el sistema ASA para comparar las complicaciones perioperatorias entre pacientes intervenidos en dos centros o por dos equipos, etc. Todos estos métodos, en tanto cuantifican el riesgo de sufrir un resultado pueden considerarse también sistemas de ajuste de riesgos y la mecánica básica de empleo será ajustar el resultado por el SAR mas adecuado al tipo de comparación que se desee utilizar.

## 1.2. Características de los sistemas de ajuste de riesgos

Los SAR más conocidos incluyen los **Grupos de Diagnósticos Relacionados** (GDR) y sus variantes, el ***Disease Staging*** (DS) y otros sistemas basados fundamentalmente en la revisión de la historia clínica, como el ***Computerized Severity Index*** (CSI) y los sistemas diseñados para pacientes en unidades de cuidados críticos (**APACHE**, **MPM**, **PRISM**, **TISS, SAPS**) pero existen otros muchos como el POSSUM, COMPLEX, SUPPORT, TISS-28, NEMS, AIM, RUGs en sus diferentes versiones, etc., así como sistemas

específicos de un procedimiento o un diagnóstico, que pueden ser de interés en determinadas áreas.

Una forma de acercarse a estos sistemas es preguntarse por:

1. El resultado que pretenden ajustar. Básicamente, puede ser un **resultado de consumo de recursos** como la estancia media (GDR) o la intensidad de recursos (TISS, RUGs), o un **resultado clínico** como la mortalidad (DS, CSI, APACHE, MPM). El concepto de gravedad en cada caso puede ser muy diferente. Así, un paciente con una puntuación MPM muy elevada (elevado riesgo de muerte temprana) puede ser clasificado en un GDR de bajo peso (bajo coste) ya que si muere muy tempranamente su consumo de recursos será escaso. Por ello, emplear estos sistemas para controlar resultados diferentes a los que fueron establecidos en su diseño (por ejemplo, GDR para ajustar mortalidad) puede ser incorrecto.

2. Los requerimientos de información de cada sistema también serán diferentes. Así, mientras que **los sistemas para ajustar resultados de consumo de recursos suelen estar construidos a partir de bases de datos clínico-administrativas** tipo Conjunto Mínimo de Datos Básicos (CMBD), **los sistemas para resultados clínicos suelen requerir datos provenientes de la *historia clínica* u otras fuentes primarias**, lo que tiene implicaciones en su coste y factibilidad, pero también en su capacidad de predicción de riesgos y en su credibilidad clínica.

3. En cuanto a los criterios de clasificación utilizados, hay que resaltar, en primer lugar, el **papel otorgado al diagnóstico** (Dx), que permite clasificar estos sistemas en *diagnóstico-dependientes* (GDR), en general vinculados a la medición de costes, y *diagnóstico-independientes* (APACHE, MPM), casi siempre vinculados a la medición de la mortalidad temprana en unidades de críticos, en las que la estabilidad de los sistemas orgánicos suele tener más importancia que el diagnóstico. El papel otorgado a los procedimientos quirúrgicos (Px) mayores es también esencial para medir el consumo de recursos.

4. La elección del momento en que se recogen los datos tiene importantes implicaciones, ya que los sistemas que usan el CMBD, que obtiene los datos al alta (retrospectivos),

no podrán fijar la secuencia temporal de parte de los eventos ocurridos durante la hospitalización, lo que conlleva importantes limitaciones para separar el riesgo debido a la comorbilidad previa del paciente, del riesgo derivado de las complicaciones adquiridas en el hospital. Este aspecto es trascendental en las comparaciones entre proveedores ya que si se ajusta el riesgo que deriva, por ejemplo, de las infecciones nosocomiales, éste no será tenido en cuenta, obteniéndose resultados similares de calidad o eficiencia en centros con tasas de infecciones nosocomiales muy diferentes.

5. Finalmente, y en cuanto al tipo de medida que ofrecen estos sistemas, algunos utilizan ***escalas continuas***, en las que se valora el riesgo frente a un riesgo promedio (p.ej., en el sistema de los GDR, un peso de 1,15 implica un riesgo de consumo de recursos un 15% mayor que el promedio de todos los pacientes) y tienen un sentido relativo obvio, mientras que otros sistemas utilizan ***escalas ordinales*** que no implican tales referencias (en el CSI una puntuación "2" no implica el doble de riesgo que una puntuación "1") o construyen **agrupaciones de base diagnóstica (GDR) estrictamente *categóricas***.

En la **tabla 4** se presenta una generalización de las diferencias esenciales entre los sistemas para ajuste de resultados clínicos y de costes.

*TABLA 4. Características de los sistemas de ajuste de riesgos*

| | CLÍNICOS | COSTES |
|---|---|---|
| Requerimientos de información | ++++ | + |
| Fuente usual | Historia clínica | CMBD |
| Coste del sistema | ++++ | + |
| Capacidad ajuste | +++ | + |
| Credibilidad clínica | +++ | + |
| Desarrollo | Juicios clínicos | Empírica |
| Papel del diagnóstico | ++ | ++ |
| Papel del procedimiento | + | ++++ |
| Toma de datos | Variable (concurrente) | Al alta (retrospectiva) |

## 2. Aspectos metodológicos de los sistemas de ajuste de riesgos

Los procesos de desarrollo y construcción de los SAR son esenciales para reconocer las fortalezas y debilidades de estos métodos. Un sistema de ajuste de riesgos se construye a partir de los siguientes elementos:

1) Definición del objetivo del sistema.

2) Selección de los factores de riesgo que definirán la gravedad.

3) Desarrollo del sistema de ajuste (modelos de desarrollo).

4) Validación del sistema (modelos de validación).

### 2.1 Definición del objetivo del sistema

El primer paso en la construcción de un SAR es la definición de su propósito, esto es, qué tipo de organizaciones se pretende evaluar y qué resultado se quiere aislar de la influencia de los factores de riesgo.

Respecto al objeto de evaluación, las posibilidades básicas son cualquier tipo de centro (hospitales generales, centros de salud, centros de larga estancia), unidad de trabajo (servicios clínicos) o incluso médicos individuales.

Selección del resultado y ventana de observación. Cuando la fuente de información es el CMBD los ***resultados*** quedan limitados a las variables disponibles en el mismo: mortalidad intrahospitalaria, duración de la estancia, reingresos o la presencia de determinados códigos que indican complicaciones.

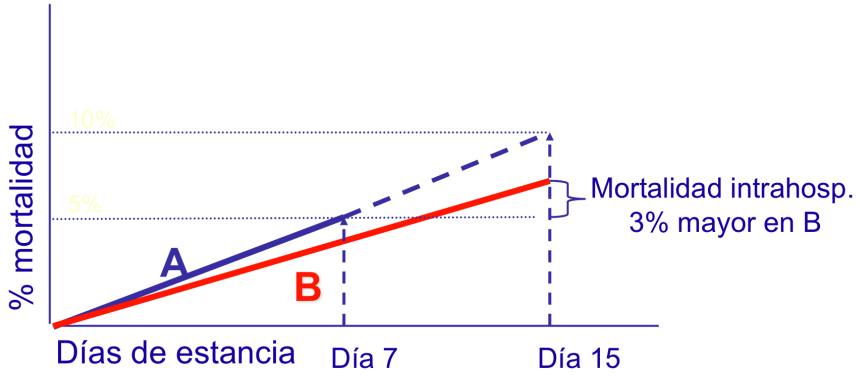
El ***tiempo de observación*** es un elemento clave en este marco. De un lado la ventana temporal para la identificación del resultado tiene que ser adecuada al propósito del sistema. Por ejemplo, si el objetivo es evaluar la efectividad de las UCI es más razonable utilizar la mortalidad en tiempos de observación cortos (hasta los 30 días) que el reingreso al año, ya que el primer resultado tendrá una mayor capacidad de inferencia sobre la calidad de estas unidades, mientras que el segundo vendrá muy afectado por todo el tratamiento posterior al alta. En segundo lugar, se

deberían garantizar tiempos de seguimiento homogéneos para todos los casos. Así, la mortalidad intrahospitalaria es una medida defectuosa al no recoger a los pacientes dados de alta y fallecidos en su domicilio. Ello implica que en los proveedores con estancias menos prolongadas (por ejemplo, los que tengan una política de altas de terminales) existirá un menor tiempo a riesgo de muerte intrahospitalaria y tendrán tasas de mortalidad artificialmente menores. Finalmente, otro elemento clave, especialmente en los SAR para intensivos, es el momento en que se recogen los datos, aspecto ya comentado en otro apartado.

## *Figura 2.*
## *Efecto de la variabilidad en la ventana temporal de seguimiento de los resultados*

**Hosp A:**
**Estancia Media = 7 días;**
**Mortalidad al 7 día= 5%;**
**Mortalidad al 15 día= 10%**

**Hosp B:**
**Estancia Media = 15 días;**
**Mortalidad al 7 día= 4%;**
**Mortalidad al 15 día= 8%**




La definición de la unidad de análisis tiene 2 vertientes de interés: el individuo y el ámbito del análisis. Respecto al individuo de análisis, las posibilidades son variadas: episodios de hospitalización (ingresos) o de cuidados (visitas médicas), procesos completos que pueden incluir uno o más episodios de hospitalización, con o sin atención ambulatoria, personas-año, etc. Este aspecto tiene importantes implicaciones analíticas y su elección depende fundamentalmente del propósito de la evaluación. Respecto al ámbito del análisis, las posibilidades son sistemas generales, comprensivos de todo tipo de pacientes, o sistemas específicos para una enfermedad o en un procedimiento, o un grupo de

enfermedades o procedimientos. Los primeros han sido diseñados para evaluar la efectividad o eficiencia de hospitales y su prototipo serían los GDR y sus variantes, mientras que los segundos tienen como objeto valorar algún tipo concreto de atención.

*Un marco conceptual sólido debe relacionar las variables a ajustar con cada resultado particular*

## 2.2. Selección de los factores de riesgo que definirán la gravedad

No todas las características de los pacientes influyen de modo relevante en todos los resultados, lo que explica la diversidad de las dimensiones consideradas por cada SAR en función de su objetivo. Un ***marco conceptual sólido debe relacionar las variables a ajustar con cada resultado particular***, aspecto no necesariamente fácil dada la existencia correlaciones múltiples entre muchos factores de riesgo. Las mismas variables puedan ser factores de ajuste de resultados diferentes -como mortalidad y duración de la estancia o costes- pero esto no siempre sucede, siendo deseable ajustar modelos por separado para cada resultado de interés. No obstante, en algún caso puede ser útil la combinación de sistemas para mejorar la explicación de la varianza del resultado inicial (esta es la base de algunos SAR, como los *Refined-DRG* que crean grupos dentro de cada GDR para mejorar la capacidad explicativa sobre los costes).

*Asegurar que las variables de ajuste dependan del riesgo del paciente y que no sean un componente de los propios cuidados desarrollados.*

Para la selección de factores de riesgo, las formulas pasan por la revisión de la literatura clínica y el juicio de los expertos, con el problema de traducir algunos factores pronósticos de difícil medición a *proxys* con los que mantengan una buena correlación. En todo caso, dado que el objetivo de todos estos sistemas es la evaluación de la calidad y/o eficiencia de la atención hospitalaria, parece importante **asegurar que las variables de ajuste *dependan del riesgo del paciente y que no sean un componente de los propios cuidados desarrollados***.

En este apartado se realizará un breve comentario de las variables más utilizadas para ajustar gravedad. Nótese, que algunas de estas variables pueden utilizarse como "resultados" (por ejemplo, el estado funcional se puede utilizar como factor de riesgo respecto al resultado mortalidad, pero también se emplea en otros sistemas como resultado a explicar por otros factores).

La **edad** es una característica predeterminada de los pacientes,

fácilmente disponible, que puede ser un importante predictor de riesgos, en parte porque se asocia a otras características del paciente (determinados diagnósticos y comorbilidad) y en parte porque es un predictor *independiente* de peores resultados (muerte, complicaciones, duración de la estancia) ya que los ancianos pueden requerir mayores períodos de recuperación y tener mayor riesgo de complicaciones. Estos efectos, importantes a nivel global, pueden ser muy modestos cuando se estudian grupos con edades similares o cuando se analiza algún tipo de patología especialmente prevalente en algún grupo de edad específico. Un grupo de edad de especial interés son los mayores de 79 años que suelen presentar más complicaciones, gravedad, etc., que otros pacientes pero recibir un esfuerzo terapéutico menor que los pacientes más jóvenes con cuadros clínicos similares, implicando la existencia de un grupo de pacientes que -por su mayor gravedad- son modelizados como pacientes de alto coste, pero en los que los clínicos adoptan una actitud expectante de bajo coste.

El **sexo**, otra variable fácilmente disponible y de gran importancia en los estudios epidemiológicos, ha mostrado escasa relevancia como predictor de resultados hospitalarios a corto plazo, y prácticamente no está incluido en ningún sistema general de ajuste de riesgos.

El **diagnóstico principal** (DxP), la casuística en sentido estricto, presenta pocas dudas acerca de su importancia como predictor de riesgos, tanto sobre los resultados clínicos como sobre la utilización de recursos. Los criterios esenciales para realizar un Dx exacto son la etiología, la localización y las manifestaciones fisiopatológicas (por ejemplo: neumonía de lóbulo superior izquierdo por *Sp. Pneumoniae* con insuficiencia respiratoria). No obstante, en la práctica clínica no siempre es posible o deseable especificar completamente el Dx y, en muchas ocasiones, este no va más allá de un diagnóstico genérico o sindrómico. Otras veces no es posible establecer la línea divisoria entre enfermedad y resultados anormales de una prueba, o las mismas manifestaciones provienen de diferentes enfermedades y viceversa. La adaptación a esta realidad clínica de alta incertidumbre y dependencia de la intensidad diagnóstica ha llevado al desarrollo de clasificaciones diagnósticas no mutuamente excluyentes (pacientes iguales pueden ser clasificados en diferentes epígrafes: por ejemplo, como trombosis o accidente vásculo-cerebral o hemiplejia), lo que introduce importantes posibilidades de variabilidad en la codificación.

Identificar el DxP puede no ser suficiente para ajustar algunos resultados y determinar la **gravedad dentro de un mismo DxP** puede ser esencial en muchos casos (por ejemplo, no tienen el mismo riesgo de muerte ni la mismas necesidades de recursos un cáncer localizado que uno diseminado). La gravedad del Dxp es un concepto que varía en función de los objetivos de cada sistema de medición y que, en general, se asocia al pronóstico de muerte en una determinada ventana temporal. En oncología suelen utilizarse las clasificaciones de gravedad en estadios definidos por la clínica o la patología.

*La diferencia importante es que la primera es un factor del paciente, y la segunda un posible problema de calidad que no debemos introducir en el ajuste de riesgos.*

La **comorbilidad** se refiere a la presencia de enfermedades no relacionadas con el DxP pero coexistentes en el momento del ingreso. La comorbilidad no debe confundirse con la gravedad del DxP (patología relacionada) ni con las **complicaciones** (patología no presente en el momento del ingreso). Nótese que si un paciente ingresa con una neumonía y tiene una infección urinaria sobreañadida, ésta la calificaríamos de comorbilidad; pero si no la tenía cuando ingresó y sobreviene durante la hospitalización la calificaremos de complicación: **la diferencia importante es que la primera es un factor del paciente, y la segunda un posible problema de calidad que no debemos introducir en el ajuste de riesgos**. El prototipo de comorbilidad son las patologías crónicas (diabetes, hipertensión arterial, coronariopatía isquémica, bronquitis obstructiva crónica, etc.), aunque ocasionalmente pueden ser agudas (por ejemplo, una fractura de cadera producida por caída al sufrir un accidente vásculo-cerebral). Como sucedía con el DxP, no es suficiente identificar la existencia de comorbilidad sino que hay que valorar su gravedad y extensión en relación al resultado a medir. Existen índices específicos de comorbilidad (índice de Charlson y otros), pero la mayor parte de los SAR recogen directamente la comorbilidad para la construcción de grupos de pacientes o el ajuste de riesgos.

La **estabilidad clínica** refleja el estado fisiológico de los sistemas corporales del paciente mediante el examen de sus signos vitales (frecuencia cardiaca y respiratoria, presión arterial, etc.), bioquímica (K, Na, creatinina, ...), parámetros hematológicos, gases sanguíneos y nivel de conciencia. Es una de las variables esenciales en los resultados a corto plazo, y la base de buena parte de los SAR utilizados en las UCI.

El **estado funcional**, aproximadamente la capacidad de realizar las actividades de la vida diaria, es otra variable de interés,

no sólo porque tiene una estrecha relación con el consumo de recursos de enfermería, aspecto clave en los SAR en centros de largo tratamiento, sino porque en algunos casos tiene una fuerte asociación con el pronóstico. El **estado psicológico, cognitivo** y **psicosocial** ha sido reconocido en diversos trabajos como una variable trascendente en los resultados que, sin embargo, resulta prohibitiva para los SAR al requerir detalladas entrevistas por personal especializado para su obtención. Las **características socio-culturales y socio-económicas** también han mostrado como factores relacionados de una forma importante con los resultados. El bajo nivel socio-económico ha sido asociado a peores resultados en numerosos estudios, al igual que el grupo étnico en los estudios estadounidenses. De nuevo en este caso antes de ajustar por estos factores ha de plantearse su proceso de asociación con el resultado ¿es o no un factor independiente de la atención?. Otros factores asociados al nivel socio-económico como el alcoholismo o las toxicomanías y los riesgos ocupacionales también se han asociado a peores resultados, al igual que el nivel cultural, algunas prácticas religiosas y algunas prácticas alimentarias restrictivas. Relacionado con las anteriores, las **actitudes y preferencias del paciente** (por ejemplo, rechazando cuidados agresivos o por motivos religiosos, incumpliendo los tratamientos, ...) pueden ser un importante predictor de resultados.

## 2.3. Construcción y desarrollo de sistemas de ajuste

La necesidad de combinar un buen rendimiento estadístico con la aceptabilidad clínica ha convencido a la mayor parte de los investigadores de que la combinación de juicio clínico y modelizaciones empíricas es mejor que cualquiera de las dos aproximaciones por separado. Incluso en el supuesto de modelos estrictamente empíricos, el juicio clínico es útil para decidir los puntos de corte en los valores de las variables o para la revisión posterior de los modelos.

Salvo que el sistema se base exclusivamente en juicios clínicos, se requiere una base de datos para su construcción y el primer paso es la "limpieza" de esta base: 1) excluir el tipo de casos que no se desea considerar el análisis, 2) excluir los centros que no cumplan una serie de criterios de calidad de la información, 3) depurar los valores no plausibles o sin significado, 4) valorar

la especificidad y congruencia de la información, 5) valorar la posibilidad de construir enlaces entre registros de la misma base de datos, cuando el resultado elegido lo requiera.

El siguiente problema es el tratamiento de los valores perdidos, desconocidos, imposibles o incongruentes, que va desde la eliminación de los casos con algún valor *missing* a su imputación mediante algún método estadístico. En ambos casos, el sesgo puede ser importante si los *missing*, como es probable, no se distribuían al azar.

El siguiente paso será la reducción de variables, especialmente de aquellas que presenten un alto grado de correlación entre si. Lo usual es el empleo de diversas técnicas de análisis, especialmente los modelos multivariantes para resultados continuos o dicotómicos, y los modelos de reducción de datos, existiendo diversas revisiones que valoran las limitaciones de estos modelos. En esta fase tiene interés valorar la inclusión de variables con escasa frecuencia o que ofrecen resultados ilógicos, y la posibilidad de incorporarlas agrupadas; por ejemplo, es posible que algunas comorbilidades de baja frecuencia no sean utilizables en la modelización, pero pueden ser incorporadas en un índice global de comorbilidad.

A continuación se debería explorar la estructura de las variables independientes a incluir (factores de riesgo) y la forma de su relación con la dependiente (resultado). Ello orientará sobre la posibilidad de introducir las variables como continuas o la necesidad de categorizarlas (y en que rangos) o de incorporar interacciones. El paso final será utilizar las técnicas estadísticas más adecuadas para efectuar la modelización, aspecto que se tratará posteriormente como parte del rendimiento de los modelos.

## 2.4. Validación

La noción de medición es la clave de los procesos comentados, ya se trate de mediciones de sucesos simples, como la mortalidad, o de conceptos abstractos, como la gravedad o la calidad. Para describir la calidad de las mediciones realizadas por un instrumento se emplean dos conceptos: **validez** y **fiabilidad**. La validez denota el grado en que un instrumento de medida mide lo que intentaba medir, o inversamente, el grado en que la medición se distancia sistemáticamente de su objetivo. Estrictamente hablando ***no se***

***valida un instrumento en si, sino el propósito para el que es usado***. ***La validez es, por otro lado, una cuestión de grados, no existe en términos absolutos*** y, en el mismo sentido, es un concepto dinámico.

En su aplicabilidad a los SAR, en general, la medida de resultado opera como criterio (***validez de criterio***) sobre el que validar la modelización del riesgo y la validación consiste en responder a la cuestión: *¿en qué medida los indicadores de riesgo elaborados predicen el resultado que les da sentido?*

En su aplicación es importante distinguir entre los datos usados para desarrollar el modelo y los empleados para su validación, pues si sólo se atiende a las predicciones en los primeros se sobrevaloraría la validez del modelo. En general, se aborda mediante la llamada ***validación cruzada*** (*cross-validation*), consistente en partir aleatoriamente la base de datos y utilizar una mitad para la construcción del modelo (base de datos de desarrollo) y la otra mitad para comprobar que los factores de riesgo obtenidos en la primera predicen bien los resultados también en esta segunda base (base de datos de validación). La validación puede repetirse invirtiendo la utilidad de las bases de datos, o bien pueden utilizarse otros métodos como el *bootstrapping* (obtener repetidas muestras aleatorias sin reemplazamiento sobre las que validar el modelo).

Otros tipos de validación de interés son la **validez de contenido** (¿en qué medida los factores de riesgo incorporados representan el universo de factores de riesgos relevantes? ¿son importantes los factores de riesgo no incluidos?), la **validez de constructo** (partir de la relación teórica entre conceptos para examinar la relación empírica entre sus mediciones), y la **validez aparente** (¿a los que tengan que usar el SAR les parecerá válido?.

Cuando los 'resultados' se utilizan para evaluar la calidad de la asistencia, es necesario establecer previamente que los resultados pueden ser atribuidos a la asistencia recibida. El término **validez atribucional** se ha empleado para resaltar este aspecto, y examina el problema de si el ajuste de riesgos efectuado es suficiente para considerar que las variaciones en resultados observadas no están relacionadas con las características intrínsecas de los pacientes, esto es, para que puedan ser atribuidas a otras causas, como la diferencia en la efectividad del tratamiento o en la calidad de la asistencia, siendo uno de los aspectos más discutidos en todos los sistemas de ajuste de riesgos.

## 2.5. Rendimiento de los sistemas de ajuste de riesgos

*La validez final y la utilidad práctica de los ajustes viene dada por su capacidad para capturar las diferencias entre pacientes, problema que es abordado usualmente comparando las predicciones realizadas por los modelos ajustados con los resultados reales, la comentada validez predictiva.*

La **validez final y la utilidad práctica de los ajustes viene dada por su capacidad para capturar las diferencias entre pacientes, problema que es abordado usualmente comparando las predicciones realizadas por los modelos ajustados con los resultados reales, la comentada validez predictiva**. El término ajuste de riesgos sugiere que un modelo proporciona una estimación para cada paciente del resultado esperado en función de la valoración de sus factores de riesgo. En la práctica, esto no siempre es así. En algunos sistemas, transformar puntuaciones de riesgo en resultados esperados es relativamente sencillo. Por ejemplo, el MPM utiliza una serie de hallazgos para asignar a cada paciente una puntuación que es una estimación de la probabilidad de muerte en 30 días para ese caso, y el número de muertes esperadas en un hospital será la suma de sus probabilidades estimadas. En este caso, puede calcularse la probabilidad de muerte para cada paciente a partir de sus datos, y construir una tasa de mortalidad esperada (sumando dichas probabilidades y dividiéndolas por el número de casos) que puede ser comparada con la tasa de mortalidad observada. En otros sistemas, sin embargo, pueden ser necesarios diversos cambios (calibración del modelo y análisis posteriores) para permitir esta transformación.

La modelización multivariante proporciona el marco para el traslado de puntuaciones de riesgo, categorías de riesgo o variables específicas en resultados esperados. Las variables independientes pueden ser cualquiera de los factores de riesgo comentados anteriormente, y la dependiente es el resultado a analizar que, básicamente, puede ser una variable continua (días de estancia, costes) o dicotómica (presencia o no de muerte, complicación, reingreso).

Estadísticamente, suelen emplearse modelos de la familia de los **modelos lineales generalizados**, siendo los más empleados el **modelo de regresión lineal múltiple**, el de **regresión logística** y el de **riesgos proporcionales de Cox**.

El **coeficiente $r^2$** es la medida resumen más utilizada del rendimiento de los modelos multivariantes cuando la variable dependiente es continua y puede ser interpretado como la fracción de la variabilidad total de la variable dependiente que

puede ser explicada por (o atribuida a) diferencias de riesgo entre los casos incluidos en el modelo. En teoría, el coeficiente $r^2$ puede alcanzar un valor máximo de 1 (predicción perfecta) pero en la práctica esta expectativa no es realista.

*El coeficiente de determinación no es apropiado para comparar modelos desarrollados a partir de bases diferentes.*

Las estrategias publicitarias de las firmas que comercializan sistemas de ajuste de riesgos tienden a presentar el $r^2$ de las bases de datos de desarrollo como sinónimo de alta capacidad explicativa sobre la variable dependiente, asumiendo que ésta será la misma al aplicarse sobre otros datos cuando cambios en la dispersión de la variable dependiente cambiarán la capacidad explicativa del modelo. Por ello**, *el coeficiente de determinación no es apropiado para comparar modelos desarrollados a partir de bases diferentes***. A este respecto, debería tenerse en cuenta que cuanto más difieran los datos del modelo de desarrollo de los de la organización que se quiere utilizar el sistema, menor será el valor de $r^2$. Y, en resumen, *los* $r^2$ *obtenidos en algunas bases de datos no son extrapolables a otras bases de datos*.

*Los modelos tienden a estar sesgados infraestimando el riesgo para los casos de alto riesgo y sobreestimándolo en los de bajo riesgo.*

Hay que señalar que **los modelos tienden a estar sesgados infraestimando el riesgo para los casos de alto riesgo y sobreestimándolo en los de bajo riesgo**. Por ejemplo, los pesos de los GDR están sesgados en este sentido, aunque eufemísticamente a este sesgo se le denomine "**compresión**". Esto conlleva que la interpretación de los resultados de un hospital concreto debería tomar en consideración la naturaleza del estándar, tanto por su valor relativo como porque los hospitales con casos más graves verán infraestimados sus riesgos en bases de datos integradas fundamentalmente por casos de bajo riesgo.

*el valor del r2 está en función del número de variables e interacciones introducidas en una base de datos concreta, del entorno del que proviene la muestra,*

En conjunto, **el valor del $r^2$ está en función del número de variables e interacciones introducidas en una base de datos concreta**, **del entorno del que proviene la muestra**, especialmente del tipo de pacientes incluidos, de la dispersión de las variables, **de la inclusión o no de *outliers*** y **de la transformación de variables**. Estos aspectos llevan a que, *aunque el $r^2$ sea un valor numérico, determinar su importancia requiera la evaluación subjetiva de diversos factores*. La valoración de estos factores, sin embargo, en muchos sistemas comerciales no podrá ser realizada por tratarse de datos que forman parte de la "caja negra" considerada secreto comercial y resultará en la imposibilidad de evaluar los posibles sesgos del ajuste, y la dirección de tales sesgos. En todo caso, y **a efectos prácticos,**

*es saludable no creer que los r2 que ofrecen los folletos comerciales de los SAR son un indicador de la calidad del respectivo SAR.*

**es saludable no creer que los $r^2$ que ofrecen los folletos comerciales de los SAR son un indicador de la calidad del respectivo SAR**.

La **regresión logística** es el método más utilizado para modelizar variables dependientes dicotómicas. En este modelo, la variable dependiente es el log natural de la Odds Ratio (OR) del suceso. El acuerdo respecto a cómo valorar los modelos que predicen resultados binarios es menor. En estos modelos lo que se predice es la presentación o no de un suceso en cada caso, asignándole una probabilidad de acaecer. Como en el caso de los modelos de regresión lineal ordinaria o de mínimos cuadrados, las inferencias basadas en la modelización logística son enteramente válidas cuando ciertas asunciones se cumplen. Sin embargo, tanto unos como otros modelos se emplean a menudo a pesar de violarse estas asunciones. Esto es aceptable parcialmente dado que las técnicas han mostrado ser robustas aun 'alejadas' de las condiciones ideales de uso. De hecho, los algoritmos para transformar medidas de riesgo en predicciones se juzgan sobre todo considerando el grado de aproximación entre dichas predicciones y la realidad.

*Un modelo se dice calibrado respecto a un conjunto de datos cuando el promedio de sus predicciones se aproxima al promedio de los resultados reales. Por otro lado, un modelo discrimina sí predice mayores probabilidades del suceso*

La práctica de evaluar predicciones es familiar en el contexto clínico, por ejemplo en la evaluación de pruebas diagnósticas. Cuando el resultado es dicotómico podemos calcular los índices clásicos de sensibilidad, especificidad, valor predictivo positivo, negativo: proporción –o probabilidad- de casos (verdaderos positivos) bien clasificados o detectados por la prueba (sensibilidad); la proporción de no casos (verdaderos negativos) bien clasificados (especificidad); proporción de casos entre los positivos (valor predictivo positivo), etc. Es importante señalar que la sensibilidad y la especificidad de una prueba con resultado dicotómico no se ve afectada por diferencias en la proporción de casos que existen en la población (prevalencia) pero sí su valor predictivo.

En la evaluación de estos modelos se emplean dos criterios: calibración y discriminación. **Un modelo se dice calibrado respecto a un conjunto de datos cuando el promedio de sus predicciones se aproxima al promedio de los resultados reales**. Por otro lado, **un modelo discrimina sí predice mayores probabilidades del suceso** (ej.: muerte) **a los pacientes con el suceso** (los que realmente han muerto) **que a los pacientes sin el suceso**.

Tanto la importancia relativa de uno u otro criterio, como los índices para expresarlos son motivos de controversia. Si bien es deseable una buena calibración, aquellos que priman la discriminación argumentan que una deficiencia en ésta no es subsanable, frente a la posibilidad de recalibrado en los casos de mala calibración. Para ilustrar este argumento imaginemos una población con un 10% de fallecimientos, en la que el modelo asignara a cada persona una probabilidad constante del 0,10, la calibración sería perfecta, pero inútil para diferenciar los casos que fallecerán de aquellos que sobreviven. Por contra, un modelo que sobre esta población asignara un probabilidad de 0,8 a los pacientes que sobreviven, y de 0,9 a los que fallecen, aun cuando, en términos numéricos, sus predicciones sean erróneas, puede emplearse de un modo perfecto para predecir el resultado. Un simple cambio (recalibración) puede resolver el problema de este segundo modelo, pero no existe ninguna corrección posible para el primer caso.

Otros autores sostienen que si el modelo no alcanza una buena calibración, no tiene valor estudiar su capacidad discriminativa. Si nos atenemos al uso, si el modelo se emplea para hacer predicciones individuales, la capacidad discriminativa toma relevancia; sin embargo, si el modelo pretende determinar la incidencia esperada del suceso y contrastarla con la observada, entonces la calibración merece mayor atención.

Sobre la medición de la discriminación, aun cuando no hay consenso sobre la medida más apropiada, el **estadístico C** resulta el más empleado. Varias definiciones son posibles, una ilustrativa parte de considerar todas las parejas posibles que se pueden formar entre los casos que presentan el suceso y aquellos que no lo presentan. El estadístico C equivale a la proporción de parejas para las cuales la probabilidad estimada del suceso (pe. el fallecimiento) es superior en el miembro fallecido de la pareja, que en su correspondiente vivo. Gráficamente, el estadístico C corresponde al área bajo la **curva *Receiver Operating Characteristics*** (ROC), un gráfico en el que se observan todo los pares sensibilidad/especificidad resultante de aplicar diferentes puntos de corte en todo el rango de resultados observados. En el eje de coordenadas se sitúa la fracción de verdaderos positivos y en las abscisas la fracción de falsos positivos (1-sensibilidad). Cada punto constitutivo de la curva corresponde a una decisión sobre punto de corte para la predicción. La curva, en una prueba o modelo con discriminación perfecta, sin solapamiento

*El estadístico C toma valores entre 0,5 (el modelo no funciona mejor que el azar para predecir el resultado) y 1 (el modelo predice perfectamente)*

de resultados o predicciones entre las poblaciones con y sin el suceso, alcanzará el extremo superior izquierdo, donde sensibilidad y especificidad toman su máximo valor en términos del estadístico C. ***El estadístico C toma valores entre 0,5 (el modelo no funciona mejor que el azar para predecir el resultado) y 1 (el modelo predice perfectamente)***.

En cuanto a sus características, cabe subrayar que el estadístico C depende sólo de los rangos de predicciones, no de sus valores. Así un modelo que asigne probabilidades sin ningún solapamiento entre quienes desarrollan y no el suceso, por ejemplo 0,20 para los que fallecen y 0,19 para los que sobreviven, tendrá un estadístico C de 1.

La ***curva de calibración*** compara gráficamente la probabilidad estimada con el resultado real. Para construirla, los casos se dividen en grupos según la probabilidad estimada. El promedio de la probabilidad estimada en cada grupo se compara con la frecuencia real de suceso en el grupo. Lemeshow y Hosmer propusieron aplicar la prueba de chi cuadrado a los datos organizados de modo similar a la curva de calibración. Básicamente consiste en la división de los datos en deciles según nivel creciente de riesgo, y la comparación en cada decil de la frecuencia observada y esperada. El estadístico, obtenido a partir de las desviaciones de los deciles y una estimación de su varianza, se distribuye como una chi cuadrado con 8 grados de libertad.

Todas las cuestiones comentadas relativas al coeficiente de determinación y su uso en la valoración del rendimiento de modelos con variables continuas, son aplicables cuando se plantea su aplicación en modelos con variable respuesta dicotómica. Sin embargo, cuando la variable es dicotómica y ha sido modelizada por un procedimiento diferente al de mínimos cuadrados, existen pocos precedentes que soporten el uso e interpretación del $r^2$. Los investigadores han encontrado **que los modelos que predicen resultados del tipo 0/1 raramente alcanzan $r^2$ del 0,30**. Además se ha observado que su tamaño se ve afectado por la frecuencia del resultado. Como ventaja del $r^2$ se ha observado que resulta más sensible que el estadístico C cuando se quiere comparar el cambio en el rendimiento de sucesivos modelos sobre la misma población.

Las cautelas aplicables a estos modelos e índices de rendimiento

son todas las planteadas anteriormente, en particular sobre las comparaciones de los valores de $r^2$ o C, cuando son obtenidas de diferentes grupos de datos, o existen diferentes criterios de definición de la población, protocolos en la depuración de datos, o están sobre-representados casos difíciles, en definitivas situaciones que hacen variar sustancialmente la variabilidad que el modelo ha de explicar.

## 3. Sistemas de ajuste de riesgos: uso practico

El único sistema de ajuste ampliamente utilizado en la gestión hospitalaria en España son los GDR[3]. En el Sistema Nacional de Salud (SNS) se utilizan, sobre todo, los AP-DRG (3M©) y los CMS-DRG y su uso más extendido es ajustar la comparación de la estancia media (EM) entre hospitales. La EM de un hospital depende del tipo de pacientes que trata y cuando se evalúan comparativamente las EM de diversos hospitales, las diferencias provendrán tanto de una mayor o menor eficiencia de los centros, como de su diferente casuística, aspecto que se controla ponderando la duración de la estancia por el peso de cada GDR atendido. Este uso permite el desarrollo de algunos indicadores. A partir de estos indicadores se pueden revisar los procesos en los GDR en que el hospital se muestra ineficiente para su mejora.

Otros usos incluyen los reembolsos cruzados entre hospitales o comunidades autónomas de los pacientes que son atendidos en otra comunidad u otra área si funcionan bajo presupuesto capitativo. En Cataluña se ha empleado, parcialmente, para el reembolso a los hospitales de la *Xarxa Hospitalaria d'Utilització Publica*.

Las formas típicas de utilización incluyen:

1. Cálculo de la **estancia media ajustada por funcionamiento** (EMf), esto es, la EM que tendría un hospital con los pacientes (GDR) que ha atendido si mantuviera en cada grupo la EM de un patrón de referencia, habitualmente el promedio de una base de datos con diversos hospitales similares.

*3 Las unidades de cuidados intensivos utilizan varios sistemas (APACHE, SAPS, MPM) con fines pronósticos. Igualmente en algunas especialidades se utilizan numerosas escalas habituales en la práctica clínica: el ASA por los anestesistas, el POSSUM en cirugía, la escala de coma de Glasgow y otras son ejemplos de estas escalas.*

2. Una segunda opción, es calcular la **estancia media ajustada por casuística** (EMc), aquella que tendría un hospital si tratará la misma casuística que el patrón de referencia, pero con sus EM en cada GDR. En el primer caso se ofrece una idea de la complejidad del hospital, mientras que en el segundo se controla la casuística como factor de confusión en la prolongación de la estancia media.

El producto más típico en gestión hospitalaria es un listado que incluye todos los GDR, ordenados por volumen, la EMc del hospital y la EMf (estancia media ajustada por funcionamiento), y la estancias de más o de menos que ha tenido un hospital respecto al patrón de referencia. Esto permite incidir en el manejo de aquellos GDR con mayor estancia ajustada que el estándar de la base de datos.

*el GDR es hospitalario –no de un determinado servicio- y que valora el conjunto de cuidados del hospital y su coordinación*

Nótese, que **el GDR es hospitalario –no de un determinado servicio- y que valora el conjunto de cuidados del hospital y su coordinación**. Así, un paciente con un infarto agudo de miocardio puede haber pasado por urgencias, UCI, cardiología, cirugía cardiovascular y reanimación, y haber recibido servicios de radiodiagnóstico, análisis clínicos y otros. Considerar que un posible exceso de estancias depende del servicio que le dio el alta (cardiología) y no del conjunto de la atención es un error muy extendido en la gestión hospitalaria del SNS que debería ser evitado.

Debe hacerse notar que los GDR son una buena aproximación a la eficiencia productiva (contestan a la pregunta de si "producir" un alta hospitalaria resulta mas o menos costoso en uno u otro centro) y tienen su lógica en los sistemas de salud que reembolsan a sus hospitales por alta (pago prospectivo por caso de EEUU). Los GDR no informan de otros aspectos importantes de la eficiencia, como si un centro opera más o menos que otros, o si realiza más o menos ingresos o intervenciones innecesarias. Dado que existen relaciones entre el volumen producido y los costes unitarios (los costes fijos se reparten entre una mayor cantidad de altas y cada una de ellas es mas barata, aunque el agregado implique un mayor gasto), el reembolso por GDR incentiva la utilización de servicios y la eficiencia de los proveedores no debería ser valorada con solo este indicador.

## 4. Sistemas de ajuste de riesgos: perspectivas

Pese a que los sistemas de ajuste tienen importantes limitaciones, la comparación *ajustada* de organizaciones puede servir a los objetivos básicos de:

- ✓ Proporcionar a gestores, compradores, usuarios y responsables de la toma de decisiones, información, ajustada por gravedad, para comparar el coste, la utilización y la calidad de los proveedores de servicios sanitarios,

- ✓ Proporcionar a los proveedores de servicios una referencia de sus costes y calidad respecto a otros proveedores, ajustada por la gravedad de sus pacientes.

- ✓ Proporcionar a los clínicos una referencia sobre la efectividad de un tratamiento/prueba, y sobre que grupos de pacientes tienen una mayor probabilidad de obtener un beneficio o de recibir un daño concreto.

La utilidad de estas comparaciones para los centros sanitarios es obvia. Si una UCI tiene una mortalidad intrahospitalaria que duplica la de otras unidades de intensivos tras su ajuste por el APACHE III o el MPM, este servicio debería revisar sus procesos de atención para identificar posibles problemas de calidad. Este caso es extensivo a muchas otras situaciones (infecciones nosocomiales, complicaciones quirúrgicas, mortalidad intraoperatoria, reingresos por determinadas causas, etc.). Desde otro punto de vista, si los pacientes intervenidos de reparación de hernia inguinal tienen una tasa de recidivadas al año que quintuplica la de las intervenciones con refuerzo de pared ¿valdría la pena realizar un estudio causal? ¿estos resultados se dan en todos los grupos de pacientes? ¿hay grupos en que no hay diferencias?.

Para los compradores de servicios sanitarios las cuestiones de fondo son del tipo: si un hospital tiene mayores costes para los mismos procesos derivados de que su tasa de infecciones nosocomiales triplica la de otros hospitales ¿por qué reembolsar estos costes y financiar los costes de no-calidad? ¿Y los mayores costes derivados del uso innecesario de estancias o pruebas diagnosticas? ¿Porque la población para la que compran servicios debe estar sujeta a un mayor riesgo de mortalidad?. Aunque la utilidad de las comparaciones para los compradores de servicios

deriva -en buena parte- de su capacidad real de compra (en función de la existencia de un mercado de atención sanitaria y de usuarios que eligen uno u otro comprador según el coste de las primas y calidad de la atención, aspectos no tan evidentes en nuestro país), la coincidencia de facto entre comprador y administración pública, con la exigencia de gestión eficaz y atención de calidad, permite extender la utilidad de estos ajustes al Sistema Nacional de Salud.

Las principales limitaciones de esta aproximación derivan de la disponibilidad de las variables esenciales para el ajuste, la calidad de las fuentes de datos y del rendimiento del sistema de ajuste de riesgos empleado. Estas limitaciones no son fácilmente obviables (y tampoco la tentación de desarrollar comparaciones mecanicistas). No debe olvidarse que los resultados a partir de bases de datos son siempre retrospectivos y es poco probable que, incluso en las bases con mayor información, puedan ser ajustadas todas las variables relevantes y tratarlas "como si no hubiera existido selección".

El desarrollo -y utilización práctica- de perfiles de práctica a partir del CMBD es uno de los desafíos que enfrenta el Sistema Nacional de Salud en la medida que la monitorización de resultados ajustados por gravedad es una necesidad para la práctica clínica, la gestión sanitaria y cualquier política de incentivos.

Frente a cualquiera de sus limitaciones, el ajuste de gravedad -incluido costes- a partir del CMBD, tiene las ventajas de su bajo coste relativo y de estar disponible o fácilmente disponible. Utilizados creativamente y no sin cierto sentido crítico, los sistemas orientados hacia los resultados son un instrumento central y de incuestionable utilidad para la toma de decisiones en gestión sanitaria, para la evaluación de la calidad asistencial y para la monitorización de los comportamientos de los proveedores de servicios, lo que no implica que sean el único instrumento, que vayan a desplazar a los indicadores de productividad clásicos o que no puedan ser usados combinadamente con otros indicadores.

En muchos casos, las aproximaciones realizadas a partir de sus datos pueden no ser suficientes para juzgar inequívocamente la calidad o la eficiencia de un hospital o un servicio médico, pero casi siempre pueden ser útiles para identificar problemas que requieran un posterior estudio. La combinación de CMBD

y revisión de historias clínicas (audit médico) puede ser aquí la clave de las mejoras en calidad y eficiencia de la atención hospitalaria.

## Referencias bibliográficas


1. Blumberg MS. Risk adjusting health care outcomes: a methodologic review. ***Med Care Rev. 1986***; 43:351-393.

2. Casas M. Los grupos relacionados de diagnóstico: experiencia y perspectivas de utilización. Barcelona: Masson y SG editores, 1991.

3. Hornbrook MC. Techniques for assessing hospital case mix. ***Annu Rev Public Health. 1985***; 6:295-324.

4. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. ***Ann Intern Med. 1995***;123:763-70. [http://www.annals.org/content/123/10/763.full.pdf+html?sid=2507d74d-1026-491d-a728-3dc57bea2e63]

5. Iezzoni LI, Ed. Risk adjustment for measuring health care outcomes (2nd edition). Ann Arbor, Michigan: Health Administration Press, 1996.

6. Iezzoni LI. The risks of risk adjustment. ***JAMA. 1997***;278:1600-7. [http://jama.ama-assn.org/content/278/19/1600.abstract]

7. Librero J, Peiró S, Ordiñana R. Comparación de resultados, calidad y costes usando bases de datos. Valencia: Instituto Valenciano de Estudios en Salud Pública, 1998.

8. Peiró S, Casas M. Análisis comparado de la actividad y resultados de los hospitales. Situación en España y perspectivas. En: Cabasés JM, Villalbí JR, Aibar C, eds. Invertir en Salud. Prioridades para la salud pública en España-Informe SESPAS, 2002. Valencia: SESPAS y Escuela Valenciana de Estudios para la Salud, 2002; 511-29. [http://www.sespas.es/informe2002/cap24.pdf]

9. Peiró S, Librero J. Evaluación de calidad a partir del conjunto mínimo de datos básicos al alta hospitalaria. ***Rev Neurol.***

***1999***;29:651-61. [http://www.revneurol.com/sec/resumen.php?id=99330#]

10. Peiró S. Los mejores hospitales. Entre la necesidad de información comparativa y la confusión. ***Rev Calidad Asistencial. 2001***;16: 119-30. [http://www.elsevier.es/sites/default/files/elsevier/pdf/256/256v16n02a13028294pdf001.pdf]