Recibido: domingo, 15 abril 2007

economía ::
matematicalia
revista digital de divulgación matemática
Vol. 3, no. 2 (abr. 2007)

Las "ligas" en Sanidad: cómo usamos la Estadística para comparar la calidad de los hospitales y el coste-efectividad de las políticas de salud (*)

Beatriz González López-Valcárcel
Universidad de Las Palmas de GC
e-mail: bvalcarcel@dmc.ulpgc.es
páginas web: http://www.ulpgc.es/index.php?pagina=bvalcarcel.dmc&ver=inicio, http://www.gi.ulpgc.es/aecrss

Carles Murillo Fort
Universidad Pompeu Fabra
e-mail: carles.murillo@upf.edu
página web: http://www.upf.edu/cres/es/qui_som/investigadors1/murillo.html

Pinche sobre una fórmula para ampliarla. Vuelva a pinchar sobre ella para reducirla, o pinche manteniendo pulsada la tecla [shift] para reducir todas las que permanezcan ampliadas.

1. Introducción. El contexto. Los métodos cuantitativos y las políticas públicas.

El diseño y la evaluación de las políticas públicas es objeto de una intensa atención científica. Una política pública viene definida por uno o por varios programas, simultáneos o secuenciales, con objetivos que pueden ser complementarios o no. La problemática estadística que suscitan tales problemas entronca directamente con la posibilidad de experimentación (si se asignan los participantes a los programas aleatoriamente, de una población homogénea, la inferencia sobre los efectos medios de cada programa es inmediata). Dado que en ciencias sociales en general, y en políticas públicas en particular, no se puede experimentar, los métodos cuantitativos se han empeñado en identificar y estimar los resultados mediante diseños semi-experimentales, y mediante el uso de información exógena en estudios observacionales.

Entendemos los métodos cuantitativos en sentido amplio, incluyendo los métodos estadísticos, econométricos, los algoritmos de computación, las matemáticas y la investigación operativa. Todos ellos aportan instrumentos válidos para diseñar y evaluar políticas públicas.

Dentro de los métodos cuantitativos, la microeconometría ha experimentado un gran desarrollo en los últimos años en el campo de la salud (Jones, 2000). El interés de los investigadores, movido por la necesidad de resolver cuestiones pero también por la oportunidad de los datos y por la capacidad de cálculo de los ordenadores, se ha desplazado desde los modelos para datos transversales o para series temporales hacia los modelos para datos longitudinales (de panel) y los modelos para datos jerárquicos. Ahora que se dispone de registros poblacionales y grandes bases de datos, se trabaja con poblaciones completas, y los elementos de error aleatorio de los modelos reflejan errores de medida de las variables y errores de especificación de los modelos (la heterogeneidad no observable), más que errores de muestreo.

La investigación biomédica y en servicios sanitarios tiende a generar datos experimentales y a aplicar una metodología estadística que ya está bastante estandarizada (casos-controles, estudios de cohortes). Hay protocolos metodológicos para los ensayos clínicos (criterios de selección, diseño estadístico del experimento) bien desarrollados. Pero en econometría generalmente seguimos siendo usuarios pasivos de datos no experimentales. No pudiendo experimentar con la población, se experimenta con la muestra, por ejemplo mediante los métodos bayesianos de estimación (*Gibbs sampling, Bootstrapping, MCMC*).

La estandarización metodológica de los métodos cuantitativos en el ámbito de las políticas públicas responde a un nuevo paradigma, el de la Política Basada en la Evidencia (PBE), derivado por contagio de la Medicina Basada en la Evidencia (MBE). Se practica el "*Benchmarking*" para aprender de los demás, comparar y estandarizar reformas y para asignar recursos centralizados a las unidades prestadoras de los servicios asistenciales.

En este artículo presentamos una visión general del uso de métodos cuantitativos para construir rankings que orienten las políticas públicas de salud, y cómo emplearlos. Se hace énfasis en los problemas metodológicos y en las

dificultades para mejorar las políticas a partir de los análisis, más que en lo intrincado de la metodología estadística o en los resultados concretos de los ejemplos que se presentan.

2. El paradigma del *Benchmarking*. Tipos de métodos.

Vivimos en un mundo obsesionado por la *Evidencia* que busca practicar la MBE y hacer Políticas Basadas en la Evidencia (PBE). El *Benchmarking* se practica tanto a nivel de mesogestión de centros -Top20 Hospitales (Garcia-Eroles et al., 2001; Iasist, 2006)- como para los países del mundo. La OMS compara los sistemas sanitarios en cuanto a los recursos que emplean y a los resultados que consiguen. La OCDE ha puesto en marcha sistemas de indicadores, y bases de datos internacionales, para los ámbitos sanitario, educativo y otros. Las Naciones Unidas elaboran y difunden, ya desde hace años, el Índice de Desarrollo Humano que combina varias dimensiones del bienestar -PIB, educación, salud-. Gracias a los indicadores homogéneos internacionalmente de la OCDE (Kelly et al., 2006; OECD, 2003) podremos, por ejemplo, comparar el "tiempo puerta-aguja" en caso de infarto en España con el de los vecinos, y orientarnos por *Benchmarking* para encontrar el camino de las intervenciones eficientes. El *Benchmarking* se usa también ampliamente para diseñar sistemas de incentivos a los gestores; para ajustar fórmulas de pago a proveedores (separar el grano de la paja, no remunerar la parte de los costes que corresponde a la ineficiencia); para elaborar ligas de hospitales o de universidades que informen al público y permitan elegir con criterios de calidad; para ordenar posibles intervenciones públicas según criterios de coste-efectividad (ligas de AVACs), entre otras aplicaciones.

Los métodos para obtener un ranking difieren según los objetivos, los datos y el modelo subyacente (grado de incertidumbre y forma de incorporar, en su caso, juicios de valor para ponderar diferentes objetivos). Desde el punto de vista estadístico es relevante la diferenciación entre modelos deterministas y aleatorios, pero desde la perspectiva de las políticas es más interesante la clasificación de la Figura 1. En ella diferenciamos los ranking que corresponden a uno y a múltiples objetivos, y entre éstos separamos los métodos que ponderan exógenamente los objetivos, a criterio del modelizador, y los que obtienen pesos óptimos de cada uno de los objetivos -u outputs- como resultado, y no como input, del método.

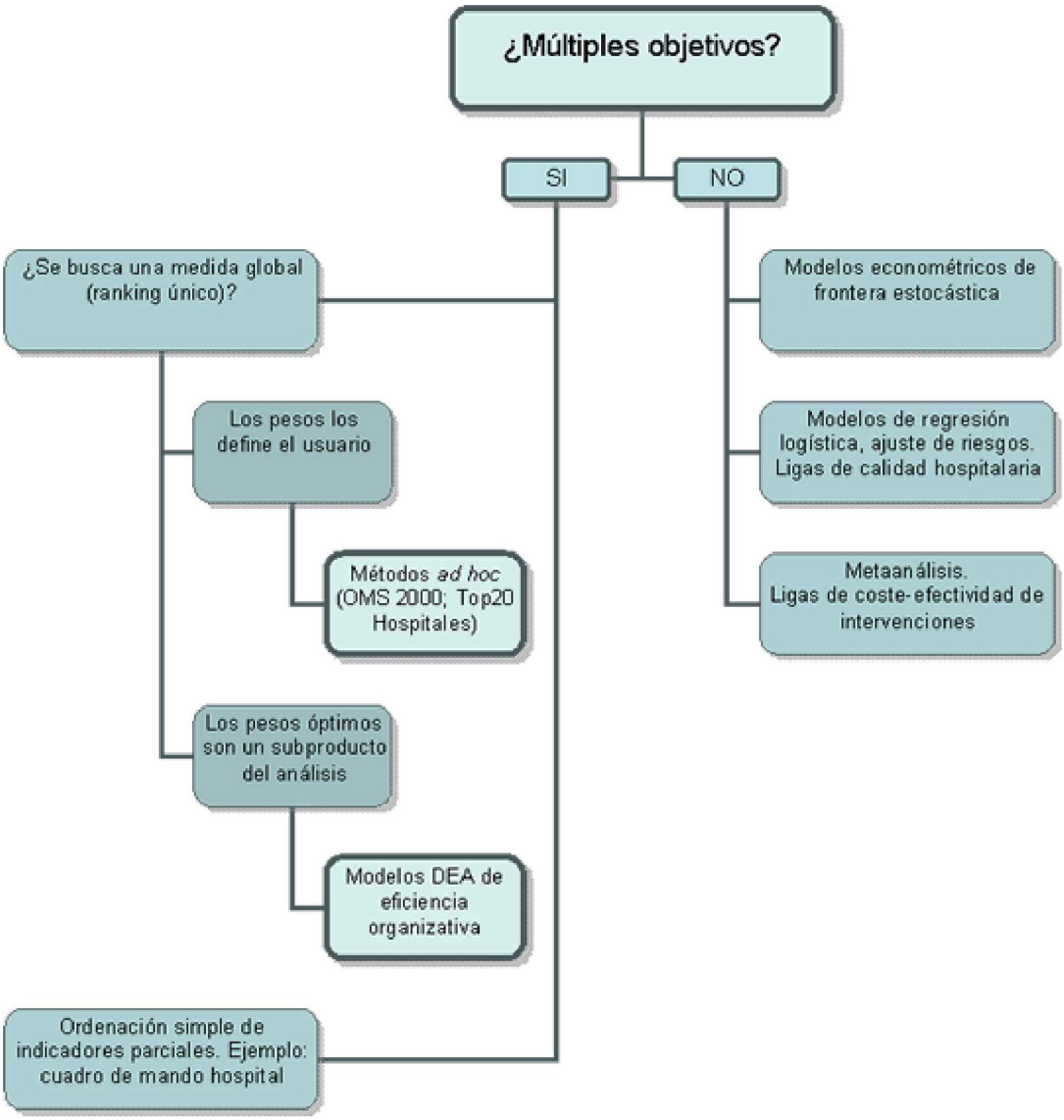

Figura 1. Los ranking en sanidad. Una clasificación de los métodos orientada a las políticas.

3. Ranking unidimensionales de la calidad de los hospitales y del coste-efectividad de las políticas

Los ranking de criterio único incluyen tres tipos de aproximaciones bien diferenciadas:

a) los modelos econométricos de frontera estocástica de producción o de costes, que tratan de medir comparativamente la eficiencia de las unidades productivas, en el marco teórico de la teoría económica de la producción;

b) los modelos de regresión logística y similares, que estiman la calidad (por ejemplo, mortalidad hospitalaria - complicaciones, reingresos...- esperada de cada centro, ajustando por severidad de los casos tratados, y en su caso por otros factores no controlables de entorno);

c) los ranking de intervenciones según coste-efectividad, que tratan de comparar el coste de oportunidad de programas alternativos, para ayudar a los financiadores públicos en la priorización informada de los problemas y de las políticas e intervenciones sobre ellos.

Los **modelos econométricos de frontera estocástica** estiman el grado de ineficiencia productiva de un conjunto de unidades de producción. Estiman la ineficiencia técnica y/o asignativa, con fronteras de producción o de costes, suponiendo que la propia frontera es estocástica, por lo que son modelos con un "error compuesto" de dos componentes. Uno de ellos forma parte de la frontera y representa los errores de medida, omisión de variables y presencia de acontecimientos no predecibles y fuera de control que afectan a la producción. El segundo componente, asimétrico (positivo en modelos de fronteras de costes, y negativo en modelos de fronteras de producción), capta la ineficiencia.

Un modelo de frontera estocástica de producción para un único output (y) de la unidad de producción i-ésima se formularía así:

$$\log(Y_i) = x'_i\,\beta + z'_i\,\gamma + v_i + u_i\ ,$$

donde x' es un vector de k factores de producción (o una función de ellos, por ejemplo, el logaritmo), z' es un vector de variables de entorno no controlables que influyen en la producción, v es un error aleatorio que se supone distribuido con una función de distribución continua de media cero independiente de las x y de las z, y u es la ineficiencia, variable aleatoria negativa o nula, distribuida independientemente de las x y de las z. El modelo de frontera de costes se formularía de forma similar, salvo que $u_i \geq 0$. Estos modelos admiten diversas formas funcionales (González-López-Valcárcel et al., 1996). Aunque se han sofisticado en varias direcciones, no están todavía listos para ser empleados con propósitos prácticos de medir la eficiencia con fines remunerarativos. Su atractivo, hasta ahora, es más académico que profesional (Chirikos et al., 2000; González-López-Valcárcel et al., 1996; Jacobs, 2001; Worthington, 2006). En su aplicación a sanidad, una parte del problema es que están desvinculados de los resultados clínicos y de la calidad asistencial. Los ajustes por calidad son un problema común en la evaluación de servicios públicos de múltiples outputs, donde el mercado no emite señales porque no se comercializan (las universidades, o la policía serían equiparables a los hospitales).

Los ranking unidimensionales de la calidad (clínica) hospitalaria se emplean para emitir señales de calidad y orientar la elección de los pacientes, en un contexto de política sanitaria de autorregulación por el mercado. En Estados Unidos, ya desde hace años, se elaboran y difunden ranking de calidad de hospitales para determinadas intervenciones; es el caso del tratamiento del infarto y de la cirugía cardíaca. Incluso se comparan los resultados de cirujanos individuales (Green et al., 1995). En Nueva York se publicó el ranking de cirugía cardiaca por primera vez en 1997. Algunos de los servicios que quedaron en los últimos puestos cerraron; el resto mejoró su tasa de mortalidad en los años sucesivos (Cutler et al., 2004). Se demostró que el efecto experiencia influye en la mortalidad (Wu et al., 2004).

La elección del criterio de calidad (¿mortalidad, complicaciones, reingresos?) no es trivial. El indicador elegido debería ser fiable, válido, sensible, preciso, con interpretación clínica, útil para la elección informada de hospital por los pacientes y para disponer de estándares de buena práctica los profesionales; fácil de obtener, y difícil de manipular.

El proceso de construcción de un ranking de indicador único de hospitales discurre por las siguientes fases:

1. Seleccionar el indicador (por ejemplo, la tasa de mortalidad intrahospitalaria tras un infarto o la satisfacción de los pacientes atendidos) y el periodo de referencia (por ejemplo, el último año).
2. Calcular el indicador bruto para cada centro, servicio o persona evaluada (i) con los datos registrados de sus n_i pacientes, a partir de una muestra representativa de los pacientes atendidos en el periodo de referencia.
3. Estandarizar el indicador. El proceso de estandarización (ajuste por riesgo, por gravedad del paciente y por condiciones del entorno del hospital independientes de su práctica o, en el caso de centros de atención primaria, por las características personales del médico sujeto a valoración) requiere un modelo estadístico de

ajuste que a su vez introduce incertidumbre adicional, por los errores de medida, de especificación y de estimación.

4. Ordenar los centros, servicios o personas según el indicador estandarizado, calcular los intervalos de confianza y contrastar la significación de las diferencias.

Cada una de estas fases es susceptible de problemas estadísticos que pueden invalidar los resultados, como veremos a continuación.

4. Problemas de interés estadístico en la construcción de los ranking y diseminación de resultados de las ligas

1. ¿Qué queremos comparar? Método de cálculo de los intervalos de confianza y significación de las diferencias

Según el objetivo de las comparaciones, habrá que elegir el método adecuado de contraste de la significación de las diferencias en el valor del indicador. Concretamente, si queremos hacer comparaciones múltiples debemos construir intervalos de confianza múltiples o simultáneos, para lo cual puede usarse un método basado en la desigualdad de Bonferroni.

Otro enfoque clásico es el de los gráficos p, o de fracción defectuosa, que se usan en el control estadístico de calidad de procesos. La diferencia básica es que estos últimos construyen los intervalos de confianza centrados en la tasa p "estándar", única para todos los hospitales, basada en la muestra total, en datos históricos, o en estándares prefijados. La amplitud del intervalo de confianza, pues, sólo depende del tamaño del centro.

2. Problemas de tamaño y efectos de la incertidumbre

Veamos un ejemplo ilustrativo sencillo. Sean dos hospitales, uno es un pequeño hospital con 40 intervenciones anuales de cirugía cardíaca y el otro es un complejo hospitalario que hace 1000 intervenciones al año. Para estimar la tasa de complicaciones (o de mortalidad, o reingresos) se calcula la frecuencia muestral (es decir, la proporción de pacientes intervenidos que sufrieron el suceso adverso). Suponiendo pacientes homogéneos, que no hay sesgo de selección y que ambos hospitales tienen una probabilidad idéntica de que ocurra el suceso adverso (igual calidad), los números de sucesos adversos siguen sendas distribuciones binomiales con parámetros (n,p), donde n es 40 para el hospital pequeño y 1000 para el grande, y p es idéntica.

Dados n y p podemos calcular los intervalos simétricos respecto a p que contienen una proporción determinada $(1-\alpha)$, por ejemplo el 95%, de la masa de probabilidad de la distribución respectiva. Los límites inferior y superior de las tasas del efecto adverso, p_1 y p_2, se determinarán para cada hospital mediante las expresiones siguientes:

$$\Pr(p_1 \le p \le p_2) = 1-\alpha,$$
$$\sum_{x=0}^{np_1} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\alpha}{2},$$
$$\sum_{x=np_2}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\alpha}{2}.$$

Hemos representado en la Figura 2 dichos límites (p_1 y p_2) para ambos hospitales, para una probabilidad del 95% y niveles de calidad (tasas de efectos adversos) entre el 5% y el 25%. Como puede observarse, los intervalos son mucho más amplios para el pequeño hospital que para el grande. Además, las distancias se acrecientan para sucesos adversos más probables (al aumentar p).

Si suponemos que el número de pacientes intervenidos (n_i) en cada hospital (i) es un parámetro fijo, el número de fallecidos en cirugía cardiaca es una binomial $B(n_i,p_i)$. El parámetro de interés es p_i (tasa de fracasos, por mil intervenciones). Su IC dependerá del número de casos (n_i) y de la probabilidad de éxito y de fracaso (p_i y $1-p_i$). Cuanto menor sea el tamaño o nivel de actividad del un hospital, menor será la potencia de los contrastes sobre su calidad. Los hospitales pequeños tienen intervalos de confianza amplios, compatibles con un amplio rango de calidad asistencial. En la práctica, pues, tendrán mayor probabilidad de ocupar los puestos en ambos extremos del ranking. Si las listas no informan sobre intervalos de confianza, están contaminadas por el efecto tamaño y las comparaciones pueden no ser relevantes.

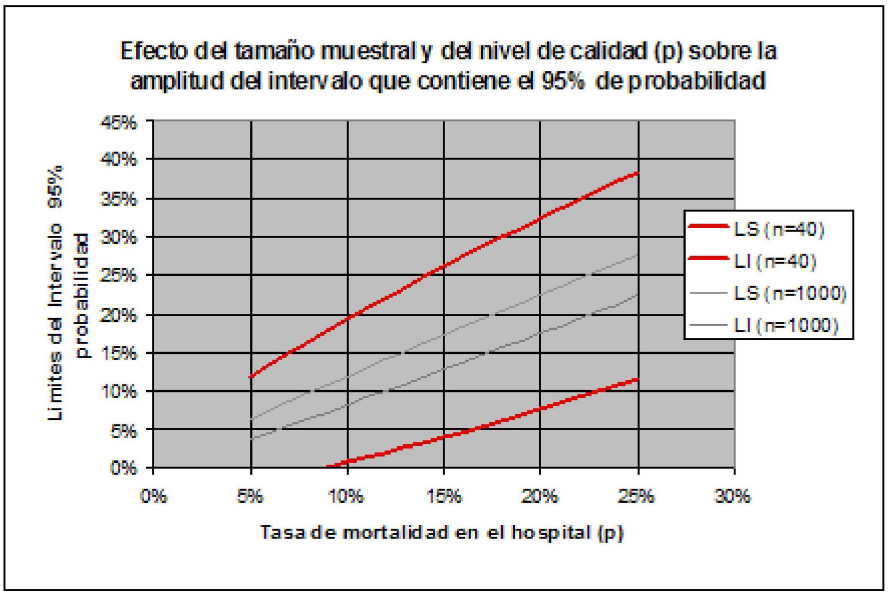

Figura 2. Los límites se han calculado aproximando la distribución binomial a la normal.

3. Sesgo de selección y factores de confusión

Pero no sólo hay contaminación por el efecto tamaño. Es más grave el sesgo de selección y la heterogeneidad no observable entre pacientes (factores de confusión). Por ejemplo, la mortalidad de los pacientes ingresados con infarto de miocardio puede depender más del estado de salud en que llegan al hospital que de la calidad de la atención sanitaria que éste les presta. Puesto que hay sesgo de selección (los mejores hospitales reciben a los pacientes más graves), si no se ajusta por gravedad se encuentra mayor mortalidad asociada a los mejores hospitales.

Para ajustar por dichos factores se emplean modelos de regresión, generalmente logística, que predicen la mortalidad de diferentes servicios hospitalarios homogéneos (por ejemplo, cirugía cardiaca) ajustando por gravedad y por otras variables de confusión, y ordenan a los hospitales en orden creciente o decreciente de calidad. Una solución elegante al sesgo de selección -hospitales más prestigiosos reciben pacientes más graves- proviene de los modelos econométricos bayesianos, que emplean la información exógena sobre distancia del domicilio del paciente a los distintos hospitales para estimar el sesgo de selección (Geweke et al., 2001). Mediante métodos de estimación bayesiana, por simulación (MCMC) Geweke y colaboradores obtienen distribuciones *a posteriori* de calidad y consiguen comparar la calidad de 114 hospitales americanos en el tratamiento de la neumonía, comparaciones bilaterales y entre grupos de hospitales. El sesgo de selección no es independiente del efecto tamaño, y esto complica las cosas: hospitales más grandes y por tanto con mayor experiencia tienden a admitir los casos más complicados o de mayor riesgo, y por eso presentarán mayores tasas brutas de mortalidad.

Así pues, un método de ajuste por riesgo, severidad o gravedad es esencial para que las ligas sean creíbles.

4. Problemas de datos

La precariedad de los datos y los errores en las variables pueden invalidar por completo los resultados. Estos errores pueden deberse a problemas de los sistemas de información estadística, pero también ocurre que los propios hospitales manipulan a su favor la información que difunden, o que actúan para optimizar el comportamiento del indicador de la calidad y no la calidad misma. La calidad de los datos puede ser muy variable entre centros (Wolff y Helminiak, 1996), lo que dificulta todavía más la inferencia.

5. Problemas de difusión de los resultados

Metodología compleja, los resultados que se difunden se reducen a una tabla, sólo en letra pequeña se aclaran las limitaciones metodológicas. La simplificación induce a interpretaciones erróneas sobre la significatividad de las diferencias en los puestos de la clasificación, sobre las distancias entre las posiciones de los centros...

5. Las ligas de coste-efectividad y los problemas relacionados con los metaanálisis

Los ranking de coste-efectividad comparan los costes por Año de Vida Ganados Ajustado por Calidad (AVAC) mediante intervenciones o programas de salud alternativos. Los costes se convierten, con propósitos comparativos, a unidades monetarias homogéneas (euros de 2001, por ejemplo), pero la efectividad -ganancia de salud- se puede medir de muchas maneras, dependiendo del programa que estemos evaluando. Unos salvan vidas, otros alargan la esperanza de vida, otros mejoran su calidad. Para dotarse de una métrica común, una unidad de medida de la efectividad que incorpore las dos dimensiones, cantidad y calidad de vida, se han definido los AVAC, que intentan valorar la utilidad de

diferentes estados de salud y así reducir a una dimensión la salud ganada. La ratio coste-efectividad es un criterio de eficiencia que se define como lo que cuesta conseguir una mejora unitaria de salud: cuánto cuesta ganar un AVAC.

Se han publicado varias de esas listas ordenadas en diversos textos, y la Universidad de Harvard mantiene una página web con la ordenación de intervenciones sanitarias, clasificadas por problemas de salud[1]. Las ligas de coste-efectividad han gozado de la atención de los investigadores y de los políticos porque cubren una necesidad y por su aparente simplicidad. No obstante, sufren limitaciones que previenen contra su uso indiscriminado y simplista. No han de ser tomadas como recetas a aplicar ciegamente en las decisiones de financiar o no los tratamientos médicos y tecnologías sanitarias, sino como una orientación para no actuar a ciegas. Muchos de los problemas son similares a los de las ligas de hospitales (necesidad de ajustes, gestión científica de la incertidumbre, contrastes de comparaciones múltiples).

Las limitaciones de las ligas AVAC son de tres tipos:

1. Limitaciones inherentes a cada uno de los estudios coste-utilidad y al propio instrumento AVAC.
2. Problemas relacionados con el metaanálisis.
3. Problemas para tomar decisiones de política sanitaria sobre la base de la información de la liga.

Nos centramos en el segundo tipo de problemas, que son más estadísticos.

Las tablas que ordenan las intervenciones según coste-efectividad se elaboran mediante *metaanálisis* que comparan (o "agregan") resultados de múltiples estudios de evaluación realizados en diferentes espacios y tiempos. Suele haber *problemas de selección* de los estudios a incluir (falta de criterio científico explícito de inclusión; no se garantiza que los estudios de base hayan cumplido los estándares protocolizados de calidad de las evaluaciones económicas; se comparan estudios referidos a distintos momentos tecnológicos y a distintos países, que pueden diferir en costes y en efectividad).

También hay un serio problema de sesgo de selección derivado del sesgo de publicación: es mucho más probable que se publiquen resultados favorables a los nuevos tratamientos y que los estudios financiados por la industria resulten en recomendaciones favorables.

Además, se excluyen de la liga los numerosos estudios que no emplean el AVAC como medida de utilidad, que están desigualmente distribuidos entre enfermedades, según un análisis comparativo de 455 trabajos de evaluación económica incluidos en la base de datos Health Economic Evaluations Database (HEED) (Anell et al., 2000).

Otra fuente de sesgo de selección es que las políticas de salud pública, basadas en intervenciones comunitarias para cambiar estilos de vida individuales, no se suelen someter a evaluación económica tanto como los tratamientos médicos individuales, animados y financiados por la industria.

Otro problema fundamental es que sólo se comparan promedios, sin considerar la incertidumbre implicada en los estudios originales ni los respectivos análisis de sensibilidad. Los intervalos de confianza de dos intervenciones pueden superponerse pero la tabla no informa al respecto. Aunque la Medicina se Base en la Evidencia (MBE), esa evidencia es fragmentaria, incompleta e incierta. Hay incertidumbre diagnóstica y sobre la efectividad de muchos tratamientos médicos a corto y a largo plazo, y sobre los costes.

Incluso salvando las limitaciones anteriores, la ratio coste por AVAC ganado no debe ser el único criterio de decisión o priorización, porque la disposición social a pagar por ganar salud depende de quiénes sean los beneficiarios y de cómo se distribuyan los AVACs ganados. Las tablas nada dicen sobre cómo se concentran las ganancias de salud y quiénes son los beneficiarios. A sus efectos es lo mismo alargar 50 años la vida de una sola persona que alargar un año la vida de cada una de las 50 que sufren una enfermedad. Para la sociedad no es lo mismo (Nord et al., 1999).

6. Ranking multidimensionales. ¿Cómo ordenar cuando las políticas tienen múltiples objetivos?

Generalmente las políticas tienen objetivos múltiples, complementarios o no. La cuestión es si ponderar o no ponderar. Cuanto más local sea el nivel de la gestión, se requiere mayor desagregación de indicadores, y monitorizar los objetivos uno a uno. Frente a la ventaja de la simplicidad, los indicadores unidimensionales no ayudan a definir estrategias globales de regulación. Para asignar fondos entre unidades con las mismas reglas del juego, el regulador central requerirá una medida sintética de *performance* que compense las pérdidas respecto a la media de unas dimensiones con las ganancias de otras dimensiones. Cuanto más centralizado es el organismo que necesita el ranking, más necesidad suele haber de sintetizar las múltiples dimensiones del éxito en un único indicador global de "*performance*" o eficiencia organizativa. El gerente de un hospital o de una universidad, por ejemplo, necesita en su cuadro de mando indicadores separados de productividad y actividad en cada uno de los servicios y unidades, para tomar medidas específicas que resuelvan los problemas. Por ejemplo, necesitará conocer la lista de espera de cada servicio y prueba diagnóstica. En cambio, la dirección regional del sistema de salud y la autoridad sanitaria central demandarán un indicador global de eficiencia.

Los métodos que buscan el "orden global" perfecto de un grupo de unidades (sean países, hospitales o problemas de salud) son complejos y requieren ponderaciones de los diferentes outputs o dimensiones. Esas ponderaciones pueden ser establecidas *ex ante* por el investigador, o bien resultar del propio análisis.

7. Las experiencias con ponderaciones preestablecidas

El ranking de sistemas de salud del mundo del informe 2000 de la OMS (WHO, 2001) resultó de la ponderación *ad hoc* (con criterios OMS) de las dimensiones que, según un panel de expertos, definen la *performance* de los sistemas de salud. Ha recibido muchas, y acertadas, críticas porque uniformiza con supuestos objetivos comunes a los países del mundo que pueden tener, y tienen, sus propios objetivos distintos del resto (Williams, 2001). Confunden, pues, heterogeneidad e ineficiencia (Greene, 2003). Sin embargo, tiene el mérito del pionero, cuya estela han seguido muchos trabajos, organismos e investigadores.

En septiembre de 2001, el NHS británico empezó a publicar, entre otros, un ranking de los consorcios que proveen asistencia hospitalaria aguda, clasificándolos con estrellas de excelencia, entre cero y tres según un amplio conjunto de indicadores de cumplimiento de objetivos clínicos, de gestión, de calidad, de tiempos de espera y otros que interesan al paciente. En 2004 el Ministerio de Salud encargó a un grupo de investigación del *Centre for Health Economics* de la Universidad de York y del *National Institute for Economic and Social Research* el diseño de una nueva metodología para medir la productividad y resultados globales del NHS. El informe, publicado en 2005 (Dawson et al., 2005), propone un índice global de output del NHS ajustado por calidad (el "*value weighted output index*"), que se calcularía con la siguiente fórmula:

$$I_{yt}^{xq} = \frac{\sum_j x_{jt+1} \sum_k \pi_{kt} q_{kjt+1}}{\sum_j x_{jt} \sum_k \pi_{kt} q_{kjt}},$$

donde x_{jt} es la cantidad del output j en el periodo t, q_{kjt} es la cantidad del atributo, resultado o característica k que produce la unidad j, y $\prod_{kt}$ es el valor social marginal de ese atributo. El índice requiere datos de actividad (x) y de resultados en términos de salud y satisfacción de los pacientes (q), por ejemplo tiempos de espera, que afectan a la utilidad, o la incertidumbre asociada a esa espera. Además, es preciso tener datos de la valoración social de cada uno de esos resultados ($\prod$), que son las ponderaciones incorporadas en la fórmula.

Como para su cálculo habría que disponer de datos inexistentes hoy por hoy en el sistema, proponen también, a modo de solución provisional, índices basados en costes (CWOI), que incorporan distintas combinaciones de cambios en supervivencia, efectos sobre la salud, tiempos de espera, satisfacción de los pacientes, reingresos y MRSA.

El informe británico abre nuevos horizontes y cambia perspectivas, sobre todo porque es paciente-céntrico. La unidad de medida es el paciente tratado por el NHS (no el proveedor ni el comprador de los servicios); y la calidad se define en función de las dimensiones de los resultados que los pacientes valoran.

Una aproximación alternativa consiste en especificar los múltiples objetivos del sistema de salud como variables dependientes en un modelo multivariante de logros, permitiendo correlaciones entre ellos. Siguiendo este enfoque, una aplicación reciente para las autoridades sanitarias del Reino Unido (Hauck et al., 2006) emplea un modelo multinivel multiecuacional con 13 objetivos, en el que las unidades de nivel I son los distritos electorales y las de nivel II las autoridades sanitarias.

Las reglas de priorización, para ordenar distintos problemas dentro de un país, y las intervenciones políticas dirigidas a mejorarlos, constituyen otro ámbito de aplicación del "método del ranking". En España es notable a este respecto el proyecto de investigación sobre identificación y priorización de necesidades de salud, integrado en la red de Investigación en Resultados y Servicios de Salud (IRYSS).

8. Métodos que obtienen ponderaciones como resultado del análisis: los modelos DEA

Entre los métodos que evalúan la *eficiencia organizativa global* de unidades prestadoras de servicios homogéneos no comercializados, donde no hay precio de mercado que oriente sobre el "valor" de los bienes o servicios provistos, el Análisis Envolvente de Datos (*Data Envelopment Analysis*, DEA) tiene gran aceptación y es un punto de encuentro entre académicos y gestores. Los resultados sirven para diseñar incentivos, pagar a proveedores o hacer un seguimiento temporal del desarrollo organizativo. Se aplican sobre todo a servicios públicos que producen múltiples outputs sin precio de mercado que refleje su valoración relativa: educación, sanidad, juzgados. En sanidad, se han aplicado modelos DEA para tantear la posibilidad de medir la eficiencia comparativa de Autoridades Sanitarias, de hospitales y de Equipos de Atención Primaria, también en España (Puig-Junoy et al., 2004). Generalmente, se plantean como métodos sustitutivos de los modelos econométricos de frontera estocástica. Como ellos, todavía tendrán que resolver algunas cuestiones metodológicas pendientes y someterse al tribunal de los hechos para demostrar que son instrumentos útiles, precisos y robustos, para orientar el reparto de fondos entre las Unidades. Cómo gestionar científicamente la incertidumbre, cómo ajustar por factores "inevitables" condicionantes del entorno fuera de control de

los gestores, o cómo tener en cuenta la dinámica de la eficiencia, son algunas de las cuestiones pendientes (Smith et al., 2005).

La ciencia todavía no garantiza que se obtengan medidas objetivas de eficiencia organizativa independientes de los artefactos estadísticos. Hay demasiada sensibilidad de los ranking a pequeños cambios en el modelo o en los datos. Los resultados varían radicalmente, según se ajuste o no por factores que en principio se pueden considerar de entorno fuera de control para los gestores de los hospitales (Street, 2003). También suele ser difícil argumentar los resultados (justificar el ranking resultante con la lógica de la organización, y no con la lógica matemática). Significación estadística y significación socio-política ni son sinónimos ni siempre concuerdan.

9. Síntesis y conclusión

El diseño y la evaluación de las políticas públicas es objeto de una intensa atención científica. Los métodos cuantitativos (entendidos en sentido amplio) contribuyen a este fin con diversos instrumentos. La estandarización metodológica en curso responde a un nuevo paradigma, el de la Política Basada en la Evidencia (PBE), derivado por contagio de la Medicina Basada en la Evidencia (MBE). Se practica el "*Benchmarking*" tanto a nivel de mesogestión de centros como entre países del mundo, para aprender de los demás, comparar y estandarizar reformas y para asignar recursos centralizados a las unidades prestadoras de los servicios asistenciales.

En este artículo presentamos una visión general del uso de métodos cuantitativos para construir *rankings* que orienten las políticas públicas de salud, y cómo emplearlos. Los métodos para obtener un ranking difieren según los objetivos, los datos y el modelo subyacente (grado de incertidumbre y forma de incorporar, en su caso, juicios de valor para ponderar diferentes objetivos).

Los ranking de criterio único incluyen los modelos econométricos de frontera estocástica, que miden comparativamente la eficiencia de las unidades productivas en el marco teórico de la teoría económica de la producción; los modelos de regresión logística para estimar la calidad, que ajustan por factores de confusión y riesgos; y los ranking de programas públicos según coste-efectividad.

Generalmente las políticas tienen objetivos múltiples, complementarios o no. Los métodos que integran múltiples objetivos en un indicador global son más complejos que los de criterio único. La elección del método debe estar en función de su uso. Cuanto más local sea el nivel de la gestión, se requiere mayor desagregación de indicadores, y monitorizar los objetivos uno a uno. Los ranking basados en criterios múltiples buscan el "orden global" de un grupo de unidades (sean países, departamentos públicos o problemas sociales), y requieren ponderaciones, establecidas *ex ante* por el investigador, o bien resultar del propio análisis, como en el Análisis Envolvente de Datos (DEA). El análisis DEA es un punto de encuentro entre académicos y gestores, para diseñar incentivos, pagar a proveedores o hacer un seguimiento temporal del desarrollo organizativo.

Estamos asistiendo a importantes avances metodológicos, incluyendo los que provienen de la estadística bayesiana y los métodos para datos de panel y para datos jerárquicos.

En cualquier caso, debe imponerse la prudencia al leer los resultados de las "ligas", que a veces se presentan de forma excesivamente simplista, haciendo un uso inadecuado de la estadística, sin informar de los intervalos de confianza, o sin el necesario ajuste por riesgos. La evidencia disponible sugiere que publicación de estos datos contribuye a mejorar la calidad de los servicios cuando tienen un nivel basal de calidad muy mejorable. La ciencia todavía no garantiza que se obtengan medidas objetivas de eficiencia organizativa independientes de los artefactos estadísticos. Hay demasiada sensibilidad de los ranking a pequeños cambios en el modelo o en los datos. Los resultados varían radicalmente, según se ajuste o no por factores que en principio se pueden considerar de entorno fuera de control para los gestores. También suele ser difícil argumentar los resultados (justificar el ranking resultante con la lógica de la organización, y no con la lógica matemática). Significación estadística y significación socio-política ni son sinónimos ni siempre concuerdan.

Referencias

A. Anell, A. Norinder: Health outcome measures used in cost-effectiveness studies: a review of original articles published between 1986 and 1996. *Health Policy* 51(2) (2000), 87-99.

J.D. Angrist, A.B. Krueger: *Empirical strategies in labor economics*, 1998.

J.D. Angrist, A.B. Krueger: *Instrumental variables and the search for identification from supply and demand to natural experiments*. National Bureau of Economic Research, Cambridge, MA, 2001.

T. Arnesen, M. Trommald: Roughly right or precisely wrong? Systematic review of quality-of-life weights elicited with the time trade-off method. *J. Health Serv. Res. Policy* 9(1) (2004), 43-50.

X. Badia, M. Roset, M. Herdman: Inconsistent responses in three preference-elicitation methods for health states. *Soc. Sci. Med.* 49(7) (1999), 943-950.

H. Bleichrodt, C. Herrero, J.L. Pinto: A proposal to solve the comparability problem in cost-utility analysis. *J. Health Econ.* 21(3) (2002), 397-403.

H. Bleichrodt, M. Johannesson: Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *J. Health Econ.* 16(2) (1997), 155-175.

A.D. Brown, M.J. Goldacre, N. Hicks, J.T. Rourke, R.Y. McMurtry, J.D. Brown, G.M. Anderson: Hospitalization for ambulatory care-sensitive conditions: a method for comparative access and quality studies using routinely collected statistics. *Can. J. Public Health* 92(2) (2001), 155-159.

T.N. Chirikos, A.M. Sear: Measuring hospital efficiency: a comparison of two approaches. *Health Serv. Res.* 34(6) (2000), 1389-1408.

D.M. Cutler, R.S. Huckman, M.B. Landrum, National Bureau of Economic Research: *The role of information in medical markets. An analysis of publicly reported outcomes in cardiac surgery.* National Bureau of Economic Research, Cambridge, MA, 2004.

D. Dawson, H. Gravelle, M. O'Mahony, A. Street, M. Weale, A. Castelli, R. Jacobs, P. Kind, P. Loveridge, S. Martin, P. Stevens, L. Stokes: Developing New Approaches to Measuring NHS Outputs and Activity. *CHE Research Paper* 6 (2005).

P. Dolan, R. Shaw, A. Tsuchiya, A. Williams: QALY maximisation and people's preferences: a methodological review of the literature. *Health Econ.* 14(2) (2005), 197-208.

M. Frölich: Programme Evaluation with Multiple Treatments. *Journal of Economic Surveys* 18(2) (2004), 181-224.

L. García-Eroles, A. Arias, M. Casas: Los Top20 2000: objetivos, ventajas y limitaciones del método. *Rev. Calidad Asistencial* 16(2) (2001), 107-116.

J. Geweke, G. Gowrisankaran, R.J. Town: *Bayesian inference for hospital quality in a selection model.* National Bureau of Economic Research, Cambridge, MA, 2001.

H. Goldstein: *Multilevel statistical models.* E. Arnold, London, 2003.

H. Goldstein, D.J. Spiegelhalter: *League tables and their limitations: statistical issues in comparisons of institutional performance*, 2003.

B. González-López-Valcárcel, P. Barber: Changes in the efficiency of Spanish public hospitals after the introduction of Program-Contracts. *Investigaciones Económicas* 20(3) (1996), 377-402.

J. Green, N. Wintfeld: Report cards on cardiac surgeons: assessing New York State's approach. *N. Engl. J. Med.* 332(18) (1995), 1229-1232.

W. Greene: *Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems*, 2003.

M.K. Gusmano, V.G. Rodwin, D. Weisz: A new way to compare health systems: avoidable hospital conditions in Manhattan and Paris. *Health Aff. (Millwood)* 25(2) (2006), 510-520.

K. Hauck, A. Street: *Performance Assessment in the Context of Multiple Objectives: A Multivariate Multilevel Analysis.* Center for Health Economics, University of York, 2006.

J.H. Hibbard, J. Stockard, M. Tusler: Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff. (Millwood)* 22(2) (2003), 84-94.

Iasist. Top20: *Benchmarks para la Excelencia 2005*, http://www.iasist.com/top20/Top20_2005/Resultados/publicacion.pdf [31-03-2006].

R. Jacobs: Alternative Methods to Examine Hospital Efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis. *Health Care Management Science* 4(2) (2001), 102-115.

M. Johannesson: QALYs, HYEs and individual preferences-a graphical illustration. *Soc. Sci. Med.* 39(12) (1994), 1623-1632.

A. Jones: *Health Econometrics.* En A.J. Culyer, J.P. Newhouse (eds.): *North-Holland Handbook of Health Economics.* Elsevier, 2000.

N. Kahn III, T. Ault, H. Isenstein, L. Potetz, S. Van Gelder: Snapshot Of Hospital Quality Reporting And Pay-For-Performance Under Medicare. *Health Affairs* 25(1) (2006), 148-162.

E. Kelly, J. Hurst: Health Care Quality Indicators Project. Conceptual Framework Paper. *OECD Health Working Papers*, 2006.

M. Lubrano: Density inference for ranking European research systems in the field of economics. *Journal of Econometrics* 123(2) (2004), 345-369.

D.E. Nelson, D.W. Fleming, J. Grant-Worley, T. Houchen: Outcome-based management and public health: the Oregon Benchmarks experience. *J. Public Health Manag. Pract.* 1(2) (1995), 8-17.

E. Nord, J.L. Pinto, J. Richardson, P. Menzel, P. Ubel: Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Econ.* 8(1) (1999), 25-39.

OECD: *A disease-based comparison of health systems: what is best and at what cost?* Paris, 2003.

J. Puig-Junoy, V. Ortun: Cost efficiency in primary care contracting: a stochastic frontier cost function approach. *Health Econ.* 13(12) (2004), 1149-1165.

C. Sanderson, J. Dixon: Conditions for which onset or hospital admission is potentially preventable by timely and effective ambulatory care. *J. Health Serv. Res. Policy* 5(4) (2000), 222-230.

P. Smith, A. Street: Measuring the efficiency of public services: the limits of analysis. *J.R. Statist. Soc. A* 168(2) (2005), 401-417.

T.A.B. Snijders, R.J. Bosker: *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* Sage Publications, London, 1999.

A. Street: How much confidence should we place in efficiency estimates? *Health Econ.* 12(11) (2003), 895-907.

US Dept. Health and Human Services HHS: *Hospital Compare. A quality tool for adults, including people with Medicare*, http://www.hospitalcompare.hhs.gov [2006].

A. Vass: Doctors urge caution in interpretation of league tables. *BMJ* 323(7323) (2001), 1205.

M. Vera-Hernández: Evaluating health interventions without experiments. *Gac. Sanit.* 17(3) (2003), 238-248.

WHO: *The world health report 2000 - health systems: improving performance*, 2001.

A. Williams: Science or marketing at WHO? A commentary on 'World Health 2000'. *Health Econ.* 10(2) (2001), 93-100.

N. Wolf, T.W. Helminiak: Nonsampling measurement error in administrative data: implications for economic evaluations. *Health Econ.* 5(6) (1996), 501-512.

A. Worthington: *An empirical survey of frontier efficiency measurement techniques in healthcare services*, 2006.

C. Wu, E.L. Hannan, T.J. Ryan, E. Bennett, A.T. Culliford, J.P. Gold, O.W. Isom, R.H. Jones, B. McNeil, E.A. Rose, V.A. Subramanian: Is the impact of hospital and surgeon volumes on the in-hospital mortality rate for coronary artery bypass graft surgery limited to patients at high risk? *Circulation* 110(7) (2004), 784-789.

---

[1] http://www.tufts-nemc.org/cearegistry.

Sobre los autores

Beatriz González López-Valcárcel es catedrática de Métodos Cuantitativos en Economía y Gestión de la Universidad de Las Palmas de GC y co-directora del Master Interuniversitario en Economía de la Salud y Gestión Sanitaria, de las Universidades de Las Palmas y La Laguna. Investiga en Economía de la Salud, principalmente: uso de medicamentos, recursos humanos y mercados laborales, economía del tabaco, y nuevas tecnologías y salud. Es autora de decenas de publicaciones en libros y revistas científicas, nacionales e internacionales. Cuenta con cuatro sexenios de investigación reconocidos por la CNEAI y es consultora internacional en México, Brasil, Argentina, Chile, Costa Rica y Mozambique.

Carles Murillo Fort es catedrático de Economía Aplicada de la Universidad Pompeu-Fabra, investigador permanente del Centro de Investigación en Economía y Salud y Director del Observatorio de Relaciones con Latinoamérica de dicha universidad. Sus principales intereses investigadores se enmarcan en Economía de la Salud, Gestión de Empresas de Servicios, Comercio Internacional y Econometría Aplicada.

matematicalia
revista digital de divulgación matemática

(*) Este texto es un resumen de los dos siguientes artículos: B. González: Métodos cuantitativos y "*Benchmarking*": utilidad para orientar las políticas públicas. *Ekonomiaz* 60(1) (2005), 123-139; C. Murillo, B. González: Potencialidades y limitaciones de los ranking de calidad de proveedores sanitarios. *Revista de Estudios de Economía Aplicada* 24(3) (2006), 777-788.

Cerrar ventana