# *Métodos cuantitativos y «benchmarking»: su utilidad para orientar las políticas públicas*



Un área en la que los métodos cuantitativos, entendidos en sentido amplio, han demostrado ser útiles por su contribución al diseño y evaluación de las políticas públicas es la elaboración de «ligas» o ranking, bien sean de criterio único o múltiple. Investigadores académicos y responsables de las políticas públicas están alcanzando puntos de encuentro en su aplicación a países, a unidades de producción y en la priorización de problemas. El artículo ofrece una tipología de los métodos y revisa su aplicabilidad, limitaciones y retos para el futuro.

*Metodo kuantitatiboak, zentzu zabalean hartuta, baliagarriak izan dira arlo batean, politika publikoak diseinatzeko eta ebaluatzeko egin duten ekarpena dela-eta, hain zuzen, irizpide bakarreko edo askotako «ligak» edo rankingak egiteko arloan. Ikertzaile akademikoak eta politika publikoen arduradunak elkarguneak aurkitzen ari dira horiek herrialdeei eta ekoizpen-unitateei aplikatzeari dagokionez, arazoen lehentasunak ezartze aldera. Artikuluak metodoen tipologia eskaintzen du, eta haiek aplikatzeko modua eta haien mugak eta etorkizunerako erronkak berrikusten ditu.*

The elaboration of ranking or league tables is a major area in which quantitative methods contribute to public policies design and evaluation. These rankings may be based on one or on multiple criteria. Academic researchers and those responsible for public policies are collaborating in the application of these methods to countries, departments and to problems priorization. This paper classifies the methods and reviews their applicability, their limitations and the challenges for the future.

Beatriz González López-Valcárcel
*Universidad de Las Palmas de GC*



ÍNDICE



Palabras clave: evaluación, métodos cuantitativos, política pública

N.º de clasificación JEL: C10, I18, H5

## 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, particularmente en los tiempos que corren, en los que el concepto de coste de oportunidad ha trascendido más allá de la academia y las sociedades son conscientes de la limitación de los recursos públicos y de la necesidad de elegir. Una política pública viene definida por uno o por varios programas, simultáneos o secuenciales, con objetivos que pueden ser complementarios o no. Está surgiendo un cuerpo de literatura científica dedicada a evaluar las políticas, con estrategias e instrumentos microeconométricos de identificación y estimación (Frölich, 2004). La problemática estadística que suscitan tales problemas entronca directamente con la posibilidad de experimentación (si se asignan los participantes a los programas aleatoriamente, sobre 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 tanto como los estadísticos desearían, por problemas políticos —asignaciones condicionadas por elementos contextuales—, por no perder eficiencia —conviene dar la beca al candidato con mejores expectativas de rendimiento— o por restricciones éticas —para probar la efectividad de un tipo de cirugía no es ético crear un grupo de control formado por pacientes que se so-

meten a falsa cirugía o cirugía placebo—, los métodos cuantitativos se han empeñado en identificar y estimar los resultados de las políticas públicas mediante diseños semi—experimentales, y mediante el uso de información exógena en estudios observacionales.

Se aplican métodos cuantitativos a todo tipo de políticas públicas. En este artículo, presentamos ilustraciones y casos provenientes mayoritariamente del ámbito de las políticas de salud, aunque hacemos referencia a otros ámbitos a los que la metodología puede extenderse de forma inmediata y sencilla.

Entendemos los métodos cuantitativos en sentido amplio, incluyendo los métodos estadísticos, econométricos, los algoritmos de computación, y la investigación operativa. Todos ellos aportan instrumentos válidos para diseñar y evaluar políticas públicas. En su aplicación al ámbito sanitario, se benefician de la simbiosis entre la economía y otras disciplinas (ciencia política, epidemiología,…), integrados en la Investigación en Servicios Sanitarios (ISS), la cual está tomando cuerpo y reconocimiento oficial como ciencia híbrida sin filiación disciplinar, ocupada de aunar perspectivas múltiples, académicas y profesionales, en un esfuerzo por mejorar los servicios sanitarios. Un proceso similar tuvo lugar después de la II Guerra Mundial con la Investigación Operativa.

Dentro de los métodos cuantitativos, la microeconometría ha experimentado un gran desarrollo en los últimos años, aportando a la econometría general y recibiendo de ella, con aportaciones específicas para el ámbito de la salud (Jones, 2000), de la economía laboral, de la educación, y de otras políticas públicas. El interés de los investigadores, movido por la necesidad de resolver cuestiones pero también por la oportunidad de obtención de 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.

El uso generalizado de modelos para datos longitudinales con muestras grandes —el panel de hogares, por ejemplo— y de modelos multinivel con datos jerárquicos implica que se trabaja con varias dimensiones, cada una con su correspondiente tamaño muestral. Generalmente, los paneles tienen tamaños grandes en el espacio (n grande) pero seguimiento temporal corto (t pequeño). Paradójicamente, con los grandes paneles de microdatos seguimos sufriendo el problema de muestras pequeñas, aunque ahora lo son en la dimensión temporal (seguimiento corto de la población), por lo cual hay que buscar métodos de estimación consistentes en la dimensión temporal.

Se plantea un problema similar con los modelos multinivel, en los que hay tantos tamaños muestrales como niveles. Por ejemplo, si el nivel 1 son los pacientes y el nivel 2 los hospitales en donde ingresan, los tamaños muestrales relevantes son el número de pacientes ingresados en cada hospital y el número de hospita-

les. Lo propio ocurre en los modelos que evalúan el rendimiento educativo bajo diferentes políticas o programas. El nivel 1 es el alumno, el nivel 2 es el centro donde estudia. De hecho, los modelos multinivel nacieron en el ámbito de la educación, para evaluar comparativamente el efecto de las capacidades individuales y de la influencia de la escuela en el éxito educativo (Goldstein, 2003).

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) que están bien desarrollados. En econometría seguimos siendo generalmente 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*).

Por otra parte, hay desarrollos recientes de métodos microeconométricos para evaluar los resultados de programas públicos, que fundamentalmente se están aplicando en economía laboral (Angrist et al., 1998;Angrist et al., 2001), aunque tienen un gran campo abierto de aplicación en economía de la salud (Vera-Hernandez, 2003). Tratan de estimar el efecto medio de cada tratamiento (resultado esperado en caso de participación en el programa menos resultado esperado en caso de no participación; por ejemplo, el resultado es el salario y el tratamiento un programa formativo de fomento del empleo) con modelos paramétricos o no paramétricos que resuelvan, entre otros, los problemas de sesgo de selección (la adscripción al programa no es aleatoria), y descuentan el efecto de los factores de confusión. En la actualidad se están desarrollando modelos microeconométricos para evaluar programas que emplean múltiples tratamientos (o intervenciones políticas simultáneas) (Frölich, 2004). Esperamos que en los próximos años la economía laboral y la economía de la salud converjan y se polinicen mutuamente, al menos en lo que atañe a ese tipo de modelos.

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.

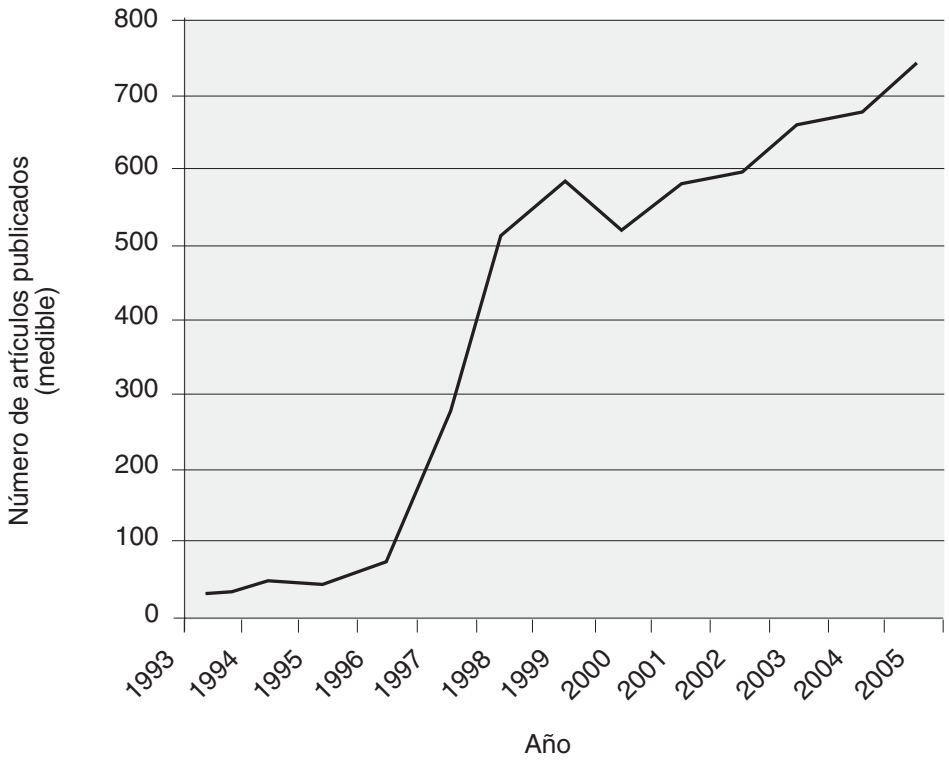
En este artículo presentamos una visión general del uso de métodos cuantitativos para construir *ranking* que orienten las políticas públicas, con especial referencia a las de salud, y cómo emplearlos. Se hace hincapié en los problemas metodológicos y en las dificultades para mejorar las políticas a 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*

Vivimos en un mundo obsesionado por la e*videncia* que busca hacer Políticas Basadas en la Evidencia (PBE). El *bench-*

Figura n° 1

**El *benchmarking* en las publicaciones médicas 1993-2005**



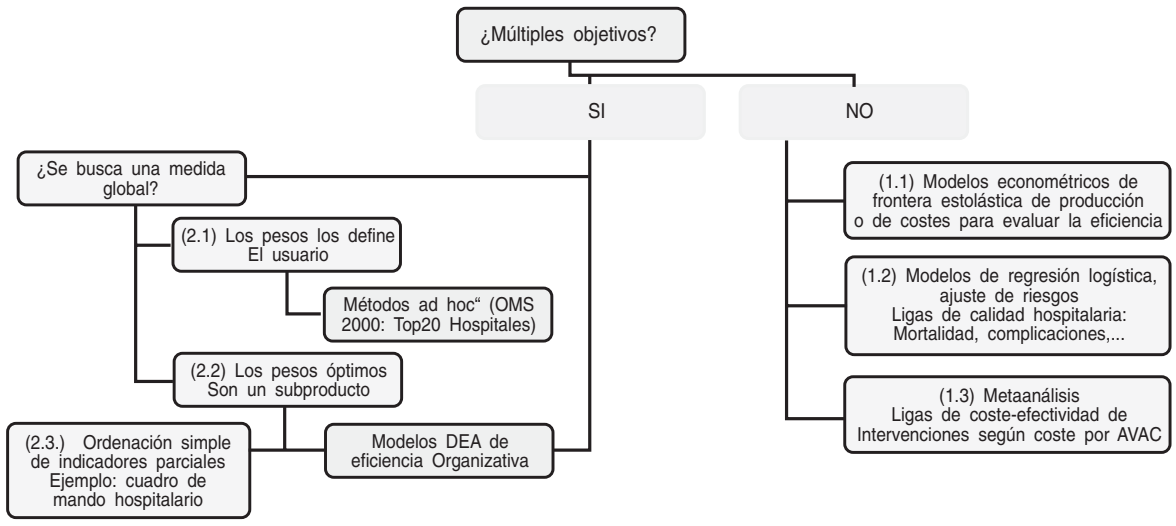

*marking* se practica tanto a escala de mesogestión de centros- Top20 hospitales(Garcia-Eroles et al., 2001;Iasist, 2006)- como a escala mundial. La Organización Mundial de la Salud (OMS) compara los sistemas sanitarios en cuanto a los recursos que emplean y a los resultados que consiguen. La OCDE ha 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 Indice 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 las siguientes actividades: diseñar sistemas de incentivos a los gestores; ajustar fórmulas de pago a proveedores (separar el grano de la paja, no remunerar la parte de los costes que corresponde a la ineficiencia); elaborar ligas de hospitales o de universidades que informen al público y permitan elegir con criterios de calidad; ordenar posibles intervenciones públicas según criterios de coste-efectividad (ligas de AVACs), entre otras aplicaciones.

El *benchmarking* como instrumento de mejora sistemática de las organizaciones procede de la literatura de la Gestión de Calidad Total. Empezó a extenderse hacia las publicaciones científicas del ámbito sanitario a mediados de los noventa, aunque no se incluye como término de referencia en *medline* (*Medical Subject Headings*) hasta 1998, con la siguiente definición: «método para medir el desempeño (*performance*) contra estándares establecidos de la mejor práctica». En ese período se iniciaron algunos experimentos puntuales (Nelson et al., 1995). La progresión de artículos publicados en los últimos diez años en el ámbito sanitario (figura 1) ha sido notable. El gran salto se ha producido entre 1996 y 1998.

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 2. 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 nº 2

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




Fuente: Elaboración propia

## 3. RANKING UNIDIMENSIONAL: DE LA CALIDAD DE LOS PROVEEDORES A LA COMPARACIÓN DE «CIUDADES GEMELAS» PASANDO POR LA LIGA AVAC

Los *ranking* de criterio único incluyen tres tipos de teorías bien diferenciadas: 1) 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; 2) los modelos de regresión logística y similares, que estiman la calidad (por ejemplo, mortalidad hospitalaria, implicaciones, reingresos,… esperada de cada centro, ajustando por severidad de los casos tratados, y en su caso por otros factores no controlables de entorno); 3) 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 intervenciones sobre ellos.

Cada vez con mayor frecuencia y trascendencia financiera se elaboran, publican y utilizan ordenaciones de unidades proveedoras de servicios públicos, por ejemplo hospitales, o centros de atención primaria, basados en un único criterio u objetivo.

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:

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

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 repartido 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, distribuída independientemente de las x y de las z. El modelo de frontera de costes se formularía de forma similar, salvo que ui≥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 eva-

luación 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).

## Ranking unidimensional de la calidad (clínica) hospitalaria

Los *ranking* unidimensionales de calidad clínica de los hospitales, por el contrario, 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 factor experiencia influye en la mortalidad (Wu et al., 2004).

Desde la perspectiva de las políticas, son muy interesantes los experimentos en curso de pago por resultados (*pay-for-performance*, P4P) a los hospitales, que se decidió en EEUU en 2003 y se inició experimentalmente en 2005, para incentivar financieramente a los hospitales, dentro de los programas de aseguramiento público Medicare y Medicaid, por proveer atención sanitaria de alta calidad a sus pacientes ingresados. La base de datos, mantenida por los *Centers for Medicare and Medicaid Services* (CMS), contiene datos para unos 4.200 hospitales de 17 indicadores de calidad clínica del tratamiento del infarto agudo de miocardio, fallo cardíaco, neumonía y prevención de infección quirúrgica. La base de datos es pública y de libre acceso, permitiendo al usuario comparar la calidad de los hospitales de su zona (US Dept.Health and Human Services HHS, 2006). El experimento financiero, de momento, se ha limitado a unos 270 hospitales con participación voluntaria y los primeros resultados son alentadores (Kahn III et al., 2006).

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 tanto para la elección informada de hospital por los pacientes como para que los profesionales dispongan de estándares de buena práctica. Así mismo debería ser fácil de obtener y difícil de manipular.

Las ligas de calidad clínica de hospitales tienen detractores y entusiastas y han sido discutidas por dos tipos de problemas. En primer lugar, los asociados a la propia medida de calidad y su ajuste por factores de confusión, es decir, problemas de posibles sesgos sistemáticos relacionados con el modelo estadístico o econométrico que generó el *ranking* y los ajustes por riesgo. En segundo lugar, por problemas de significación estadística de las diferencias en las posiciones dentro del *ranking*. Puesto que los indicadores medidos son al fin y al cabo extracciones aleatorias de una distribución de probabilidad, hay errores de muestreo que además varían entre unidades. Puede ocurrir que la incertidumbre implícita en los datos sea tan elevada que de hecho no

haya diferencias significativas entre centros que ocupan distintas posiciones en el *ranking*. Aunque hay contrastes diseñados para comparaciones múltiples (Bonferroni), la mayor parte de los *ranking* publicados no informan al respecto.

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 1.000 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; se suponen pacientes homogéneos, que no hay sesgo de selección y que ambos hospitales tienen una probabilidad idéntica de que ocurra el suceso adverso (igual calidad) y los números de sucesos adversos siguen sendas distribuciones binomiales con parámetros (n,p) donde n es 40 para el hospital pequeño y 1.000 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:

$$\Pr(p_1 \le p \le p_2) = 1-\alpha$$

$$\sum_{x=0}^{np_1} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\alpha}{2}$$

$$\sum_{x=np_2}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\alpha}{2}$$

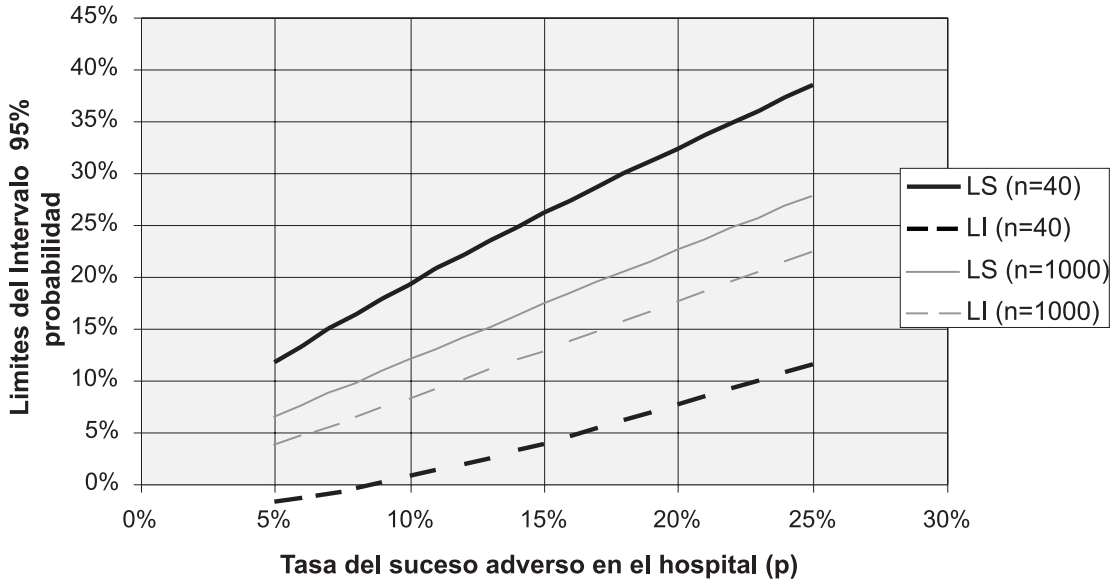
Hemos representado en la figura 3 dichos límites ($p_1$ y $p_2$) para ambos hospitales, para una probabilidad del 95% y niveles de calidad (tasas de efectos adversos) entre el 5% y el 25%. Como puede observarse, los intervalos son mucho más amplios para el pequeño hospital que para el grande. Además, 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 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 de 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.

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.

Figura nº 3

**Efecto del tamaño muestral y del nivel de calidad (p) sobre la amplitud del intervalo que contiene el 95% de probabilidad**




Nota: los límites se han calculado aproximando la distribución binomial a la normal

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.

En las ordenaciones subyace un proceso anidado, con al menos dos niveles: pacientes y hospitales. Los modelos multinivel son adecuados para este fin porque separan los efectos del grupo (hospital) de los efectos individuales (Goldstein et al., 2003). Ya se ha mencionado que su generalización a otros ámbitos de las políticas públicas es casi inmediata (alumnos y colegios, candidatos y programas, o procesados y jueces, por ejemplo).

Desde la perspectiva econométrica, los modelos paramétricos de frontera estocástica y los modelos multinivel tienen en común que, frente a los econométricos tradicionales, consideran que la parte aleatoria del modelo (el «ruído») es información, y de hecho lo que interesa a efectos de políticas no es tanto el valor estimado de los parámetros fijos, coeficientes de las x, que son meros ajustes, sino las varianzas de los componentes del error, en los modelos de regresión multinivel (Goldstein, 2003;Snijders et al., 1999), o estimar (predecir) a partir de los residuos cada uno de los valores de los errores aleatorios de las unidades muestrales (ineficiencia). Este interés en estimar consistentemente los errores aleatorios a partir de los residuos se ha beneficiado de la metodología bayesiana.

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 (Vass, 2001). Por ejemplo, sin informar de los intervalos de confianza, o sin el necesario ajuste por riesgos. La publicación de estos datos contribuye a mejorar la calidad de los servicios cuando su nivel de calidad es bajo, como ha demostrado un estudio experimental sobre la seguridad de los hospitales de Wisconsin (Hibbard et al., 2003).

Los modelos de regresión logística que ajustan por riesgos una característica dicotómica de calidad con objeto de comparar hospitales se emplean también para comparar zonas geográficas. Puesto que la variabilidad intra país es mayor que la variabilidad entre países, se han comparado ciudades «gemelas» de distintos países (París y Manhattan, por ejemplo). Comparando las tasas de altas hospitalarias potencialmente evitables por la atención primaria(Brown et al., 2001;Sanderson et al., 2000) mediante sendos modelos de regresión logística para Paris y Manhattan, se ha determinado que el riesgo relativo de ser hospitalizado por un diagnóstico potencialmente evitable por la atención primaria es 2,5 veces mayor en Manhattan que en París (Gusmano et al., 2006), concluyendo, por tanto, que en esta última ciudad hay más acceso a una atención primaria de calidad, con capacidad resolutiva.

Una línea de avance a mi juicio prometedora es la ordenación basada en criterios de dominación estocástica bayesiana, que se ha empleado, por ejemplo, para ordenar países según la producción científica en Economía (Lubrano, 2004). El concepto es antiguo, pero en los años recientes se ha empleado para construir diferentes *rankings* de unidades productivas. Hay dominación de primer grado de una distribución de probabilidad F sobre otra distribución G si $F(x) \leq G(x) \forall x \in (0,+\infty)$, es decir, si la probabilidad de obtener al azar el valor x o menor no es menor con F que con G, independientemente del valor de x que se tome. Aplicado, por ejemplo, a la distribución de la renta de dos regiones significa que en una de ellas la renta está más concentrada que en la otra (las curvas de Lorenz respectivas no intersectan). En la ordenación de unidades productivas en vez de trabajar con distribuciones de frecuencias lo haremos con estimaciones de distribuciones de probabilidad, a partir de muestras. La dominación resuelve elegantemente el problema del solapamiento de los intervalos de confianza en estadística frecuencialista clásica.

## Las ligas de coste-efectividad

Los *ranking* de coste-efecividad 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 con aparente simplicidad. No obstante, sufren limitaciones que previenen contra su uso indiscriminado y simplista. No han de ser tomadas como recetas a aplicar automáticamente 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 los hospitales (necesidad de ajustes, gestión científica de la incertidumbre, contrastes de comparaciones múltiples).

Categorizamos las limitaciones de las ligas AVAC en tres tipos: a) limitaciones inherentes a cada uno de los estudios coste-utilidad y al propio instrumento AVAC (Arnesen et al., 2004; Badia et al., 1999; Bleichrodt et al., 1997; Bleichrodt et al., 2002a; Bleichrodt et al., 2002b; Dolan et al., 2005; Johannesson, 1994; Nord et al., 1999a); b) problemas relacionados con el metaanálisis; c) 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 el metaanálisis que compara (o «agrega») 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 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 terminen en conclusiones positivas.

[1] http://www.hsph.harvard.edu/cearegistry

Además, se excluyen de la liga los numerosos estudios que no emplean el AVAC como medida de utilidad, que están desigualmente distribuídos entre enfermedades, según un análisis comparativo de 455 trabajos de evaluación económica incluídos 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 se utilice en la Medicina Basada en la Evidencia (MBE), esa evidencia es fragmentaria, incompleta e incierta. Hay incertidumbre diagnóstica, 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 quienes 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., 1999b).

## 4. RANKING MULTIDIMENSIONAL. ¿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 ejecución 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 derivarse del propio análisis.

## 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 ejecución de los sistemas de salud. Ha recibido muchas y acertadas críticas porque uniformizan con supuestos objetivos comunes a todos los países del mundo que, sin embargo, pueden tener (tienen de hecho sus propias metas distintas del resto(Williams, 2001). Confunden, pues, heterogeneidad e ineficiencia (Greene, 2003). Sin embargo, tiene el mérito de ser pionero, marcar una estela que han seguido muchos trabajos, organismos e investigadores.

En septiembre de 2001, el *Natural Health Service* (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:

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

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 $\pi_{kt}$ 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 ($\pi$), que son las ponderaciones incorporadas en la fórmula.

Como para su cálculo se debiera 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. A título ilustrativo, ofrecen resultados anuales para 1998-2004. El *output* hospitalario ajustado por calidad y valorado según costes de producción aumentó a una tasa media del 3,6% anual (3,35% sin ajustar, pag.200).

El informe británico abre nuevos horizontes y cambia perspectivas, sobre todo porque el enfoque 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 IRYSS[2].

## 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 sistemas de 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. Al igual que estos últimos, 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 instrumentos 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 perteneciente a un 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 lógi-

[2] Red de Investigación en Resultados y Servicios de Salud. http://www.rediryss.net/pub/units/rediryss/pdf/identifypriorizsns.pdf

ca de la organización, y no con la lógica matemática). Significación estadística y significación socio-política ni son sinónimos ni siempre concuerdan.

Si los efectos dinámicos son difíciles de modelizar en general, más todavía lo son en los modelos de eficiencia organizativa en sanidad. Las UTD arrastran la herencia del pasado —inercia no controlable— y toman decisiones hacia el futuro, gastando hoy para ganar mañana. La especificación dinámica de retardos y respuestas a impulsos es compleja, y requiere datos longitudinales que no siempre están disponibles o son comparables. El cambio tecnológico dificulta todavía más la comparabilidad temporal de la eficiencia organizativa.

## 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, 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 hemos presentado una visión general del uso de métodos cuantitativos para construir *ranking* que orienten las políticas públicas, con especial referencia a las 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 tres tipos de modelos: 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 se 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, mayor desagregación de indicadores se requiere, obligando a 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 derivadas 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 sistemas de 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 y evitar hacer 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 la publicación de estos datos contribuye a mejorar la calidad de los servicios cuando tienen un nivel muy bajo. La ciencia todavía no garantiza que se obtengan medidas objetivas de eficiencia organizativa independientes de los instrumentos 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 pertenecientes a un 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 BIBLIOGRÁFICAS


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

ANGRIST, J. D. AND KRUEGER, A. B. (1998): Empirical strategies in labor economics.

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

ARNESEN, T. AND TROMMALD, M. (2004): 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):43-50.

BADIA, X., ROSET, M., AND HERDMAN, M. (1999): Inconsistent responses in three preference-elicitation methods for health states. *Soc.Sci.Med* 49(7):943-950.

BLEICHRODT, H., HERRERO, C., AND PINTO, J. L. (2002a): A proposal to solve the comparability problem in cost-utility analysis. *J Health Econ* 21(3):397-403.

BLEICHRODT, H., HERRERO, C., AND PINTO, J. L. (2002b): A proposal to solve the comparability problem in cost-utility analysis. *J Health Econ* 21(3):397-403.

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

BROWN, A. D., GOLDACRE, M. J., HICKS, N., ROURKE, J. T., MCMURTRY, R. Y., BROWN, J. D., AND ANDERSON, G. M. (2001): Hospitalization for ambulatory care-sensitive conditions: a method for comparative access and quality studies using routinely collected statistics. *Can.J Public Health* 92(2):155-159.

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

CUTLER, D. M., HUCKMAN, R. S., LANDRUM, M. B., AND NATIONAL BUREAU OF ECONOMIC RESEARCH (2004): *The role of information in medical markets an analysis of publicly reported outcomes in cardiac surgery*. Cambridge, MA: National Bureau of Economic Research.

DAWSON, D, GRAVELLE, H, O'MAHONY, M, STREET, A, WEALE, M, CASTELLI, A, JACOBS, R, KIND, P, LOVERIDGE, P, MARTIN, S, STEVENS, P, AND STOKES, L. (2005): Developing New Approaches to Measuring NHS *Outputs* and Activity. CHE Research Paper 6.

DOLAN, P., SHAW, R., TSUCHIYA, A., AND WILLIAMS, A. (2005): QALY maximisation and people's preferences: a methodological review of the literature. *Health Econ* 14(2):197-208.

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

GARCIA-EROLES, L., ARIAS, A., AND CASAS, M. (2001): Los Top 20 2000: objetivos, ventajas y limitaciones del método. *Rev Calidad Asistencial* 16(2):107-116.

GEWEKE, J., GOWRISANKARAN, G., AND TOWN, R. J. (2001): *Bayesian inference for hospital quality in a selection model*. Cambridge, MA: National Bureau of Economic Research.

GOLDSTEIN, H. (2003): *Multilevel statistical models*. London: E. Arnold.

GOLDSTEIN, HARVEY AND SPIEGELHALTER, DJ. (2003): League tables and their limitations: statistical issues in comparisons of institutional performance. .

GONZÁLEZ-LÓPEZ-VALCÁRCEL, B. AND BARBER, P. (1996): Changes in the efficiency of Spanish public hospitals after the introduction of Program-Contracts. *Investigaciones Económicas* XX(3):377-402.

GREEN, J. AND WINTFELD, N. (1995): Report cards on cardiac surgeons. Assessing New York State's approach. *N.Engl.J Med* 332(18):1229-1232.

GREENE, W. (2003): Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems . GUSMANO, M. K., RODWIN, V. G., AND WEISZ, D. (2006): A new way to compare health systems: avoidable hospital conditions in Manhattan and Paris. *Health Aff (Millwood.)* 25(2):510-520.

HAUCK, K AND STREET, A. Performance Assessment in the Context of Multiple Objectives: A Multivariate Multilevel Analysis (2006): Center for Health Economics University of York. HIBBARD, J. H., STOCKARD, J., AND TUSLER, M. (2003): Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood.)* 22(2):84-94.

IASIST. Top20. Benchmarks para la Excelencia 2005. http://www.iasist.com/top20/Top20_2005/Resultados/publicacion.pdf. 31-3-2006.

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

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

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

KAHN III, N., AULT, T., ISENSTEIN, H., POTETZ, L., AND VAN GELDER, S. (2006): Snapshot Of Hospital Quality Reporting And Pay-For-Performance Under Medicare. *Health Affairs* 25(1):148-162.

KELLY, E. AND HURST, J. (2006): *Health Care Quality Indicators Project. Conceptual Framework Paper*. OECD Health Working Papers.

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

NELSON, D. E., FLEMING, D. W., GRANT-WORLEY, J., AND HOUCHEN, T. (1995): Outcome-based management and public health: the Oregon Benchmarks experience. *J Public Health Manag.Pract.* 1(2):8-17.

NORD, E., PINTO, J. L., RICHARDSON, J., MENZEL, P., AND UBEL, P. (1999a): Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Econ* 8(1):25-39.

NORD, E., PINTO, J. L., RICHARDSON, J., MENZEL, P., AND UBEL, P. (1999b): Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Econ* 8(1):25-39.

OECD (2003): *A disease-Based Comparison of Health Systems. What is best and at what Cost?* Paris.

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

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

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

SNIJDERS, T. A. B. AND BOSKER, R. J. (1999): *Multilevel analysisan introduction to basic and advanced multilevel modeling*. London: Sage Publications.

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

US DEPT.HEALTH AND HUMAN SERVICES HHS. (2006): Hospital Compare. *A quality tool for adults, including people with Medicare*. http://www.hospitalcompare.hhs.gov .

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

VERA-HERNANDEZ, M. (2003): [Evaluating health interventions without experiments]. *Gac.Sanit.* 17(3):238-248.

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

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

WORTHINGTON, A. (2006): An empirical survey of frontier efficiency measurement techniques in healthcare services. http://www.bus.qut.edu.au/schools/economics/documents/disc_papers_pre2001/Worthington_67.pdf

WU, C., HANNAN, E. L., RYAN, T. J., BENNETT, E., CULLIFORD, A. T., GOLD, J. P., ISOM, O. W., JONES, R. H., MCNEIL, B., ROSE, E. A., AND SUBRAMANIAN, V. A. (2004): 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):784-789.