Rev Calid Asist.2010;**25(1)**:1–3

Revista de Calidad Asistencial

www.elsevier.es/calasis

EDITORIAL

Validez de los indicadores agregados de desempeño

Validity of synthetic indicators for measuring performance

J.R. Repullo

Departamento de Planificación y Economía de la Salud, Escuela Nacional de Sanidad, Instituto de Salud Carlos III, España

Recibido el 3 de noviembre de 2009; aceptado el 5 de noviembre de 2009
Disponible en Internet el 16 de diciembre de 2009

Primero, lo obvio: los humanos contemporáneos tenemos una clara tendencia a *rankings*, *tops*, estrellas o listas ordenadas que expresen de forma sintética lo bueno o lo malo de una realidad compleja. Así como el precio se traduce en una cantidad inequívoca, queremos que la cualidad y la calidad de un producto pueda convertirse, si no en una cifra, al menos en un listado de mejor a peor.

Segundo, lo menos obvio: parece que esta tendencia contrasta con la escasa capacidad del ser humano de disponer y de procesar información relevante para la toma de decisiones, lo que el nobel Herbert Simon llamó "racionalidad limitada". Para ello, se contaba clásicamente con el intermediario (un *broker*, como por ejemplo el médico de cabecera) que nos aconsejaba en calidad de agente personal lo que más nos convenía una vez que conocía nuestras necesidades y preferencias genéricas.

Ahora queremos saltarnos a este proveedor de información y consejo y buscar directamente la información para decidir por nosotros mismos sin ninguna interferencia. Esto nos coloca cada vez más frecuentemente en situaciones en las que tenemos mucha más información de la que podemos entender, interpretar y utilizar. De ahí la demanda de que alguien nos ordene el mundo caótico y haga de "agente multiuso", es decir, que empaquete todos los criterios para cualificar las opciones de decisión en indicadores sintéticos o agregados que nos saquen de la ansiedad que produce la incertidumbre.

Cuando hay demanda de este tipo de productos, se abren enormes posibilidades de generar oferta. Sabemos, además, que existen cada vez más datos sobre el ámbito de la salud y los servicios sanitarios: producir estos datos es tan caro o más que antes, pero reproducirlos es virtualmente gratis. De hecho, la necesidad de difusión intraorganizativa e interorganizativa de estos datos lleva a que las organizaciones los publiquen en Internet, con lo que cualquiera que tenga cierto ingenio y dedicación puede utilizarlos como maná que cae del cielo estadístico para construir excelentes y provechosos menús comerciales.

Nada que objetar si se trata de buenos cocineros. Pero no parece que haya tanto talento o recursos como para que éstos proliferen. Es demasiado fácil poner en circulación platos defectuosos que son fácilmente devorados por la demanda insaciable de los medios de comunicación generales o profesionales y que, además, producen rendimientos inmediatos en clave de notoriedad, publicidad y atracción hacia otros productos asociados.

Un caso particularmente llamativo, a nivel europeo, es el *Health Consumer Powerhouse* que produce un índice con una sigla equívoca, EHCI, similares a los European Community Health Indicators lo que le otorga por resonancia más visibilidad y presencia. Este informe produce un ranquin que nos visita todos los años con titulares llamativos que coloca al Sistema Nacional de Salud español en el furgón de cola de los países europeos. No es cuestión aquí de abordar la crítica a este modelo de comparación de desempeños de sistemas, cualquiera que lo revise siquiera superficialmente verá cómo se cae de las manos por su insoportable levedad técnica. En octubre de 2009 señalé algunas críticas a este índice, que

Correos electrónicos: jrepullo@isciii.es, jrepullo@gmail.com.

1134-282X/$ - see front matter
doi:10.1016/j.cali.2009.11.003

pueden refrescarse en la revista electrónica de la Organización Médica Colegial[1]. Como suponemos que a mediados de 2010 volverán a las portadas de las revistas, es útil estar ya vacunados para anular con anticuerpos este petulante e inútil informe.

Sin embargo, queda el poso del interés público por la comparación de desempeños de sistemas y servicios sanitarios. Según avanzan las bases de datos que gestiona el Ministerio de Sanidad y Política Social a través del Instituto de Información Sanitaria, se abren oportunidades y se crean expectativas de buenas comparaciones entre comunidades autónomas, áreas de salud u hospitales[2]. El bloqueo político existente para facilitar datos comparados por regiones o centros es un obstáculo mayor para crear buenos productos sistémicos de análisis de desempeño, pero también es un aliciente para que aparezcan iniciativas comerciales de muy variada validez metodológica y calidad técnica. Por eso, conviene afilar el lápiz de la metodología para saber construir o interpretar este tipo de instrumentos.

En particular, interesa destacar 3 campos que afectan a la construcción de indicadores agregados de desempeño:

a) Los sistemas sanitarios son organizaciones complejas y multidimensionales que devuelven una imagen cambiante según la perspectiva (¿salud o satisfacción?) y los valores del observador (¿libertad irrestricta de elección o planificación de oferta de servicios?). No hay alternativa metodológica para superar este escollo, pero sí existe la exigencia de transparencia (explicitar perspectivas y valores) o la posibilidad de crear distintos índices que reflejen perspectivas con imposibles cualidades aditivas.
b) Para hacer un indicador sintético (listado o ranquin), hay que hacer 2 cosas que metodológicamente no están bien resueltas: medir los ítems (otorgar puntuación al grado de cumplimiento de los criterios) y agregar las puntuaciones de los ítems y las dimensiones (si no ponderamos los ítems les otorgamos el mismo peso, y si otorgamos pesos basados en expertos, estamos incorporando las preferencias de los expertos al modelo y no necesariamente las de la población a la que va destinada la comparación).
c) En el caso de análisis de desempeño, se añade la necesidad de relacionar resultados (puntuación) con costes, por ejemplo, resultados (numerador) por unidad de coste (costes en el denominador); el mejor desempeño será el que tenga más alto este índice (más resultados para el mismo gasto). Al hacer esto, lo que ocurre es que muchos países/regiones pobres (por mala que sea su puntuación) se alzan a los puestos de cabeza de la tabla (por la miseria de su gasto o el coste en el denominador).

No se trata de impugnar estas herramientas, pero sí de señalar su debilidad metodológica; pueden servir para empezar a pensar, pero no deben usarse para dejar de pensar. En todo caso deberían seguirse estas 3 normas de prudencia, exigibles a los que las desarrollen o las comercialicen:

a) Transparencia: la información básica que alimenta los distintos indicadores debe ser clara y trazable hasta la fuente estadística que la origina. Si se introducen elementos no estadísticos (valoraciones de expertos, encuestas a personas o asociaciones, existencia de reglamentación o cualquier variable que suponga interpretar grados de cumplimiento), debería llevar un detallado anexo metodológico con las mismas normas requeridas en una publicación científica (que en definitiva permitan replicar el estudio en otro tiempo o lugar y por otros investigadores).
b) Fiabilidad de indicadores: no siempre sabemos qué hay en numeradores y denominadores, y es fundamental que se explicite y se detalle la forma en la que se construyen los ítems de información (en realidad, incluido en lo que antes decíamos de anexo metodológico).
c) Versatilidad, es decir, capacidad de reordenar microdatos: el tratamiento de los indicadores no debe hacer desaparecer la información básica. Otros investigadores deberían poder reordenar estos datos que componen los índices sintéticos para analizar el efecto que estos cambios pueden tener en las distintas dimensiones que pretende capturar el sistema de análisis de desempeño. Por eso, se precisan tablas no cocinadas con los valores de los distintos indicadores y unidades analizadas (aunque éstas puedan anonimizarse) o bien la cortesía de brindar los datos en una *web* con la posibilidad de que los lectores interesados los descarguen.

Queda el tema principal: la validez científica de estos indicadores sintéticos. Si revisamos la teoría de medición en ciencias sociales (Trochim tiene una excelente revisión disponible en Internet[3]), encontramos que la validez de mayor nivel, la llamada ''*validez de constructo*'', es menos alcanzable cuanto más abstracto sea el concepto que queremos medir. Una báscula permite medir razonablemente el concepto ''masa'' a través de la medida del ''peso'' (recordemos que no pesamos lo mismo en todos los puntos de la superficie terrestre, ya que varía la gravedad). Pero el concepto de ''inteligencia'' o de ''desempeño'' nos adentra en niveles de abstracción en los que la validez del concepto va a ser difícil de alcanzar. Para conseguirlo, contamos con 2 tipos de instrumentos:

1. La validez ''traslacional'' indaga en qué medida en el término ''desempeño de sistemas o servicios sanitarios'' hemos incluido todos los elementos adecuados (en función de los marcos teóricos y conceptuales en los que nos apoyamos). De ahí vienen la *face-validity* (validez de apariencia, juzgada por expertos) y la *content validity* (validez de contenido, apoyada por listas de comprobación de aspectos que incluyen el dominio del concepto a través de revisiones de la literatura médica).
2. Además de indagar en la riqueza y la exhaustividad de la definición del concepto, existe otra forma de acercarnos a la validez: a través de su aplicación; por tanto, se dirigiría a las mediciones que resultan de su operativización. Básicamente tenemos 4 opciones:
   - demostrar su capacidad predictiva (por ejemplo, un test de inteligencia matemática que puede predecir al principio del curso las notas que tendrán los alumnos en esta asignatura al final del curso);

- demostrar por aplicación concurrente (simultánea) que el criterio es capaz de distinguir entre grupos que a priori sabemos que se comportarán de forma diferente al aplicarse la medición (por ejemplo, ver si los países pobres frente a los ricos muestran un desempeño muy diferente de su sistema sanitario);
- comprobar si otro concepto similar o cercano está alineado y actúa de forma convergente al aplicarlo a la misma población (imaginemos instrumentos para medir 2 conceptos: ''integración de procesos asistencial'' y ''coordinación de niveles de atención''),
- y en contraste con lo anterior, asegurarnos que si aplicamos 2 conceptos diferentes a la misma población, van a tener un efecto discriminante al no presumirse alineación de sus contenidos (por ejemplo, competencia técnica y motivación trascendente de los médicos).

Si queremos crear una buena medición del desempeño de los sistemas sanitarios, lo primero sería una buena definición del propio concepto de ''desempeño''; la OMS lo intentó en el año 2000[4] con un modelo en el que relacionaba *inputs* (gasto sanitario) con una serie de *outputs*: nivel medio y distribución de la salud medida en DALYS, nivel medio y distribución de la *responsiveness* (capacidad de respuesta a preferencias y trato digno del paciente) y equidad en la financiación. Esta validez traslacional, apoyada por expertos y trabajos de revisión de la literatura médica, no es suficiente. Hay que asomarse a la forma de hacer operativas las dimensiones y las medidas en las que se sustenta el concepto.

Por ejemplo, la validez predictiva puede ensayarse en la medida en la que usemos el concepto para identificar resultados de cambios organizativos o de nuevas políticas. La comparación de reformas (predicción de mejoras) sería el campo natural de utilización de estos análisis de desempeño y su ámbito natural de validación. Con sólo mencionarlo ya anticipamos la enorme dificultad para atribuir resultados a políticas cuando el marco exploratorio es demasiado amplio (para ello se activan otros modelos más cualitativos de análisis, como los que desarrolla el Observatorio Europeo de Sistemas y Políticas de Salud o, de forma más específica, la red Health Policy Monitor[5]).

La aplicación concurrente en países pobres, ricos, grandes y pequeños en el precitado informe 2000 la OMS produjo los principales cuestionamientos de validez: sólo leer los 8 primeros países del ranquin hace surgir dudas sobre lo que estamos realmente midiendo como *desempeño*: 1) Francia; 2) Italia; 3) San Marino; 4) Andorra; 5) Malta; 6) Singapur; 7) España, y 8) Omán.

La validez convergente obligaría a indagar sobre aquellos resultados en salud de carácter más general en relación con otros más específicos que podemos vincular de forma más directa con la acción de los servicios sanitarios (en línea, por ejemplo, con la mortalidad innecesariamente prematura y sanitariamente evitable).

Quizás, la validez divergente podría tener un campo particular de escrutinio para delimitar la contribución de la efectividad clínica de los servicios (en función del criterio de necesidad) y la capacidad de respuestas a preferencias y a demandas de los pacientes.

Por tanto, la tarea para fundamentar la validez de concepto del ''desempeño de los sistemas sanitarios'' es una ruta larga y no exenta de problemas y controversias, pero un camino que merece la pena ser recorrido, contando con un buen arsenal de métodos y personas que apoyen su desarrollo. Es, quizás, una gran tarea a la que la investigación de servicios sanitarios está claramente convocada, y los poderes públicos, con una historia de desatención a esta disciplina, están emplazados a fomentar y a apoyar.

Quizás no sea posible al final mantener este complejo constructo unitario de *desempeño*, pues las dimensiones que definimos y exploramos pueden acabar componiendo un pequeño monstruo de Frankenstein escasamente armonioso y funcional. La utilización de métodos de análisis envolvente de datos para integrar dimensiones diversas de desempeño podría ser una alternativa con mayor robustez matemática, pero podría no resolver el problema de fondo[6].

En todo caso, cabría formular una validez utilitaria o pragmática como guía para impulsar y animar estos procesos: un sistema de evaluación de desempeño será válido si consigue que los sistemas de salud a través de la comparación y la emulación consiguiente dirijan sus pasos en la resolución de problemas, en la mejora de la calidad de los servicios que prestan y en la propia superación de las lagunas y los errores de información y construcción de indicadores que van siendo evidenciados por el propio proceso de aplicación.

Por esto, demos una esperanzada bienvenida a los conceptos del desempeño, pero preparémonos para una larga travesía, en la que habrá tanto leales compañeros de viaje como piratas y oportunistas que querrán aprovecharse de la facilidad de producir titulares en medios de comunicación y de conseguir no sólo el minuto de gloria, sino filones de valioso metal que explotar.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Repullo JR. ¿Somos de los mejores o de los peores? Sobre los ''rankings'' que tanto nos apasionan para compararnos con otros europeos [consultado 11/2009]. Disponible en: http://www.medicosypacientes.com/colegios/2009/10/09_10_05_sns.
2. Indicadores clave del Sistema Nacional de Salud. Instituto de Información Sanitaria, Ministerio de Sanidad y Política Social. Disponible en: http://www.msps.es/estadEstudios/estadisticas/sisInfSanSNS/pdf/relIndicINCLASNS.pdf.
3. Trochim W. Research methods knowledge base. Web Centre for Social Research Methods. Disponible en: http://www.socialresearchmethods.net/kb/index.php.
4. The world health report 2000 - Health systems: Improving performance. Geneva: OMS [consultado 10/2009]. Disponible en: http://www.who.int/whr/2000/en.
5. Observatorio Europeo de Sistemas y Políticas Sanitarias. Disponible en: http://www.euro.who.int/observatory. Base de datos de Monitorización de Políticas de Salud (Health Policy Monitor). Disponible en: http://www.hpm.org/en/Search_for_Reforms/Search.html.
6. Rubio B, Rubio S, Repullo JR. En busca de nuevas herramientas de análisis de la eficiencia en el sector público sanitario. Revista de Administración Sanitaria. 2007;5:659–72.