Departamento de Psicología Social y Metodología

Facultad de Psicología

Tesis doctoral

Programa de doctorado en Psicología Clínica y de la Salud

Área Metodología

Valoración de procedimientos de "Mapping" de instrumentos específicos sobre genéricos en Calidad de Vida Relacionada con la Salud (CVRS)

Doctorando: Manuel Monroy Vega

Director: Miguel Ruiz Díaz

AGRADECIMIENTOS

Esta investigación ha sido posible gracias a la aportación de muchas personas que no han dudado en compartir su conocimiento y experiencia, brindándome todo su apoyo y colaboración para que esta tesis doctoral se convierta en realidad.

En primer lugar, me gustaría agradecer a mi director, Dr. D. Miguel Ruiz Díaz su dedicación, paciencia y valiosos consejos aportados en todo el recorrido de este proyecto, y que gracias a ellos ha sido posible elaborar este trabajo.

A Pfizer España y a la Asociación madrileña para la Lucha Contra las Enfermedades del Riñón (ALCER) por su inestimable ayuda en facilitarme la información de las patologías analizadas.

A la Universidad Autónoma de Madrid, a sus profesores, que me han brindado su apoyo e interés en la realización de esta tesis. Y a todo el personal, que tan amablemente siempre me han atendido y ayudado en la resolución de todo el trámite administrativo.

Al grupo "miércoles al sol" (Ludgerio, Rafael, Antonio, Jesús, Ricardo, Miguel, José Manuel y Pablo) por compartir conmigo este camino de iniciación en la investigación.

A mis amigos y compañeros del Centro Geográfico del Ejército, por su apoyo en perseguir mi objetivo.

Y en especial, a Elena, Martina y Pablo por soportar sin queja alguna no haberles dedicado todo el tiempo que se merecen y transmitirme fuerza para seguir y no parar hasta lograr este reto.

A todos, muchas gracias por compartir conmigo esta ilusión.

ÍNDICE

RESUMEN GENERAL DE LA TESIS

El constructo calidad de vida (CV) tiene como características básicas su carácter de percepción individual, su variabilidad en el tiempo y su multidimensionalidad estructural. Pero el hecho de que la CV sea subjetiva no quiere decir que no pueda ser objetivable y, en consecuencia, medible. El concepto CV se utiliza indistintamente en el mundo de la economía, de la política y de la salud, entre otros, pero es en la disciplina de la salud en la que ha suscitado mayor interés. De hecho, en el ámbito de la salud, se ha acuñado el término Calidad de Vida Relacionada con la Salud (CVRS), que es un concepto más restringido.

El continuo aumento de la demanda de una mayor salud y calidad de vida por parte de la población en general, el aumento considerable de la esperanza de vida y la disponibilidad de unos recursos sanitarios limitados, han hecho que la comunidad científica preste más atención a lo que se ha venido a llamar "Outcomes" (resultados) en salud, dando lugar al desarrollo de un área nueva de investigación denominada investigación de resultados de la salud (IRS).

Hay que resaltar que, en la mayoría de las enfermedades, la mejoría de los pacientes está relacionada habitualmente con aspectos del enfermo que los tratamientos no contemplan, como su estado de ánimo, su capacidad de afrontamiento o el apoyo que recibe de familiares y amigos. La CVRS es una medición de la salud (cuantificación numérica) en la que se debe tener muy en cuenta el punto de vista del propio paciente.

Los órganos gubernamentales responsables de los asuntos relacionados con la salud definen tres conceptos que el clínico ha de tener presentes para la utilización correcta de los medios disponibles. Se define la "eficacia" como la probabilidad ideal de que un individuo se beneficie de una tecnología para resolver un problema médico. Se define la "efectividad" como la probabilidad real de que el paciente se beneficie de dicha tecnología. Por último, se define la "eficiencia" como la relación entre los beneficios obtenidos por la efectividad y los costes que supone obtener dichos beneficios. Hoy en día, es posible y necesario valorar la eficiencia con la finalidad de reducir al máximo los costes de la intervención, en la medida de lo posible.

Dependiendo de las fuentes de información existentes, se puede valorar la eficiencia de una intervención de distintas maneras. Cuando el informante es el propio terapeuta/médico se estima como "coste-efectividad", si es el paciente el que informa, se estima como "coste-utilidad", y cuando el informante es el gestor económico, se estima como "coste-beneficio". Puesto que la salud es en sí misma un valor, y no se puede permitir su pérdida, aun cuando el coste sea elevado, serán el coste-efectividad y el coste-utilidad los que se tengan en mayor consideración. En esta tesis analizaremos el coste-utilidad.

El indicador de coste-utilidad más extendido es el de "años de vida corregidos por la calidad" (AVAC- QALY), que se define como la relación entre el aumento de esperanza de vida y el aumento en la calidad de vida, para un paciente o un conjunto de pacientes, alcanzado por una intervención o tratamiento. Los años de vida ganados se obtienen mediante estudios longitudinales prospectivos de seguimiento y se acostumbra a utilizar estadísticos como la mediana de supervivencia. La calidad de vida suele medirse mediante cuestionarios autoinformados validados, utilizando estadísticos de tendencia central.

Definimos la "utilidad" como el grado de preferencia que la sociedad elige entre una serie de alternativas de salud que se le presentan multiplicada por la probabilidad de que esta alternativa se dé.

Cuando hablamos de la medición en el ámbito de la salud, nos encontramos con un conjunto de instrumentos en los que los ítems que los componen no son homogéneos entre sí y que su objetivo no es otro que crear un índice válido que agregue la información de los dis-

tintos componentes de esa escala de salud. Es lo que denominamos un modelo formativo. En este caso, cada indicador (o variable observada) capta un aspecto diferente de la realidad que queremos medir, pudiendo ocurrir que incluso los indicadores fueran independientes entre sí. La combinación de indicadores se entiende como un agregado que capta toda la información de interés para crear el índice resumen.

La medición de la utilidad se ha realizado tradicionalmente con instrumentos denominados genéricos por varios motivos. En primer lugar, el valor de utilidad de un estado de salud particular es el resultado de la preferencia de la población por ese estado de salud, frente a otros resultados posibles. En segundo lugar, el interés se centra en cuantificar la salud en cada grupo de pacientes, no en realizar un diagnóstico individual. En tercer lugar, las utilidades deben ser unidimensionales, y no tiene sentido el estudio del perfil de las puntuaciones de un sujeto. Por último, los pesos dados a cada dimensión al resumir los estados de salud son específicos de cada población cultural y no se deben recalcular en función de la patología. Los instrumentos genéricos de calidad de vida se caracterizan por constar de dos elementos: un sistema descriptivo multiatributo y una regla de valoración o algoritmo que sirve para obtener los valores de utilidad (función multiatributo MAUF).

Por otro lado, existen instrumentos específicos que han sido generados para interpretar clínicamente los resultados de una actuación, analizar la evolución de los pacientes de una determinada patología a lo largo de un determinado periodo de tiempo y estudiar los efectos de los tratamientos. Su objetivo no es encontrar una puntuación global que resuma su CVRS, sino más bien el diagnóstico por dimensiones. De ellos lo que podemos obtener es el índice de severidad en ese paciente.

El procedimiento de mapping consiste en llevar a cabo una proyección matemática de la puntuación generada por el instrumento específico a la métrica de la función creada por el instrumento genérico.

En primer lugar, debemos determinar el procedimiento para obtener el índice de severidad de los sujetos en el instrumento específico. Puesto que el mapping final se realiza sobre una única puntuación de utilidad, será necesario determinar la manera de agregar las puntuaciones en las distintas dimensiones cuando el instrumento no sea unidimensional.

Una vez decidido el tratamiento que se dará al cuestionario específico, habrá que decidir el procedimiento más idóneo para trasladar las puntuaciones de severidad a la métrica de la utilidad.

El objetivo de la presente tesis es obtener distintas funciones de traslación métrica de diversos instrumentos de CVRS de patologías específicas sobre instrumentos genéricos de utilidad y valorar el método de proyección más adecuado, dadas las características no lineales de la función multiatributo. Se ha llevado a cabo un estudio bibliográfico para determinar las posibles estrategias de traslación (mapping) entre instrumentos específicos y genéricos en diferentes patologías. Analizadas las mismas, se ha optado por llevar un proceso gradual de mapping cambiando la patología a fin de tomar las menos graves, graves y muy graves teóricamente, y realizar el proceso de mapping con uno, dos y tres instrumentos genéricos.

En el primer estudio se ha tratado una patología considerada menos grave en la cual hemos trabajado con los criterios establecidos por un estudio anterior con la finalidad de replicar el procedimiento de obtención de su índice de severidad y su posterior traslación a un instrumento genérico.

En un segundo estudio se ha tratado una enfermedad grave en la cual no está definido el procedimiento de obtención del grado de severidad, lo que ha supuesto analizar un procedimiento de obtención del mismo y un posterior procedimiento de mapping con dos instrumentos genéricos, determinando cuál de ellos es el más idóneo respecto a la patología estudiada.

En el tercer estudio hemos analizado una patología grave de la cual tampoco estaba definido el proceso de obtención del índice de severidad y se ha realizado un procedimiento de mapping sobre tres instrumentos genéricos. En este último estudio, además, hemos realizado estudios de análisis de perfiles latentes de pacientes.

Como conclusiones generales debemos indicar las siguientes. Uno de los principales problemas que encontramos es la forma de definir el índice de severidad del instrumento específico, y se considera que la opción más favorable es establecer una estructura unidimensional, de manera que cada ítem aporte su peso y una función aditiva que nos permita obtener el valor de la severidad, aun cuando el instrumento tenga su procedimiento de obtención. Respecto a

los instrumentos genéricos, debemos indicar que no miden la utilidad teniendo en cuenta las mismas dimensiones, por lo que la traslación entre ellos es muy relativa al tener funciones multiatributo diferentes. Llevar a cabo un proceso de agregación por la media de los valores de un instrumento respecto al otro desvirtúa en gran medida el valor de la utilidad, ya que los valores extremos distorsionan su valor; es preferible eliminar aquellos sujetos que han puntuado de manera no coherente entre los diferentes instrumentos. No todos los instrumentos genéricos son igual de sensibles a cada patología, por lo que hay que especificar cuál de ellos es el más idóneo en cada caso. En los instrumentos específicos no se tienen en cuenta covariables que sí tienen impacto en la CV del paciente y que se reflejan en el índice de utilidad del instrumento genérico, por lo que deben ser consideradas en los instrumentos específicos sabiendo que esta estrategia alterará las propiedades psicométricas del instrumento original, siendo una cuestión que quedará pendiente de evaluación. El análisis de los índices de ajuste por quintiles resulta mucho más adecuado, ya que nos permite detectar en qué parte de la curva el ajuste es mejor o peor. El procedimiento de análisis de perfiles latentes nos permite encontrar perfiles comunes dentro de la patología y en cada uno de los instrumentos genéricos, si bien hay que considerar que los mismos se han obtenido a partir de valores medios de severidad y utilidades.

ANTECEDENTES

El campo de la salud y de las tecnologías sanitarias siempre ha estado interesado en evaluar la calidad de sus productos y servicios, su accesibilidad, y la distribución de los recursos económicos disponibles. Para ello, tradicionalmente ha utilizado resultados *(outputs)* como la mortalidad, la morbilidad, la expectativa de vida, las secuelas de las enfermedades y la efectividad de los tratamientos. El área de la salud ha sido regularmente criticada por evaluar únicamente estos resultados y olvidar que uno de los principales objetivos de la medicina, si no el más importante, es conseguir el bienestar de los pacientes, y no solo la lucha contra la enfermedad.

Los grandes avances tecnológicos en diagnósticos, pruebas y tratamientos a lo largo de la historia han priorizado que el paciente supere la enfermedad más que consiga su propio bienestar. Esto se aprecia claramente cuando hablamos de enfermedades crónicas, donde las nuevas tecnologías y tratamientos, capaces de prolongar la vida casi a cualquier precio, plantean una dicotomía entre la calidad de vida que puede alcanzar el paciente y la cantidad de vida (supervivencia) que permite prolongar.

En este punto hay que resaltar que, en la mayoría de las enfermedades, la mejoría de los pacientes está relacionada habitualmente con aspectos del enfermo que los tratamientos no contemplan, como su estado de ánimo, su capacidad de afrontamiento o el apoyo que recibe por familiares y amigos. Así, Meeberg (1993) nos indica que “el bienestar de los pacientes es un punto muy importante a tener en cuenta tanto en su tratamiento como en su modo de vida”.

Bungay et al. (1996) señalan que el funcionamiento biológico es solo la primera de una serie de esferas de la salud en la vida de un paciente, siendo nuestro objetivo valorar el bienestar general, si deseamos hacer una valoración global del paciente.

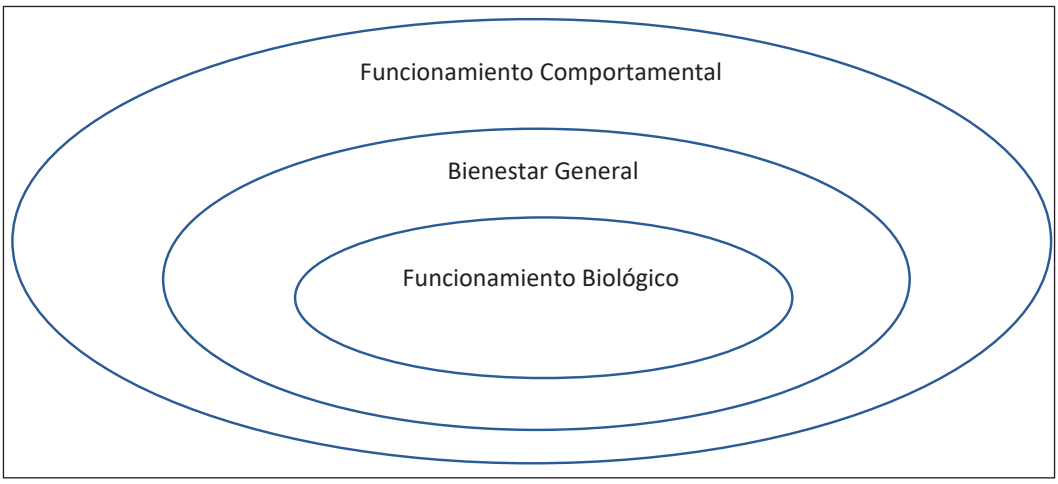


Figura 1. Esfera de salud de Bungay et al. (1996)

El constructo *calidad de vida* nace a mediados del siglo XX como consecuencia de la necesidad de valorar el impacto del crecimiento económico sobre la vida de los individuos y para determinar si dicho impacto debe ser el objetivo principal tanto del desarrollo social como de la actividad política. Rapley (2003) nos indica que se siguieron tres líneas de estudio. Plantea tres tradiciones de investigación: por un lado, la corriente escandinava (a partir de trabajos realizados por Erikson, 1993, y Uusitalo, 1994), la cual propone que la calidad de vida consiste en satisfacer las necesidades básicas; por otro, la corriente anglosajona o "American Quality of Life", que propone que la calidad de vida consiste en medir la satisfacción y la felicidad en las experiencias de los individuos (Campbell, Converse y Rodgers, 1976), y, por último, la corriente germana, que según Noll (2002) integra las dos anteriores combinando condiciones objetivas de vida y bienestar subjetivo, dando lugar a diferentes tipos de situaciones: bienestar (buenas condiciones de vida y bienestar subjetivo positivo); disonancia (buenas condiciones de vida pero bienestar subjetivo negativo); privación (malas condiciones de vida y bajo bienestar subjetivo), y adaptación (malas condiciones de vida pero alto bienestar subjetivo) (Noll y Zapf, 1994).

El interés por el estudio de la calidad de vida en el ámbito de la salud queda reflejado en el número creciente de publicaciones en los últimos años. Por ejemplo, en una búsqueda del término en inglés (HRQoL) en PubMed, desde el año 1990 hasta la actualidad, encontramos un total

de 39.455 artículos que mencionan el término en el título o en el resumen *(abstract)*. La tendencia de la serie parece ser más rápida que lineal, superando los 4.000 artículos/año en 2018. Por su parte, el término *utilidad en salud (health utility)* ha sido mencionado en el título o en el resumen en un total de 1.873 publicaciones desde 1990, con un crecimiento más lineal y menos pronunciado.

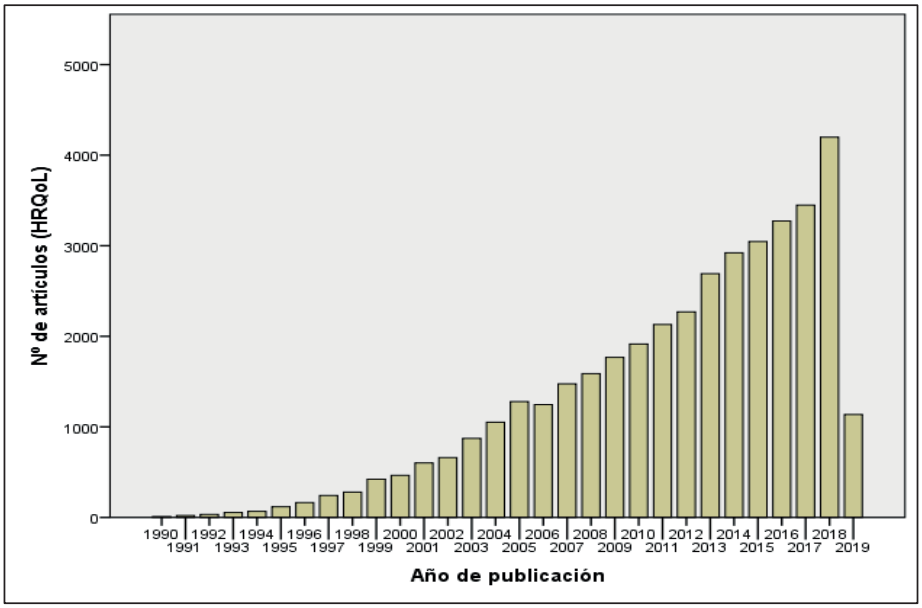


Figura 2. Número de artículos que mencionan "health related quality of life" en su título, desde 1990 (PubMed)

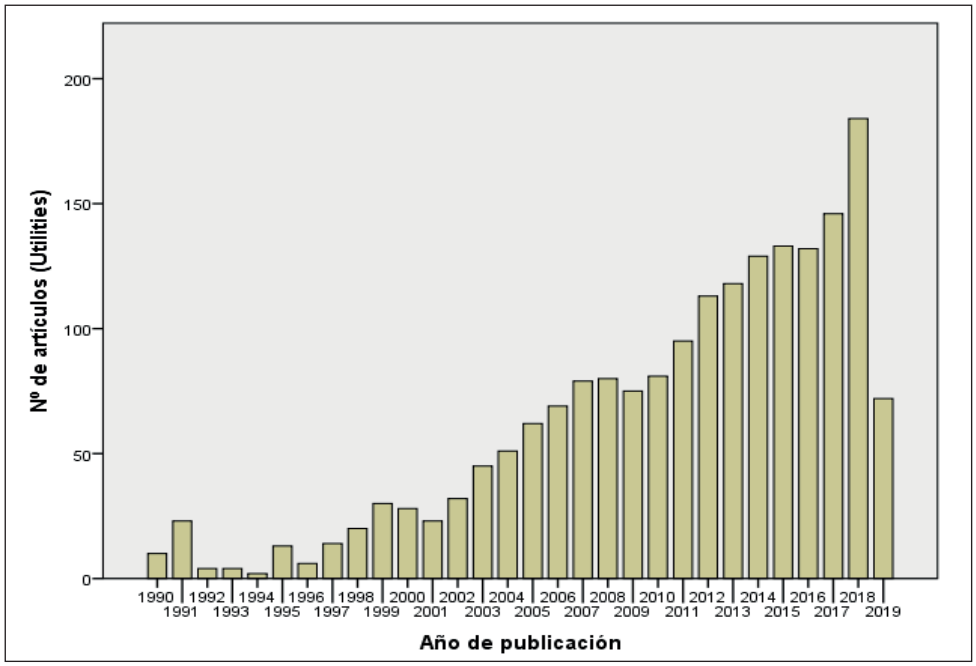


Figura 3. Número de artículos que mencionan "health utilities" en su título, desde 1990 (PubMed)

El hecho de que la calidad de vida esté íntimamente relacionada con la satisfacción implica que tiene un carácter individual y una variabilidad en el tiempo, ya que aspectos importantes para una persona pueden no serlo para otra, y estos pueden variar si cambian las circunstancias o la percepción del individuo. Por ello, Hickey et al. (1999) llegan a definir la calidad de vida como "lo que un individuo determina que es".

Actualmente, a pesar de no tener una definición formal comúnmente aceptada, existe consenso respecto a que la calidad de vida tiene que ver con las condiciones de vida y las evaluaciones subjetivas, sin existir acuerdo en el número o la naturaleza de las dimensiones que componen el constructo. Por su multidimensionalidad y complejidad se han desarrollado una gran variedad de teorías para analizarla desde distintas disciplinas. Así, podemos señalar los trabajos realizados en la literatura conocida como "Economics of Happiness" (Frey y Stutzer, 2002) en el campo de la economía, o los realizados por Segurado y Agulló (2002) en el campo laboral. Todo ello particulariza en cierto modo su medición en cada ámbito y hace muy difícil en la práctica formular una teoría general unificada, obligándonos a utilizar modelos particulares en cada disciplina de aplicación.

La definición de calidad de vida más respetada es la propuesta por el Grupo de Calidad de Vida de la OMS (World Health Organization Quality of Life Group, WHOQOL), que destaca su carácter subjetivo al definirla como "las percepciones de los individuos de su posición en la vida en el contexto cultural y de valores en el que viven y en relación a sus metas, expectativas, estándares y preocupaciones" (WHOQOL Group, 1995).

De lo expuesto hasta ahora podemos concluir que el constructo *calidad de vida* tiene como características básicas su carácter de percepción individual, su variabilidad en el tiempo y su multidimensionalidad estructural. Pero el hecho de que la calidad de vida sea subjetiva no quiere decir que no pueda ser objetivable y, en consecuencia, medible.

Es difícil encontrar concreción sobre las dimensiones particulares que deben ser incorporadas a la valoración de la calidad de vida, incluso cuando nos limitamos al ámbito de la salud. Un análisis bibliográfico (si buscamos en la base de datos de Medline encontramos 1.517 publicaciones, y 8.528 en la base de datos PubMed desde 2000 al 2018) ofrece un amplio listado de com-

ponentes de la calidad de vida: calidad del medio ambiente, entorno, valores, relaciones sociales, situación laboral y económica, salud, estado emocional, espiritualidad, ocio, cultura y un largo etcétera. La propia salud es una de las pocas dimensiones sobre la que parece existir consenso en cuanto a su contribución a la calidad de vida global. Diversos estudios demuestran una fuerte asociación entre salud y bienestar (satisfacción) (v. g.: Sanda, 2008; Ong, 1999; Tate, 2002).

El concepto *calidad de vida* se utiliza, como hemos indicado anteriormente, indistintamente en el mundo de la economía, de la política y de la salud, entre otros, pero es en la disciplina de la salud en la que ha suscitado mayor interés y en la que centraremos nuestros estudios. De hecho, en el ámbito de la salud, se ha acuñado el término *calidad de vida relacionada con la salud* (CVRS o, en inglés, *Health Related Quality of Life* - HRQoL), que es un concepto más restringido y en el que nos centraremos de aquí en adelante.

Pero esta diferenciación entre calidad de vida y calidad de vida relacionada con la salud no está exenta de polémica. Podemos encontrarnos con autores que consideran que la CVRS es una parte de la calidad de vida (v. g.: Awad, 1997; Etcheld, 2003), mientras que otros piensan que ambos términos pueden ser intercambiables, pues miden dimensiones similares (v. g.: Burke, 2001; Wu, 2000). Pese a ello, la gran mayoría de los autores sugieren que la calidad de vida debe ser diferenciada de la CVRS, debido a que este nuevo concepto se utiliza exclusivamente en medicina, con el objetivo de evaluar la calidad de las intervenciones, y por limitarse a la relación del paciente con su enfermedad, con los cuidados médicos que recibe o con el impacto de la enfermedad en su vida diaria.

La CVRS es, por tanto, una medición de la salud (cuantificación numérica) realizada desde el punto de vista del propio paciente. Su estudio ha sido muy importante para generar objetivos terapéuticos, guías y políticas en el ámbito de la salud, para describir la relación enfermedad-paciente, para mejorar la práctica clínica y para poder realizar estudios de eficacia, eficiencia y efectividad. Pero no todo son ventajas. Testa (2000) plantea el inconveniente de que los resultados de salud a menudo son evaluados utilizando mediciones influidas por percepciones y expectativas de los pacientes, lo cual podría alejarlas del criterio clínico. Así mismo, Lawton (1999) propone que la vida diaria es más que la salud de un sujeto, que las decisiones

no se basan solo en síntomas presentes y que no se considera la opinión de sujetos sanos a la hora de calcular la calidad de vida relacionada con la salud. Además, debe tenerse en cuenta que, pacientes con similares criterios clínicos, emiten a menudo respuestas diferentes respecto a su CVRS. Aspectos como la experiencia previa con la enfermedad, la edad, el sexo, la adaptación, la presencia de comorbilidades o los efectos indeseados del tratamiento, pueden introducir variaciones importantes en la valoración del sujeto respecto al estado de salud atribuido en una enfermedad concreta.

El continuo aumento de la demanda de una mayor salud y calidad de vida por parte de la población en general, el aumento considerable de la esperanza de vida y la disponibilidad de unos recursos sanitarios limitados, han hecho que la comunidad científica preste más atención a lo que se ha venido a llamar "Outcomes" (resultados) en salud, dando lugar al desarrollo de un área nueva de investigación denominada *investigación de resultados de la salud* (IRS). La investigación e intervenciones médicas persiguen aportar evidencias y dar apoyo a la toma de decisiones en el sector de la salud, buscando la efectividad y no tanto la eficacia en las intervenciones. Es en este contexto en el que más se ha desarrollado el uso de la CVRS como un resultado de salud de especial importancia.

CONCEPTO DE UTILIDAD

Los órganos gubernamentales responsables de los asuntos relacionados con la salud definen tres conceptos que el clínico ha de tener presentes para la utilización correcta de los medios disponibles. Se define la ***eficacia*** como la probabilidad ideal de que un individuo se beneficie de una tecnología para resolver un problema médico. Se define la ***efectividad*** como la probabilidad real de que el paciente se beneficie de dicha tecnología. Por último, se define la ***eficiencia*** como la relación entre los beneficios obtenidos por la efectividad y los costes que supone obtener dichos beneficios. Dado el alto nivel de seguridad y eficacia conquistados por la mayoría de los tratamientos habituales, es posible (y también necesario) valorar la eficiencia de posibles alternativas terapéuticas disponibles, con la finalidad de reducir al máximo los costes de la intervención, en la medida de lo posible.

Dependiendo de las fuentes de información existentes, se puede valorar la eficiencia de una intervención de distintas maneras. Cuando el informante es el propio terapeuta/médico, se estima como **coste-efectividad,** si es el paciente el que informa, se estima como **coste-utilidad,** y cuando el informante es el gestor económico, se estima como **coste-beneficio**. Puesto que la salud es en sí misma un valor, y no se puede permitir su pérdida, aun cuando el coste sea elevado, serán el coste-efectividad y el coste-utilidad los que se tengan en mayor consideración.

Para medir el coste-utilidad, en primer lugar, debemos definir la utilidad. Para ello nos basamos en el primer principio de la economía del bienestar, que establece que las personas deciden de manera que su bienestar sea el mayor posible o que «maximicen su utilidad» estando sujetos a determinadas restricciones (tiempo, dinero…), indicando de esta forma sus preferencias. Además, se asume que "cada persona es el mejor juez de su propio bienestar".

Weinstein et al. (1996) indican que el análisis de las preferencias va a depender de la perspectiva que se adopte para el análisis. Si se toma la perspectiva del paciente, y por tanto sus preferencias, lo que se persigue es valorar si el tratamiento es óptimo. Sin embargo, si lo que se pretende es establecer prioridades entre programas o recursos sanitarios públicos, entonces hay que analizar la perspectiva de la población (perspectiva social). Como esta perspectiva representa el interés colectivo, son las preferencias de la población general las que deben ser tenidas en cuenta.

Para maximizar la utilidad, es decir, que el estado de bienestar sea el mayor posible, se recurre a Morgenstern y Von Neumann (1953), que en su libro *Theory of Games and Economic Behavior* analizan situaciones donde se deben tomar decisiones bajo un cierto grado de incertidumbre. La decisión que se tome dependerá del grado de riesgo que los sujetos puedan admitir. La teoría de Morgenstern y Von Neumann sobre la utilidad esperada nos dice que "la utilidad esperada de cada una de las alternativas es igual a la suma de cada una de las utilidades estimadas previamente multiplicadas por su respectiva probabilidad".

Por todo lo anterior podemos concluir que la *utilidad* es el grado de preferencia que la sociedad elige entre una serie de alternativas de salud que se le presentan multiplicada por la probabilidad de que esta alternativa se dé.

Su valor oscilará entre 1 (salud total) y 0 (muerte), aunque pueden obtenerse valores inferiores a 0 cuando se presentan situaciones que son consideradas "peor que estar muerto".

Para generar una utilidad asociada a un estado de salud se requiere estimar un modelo. Podemos utilizar dos enfoques: uno denominado *descompuesto* y otro denominado *compuesto* o *de inferencia estadística.* El primero requiere que los encuestados valoren cada ítem de una dimensión manteniendo constantes las demás dimensiones en un determinado nivel (siempre el mismo para todas las dimensiones), creando funciones uniatributo para a partir de ellas generar las utilidades del resto de estados de salud por procedimientos econométricos. El enfoque inferencial requiere que cada encuestado valore un subconjunto de estados de salud multiatributo y, mediante regresiones lineales, por lo general de efectos aleatorios, estimar el resto.

En estos métodos su principal ventaja radica en que minimiza el número de estados de salud que es necesario valorar de manera directa para luego realizar predicciones para el resto de estados posibles. Stevens et al. (2007) analizan ambos por separado y de manera conjunta, concluyendo que el procedimiento compuesto tiene mayor validez predictiva.

La función de utilidad multiatributo (*multiatribute utility function* - MAUF) generada puede ser bien aditiva o multiplicativa respecto a los pesos obtenidos para cada uno de los niveles de las dimensiones del instrumento. Aunque resultan algo más complejas, las MAUF multiplicativas presentan el beneficio de tener en cuenta la interacción entre las distintas dimensiones, puesto que los coeficientes se multiplican ente sí.

Dicha función generará una escala de utilidad multiatributo (*multiatribute utility scale* - MAUS) o escala de utilidad, en la que cada estado de salud recibe un único valor de utilidad.

Existen varios procedimientos para la estimación de escalas de utilidades. Los más utilizados son (Baron et al. (2001).

- VAS - *Visual Analogue Scale:* los estados de salud se valoran situándolos sobre una escala visual analógica o termómetro de salud.
- SG - *Standard Gamble:* se realiza valorando cada estado de salud frente a una situación de incertidumbre en la que se varía la probabilidad de curación (p) y muerte ($1 - p$).
- TTO - *Time Trade-Off:* se valora cada estado de salud frente a una cantidad de tiempo en salud total.
- WTP - *Willingness to Pay Estimation:* se valora cada estado de salud frente a una cantidad de dinero que se desea pagar por él.
- ME - *Magnitude Estimation:* se valora la malignidad (aversión) de cada estado de salud frente a la aversión del siguiente estado de salud menos grave.
- PTO - *Person Trade Off:* se valora cada estado de salud frente a la cantidad de personas que habría que curar en ese estado de salud.

Un inconveniente importante en las escalas de utilidad es su falta de homogeneidad entre los distintos niveles de deterioro en cada atributo. La modificación (aumento o disminución de una unidad) en una de las dimensiones no conlleva que la utilidad se modifique un valor constante. Tal y como se refleja en la figura 4, si mantenemos constante todas las dimensiones menos una, la utilidad no sigue una línea recta, sino que genera una curva de evolución según aumenta la severidad del atributo.

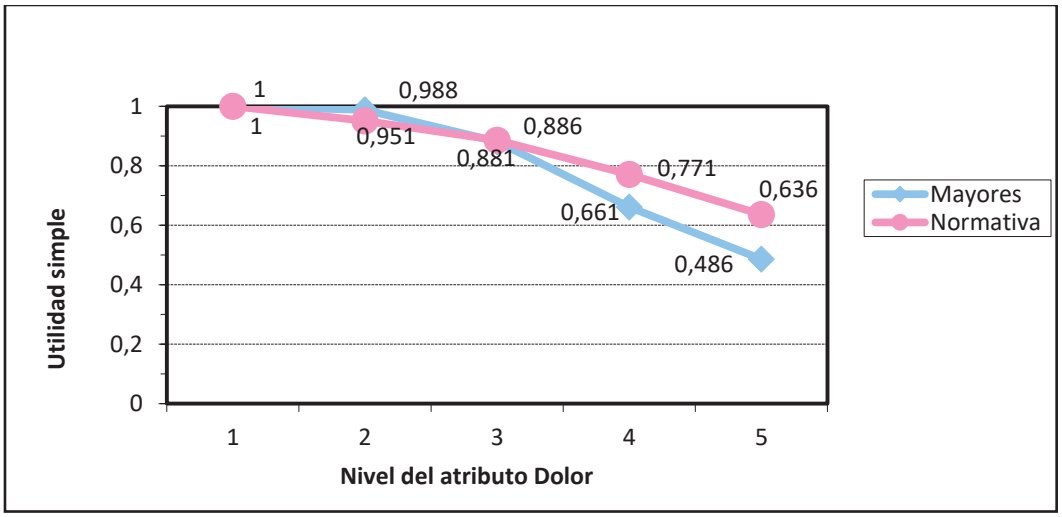


Figura 4. Peso de los niveles del atributo dolor en el HUI-III (muestra normativa y muestra de personas mayores

Además, si ordenamos todos los estados de salud valorados por un instrumento genérico de utilidad y representamos la utilidad asociada a cada estado de salud, se puede comprobar que los cambios de utilidad no son homogéneos en toda la métrica de la utilidad, de manera que pequeños cambios en el orden de los atributos situados en los extremos conllevan mayores cambios de utilidad que en la parte central

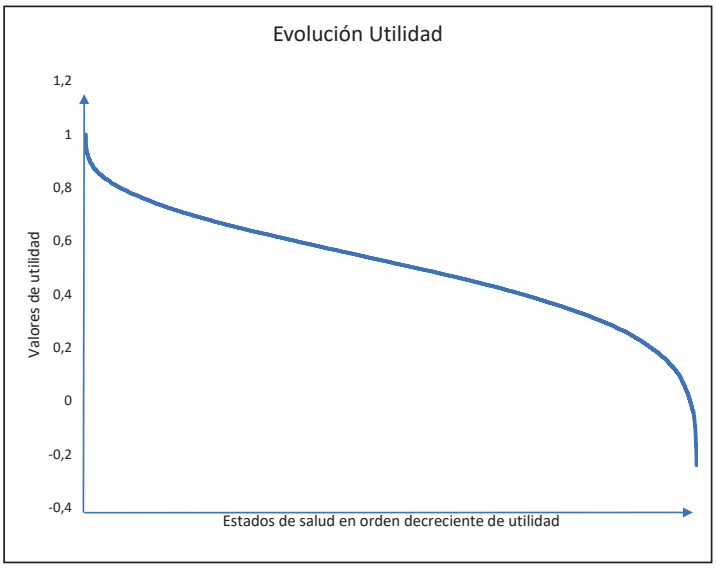


Figura 5. Evolución de la puntuación de utilidad en el instrumento SF6D

Más adelante se expone cómo se obtiene la utilidad en los instrumentos genéricos empleados en esta tesis con sus diferentes métodos de obtención de la utilidad.

Fernández López et al. (1994) nos presentan en su artículo un cuadro de referencia del valor de utilidad medio obtenido en diferentes patologías

Tabla 1. Valores de utilidad en diferentes patologías

Estado de salud	Utilidad
Sano	1,00
Menopausia	0,99
Efectos secundarios a antihipertensivos	0,95
Angina leve	0,90
Angina moderada	0,70
Algunas limitaciones físicas con dolor ocasional	0,67
Angina severa	0,50
Estar ciego, sordo o mudo	0,39
Internamiento hospitalario	0,33
Muerte	0,00
Confinado en cama con dolor severo	< 0,00
Inconsciente	< 0,00

La distribución de valores de utilidad entraña sus propias peculiaridades que limitan su tratamiento mediante distintas técnicas estadísticas, como las de modelado.

De manera característica, la forma de la distribución de las utilidades en una población general muestra una forma claramente asimétrica negativa, con acumulación de casos en la puntuación máxima, correspondientes a la salud total.

Además, debido a la fórmula de cálculo de la función multiatributo, la función de densidad presenta un desplazamiento en la distribución en el que no existen valores entre el valor de salud máxima y el siguiente valor de la distribución (debido a la presencia de la constante en la MAUF). En el caso del EQ-5D-3L y utilizando la estimación multiatributo basada en el método de escala visual analógica (VAS), el siguiente valor posible a continuación de la salud total es el valor 0,9488.

Dependiendo de las patologías, la distribución de valores no es monótona, sino que presenta acumulaciones de sujetos con frecuencia relativamente más alta en distintos puntos de la métrica. Este comportamiento ha llevado a pensar que podría ser más adecuado tratar las puntuaciones desde el abordaje de las distribuciones mixtas (una combinación de distribuciones con distintos parámetros). Además, la acumulación de sujetos en las puntuaciones más bajas de utilidad se hace más evidente en las patologías más graves (ver más abajo puntuaciones del EQ-5D-3L en pacientes con patrología renal).

A continuación, se muestran las distribuciones empíricas de diversas patologías utilizando varios instrumentos (EQ-5D-3L, SF-6D y HUI-III) y diversos métodos de corrección (VAS e intercambio temporal).

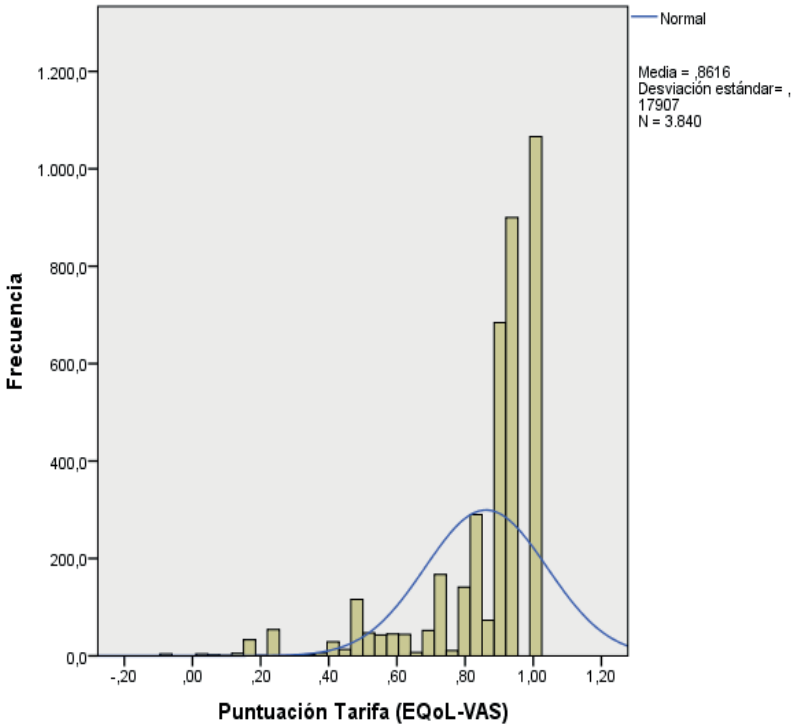


Figura 6. Distribución de utilidades del EQ-5D-3L (método VAS) en pacientes con reflujo gastroesofágico

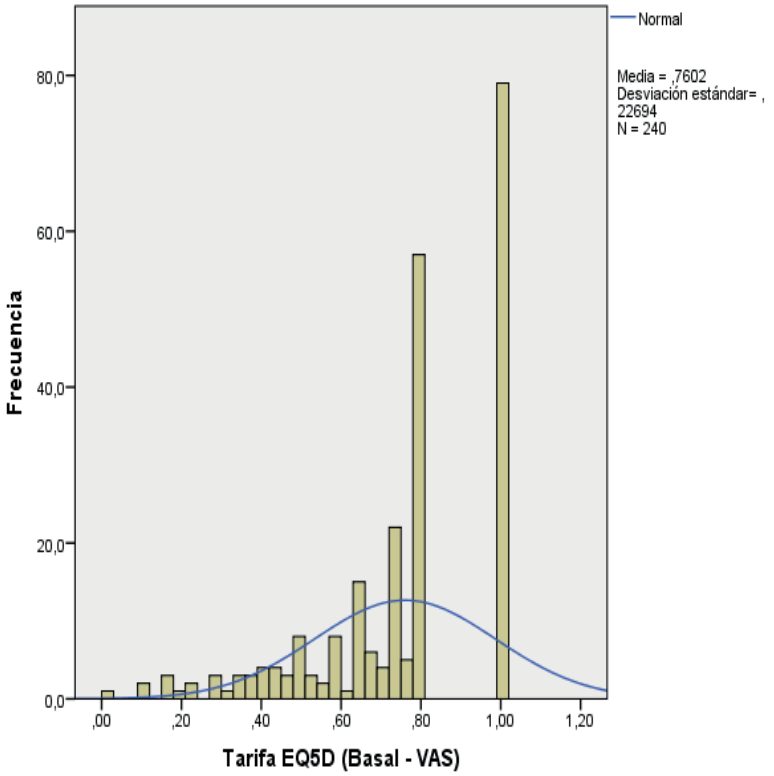


Figura 7. Distribución de utilidades del EQ-5D-3L (método VAS) en pacientes con vejiga hiperactiva

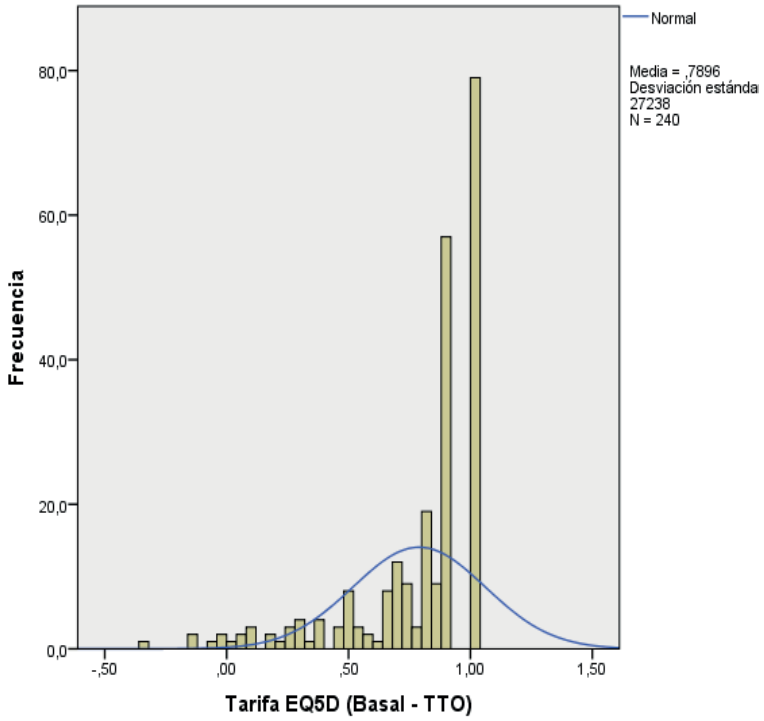


Figura 8. Distribución de utilidades del EQ-5D-3L (método TTO) en pacientes con vejiga hiperactiva

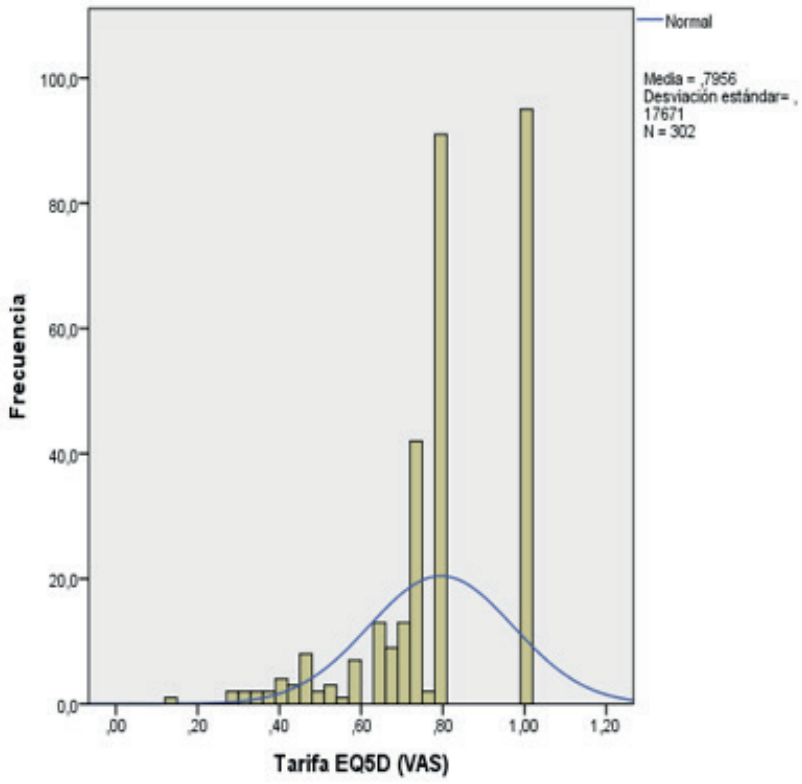


Figura 9. Distribución de utilidades del EQ-5D-3L (método VAS) en pacientes perimenopáusicas

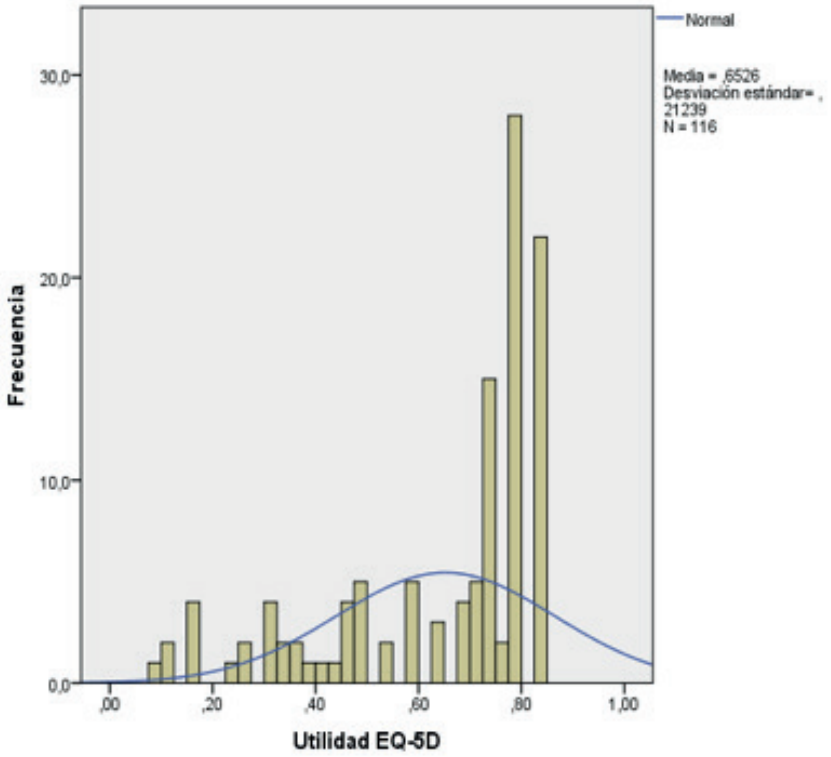


Figura 10. Distribución de utilidades del EQ-5D-3L (método VAS) en pacientes con enfermedad crónica de riñón

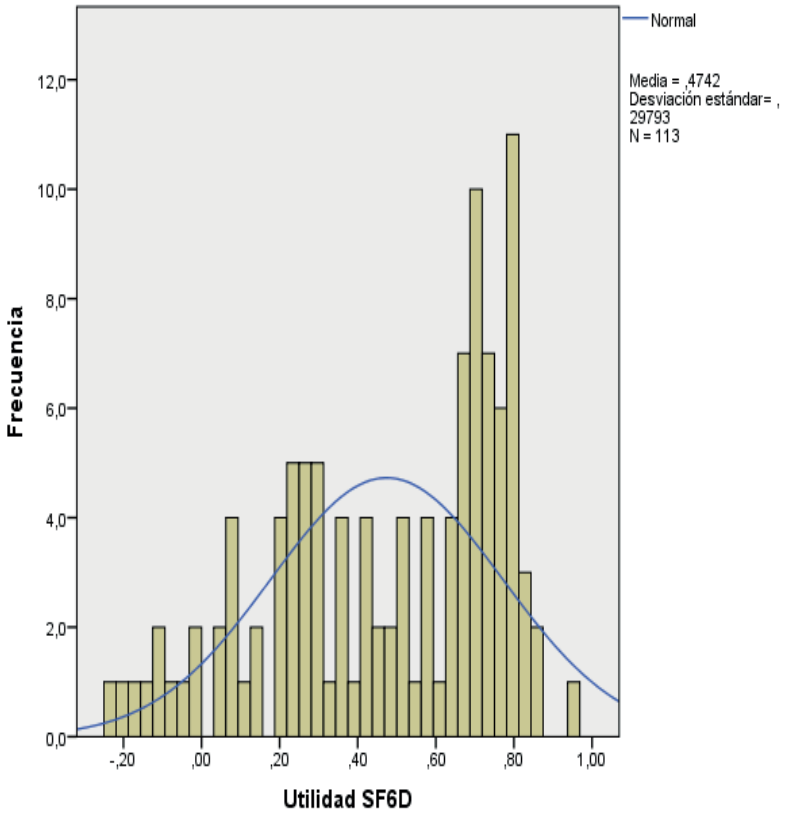


Figura 11. Distribución de utilidades del SF-6D en pacientes con enfermedad crónica de riñón

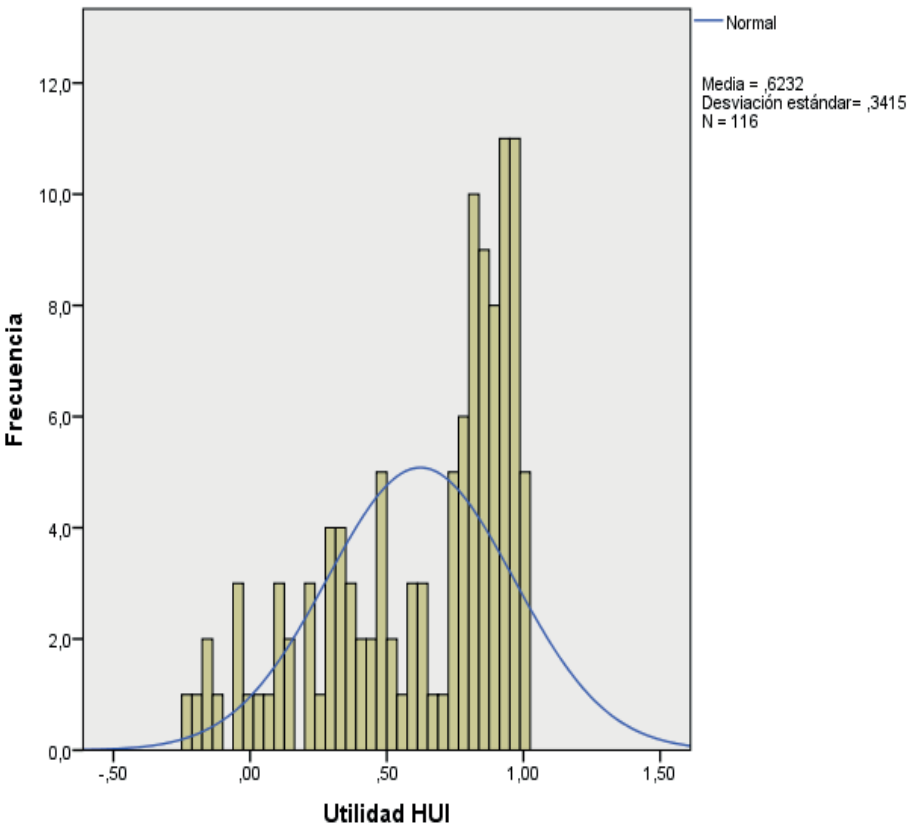


Figura 12. Distribución de utilidades del HUI-III en pacientes con enfermedad crónica de riñón

CONCEPTO DE AVAC

Para poder realizar comparaciones entre políticas sanitarias, patologías, tratamientos o intervenciones es necesario obtener una valoración numérica global que nos permita una comparación directa, simple, inequívoca, de métrica estable y estandarizada.

El indicador de coste-utilidad más extendido es el de "años de vida corregidos por la calidad" (AVAC, *quality adjusted life years* - QALY), que combina en una única cantidad los valores de mortalidad y morbilidad. El índice de AVAC se define como la relación entre el aumento de esperanza de vida y el aumento en la calidad de vida, para un paciente o un conjunto de pacientes, alcanzado por una intervención o tratamiento. La morbilidad se valora utilizando la cantidad de vida relacionada con la salud en una métrica normalizada (0-1), de manera que los AVAC se pueden obtener a partir de la calibración directa de ambas cantidades (curiosamente, al utilizar valores de utilidad en un formato similar al de las proporciones, basta multiplicar el número de años de vida por el valor de la utilidad, obteniendo un valor similar al que obtendríamos si hubiéramos realizado un cociente).

$$N.^{o}\,AVAC = Cantidad\ de\ años\ de\ vida \times Calidad\ de\ vida$$

A pesar de su capacidad de síntesis, los AVAC presentan el inconveniente de poder llevarnos a considerar similares una intervención capaz de alargar de manera considerable la vida de los pacientes con una calidad de vida inferior a la óptima, con el resultado de otra intervención capaz de alargar la vida una cantidad ínfima, pero con una excelente calidad de vida. También puede darse el caso de que la calidad de vida no sea constante durante todo el periodo de vida, que los años de vida ganados no sean igualmente valiosos en todas las cohortes de edad y que pueden generarse efectos no deseados y costes indirectos que no son contemplados en el indicador.

Un aspecto fundamental a resaltar es que los AVAC están orientados a medir preferencias por un estado de salud durante un periodo de tiempo determinado. Este índice valora la aceptación o preferencia que un paciente tiene por un estado de salud, por lo que su valoración es siempre comparativa con respecto a otros estados.

Para calcular el número de AVAC obtenidos por una intervención y así valorarla, necesitaremos estimar los años de vida ganados y la calidad de vida de los pacientes. Los años de vida ganados se obtienen mediante estudios longitudinales prospectivos de seguimiento y se acostumbra a utilizar estadísticos como la mediana de supervivencia. La calidad de vida suele medirse mediante cuestionarios autoinformados validados, utilizando estadísticos de tendencia central.

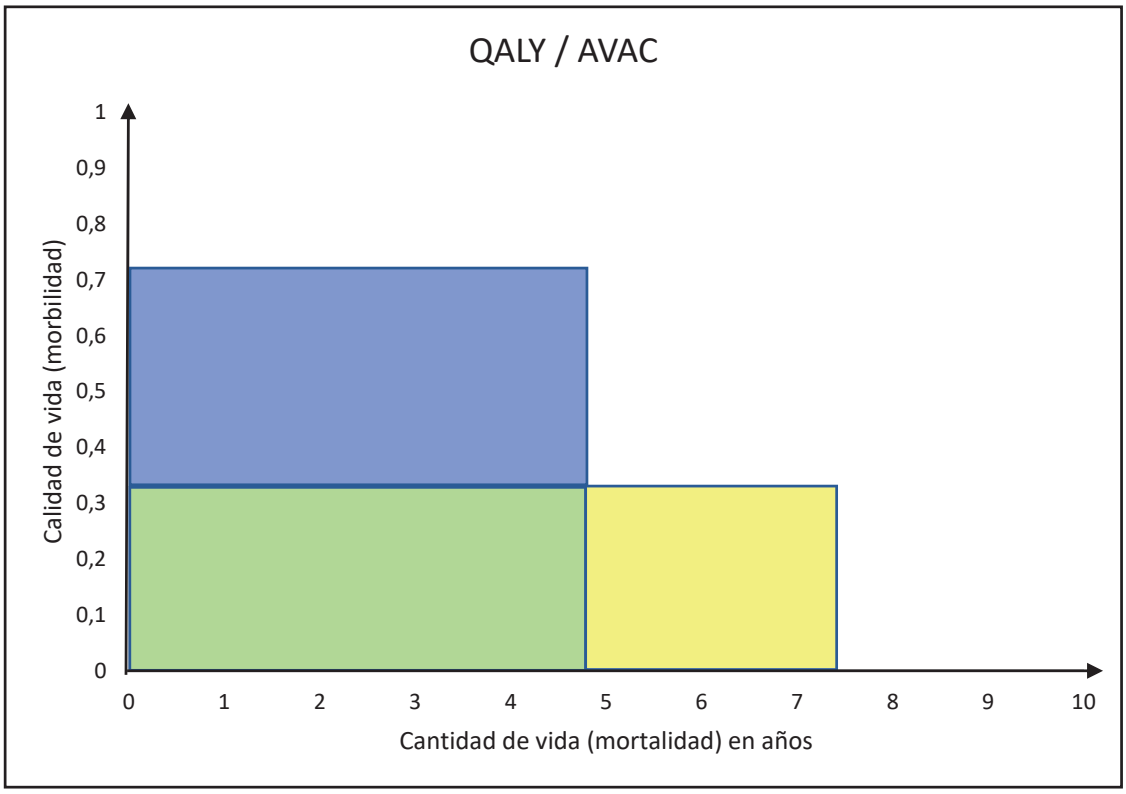


Figura 13. Esquema obtención QALY

INSTRUMENTOS PARA MEDIR LA CVRS

Cuando utilizamos instrumentos o escalas de medición en psicología o en ciencias sociales, estos tienen como característica que asumen que hay al menos un constructo subyacente que no es medible directamente, pero que es accesible utilizando un conjunto de ítems homogéneos entre sí, que reflejan la magnitud del constructo y que están relacionados con el constructo. Bajo los supuestos de la teoría clásica de los test (TCT), es lo que hemos dado en llamar escalas de tipo "reflexivo", ya que las mediciones obtenidas a partir de los ítems reflejan la magnitud del constructo que se desea medir.

Cuando hablamos de la medición en el ámbito de la salud, nos encontrarnos con un conjunto de instrumentos en los que los ítems que los componen no son homogéneos entre sí y que su objetivo no es otro que crear un índice válido que agregue la información de los distintos componentes de esa escala de salud. Es lo que denominamos un modelo formativo. En este caso, cada indicador (o variable observada) capta un aspecto diferente de la realidad que queremos medir, pudiendo ocurrir que incluso las variables observadas fueran independientes entre sí. La combinación de indicadores se entiende como un agregado que capta toda la información de interés para crear el índice resumen. Algunos autores han denominado a esta aproximación de la definición del bienestar como un síndrome (Warr, 2012), en el sentido de que la presencia de un mayor número de indicadores dará más sustento a la presencia del síndrome que los combina.

La diferencia conceptual entre escalas reflexivas y formativas es similar a la que encontramos entre el modelo de factor común y el modelo de componentes principales (PC).

En el modelo de factor común, las variables observadas reflejan la expresión del constructo (el factor común a todos los ítems), con un cierto grado de error de medición. Se intenta recoger la información compartida entre el ítem observado y el factor común (la comunalidad), despreciando la información particular del ítem (unicidad). En este modelo, las puntuaciones factoriales son desconocidas y se estiman a partir de diversos procedimientos.

En el modelo de componentes principales, la componente principal recoge el máximo de variabilidad compartida por los ítems observados, creando una combinación lineal que explique el máximo de varianza común a todos los ítems. En este modelo las puntuaciones en una componente se obtienen por combinación directa de las variables observadas.

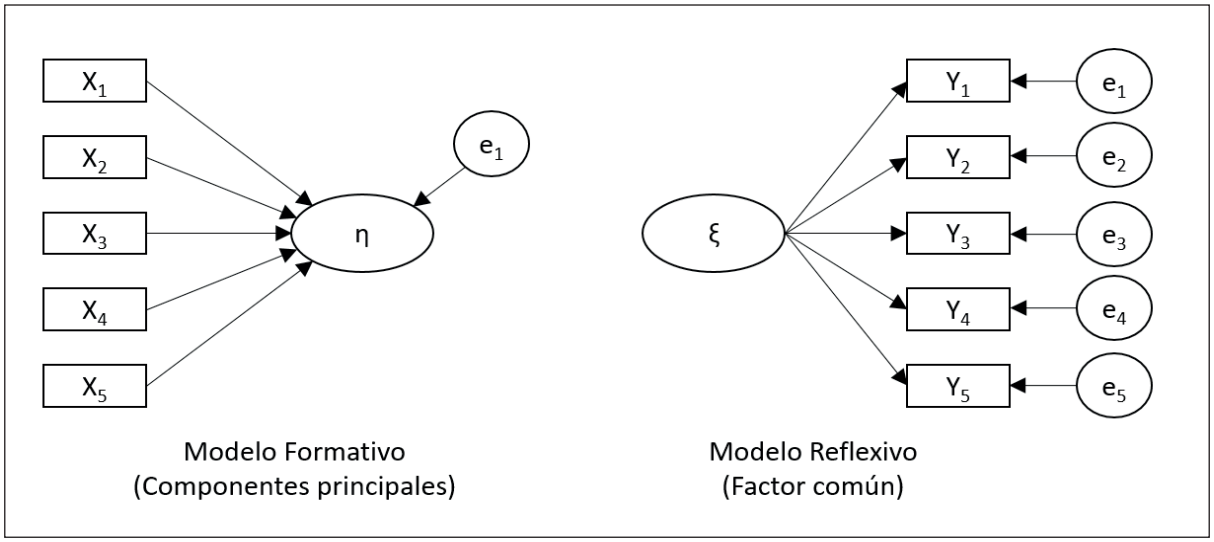


Figura 14. Representación de los modelos formativo y reflexivo

Feinstein (1987) las llamó escalas clinimétricas, y definió la clinimetría como "calificaciones, escalas, índices, instrumentos u otras expresiones arbitrarias que se han creado como 'mediciones' para aquellos fenómenos clínicos que no pueden medirse en las dimensiones habituales de los datos de laboratorio".

Por tanto, las escalas clinimétricas son diferentes de las escalas psicométricas porque, en las psicométricas, los ítems se eligen porque se cree que están relacionados con un constructo subyacente que queremos medir (son escalas reflexivas), mientras que en las clinimétricas los ítems se eligen de acuerdo con lo que queremos que mida la escala (son escalas formativas), dándose la circunstancia de que las correlaciones entre las variables "causales" con frecuencia se deben a factores extraños y varían de un estudio a otro, al carecer las escalas clinimétricas de un modelo subyacente.

En este punto debemos indicar que existe una importante controversia sobre si en realidad debe hacerse una diferenciación entre escalas clinimétricas y psicométricas, tal y como apunta Streiner (2003).

Resumiendo, muchos cuestionarios han sido creados a partir de una combinación de dimensiones (sobre las que no hay acuerdo absoluto y que pueden variar según la patología) y la mayoría no obedece a un modelo explicativo concreto y claramente especificado. Aunque existen cuestionarios cumplimentados por el profesional clínico o informados por un allegado, se prefiere aquellos que son autoevaluados por el propio sujeto (sin intervención de otras personas) reflejando las percepciones y experiencias individuales de quien responde (de ahí el nombre de Patient Reported Outcomes, PRO). Además, dado que van a ser utilizados como medidas de resultado (*output*) para valorar el impacto de las intervenciones clínicas, se prefiere aquellos cuestionarios con una alta sensibilidad a los cambios en la situación del paciente y no los que miden rasgos estables de la persona.

Es conveniente tener en cuenta que, al recoger la percepción individual o subjetiva, los instrumentos se encuentran vinculados a la cultura y al estamento social al que pertenece el sujeto, cuestiones que pueden influir en las expectativas del que responde y también pueden fijar anclajes sobre la definición social de los conceptos. "La percepción de normalidad o anormalidad de una condición física o mental está de acuerdo a un grupo de referencia, teniendo un marcado efecto la cultura sobre la salud y la enfermedad" (Thoits y Angel, 1987), por ese motivo, son cada vez más apreciados los proyectos de armonización transcultural en la creación y validación de instrumentos.

Si nos centramos en su multidimensionalidad, el principal problema que nos encontramos es que no existen normas sobre lo que debería o no incluirse a la hora de definir la calidad de vida relacionada con la salud. La mayoría de los cuestionarios coinciden en incluir las dimensiones física, psicológica y social. Por su parte, el grupo WHOQoL, de la Organización Mundial de la Salud, propone estudiar seis dimensiones o ejes: "físico, psicológico, nivel de independencia, relaciones sociales, medio ambiente y, por último, espiritualidad, religión y creencias personales". Pero este modelo no ha logrado el consenso de todos los investigadores, por lo que se han desarrollado otros modelos, fundamentalmente relacionados con enfermedades específicas, debido a las particularidades de las mismas. Uno de los esfuerzos para desarrollar un modelo capaz de integrar todas las medidas que pueden recogerse de los pacientes es el modelo de Wilson y Cleary (1995), que intenta integrar los dos marcos de referencia dominantes, uno desde el punto de vista biomédico (centrado en aspectos etiopatogénicos y fisiológicos) y el otro

desde las ciencias sociales (centrado en el desempeño, los apoyos y estrategias de afrontamiento y el bienestar subjetivo). Pero tampoco ellos han sido capaces de establecer las dimensiones mínimas del modelo de calidad de vida relacionada con la salud.

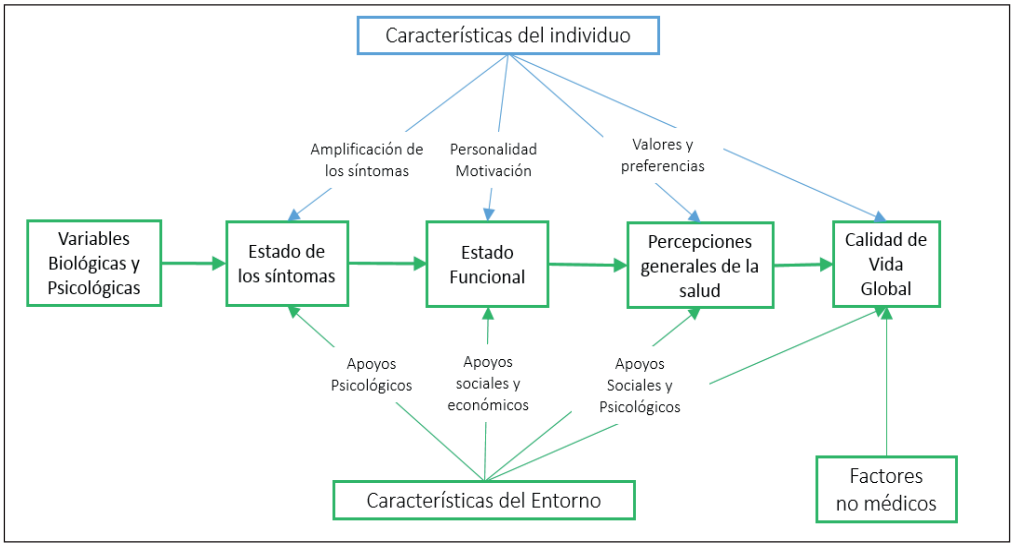


Figura 15. Modelo de CVRS de Wilson y Cleary (1995)

Vosvick et al. (2003), para el estudio de enfermedades crónicas, definen cinco dominios: funcionamiento físico, energía/fatiga, funcionamiento social, desempeño del rol y salud general, y dolor; Badía y Lizán (2003) definen tres: físico, psíquico y social; y Bishop et al. (2002) definen un modelo de siete dimensiones: frecuencia de las crisis, función física, interferencia de las crisis, apoyo social, salud general, salud mental y empleo.

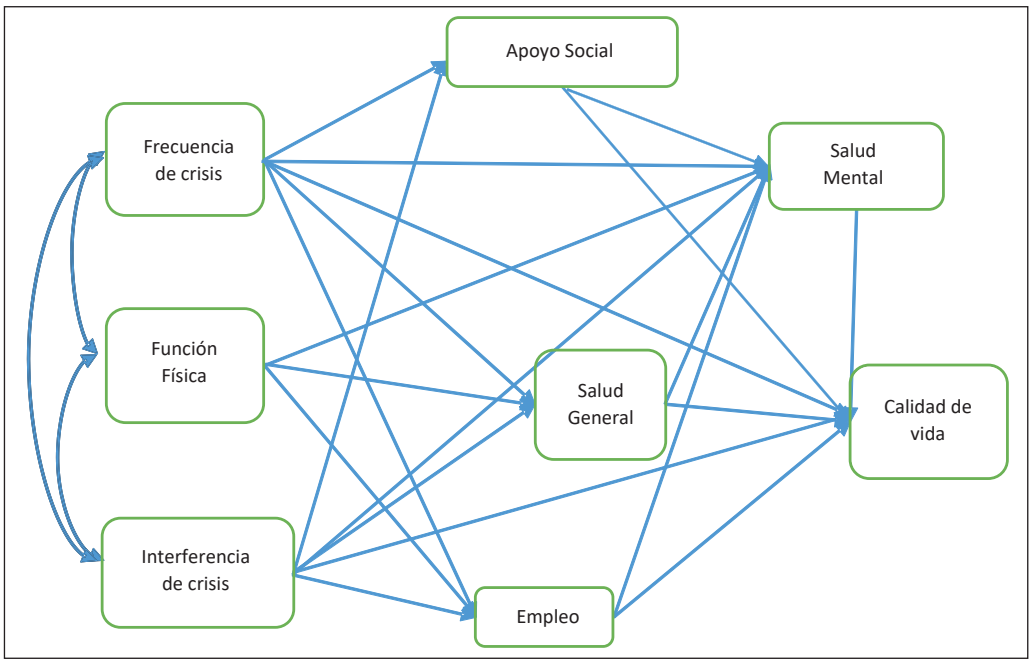


Figura 16. Modelo de CVRS de Bishop et al. (2002)

La tradición psicométrica ha dado lugar al desarrollo de múltiples instrumentos para medir la CVRS de los pacientes. Entre ellos encontramos el Medical Outomes Survey Short Form de 36 ítems (MOS SF-36), el General Health Questionnaire (GHQ), el Notingham Health Profile (NHP), la familia QLQ de la EORTC, la familia FACT de la organización FACIT, el General Health Questionnaire (GHQ-12) y un largo etcétera.

No hemos considerado aquí los instrumentos clásicos para la evaluación y diagnóstico de problemas psicológicos, ya que se centran únicamente en alguna dimensión particular de la salud mental (o bienestar psicológico). Conviene recordar algunos cuestionarios de uso muy extendido como: la escala de depresión de Beck, la escala Hamilton para el diagnóstico de la ansiedad y la depresión (HAM-A y HAM-D), la escala HADS para el diagnóstico de la ansiedad y la depresión, la escala PANAS de sentimientos positivos y negativos, la escala PANSS de pensamientos positivos y negativos, el cuestionario de ansiedad rasgo/estado STAI, etc.

Pero, mientras que la tradición psicométrica ha perseguido el escalamiento (ordenación) óptimo de los sujetos medidos a partir de la información subjetiva obtenida por autoinformes, la tradición econométrica ha apostado por una ordenación de las puntuaciones de calidad de vida que obedezca y refleje el valor social de cada estado de salud. Medición que se ha denominado medición basada en las "preferencias" y cuyas puntuaciones obtenidas finalmente son las utilidades (sociales) asociadas a los distintos estados de salud.

El primer enfoque pretende reflejar el cambio en la salud por efecto del tratamiento o discriminar entre pacientes con diferentes niveles de gravedad (severidad), permitiendo un escalamiento o diagnóstico individual de cada paciente al que se aplica el cuestionario, comparándolo normalmente con una muestra normativa. En cambio, los instrumentos dirigidos a la medición de preferencias pretenden valorar la magnitud de la preferencia (o aversión) de la población por el estado de salud en el que se encuentra el paciente, frente a otros estados de salud posibles (más abajo se discute cómo se obtienen las utilidades).

Los instrumentos con vocación psicométrica suelen ser de tipo específico, dirigiéndose a valorar una única patología, dada la variabilidad de síntomas y discapacidades impuestas por las distintas patologías. Aunque también existen instrumentos de tipo genérico que están

construidos en torno a un núcleo genérico común a varias patologías, al que se añaden módulos específicos dependiendo del tipo de patología. Por ejemplo, la familia del QLQ de la EORTC, (Organización Europea para la Investigación y Tratamiento del Cáncer), que cuenta con un módulo central para cualquier tipo de cáncer (QLQ-C30) y módulos específicos para distintos tipos de cáncer (pulmón, pecho, leucemia, vejiga, etc.). Los instrumentos específicos tienen la ventaja de ser muy sensibles a los cambios en la situación de salud del paciente.

Los cuestionarios diseñados para la medición de preferencias suelen ser de tipo genérico, ya que persiguen la comparación de patologías y grupos de pacientes. Como contrapartida suelen ser menos sensibles. Acostumbran a estar compuestos por descriptores de los posibles estados de salud (un sistema de clasificación) y un algoritmo que permite obtener la preferencia (o utilidad) asociada a cada estado de salud.

En resumen, el enfoque de medidas psicométricas reflexivo ha dado lugar tanto a instrumentos genéricos como específicos, siendo más apreciados los instrumentos específicos. Por su parte, el enfoque centrado en las medidas basadas en preferencias, suele utilizar una aproximación clinimétrica formativa y ha dado lugar a instrumentos genéricos. Los primeros se refieren a una patología o a una dimensión específica, tienen la ventaja de gozar de mayor sensibilidad y son óptimos para el diagnóstico individual, pero suelen presentar dificultad para generalizar a otras subpoblaciones (cuando son específicos de una patología); por ser multidimensionales resulta complejo hacer comparaciones entre subpoblaciones, ya que no suelen ofrecer una medida resumen global. Por el contrario, los instrumentos genéricos no están vinculados a enfermedades concretas, sino que pueden describir el estado de salud de poblaciones diversas, pero con una sensibilidad menor al cambio.

Instrumentos genéricos

La medición de la utilidad se ha realizado tradicionalmente con instrumentos genéricos. Existen varios motivos para ello. En primer lugar, como veremos, el valor de utilidad de un estado de salud particular es el resultado de la preferencia de la población por ese estado de salud, frente a otros resultados posibles, valorando la preferencia en una situación de incertidumbre en la que

cualesquiera de los otros resultados de salud serían también posibles (reflejan un proceso de elección entre alternativas bajo incertidumbre). En segundo lugar, las utilidades se aplican a la valoración de grupos de pacientes (tratamientos alternativos, patologías concurrentes, instauración de programas, etc.) y el interés se centra en cuantificar la salud en cada grupo de pacientes, para poder compararlos; no existe un interés real en realizar un diagnóstico pormenorizado de la calidad de vida de cada paciente individual. En tercer lugar, las utilidades deben ser unidimensionales y las distintas dimensiones de la salud se agregan en una única puntuación, por ello no tiene sentido el estudio del perfil de las puntuaciones de un sujeto con propósitos diagnósticos. Por último, los pesos dados a cada atributo o dimensión al resumir los estados de salud son específicos de cada población cultural y no se deben recalcular en función de la patología; es decir, se considera que las preferencias de la población por los estados de salud son estables y particulares de cada cultura o país.

Siguiendo la notación de Muthén y Muthén (1998) de su guía técnica Mplus, el modelo de medida para un instrumento genérico podría representarse como en la siguiente figura, utilizando el caso particular del EQ-5D-3L.

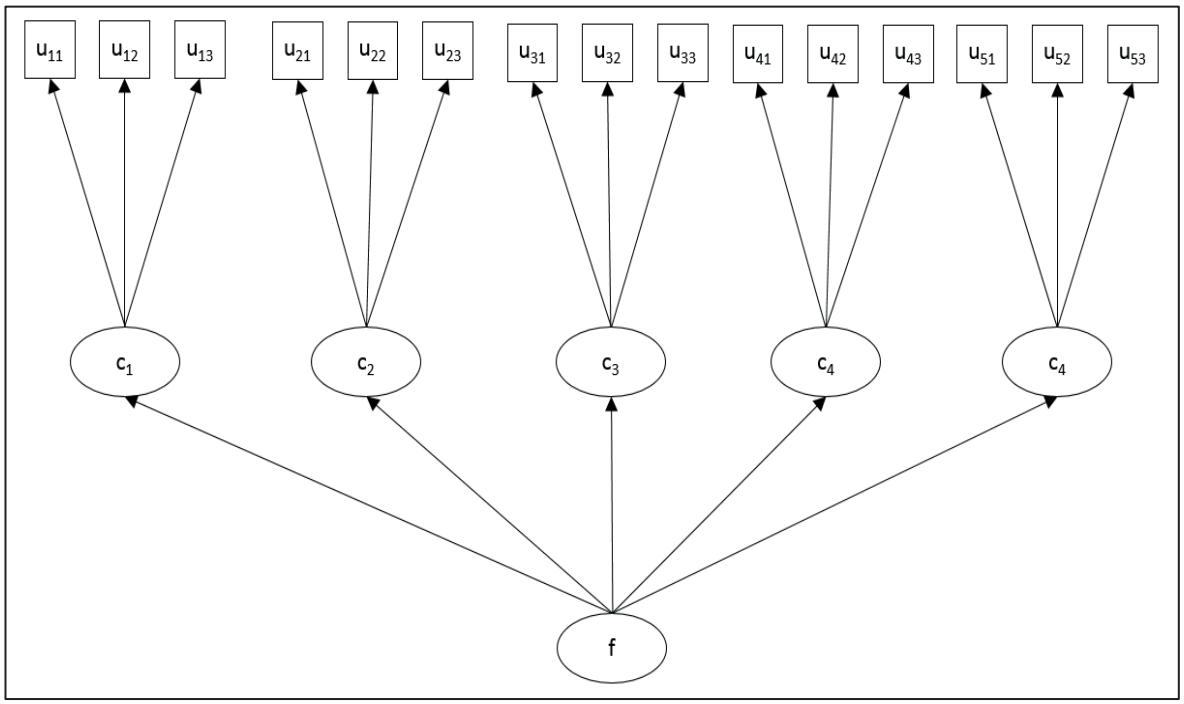


Figura 17. Representación de la utilidad como un modelo reflexivo (EQ-5D-3L)

El EQ-5D-3L está formado por cinco atributos (o dimensiones) categóricos: c_1 a c_5. Cada atributo (c_i) está medido por tres indicadores dicotómicos (u_{i1} a u_{i3}) ordinales y mutuamente exclusivos. Cada combinación de categorías de los cinco atributos constituye un estado de salud, que refleja un nivel de preferencia en la población o utilidad (*f*). Sin embargo, el modelo propuesto no debe interpretarse como un modelo de medida en el sentido de la psicometría clásica (las variables latentes c_i son categóricas), sino como un sistema de pesos que permiten obtener las puntuaciones de utilidad. Por ese motivo, el indicador observado no es necesario que vaya acompañado de un término que represente el error de medición.

Puesto que las categorías correspondientes a los niveles de respuesta de cada uno de los atributos son exclusivas entre sí (no se pueden dar respuestas simultáneas a varios niveles del atributo), optaremos por una representación simplificada como la de la siguiente figura.

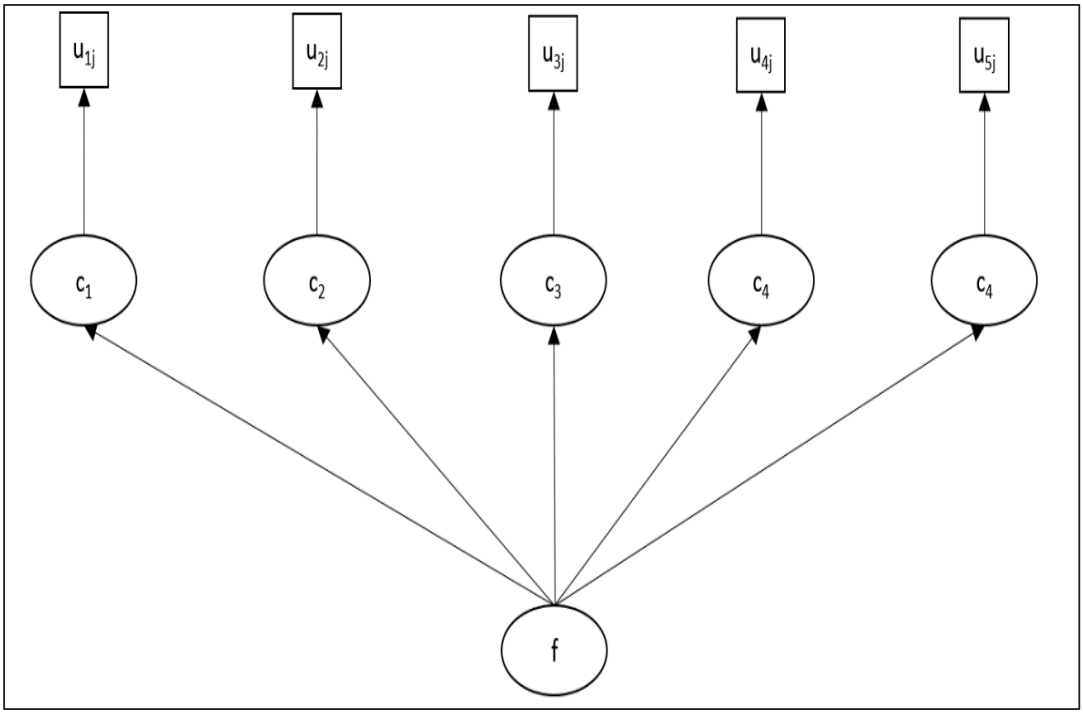


Figura 18. Modelo reflexivo simplificado (EQ-5D-3L)

Podemos asumir que un sujeto que puntúa en los valores máximos (ausencia de deterioro) en todas las dimensiones de un instrumento genérico tiene la mejor calidad de vida, pero no es fácil decidir qué sujeto tiene mejor calidad de vida cuando las puntuaciones, aun siendo parecidas, no son las máximas o cuando queremos comparar sujetos con el mismo deterioro, pero en una dimensión distinta cada uno.

Existen distintos métodos para obtener el valor de utilidad de cada uno de los estados de salud (ver más abajo) y que sean capaces de reflejar de manera precisa y adecuada las preferencias de la población por cada estado de salud. La mayoría se basan en estudios experimentales de elección forzada en situaciones de incertidumbre utilizando los estados de salud como viñetas que hay que escalar. Tras realizar las tareas para obtener las utilidades, se acostumbra a construir una expresión matemática, denominada función multiatributo (MAUF), que permite derivar la utilidad de cada uno de los estados de salud posibles. No se acostumbra a crear un listado de todos los estados de salud posibles con su utilidad asociada.

Por todo lo anterior, los instrumentos genéricos de calidad de vida se caracterizan por constar de dos elementos: un sistema descriptivo multiatributo y una regla de valoración o algoritmo que sirve para obtener los valores de utilidad (magnitud de las preferencias). Las diferencias, por tanto, entre los diferentes instrumentos estarán en la determinación de las dimensiones (o atributos) que lo integran y los niveles con que se describen esas dimensiones, y en el algoritmo utilizado para la obtención de la utilidad.

Como hemos comentado, una vez obtenida la función multiatributo, los pesos no pueden ser recalculados en las distintas aplicaciones del instrumento para hacer valoraciones de utilidad, por lo que a efectos de correspondencia con otros instrumentos debemos considerar los pesos como constantes. Este hecho nos permite adoptar una representación alternativa de la función multiatributo, como la que se muestra a continuación. En esta representación, la MAUF se considera un modelo formativo, que resume la cuantificación de los atributos, y crea la puntuación agregada correspondiente a la utilidad. Hemos preferido representar el efecto de los errores de predicción sobre la "verdadera" utilidad latente para evitar confundir el término con los errores de medición.

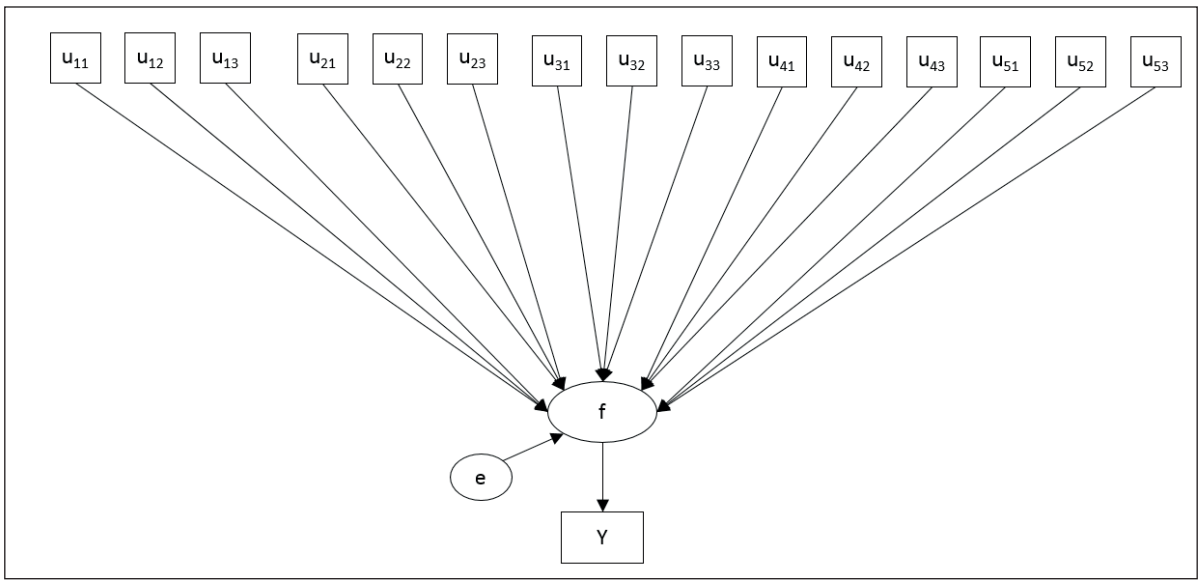


Figura 19. Representación de la utilidad como un modelo formativo (caso del EQ-5D-3L)

Alternativamente, podemos simplificar la representación del modelo mostrando solo una categoría de respuesta genérica para cada atributo (tal como hicimos en el caso anterior) y proponer la siguiente representación alternativa.

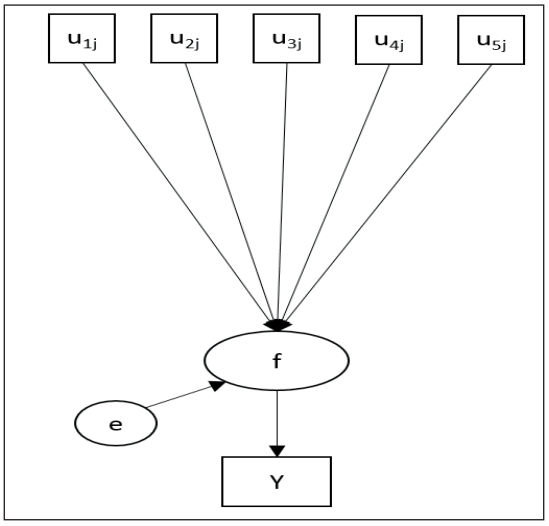


Figura 20. Modelo formativo simplificado (EQ-5D-3L)

Instrumentos específicos

Tal y como hemos indicado anteriormente, los instrumentos específicos han sido generados para interpretar clínicamente los resultados de una actuación, analizar la evolución de los pacientes de una determinada patología a lo largo de un determinado periodo de tiempo y estudiar los efectos de los tratamientos. Por el contrario, no permiten comparaciones entre intervenciones o diferentes patologías, viéndose limitada su aplicación a pacientes muy concretos.

Su objetivo no es encontrar una puntuación global que resuma su CVRS, sino más bien el diagnóstico por dimensiones. De hecho, al intentar combinar las puntuaciones de las distintas dimensiones nos encontraremos que la cuantificación de los niveles de la dimensión es arbitraria (son medidas de intervalo), ya que la diferencia entre niveles es constante; así, la diferencia entre "un poco" y "nada" es la misma que entre "un poco" y "bastante". Cuando los utilizamos para obtener un índice de severidad del estado de salud de un paciente en función de las puntuaciones en las dimensiones que componen el instrumento, nos encontramos con los inconvenientes de no tener en cuenta la preferencia del paciente respecto a otros perfiles posibles, y que perfiles distintos pueden tener el mismo valor en el índice de severidad.

Al ser específicos de una patología, no analizan aspectos de la salud que padecen los pacientes que no son propios de la patología que estudian, y que luego sí se reflejan cuando se analizan los instrumentos genéricos.

Proceso de modelización. Generación del valor utilidad

Como ejemplo de cálculo de la función multiatributo, se expone el procedimiento de los instrumentos genéricos utilizados en los estudios presentados.

EQ-5D-3L

El grupo EuroQol se reunió por primera vez en 1987 con el fin de obtener un instrumento que permitiese calcular el índice de preferencia o utilidad de la CVRS. En la actualidad existen varias versiones, debido a diferentes proyectos de mejora del instrumento.

El instrumento inicial EQ-5D-3L contiene cinco dimensiones: movilidad, cuidado personal, actividades diarias, dolor y ansiedad/depresión. En cada una de ellas se definen tres niveles (1 = sin problemas, 2 = problemas leves y 3 = problemas severos). La combinación de todos ellos genera un total de 243 (3^5) posibles estados de salud, cuya horquilla va desde el 11111 (salud perfecta) al 33333 (peor estado de salud posible).

Para la obtención del valor de utilidad se han utilizado dos métodos, el método TTO y el método VAS.

El método TTO (Time-Trade-Off) o método de intercambio de tiempos tiene por objetivo determinar el tiempo que un sujeto estaría dispuesto a dejar de vivir en un estado de salud deficiente para pasar a vivir el resto de su vida con salud plena. Se le presentan dos opciones alternativas a elegir, entre una vida más larga en el estado de salud de estudio y una vida más corta en el mejor estado salud. Dependiendo del estado de salud elegido, la cantidad de tiempo en la mejor salud se modifica hasta que se alcanza un punto de indiferencia entre ambas opciones. El método se puede aplicar de tres maneras distintas, bien en lo que se denomina procedimiento convencional, procedimiento denominado "tiempo de espera" o bien mediante un procedimiento compuesto combinación de ambos.

Si se opta por el procedimiento convencional debemos distinguir dos fases. En la primera consideramos un estado de salud deteriorado pero que el sujeto lo considera "mejor que la muerte". El individuo opta entre una vida más saludable durante X años y luego morir o una vida durante T años (con $T \geq X$) con un estado de salud deficiente (estado de salud que queremos evaluar) y luego la muerte. El tiempo T es fijo, mientras que variamos el tiempo X hasta que el

sujeto decida que le es indiferente encontrarse en cualquiera de las dos opciones descritas. El valor de la utilidad del estado de salud en estudio se determina como el cociente entre el valor de equilibrio y el valor *T* fijado.

En una segunda fase le presentamos un estado de salud que el sujeto considera "peor que la muerte". En este caso la elección está entre $T - X$ años con el estado de salud de estudio y luego la muerte o bien la muerte inmediata. Al igual que antes, variamos los valores de *X* hasta que lleguemos al punto de indiferencia. En este caso la utilidad del estado quedará definida como el cociente entre el valor de equilibrio y el valor de inicio $T - X$. Por ser un estado de salud que se considera peor que la muerte, el valor resultante ha de ser negativo.

El procedimiento convencional, a pesar de seguir utilizándose en numerosos estudios, tiene muchos detractores por utilizar dos procedimientos distintos para producir una sola escala, la cual tiene un límite superior de 1 obtenido de una manera lineal, pero un límite inferior indefinido en el caso de estudiar estados de salud peores que la muerte ($X/0$ si $X=T$). Para tal deficiencia se han propuesto tres posibles soluciones: reescalar los valores de estados considerados "peor que estar muerto" a una escala entre –1 y 0, cambiar la tendencia central de la media a la mediana o utilizar métodos basados en geometría angular (Oppe et al., 2016).

Si se opta por el segundo procedimiento, este consiste en introducir una constante denominada "tiempo de espera". Este enfoque implica agregar años de vida saludables al perfil de estudio antes de comparar las dos opciones que se plantean en el procedimiento convencional, de manera que eliminamos la incertidumbre cuando los perfiles son peores que la muerte.

Robinson et al. (2006) propusieron el enfoque alternativo llamado compuesto, que consiste en aplicar el procedimiento convencional cuando analizamos perfiles mejores que la muerte y el de tiempo de espera cuando lo hacemos con perfiles peores que la muerte. Este procedimiento es el que actualmente se utiliza en el proceso de obtención de la MAUF por parte del grupo EuroQol.

El método VAS (Visual Analogue Scale) o método de escala visual analógica se apoya en una escala gráfica cuyos valores extremos son 100 como salud perfecta y 0 como muerte. A cada entrevistado se le dan dos grupos de ocho perfiles (o estados de salud) bajo el supuesto

de que los estados duran un año y tiene que indicar en la escala el valor de utilidad que le asigna a cada uno de ellos. Una vez valorados, se les pide que designen un valor para el estado de muerte. A partir de los resultados obtenidos, una vez pasada la prueba a una muestra amplia, hay dos opciones para obtener el valor de la desutilidad: bien de manera directa según la información dada por el paciente o bien agregando en valores promedio en función del perfil de estudio. Para determinar el valor de la desutilidad aplicamos la ecuación:

$$X_{reescalada} = (X_{perfil} - X_{muerte}) / (X_{salud\ perfecta} - X_{muerte})$$

Los valores de perfiles intermedios en ambos procedimientos se conseguirían por regresiones lineales, si bien los autores no especifican el método empleado. En cierto modo, consistiría en presentar a los sujetos que realizan la encuesta el estudio de perfiles adecuados (perfiles extremos, de índices de severidad iguales y de mayor frecuencia en la población) y predecir en orden creciente del índice de severidad.

SF-6D

Ware et al. (1993) desarrollaron el instrumento MOS SF-36 (Medical Outomes Survey, Short Form 36 items). Dicho instrumento fue mejorado en el año 2000, obteniéndose la versión 2 del mismo, que permitía estandarizar las puntuaciones de cada dimensión sobre la base de normas poblacionales (Villagut et al., 2005).

Con la finalidad de poder obtener AVAC a partir del mismo, Brazier et al. (1998) generaron el instrumento SF-6D (modificado en 2002), el cual es el que actualmente se encuentra en vigor. Este instrumento define seis dimensiones (funcionamiento físico, limitaciones de rol, funcionamiento social, dolor, vitalidad y salud mental) y se obtiene a partir de un subconjunto de 11 ítems extraídos del SF-36 y con un número variable de niveles, que oscilan entre 4 y 6 por dimensión.

Para la obtención del índice de utilidad español de cada uno de los 18.000 posibles perfiles que se pueden obtener a partir del SF-6D, Abellán et al. (2012) utilizaron la técnica de valoración Standard Gamble (SG), de "lotería equivalente" o de "apuesta normalizada", por la que el entrevistado compara dos tratamientos en el que se va variando las probabilidades de uno mientras el otro permanece constante hasta que indique que la probabilidad de ambos es la misma. El motivo que expone para utilizar dicha modalidad de lotería, en lugar de la "lotería estándar", es que con ella se evita el obtener probabilidades demasiado elevadas.

Mediante procedimientos de regresión, con un procedimiento igual al del EQ5D, se estima el algoritmo SF-6D para la población española. En este punto se debe indicar que el algoritmo español, a diferencia del inglés, no incorpora ningún término de interacción entre las dimensiones.

HUI-III

El Índice de Utilidades de Salud Mark 3 (Health Utilities Index Mark III, HUI-III), fue desarrollado en Canadá por Feeny et al. (1995).

Está formado por un total de ocho atributos: visión, audición, habla, deambulación, destreza, emoción, cognición y dolor. En cada atributo se incluyen 5 o 6 niveles que varían desde la normalidad hasta el deterioro grave. Un perfil o estado de salud se identifica por el nivel seleccionado en cada dimensión, así, el estado de salud perfecta es definido por el vector (11111111) y el peor estado de salud por el vector (66566565). La combinación de todos los niveles incluidos en el HUI-III permite describir 972.000 estados de salud distintos.

Para determinar la utilidad asociada a cada estado de salud se realizan comparaciones binarias entre todas las combinaciones posibles de estados de salud. El resultado de esas comparaciones permite obtener una estimación de la utilidad asociada a cada estado de salud. Ahora bien, puesto que el número de comparaciones binarias posibles es muy elevado (972.000), se utilizan diseños incompletos que, además de reducir el número de juicios subjetivos (comparaciones binarias) necesarios para asignar valoraciones a todos los estados de salud, permiten separar el efecto individual de cada atributo sobre la valoración subjetiva (es decir, proporciona información acerca de cómo las personas combinan los atributos para llegar a un juicio global sobre un estado de salud). Tras la aplicación del método obtenemos la función de utilidad multiatributo que nos permite obtener el índice de utilidad.

Para estimar las utilidades asociadas a cada estado de salud se han utilizado dos métodos de recogida de información por escalamiento unidimensional: el de Escala Visual Analógica (EVA o VAS) y el de "apuesta normalizada" (Standard Gamble). Ambos métodos permiten obtener las valoraciones (preferencias) de los sujetos en una escala de intervalo.

En el proceso de selección de perfiles para reducir el gran número de combinaciones binarias comienza distinguiendo cuatro tipos de estados de salud. 1) "Estados de referencia" son los tres estados de salud extremos: salud total, muerte y sima; siendo este último el peor estado de salud que se obtiene al indicar la peor opción en cada una de las dimensiones. 2) "Estados metodológicos", que son tres estados de salud intermedios y muy frecuentes epidemiológicamente hablando. 3) "Estados asociados a un atributo", que se obtienen fijando en 1 (mejor estado de salud) los valores de 7 de los 8 atributos y presentando todas las variaciones posibles, salvo la opción 1, del octavo atributo, obteniéndose 37 estados de salud. 4) "Estados prevalentes", que son los estados de salud más frecuentes en la población.

Ruiz et al. (2003) utilizaron el siguiente procedimiento para la obtención de la MAUF para la población española:

1. Se divide la muestra en dos grupos en función del estado de salud más bajo (Sima o Muerte);
2. Se calcula la media recortada de las valoraciones y utilidades asociadas por estado de atributo simple en cada grupo;
3. Se corrigen las valoraciones de los estados medidos simultáneamente con el EVA por el sesgo del extremo de la escala;
4. Mediante una ecuación de regresión se transforman las valoraciones en utilidades;
5. Se reescalan las utilidades de un grupo para poner ambos en la misma escala (Sima-Salud total);
6. Se funden las utilidades de los dos grupos con el cálculo de una media ponderada;
7. Se convierten las utilidades en desutilidades para la obtención de las constantes de la MAUF;
8. Al mismo tiempo se reescalan las utilidades de los estados de atributo simple a escala estándar 0-1;
9. Las utilidades reescaladas se transforman también en desutilidades.

TÉCNICAS DE ANÁLISIS

A continuación, se describen los métodos de análisis estadístico más utilizados para la traslación métrica de un instrumento específico de CVRS a otro instrumento genérico de utilidad, como marco general para compararlos.

Concepto de mapping

Con los instrumentos genéricos obtenemos el valor de la utilidad y con los específicos el valor de severidad de la patología. Se ha de buscar un procedimiento que nos permita trasladar el valor de la severidad al de utilidad cuando, por circunstancias diversas, el paciente no haya cumplimentado los instrumentos autoinformados genéricos (PRO), pero dispongamos de información de estudios anteriores de la patología que padece. A este procedimiento de traslación de puntuaciones de severidad a utilidades lo denominaremos "mapping" (traslación).

El procedimiento de mapping consiste en llevar a cabo una proyección matemática de la puntuación generada por el instrumento específico (X_ESP) a la métrica de la función creada por el instrumento genérico (Y_MAUF).

$$Y_{MAUF} = f(X_{Esp})$$

Dicho de otra manera, consiste en trasladar o predecir los valores de la curva (normalmente de la función de densidad) que genera el instrumento genérico a partir de los valores en la curva (normalmente también de la función de densidad) que se obtiene en el instrumento específico. Existen varios procedimientos para ello

Método de cálculo del índice de severidad

Si el método de corrección del instrumento específico no es conocido, el primer paso será determinar el procedimiento para obtener un valor para la puntuación de los sujetos en el instrumento específico. Este es un tema que atañe al campo de la psicometría y existen distintas estrategias para obtener el sistema de puntuación, que dependerán, entre otras cosas, de si el instrumento específico es unidimensional o multidimensional. Puesto que el mapping final se

realiza sobre una única puntuación de utilidad, será necesario determinar la manera de agregar las puntuaciones en las distintas dimensiones cuando el instrumento no sea unidimensional. El resultado final será un índice de la patología de manera que se consiga la ordenación de los sujetos en función de su puntuación en el instrumento o la ordenación los diferentes perfiles de severidad (si consideramos el instrumento específico como un instrumento formativo). Con frecuencia es recomendable reescalar el índice de severidad específico para que las puntuaciones se encuentren entre los valores 0 y 1, de manera que se pueda llevar a cabo el mapping o traslación entre el instrumento específico y el genérico en una misma escala de valores. Aunque este procedimiento será distinto cuando la utilidad de la MAUF se encuentre en el rango [–1, 1].

Análisis factorial o Técnicas de escalamiento

Tal y como se ha indicado anteriormente, la definición del índice de severidad como suma directa de las puntuaciones dadas por los pacientes en el cuestionario o mediante el procedimiento del instrumento específico, no es un resultado que pueda considerarse válido en todos los casos. Con frecuencia, cuando el instrumento específico ha sido diseñado como un modelo formativo, puede darse el caso de que distintos perfiles de severidad (vectores de puntuaciones en los ítems) pueden obtener índices de severidad idénticos, con lo que no se conseguirá el escalamiento de los sujetos.

A la hora de calcular el valor del índice de severidad del instrumento específico podemos encontrarnos con dos opciones en función de que se haya definido o no su estructura métrica subyacente, y por tanto que sea necesario llevar a cabo un procedimiento que nos permita obtenerla. Si se dispone de un modelo métrico ya demostrado, bastará aplicar el sistema de corrección del instrumento específico para obtener el índice de severidad. Este índice debe ser unidimensional y resumir la información en un valor global, de lo contrario habrá que plantearse si se utilizan las puntuaciones de las distintas dimensiones como predictores en el mapping o si se agregan todas las dimensiones antes de realizar el mapping.

Si no se ha estudiado el modelo subyacente con antelación, podremos utilizar el análisis factorial u otras técnicas de escalamiento (por ejemplo, el modelo de Rasch) para obtener una

única puntuación global de severidad. Aunque esta es una cuestión que atañe a las propiedades psicométricas del instrumento específico, sí conviene tener en cuenta que las puntuaciones factoriales son en sí mismas indeterminadas y que su estimación se genera siempre en una métrica de media 0 y sin límites preestablecidos. Este hecho puede afectar al mapping, ya que las utilidades sí que presentan valores asintóticos para su recorrido, [0, 1] o [–1, 1], por lo que la solución más habitual es reescalar las puntuaciones factoriales a partir de la puntuaciones máxima y mínima obtenibles por el instrumento para que tengan valores límite prestablecidos.

Por tramos con ítems dicotómicos

Aun cuando exista un modelo métrico subyacente conocido, puede que existan interacciones entre los ítems o entre las dimensiones del cuestionario específico que puedan afectar a los valores de utilidad asociados. Por ejemplo, cuando algunos ítems miden frecuencia de los síntomas, otros ítems intensidad de los síntomas y otros ítems estrategias de afrontamiento para paliar el impacto de esos síntomas en la vida cotidiana. Todas ellas son dimensiones de la calidad de vida específica, pero puede que no esté claro cómo afecta cada una o su combinación a la utilidad social percibida. ¿Qué prefiere el conjunto de la sociedad, un solo síntoma episódico muy doloroso o un conjunto de síntomas crónicos poco dolorosos?

Si además de utilizar los ítems individuales deseamos considerar también los niveles de respuesta de cada ítem, es posible modelizar cambios no homogéneos en la progresión de severidad en cada ítem. Pero en esta aproximación se acostumbra a seleccionar un único ítem por atributo/dimensión (aunque se hayan utilizado varios en el cuestionario original).

Una segunda opción es estimar un modelo de variable latente considerando que cada dimensión está representada por uno de los ítems que la componen, de manera que sus diferentes niveles sean monótonamente crecientes conforme a la severidad. Ello da la opción de generar todas las tuplas posibles por combinatoria de ítems, uno por dimensión, descomponiendo cada ítem en k variables ficticias con valores 1 (nivel seleccionado por el paciente) o 0 (nivel no seleccionado), ya que no se puede suponer que el incremento es constante en una escala Likert. Posteriormente y mediante regresiones, determinar qué combinación es la que más se ajusta a la distribución de utilidad generada por el instrumento genérico.

Por tramos escalonados

Una tercera opción consiste en descomponer cada ítem *k*-categorías en una serie de *k* variables ficticias (0 = No, 1 = Sí) y codificar las categorías ficticias de nivel inferior como cumplidas (1) cuando se alcanza un nivel de ítem particular, obteniéndose un modelo de Crédito Parcial (Masters, G.N., 1982) utilizando la estimación de máxima verosimilitud, de manera que los umbrales de categoría estimados podrían compararse entre los ítems. Por umbral entendemos el valor esperado de un sujeto de pasar de un nivel a otro. Se obtiene como la distribución normal inversa cambiada de signo de la suma de las proporciones de niveles superiores al umbral a determinar. El valor de la severidad se calcula como la suma de los umbrales. Esta aproximación sería similar a la adoptada en las escalas acumulativas de Guttman.

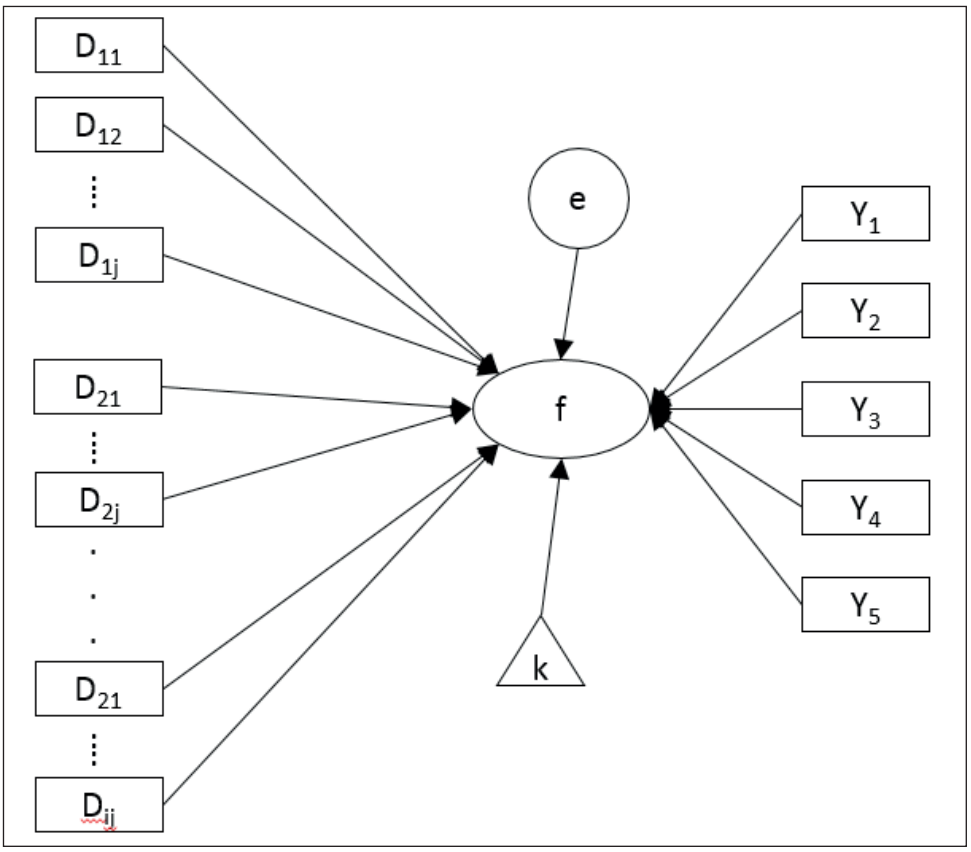


Figura 21. Representación de un modelo de mapping directo por tramos

Una vez obtenido el índice de severidad óptimo para la traslación métrica, se analizan diferentes tipos de métodos de proyección para determinar cuál es el más indicado para la patología en estudio.

Métodos de traslación (mapping)

Una vez decidido el tratamiento que se dará al cuestionario específico, habrá que decidir el procedimiento más idóneo para trasladar las puntuaciones de severidad a la métrica de la utilidad. También aquí existen varias aproximaciones.

Traslación directa

Una primera posibilidad es trasladar directamente las puntuaciones de severidad a la métrica de la utilidad seleccionando la función de vínculo más adecuada. El esquema general del modelo podría representarse según la siguiente figura. El modelo de medida de la variable exógena genera la puntuación en la severidad latente ξ del instrumento específico que sirve para pronosticar la puntuación de utilidad latente η, la cual tendrá su propio modelo de medida (no se han representado los errores).

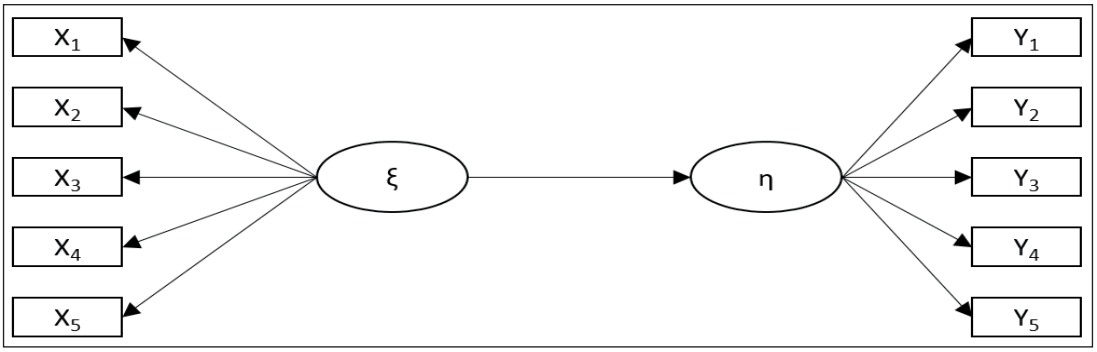


Figura 22. Mapping de la puntuación de severidad latente (ξ) sobre la puntuación de utilidad (η)

Sin embargo, el modelo de medida correspondiente a la utilidad no es estimable en el proceso, ya que el sistema de obtención de las utilidades mediante su MAUF no es modificable. Por ese motivo es mejor representar el modelo de medida de la variable endógena como un único indicador (sin error de medida) que no requiere estimación en el proceso.

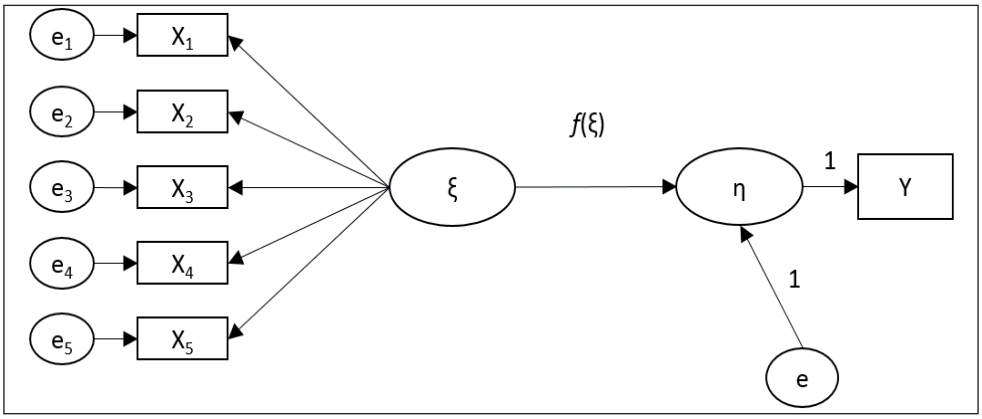


Figura 23. Modelo simplificado de mapping de la puntuación de severidad latente (ξ) sobre la puntuación de utilidad (η)

Podemos llevar a cabo procesos de traslación directa mediante procedimientos de regresión lineal, cuadrática, cúbica, Logit, Tobit, Beta, o procedimientos mixtos dependiendo de la forma empírica de las curvas obtenidas.

En la búsqueda del mejor modelo posible se suele comenzar por el modelo lineal como modelo más simple posible. Los modelos polinómicos cuadrático y cúbico pueden mejorar el ajuste por la presencia característica de un umbral superior en las puntuaciones de utilidades o la presencia de umbrales tanto en las puntuaciones superiores como inferiores. A este respecto, en nuestra experiencia, es mejor modelar las puntuaciones de desutilidad (1-utilidad) frente a la predicción directa de la utilidad. Con ello es posible anclar la ordenada en el origen en el valor (0,0), lo que tiene sentido si se quiere asociar el valor 0 de severidad con la mínima desutilidad, y también se consigue aproximar al origen la masa de datos correspondiente a los estados de mejor calidad de vida, que suele ser muy prevalente. De esa forma se minimiza el impacto de la sobrerrepresentación de los estados de salud más benignos sobre la pendiente del modelo.

La propuesta del modelo Tobit también resulta conceptualmente atractiva, ya que permite corregir el modelo respecto a la presencia de un límite de censura en las puntuaciones. Este fenómeno puede estar presente, por ejemplo, cuando el instrumento no tiene sensibilidad en uno de sus extremos y los casos se acumulan en uno de los extremos de la métrica.

Otra opción explorada es la regresión Beta, que puede ser adecuada cuando la distribución de los valores de la utilidad se caracteriza por tener una distribución inusual, con sesgos muy marcados, claramente multimodales y, a menudo, con un gran número de observaciones en los valores correspondientes a la salud plena y con una brecha entre esta y el próximo valor posible (debido al papel de la constante de la MAUF).

En la actualidad, con la proliferación de los modelos mixtos, se está explorando otra aproximación, que consiste en realizar la estimación con procedimientos que permitan estimar el recorrido de las funciones por tramos, especialmente en el recorrido de la variable criterio. No es infrecuente que nos encontrremos con una distribución multimodal de la variable criterio como la de la siguiente figura, y que nos planteemos la posibilidad de utilizar una combinación de distribuciones para distintos rangos de puntuaciones, en lugar de una distribución global única. El problema será cómo decidir el tramo de clasificación de cada paciente a la hora de hacer predicciones sin conocer el valor observado de utilidad (o desutilidad).

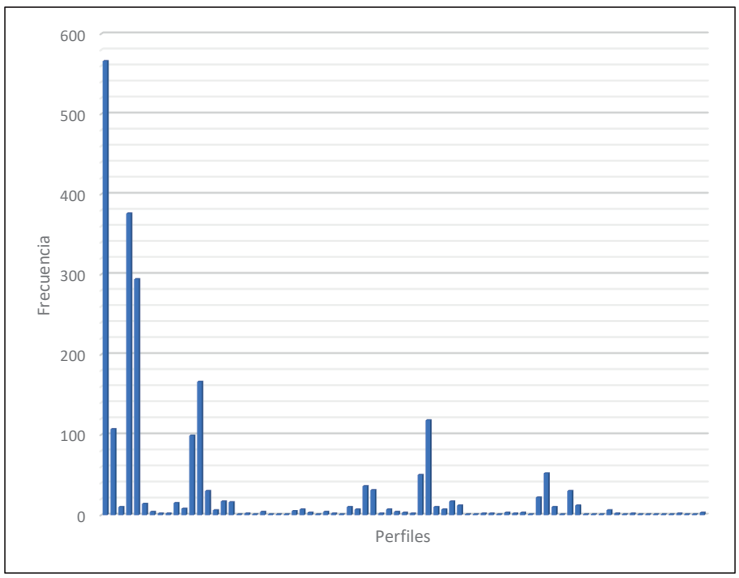


Figura 24. Distribución multimodal de valores de desutilidad del instrumento EQ-5D-3L para la patología de reflujo gastroesofágico

Una última estrategia que se ha venido utilizando es la agregación de las puntuaciones de utilidad según los valores observados de la severidad (y de posibles covariables) antes de realizar el mapping propiamente dicho. Hay que tener presente que puntuaciones iguales en el instrumento específico pueden dar lugar a puntuaciones distintas en el instrumento genérico, principalmente debido a comorbilidades que presenta el individuo y que no se ven recogidas en el instrumento específico. El problema aquí será que las estimaciones del ajuste del modelo se encontrarán sobreestimadas, ya que no se estarán tomando en consideración los residuos individuales reales.

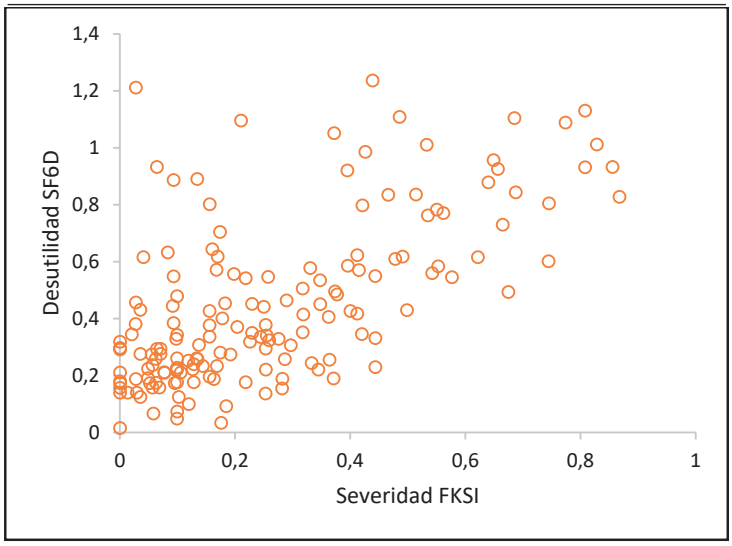


Figura 25. Distribución de índice de desutilidad SF6D observados por valor de índice de severidad (FKSI) de la patología enfermedad renal crónica (ERC)

Método de traslación en dos pasos

Este procedimiento consiste en determinar qué perfiles del instrumento genérico se relacionan con un determinado valor de severidad.

Una vez identificados los perfiles empíricos, se calcula la probabilidad asociada a la aparición de cada uno de los niveles de las dimensiones del instrumento genérico, de manera que el nivel con mayor probabilidad en cada una de las dimensiones definirá esa componente en el perfil más probable que corresponde a ese valor de severidad.

Una vez obtenido el perfil más probable podremos calcular de manera directa la utilidad correspondiente al mismo, utilizando la MAUF del instrumento genérico. Con ello habremos conseguido establecer la relación entre la severidad de la patología específica y la utilidad en el instrumento genérico.

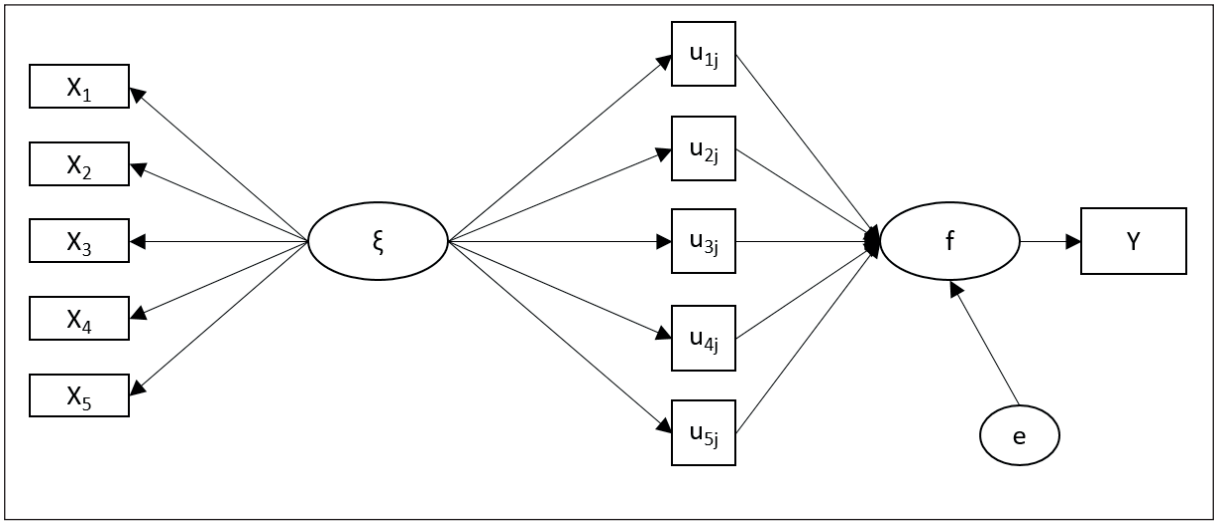


Figura 26. Representación del modelo de mapping probabilístico en dos fases

En la siguiente tabla se muestra la distribución combinada de los niveles de severidad del reflujo gastroesofágico (obtenidos con el GSFQ) y los niveles en los atributos del instrumento genérico de CVRS EQ-5D-3L, junto con el rango de desutilidades obtenidos con los perfiles observados en el EQ-5D-3L. Cada fila corresponde a un valor de severidad obtenido por el instrumento específico. Como puede apreciarse, el valor de severidad 0,17 daría como perfil más probable el definido por el vector (1, 1, 1, 1, 1), cuya desutilidad calculada es $du = 0$ y su utilidad será $u = 1$. Sin embargo, también existe cierta probabilidad no despreciable de observar el valor 2 en dolor, $P\,(eq_4 = 2) = 0{,}31$; y también cierta probabilidad de observar el valor 2 en ansiedad/depresión, $P\,(eq_5 = 2) = 0{,}15$.

Tabla 2. Ejemplo Método de tramos para traslación de severidades GSFQ a desutilidades EQ5D (listado parcial)

	Movilidad			Cuidado personal			Actividad cotidiana			Dolor			Ansiedad / Depresión			Desutilidad	Perfiles EQ5D más probables
Severidad	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3		
,06	**,96**	,04	,00	**1,00**	,00	,00	**,92**	,08	,00	**,67**	,33	,00	**,58**	**,42**	,00	,00	11111 - 11112
,17	**,91**	,09	,00	**,98**	,02	,00	**,91**	,09	,00	**,69**	,31	,00	**,81**	,15	,04	,00	11111
,23	**,87**	,13	,00	**,92**	,08	,00	**,82**	,16	,03	**,47**	**,50**	,03	**,82**	,13	,05	,00- ,07	11112 - 11111
,40	**,90**	,10	,00	**,97**	,03	,00	**,86**	,14	,00	**,44**	**,53**	,03	**,68**	,31	,02	,07 - ,13	11121 - 11121
,63	**,68**	,32	,00	**,85**	,15	,00	,54	**,41**	,05	,10	**,80**	,10	**,44**	**,44**	,12	,07 - ,18	11221 - 11222
,70	**,57**	**,43**	,00	**,77**	,23	,00	,33	**,63**	,03	,10	**,67**	,23	**,37**	**,40**	,23	,13 - ,20	11221 - 11222 21221 - 21222
,93	,00	**1,00**	,00	,33	**,67**	,00	,00	**,67**	,33	,00	,00	**1,00**	,00	**,67**	,33	,20 - ,24	22232

Análisis de perfiles latentes

Otra técnica apta para llevar a cabo la predicción de los valores de utilidad a partir de los valores de severidad son los procedimientos de análisis de perfiles de clases latentes (PCL), de manera que, una vez definidos una serie de perfiles empíricos, podamos predecir a cuál de ellos pertenece cada paciente y asignarle el valor de utilidad que le corresponda.

Este procedimiento es especialmente adecuado en el caso de estar realizando el mapping entre un instrumento específico y varios genéricos, ya que, cuando no disponemos de puntuaciones del paciente en todos los cuestionarios genéricos, nos permite inferir valor ausente de cualquier instrumento no administrado, una vez que el paciente ha sido asignado a un perfil.

Tiene el inconveniente de que el valor de la utilidad que presenta es el valor medio de los pacientes que están incluidos en un determinado perfil. Por ello es conveniente eliminar aquellos que, una vez analizados, puedan considerarse anómalos, por quedar fuera del intervalo de confianza que se establezca.

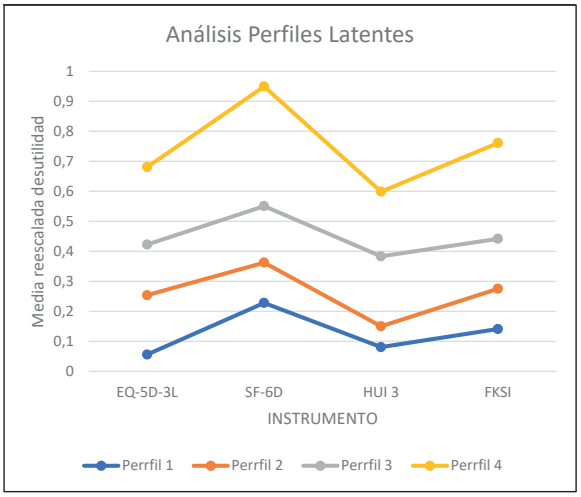


Figura 27. Análisis de perfiles latentes de la patología enfermedad renal crónica (ERC)

Valoración del ajuste

Una vez finalizado el proceso de predicción es necesario comprobar la bondad de ajuste entre la curva de utilidad real observada y la utilidad pronosticada por el proceso de mapping.

Dada la distinta sensibilidad de los modelos a los diferentes rangos de puntuaciones y la habitual acumulación de sujetos en las puntuaciones menos severas, recomendamos la comprobación del ajuste tanto de manera global como por tramos (por ejemplo, quintiles).

La bondad de ajuste se evaluará mediante el estudio de estadísticos tradicionales como el coeficiente de determinación (R^2), así como el estudio del signo y nivel de significación de los coeficientes de regresión, y el tamaño relativo de los coeficientes dentro de una dimensión dada.

También resulta interesante poder valernos de estadísticos utilizados comúnmente en la valoración de los modelos de series temporales. Las capacidades predictivas del modelo se pueden evaluar usando el error cuadrático medio (MSE), el error medio absoluto (MAE) y el error medio absoluto porcentual (MAPE).

$$MSE = \sum_{i=1}^{n} e_i^2 \quad MAE = \sum_{i=1}^{n} \frac{|e_i|}{n} \quad MAPE = \frac{100}{n} \sum_{i=1}^{n} \left|\frac{e_i}{x_i}\right|$$

Los modelos GLM y Tobit también se pueden evaluar utilizando el estadístico de chi-cuadrado (x^2), el cociente de chi-cuadrado por sus grados de libertad (x^2/df), el criterio de información Akaike (AIC) y el criterio de información bayesiano (BIC).

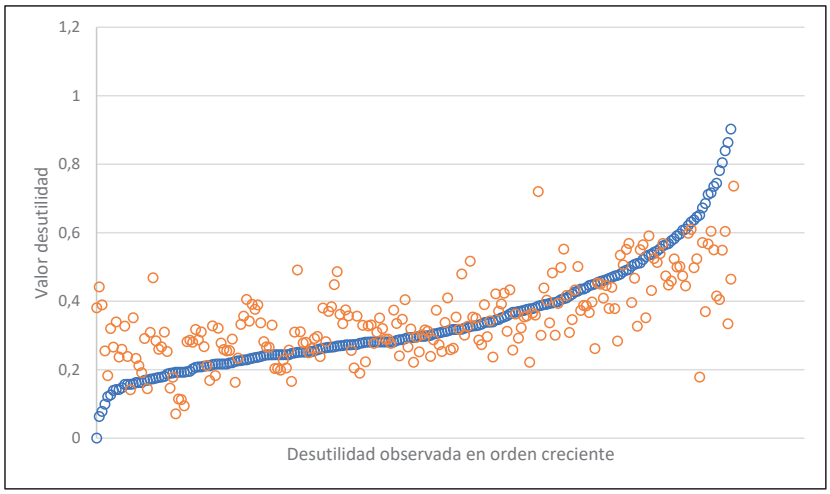


Figura 28. Representación de desutilidades observadas (azul) y pronosticadas (naranja) en la patología de reflujo gastroesofágico

OBJETIVO

El objetivo de la presente tesis es obtener distintas funciones de traslación métrica de diversos instrumentos de calidad de vida relacionada con la salud de patologías específicas sobre instrumentos genéricos de utilidad (cross-walking o mapping) y valorar el método de proyección más adecuado, dadas las características no lineales de la MAUF.

Se pretende establecer una norma de procedimiento para llevar a cabo la determinación del índice de utilidad, teniendo para ello en cuenta la diversidad de patologías.

Así mismo, se espera encontrar algún tipo de relación entre los tres índices genéricos más utilizados (EQ-5D-3L, SF-6D y HUI-III).

Se ha llevado a cabo un estudio bibliográfico para determinar las posibles estrategias de traslación (mapping) entre instrumentos específicos y genéricos en diferentes patologías. Analizadas las mismas, se ha optado por llevar un proceso gradual de mapping cambiando la patología a fin de tomar las menos graves, graves y muy graves teóricamente, y realizar el proceso de mapping con uno, dos y tres instrumentos genéricos.

En el primer estudio se ha tratado una patología considerada menos grave en la cual hemos trabajado con los criterios establecidos por un estudio anterior con la finalidad de replicar el procedimiento de obtención del índice de severidad de la misma y su posterior traslación a un instrumento genérico, que en este caso es el de referencia en la zona europea EQ-3D-3L.

En un segundo estudio, se ha tratado una enfermedad grave en la cual no está definido el procedimiento de obtención del grado de severidad, lo que ha supuesto analizar un procedimiento de obtención del mismo y un posterior procedimiento de mapping con dos instrumentos genéricos (EQ-3D-3L y SF-6D), determinando cuál de ellos es el más idóneo respecto a la patología estudiada.

En el tercer estudio hemos analizado una patología grave de la cual tampoco estaba definido el proceso de obtención de índice de severidad y se ha realizado un procedimiento de mapping sobre tres instrumentos genéricos (EQ-5D-3L, SF-6D, HUI-III). En este último artículo, además de realizar estudios correspondientes de mapping apoyados en procedimientos predictivos, hemos realizado estudios de análisis de perfiles latentes de pacientes.

MÉTODO

Sujetos

Para el desarrollo de los tres estudios que componen este trabajo de investigación se han establecido los siguientes criterios:

- Mayores de edad y de ambos sexos.
- Tener diagnosticada la enfermedad.
- Cumplimentar el consentimiento informado.
- Cumplir la declaración de Helsinki de la Asociación Médica Mundial sobre los principios éticos para las investigaciones médicas en seres humanos.
- Disponer de muestras superiores a 150 pacientes.

Resumen de los estudios

El trabajo consta de tres estudios, dos publicados y uno en revisión. A continuación, se resume cada uno de los trabajos. En los anexos se encuentran los trabajos publicados.

Primer estudio

Objetivo

Se pretende realizar un mapping (traslación) de una escala específica de la patología vejiga hiperactiva, utilizando para ello el cuestionario de salud OAB-5D derivado de la OABq-SF, en la escala basada en las preferencias, el EQ-5D-3L, en una muestra de pacientes en la población española.

Método

Diseño del estudio

Se diseñó un estudio observacional prospectivo multicéntrico, realizado en condiciones normales de la práctica clínica. Se pidió a todos los pacientes completar el consentimiento informado por escrito aprobado por el Comité de Ética en Investigación de la Universidad Autónoma de Madrid, bajo la etiqueta de "Aqua Estudio A0221077". El estudio incluyó dos visitas. En la primera se administró a los pacientes el OAB-5D y el EQ-5D-3L, y en una segunda (después de 3 meses de iniciar una nueva terapia) se le volvió a administrar el OABq-SF.

Sujetos

La muestra de trabajo fue de 246 pacientes españoles. Se obtuvieron un total de 43 de los 243 posibles estados de salud (perfiles) que se pueden generar en el instrumento EQ-5D-3L.

Instrumentos

Se utilizaron el Overactive Blader Questionnaire 5 dimensions (OAB-5D) y el EuroQol 5 dimensions 3 Levels (EQ-5D-3L).

Análisis estadísticos

Para la obtención de la utilidad se utilizaron la escala visual analógica (VAS) así como el procedimiento de método de compensación de tiempos (TTO) propios del EQ-5D-3L. Además de los mínimos cuadrados ordinarios (OLS), se estimaron modelos lineales generalizados (GLM) y Tobit. Los modelos resultantes se compararon y el mejor fue seleccionado en función de la bondad de las medidas de ajuste, atribución de signo, magnitud y la significación estadística de los coeficientes de regresión. Por último, la validez interna del mejor modelo se calculó mediante remuestreo bootstrap.

Puesto que se había demostrado que hay divergencia conceptual entre el instrumento de lengua inglesa y el de lengua castellana (Arlandis et al., 2012), para llevar a cabo el procedimiento de réplica de los resultados obtenidos por Yang et al. (2009) se mantuvo su criterio de obtención de ítems que correspondían a cada una de las dimensiones (uno por dimensión), comparando la combinación propuesta por el autor con todas las combinaciones posibles, las cuales se probaron como modelos alternativos. En los modelos alternativos, para seguir el procedimiento del autor, se fusionaron niveles consecutivos de los ítems de manera que quedasen definidos cinco niveles por dimensión, para lo cual se realizaron regresiones logísticas de manera que los pesos siguiesen una tendencia monótona creciente de severidad acorde a una desutilidad creciente en el instrumento genérico. Se seleccionaron los mejores modelos de ajuste, mientras que los modelos con pesos no significativos fueron descartados.

Los modelos mínimos cuadrados ordinarios se estimaron mediante la selección de grupos de 5 ítems, uno para cada dimensión OAB-5D, y descomponiendo cada elemento en un conjunto de 5 variables ficticias que indican el nivel de gravedad alcanzado en ese ítem en par-

ticular. Para mantener los criterios introducidos por Yang et al. (2009), se incluye un término de interacción (N2) que tendría el valor N2 = 1, cuando se alcanzó el máximo nivel de gravedad de dos o más de las dimensiones del instrumento, y N2 = 0 en caso contrario, y no se incluyó ningún término interacción.

Para las estimaciones mediante modelos de distribución Gamma y función de enlace logit se trabajó con la desutilidad VAS como variable dependiente, dado que no alcanza valores negativos (a diferencia TTO), que es una restricción para el cálculo de log ($\bar{u}$). El valor de $\bar{u} = 0$ es uno de los posibles valores observados, para el cual la función logarítmica no está definida. Para salvar dicha indeterminación, se añadió una pequeña constante arbitraria (c = 0,05).

Resultados

Una vez analizadas todas las combinaciones posibles, se obtuvo que la mejor era la compuesta por los ítems 1, 3, 9, 11 y 8 del instrumento OAB-5D, identificando cada uno y por ese orden las dimensiones urgencia, incontinencia, sueño, afrontamiento y preocupación.

Como resultado final se decidió que el mejor modelo era el obtenido por regresión OLS con las desutilidades del EQ-5D-3L agregadas por puntuación de severidad del OAB-5D. El modelo mencionado estimado alcanzó valor de R^2 de 0.89 con los datos agregados; con GLM (regresión máxima verosimilitud), se ha obtenido un ji-cuadrado de 15,3 e índices AIC = –550,914 y BIC = –475,381. Los coeficientes de las dimensiones en el modelo elegido obtienen los siguientes rangos de pesos: urgencia (0,102 - 0,216), incontinencia (0,07 - 0,171), sueño (0,071 - 0,078), afrontamiento (0,076 - 0,136) y preocupación (–0,132 - –0,028), y un valor de –0,065 para el coeficiente del término de interacción N2. Los índices de ajuste MAE, MAPE toman valores de 0,16 y 149,9%.

Discusión

Se compararon tres procedimientos de estimación (OLS, GLM y Tobit) que muestran resultados similares en términos de dimensiones significativas, niveles significativos e índices de ajuste. Los modelos se calcularon utilizando datos individuales y agregados según la severidad media del instrumento OAB-5D, seleccionándose el que tenía mejor interpretabilidad y ajuste.

Replicando el modelo propuesto por los desarrolladores del sistema OAB-5D, las estimaciones de OLS obtenidas en la muestra española fueron sistemáticamente más pequeñas que las obtenidas en la muestra anglosajona.

El ajuste del modelo (R^2) aumentó en un 47% cuando utilizamos datos agregados por severidad de OAB-5D, aunque se debe ser cauteloso con esta mejoría ya que se reduce el número de datos al llevar a cabo la agregación.

En relación con los términos de interacción N2, encontramos que no es un término esencial, ya que no alcanzó significación para cualquiera de los modelos estimados.

El procedimiento de Tobit, a nivel agregado, no mejoró las predicciones y los coeficientes fueron de signo y tamaño similares a los de las estimaciones de OLS.

Si incluimos otras variables explicativas como la edad en el algoritmo de mapeo, para tener en cuenta la gravedad de los síntomas relacionados con las condiciones de salud concomitantes, se mejoraría el ajuste y se explicaría por qué los pacientes sin síntomas de OAB obtienen altos valores de desutilidad, pero también se desdibujaría la capacidad del sistema OAB-5D para dar cuenta de la discapacidad de salud.

Una limitación principal de nuestro trabajo es el hecho de que estamos mapeando un instrumento de calidad de vida autoinformado en un instrumento de preferencia de la población general. Se sabe que existen diferencias entre estos dos enfoques para evaluar la salud, especialmente cuando los pacientes pueden adaptarse a su condición y subestimar la gravedad o el deterioro.

Conclusiones

Es posible llevar a cabo el proceso de traslación de puntuaciones del instrumento específico OAB-5D al genérico EQ-5D-3L, pudiendo concluir que existe una asociación entre ambos instrumentos.

Segundo estudio

Objetivo

Se pretende realizar un mapping (traslación) de las puntuaciones específicas obtenidas en el instrumento GSF-Q en dos instrumentos genéricos: SF-6D y EQ-5D-3L, para obtener estimaciones de utilidad derivadas de la patología de reflujo gastroesofágico (ERGE).

Método

Diseño del estudio

El presente estudio es un análisis secundario llevado a cabo utilizando los datos recopilados para la validación cultural del GSF-Q en español. El diseño de la muestra original aseguraba la representatividad de tres estratos: sexo, edad (<45, ≥45 años) y gravedad de los síntomas (Savary-Miller: 0-I, ≥II). Los pacientes fueron seleccionados al azar por demanda de atención y cubriendo cada estrato de muestra. Las escalas se administraron en una sola visita y firmaron un formulario de consentimiento informado, el cual lo validó el Comité de Ética de uno de los centros participantes. El estudio reclutó la participación de 510 gastroenterólogos.

Sujetos

En el estudio participaron 2.251 pacientes, de los que un 80% estaba diagnosticado de ERGE.

Instrumentos

Se utilizaron los instrumentos Gastrointestinal Short Form Questionnaire (GSF-Q), el EuroQol 5 dimensions 3 Levels (EQ-5D-3L) y el Medical Outcomes Survey Sort Form 6 dimensions (SF-6D).

Análisis estadísticos

El primer paso consistió en verificar la unidimensionalidad de los ítems de GSF-Q y, de cumplirse, derivar un índice de severidad global debido a la condición de paciente de ERGE. Un primer enfoque fue estimar un modelo de variable latente unidimensional asumiendo que la variable latente era continua y los ítems/indicadores eran ordinales, usando el método de estimación WLSMV. Un segundo enfoque se estableció con un modelo de Crédito Parcial utilizando máxima verosimilitud, descomponiendo cada ítem *k*-categorías en una serie de *k* variables ficticias (0 = No, 1 = Sí), y codificar las categorías ficticias de nivel inferior como cumplidas (1) cuando se alcanzó un nivel de ítem particular. De esta manera, los umbrales de categoría estimados podrían compararse entre los ítems.

Una vez que se obtuvo un índice de severidad específico de ERGE resumido, este índice se mapeó en cada uno de los dos valores de utilidad (por separado) y se probaron varios modelos para predecir el valor de utilidad asociado a cada condición de gravedad de ERGE.

Se modelaron valores de desutilidad en lugar de valores de utilidad, por varias razones. En primer lugar, debido a que la masa de datos generalmente se concentra en torno a estados de salud más indulgentes (desutilidades de valor 0) más próximas al origen del eje. En segundo lugar, siempre es posible estimar un modelo sin el término de intercepción, al anclar desutilidades de valor 0 (salud perfecta) en el valor de severidad 0. Dado que la ERGE no es necesariamente una condición incapacitante, y para atenuar el impacto de posibles comorbilidades en el valor de desutilidad para cada individuo, las irregularidades se agregaron, utilizando el valor medio, por gravedad de la ERGE, antes del modelado.

Se estimaron modelos de regresión lineal, cuadrático, cúbico y Tobit. Se analizaron covariables para su inclusión en los modelos. Para anclar los mejores estados de salud posibles en ambos instrumentos, las puntuaciones del factor de gravedad del ERGE se reescalaron en el rango 0-1, y los modelos de regresión se ajustaron a través del origen.

Junto con la significación estadística de los coeficientes de regresión, se evaluó la bondad de ajuste (GOF) del modelo utilizando R^2, error absoluto medio (MAE) y error porcentual absoluto medio (MAPE). MAE y MAPE se calcularon en general y por grupo de quintiles en función de los puntajes de gravedad para evaluar el GOF local en los diferentes niveles de gravedad.

Resultados

Las puntuaciones GSF-Q variaron entre 0 y 30 con un valor promedio de 10,54 (SD = 5,94). Las puntuaciones de severidad variaron entre –1,40 y 1,88 con un valor promedio de 0 (SD = 0,636), con una distribución simétrica (Asimetría = 0.021, SE = 0.052).

A nivel individual, las puntuaciones de utilidad promedio del instrumento SF-6D (MSF = 0,656, SDSF = 0,207) fueron significativamente menores que las puntuaciones del EQ-5D-3L (MEQ = 0,744, SDEQ = 0,206). Ambas puntuaciones de utilidad mostraron un marcado sesgo negativo, SF-6D: Sesgo Skewness = –0,784, SESF = 0,052; EQ-5D-3L: Skewness EQ = –1,049, SEEQ = 0,052, con una alta correlación entre ellos ($r = 0.733$, $p < 0.001$).

El análisis factorial confirmatorio para la solución de 1 dimensión (suponiendo que las variables fueran ordinales) alcanzó buenos índices GOF con CFI = 0,951 y TLI = 0,918. Se obtuvo una ecuación resultante aditiva.

El instrumento EQ-5D-3L se mostró menos sensible a la gravedad de la ERGE. Solo se observaron 78 (32%) de los 243 posibles perfiles EQ-5D-3L, y 17 (7%) de ellos reunieron a más del 90% de los pacientes. En el instrumento SF-6D, se observaron 975 (5,4%) de los 18.000 posibles estados de salud, 35 (0,2%) perfiles presentaron una prevalencia superior a 5/1000, y solo el 25,5% de los casos.

El mejor modelo para mapear GSF-Q en desutilidades SF-6D y EQ-5D-3L fue un modelo cúbico incluyendo covariables sociodemográficas diferentes en cada uno de los casos. El modelo de traslación GSFQ-SF6D obtuvo un GOF bueno (R^2 = 0,888), con MAE = 0,092 y MAPE = 27,9%. El modelo de traslación GSFQ EQ-5D-3L obtuvo un GOF bueno (R^2 = 0,831), MAE = 0,086 y MAPE = 37,0%.

Discusión

A pesar de las buenas propiedades psicométricas de instrumentos específicos y genéricos, al tratar una enfermedad que genera una incapacidad leve en la salud de los individuos, es difícil evitar que cuando estos cumplimentan los instrumentos genéricos limiten su pensamiento a una enfermedad específica, aislando sus juicios de otras comorbilidades que puedan estar presentes o del impacto de las discapacidades normales asociadas con el envejecimiento. El resultado final es que los instrumentos genéricos pueden captar los efectos de otras discapacidades y limitaciones que no están directamente relacionadas con la enfermedad específica que se está mapeando.

Una posible estrategia para evitar estos problemas sería diseñar un experimento de elección de preferencias teniendo en cuenta las condiciones de salud derivadas del instrumento específico. Otra posibilidad podría ser describir condiciones de salud específicas solo por el conjunto de perfiles de salud genéricos que prevalecen y son significativos en la enfermedad en particular, y solo se mapean esas condiciones.

En este momento, la asignación directa de estados de salud específicos a valores de utilidad genéricos parece seguir siendo la mejor opción, si bien se debe tener especial cuidado al agregar valores de utilidad genéricos sobre puntuaciones de severidad específicas para suavizar el impacto de efectos que no son propios de la patología.

Las puntuaciones de utilidad SF-6D obtenidas fueron más sensibles a la gravedad de la patología que las obtenidas en EQ-5D-3L. La distribución de las primeras fue más amplia, con menos probabilidad de efectos de techo, y no mostró una brecha entre la salud perfecta y el siguiente valor mayor. La función de distribución acumulada observada de las puntuaciones de desutilidad SF-6D fue más uniforme que la del EQ-5D-3L, que fue más pronunciada (especialmente en los estados de salud más leves).

Las puntuaciones del GSF-Q mostraron un buen comportamiento unidimensional que permitió resumir la severidad en una única puntuación utilizando las ponderaciones del análisis factorial. La estrategia de obtener puntuaciones mediante análisis factoriales debería preferirse a una que utilice la codificación ficticia de ítem-respuesta en los modelos de regresión, ya que evita decidir cómo agregar niveles de respuesta de ítem y minimiza el posible impacto de las covariables en los niveles de respuesta particulares.

Para cada uno de los instrumentos genéricos, se seleccionó el mejor modelo. En ambos casos fue el cúbico el que obtuvo mejor ajuste. La inclusión de covariables importantes en todos los modelos sugiere que la pérdida en la CVRS puede verse influida no solo por los síntomas de ERGE, sino también por las comorbilidades personales presentes. Esto quiere decir que los síntomas de ERGE pueden no ser muy importantes cuando se evalúa la CVRS usando un instrumento genérico si se presentan otras afecciones de salud, como el envejecimiento, el tratamiento de comorbilidades y el sobrepeso.

Se demostró que el mapeo de las puntuaciones GSF-Q específicas de GERD en utilidades genéricas (SF-6D y EQ-5D-3L) fue posible, alcanzando valores de bondad de ajuste adecuados.

El uso de modelos de predicción cúbicos requiere un cuidado especial, ya que pequeñas variaciones en los predictores pueden conllevar valores predictivos excesivamente grandes, que pueden producir predicciones no razonables.

Dada la cantidad reducida de estados de salud prevalentes obtenidos para los instrumentos genéricos (especialmente para EQ-5D-3L), surge la pregunta de si se pueden identificar algunos estados de salud característicos o "naturales" relacionados con la enfermedad para cada instrumento genérico, descartando otros estados de salud influenciados por la comorbilidad. Desde la lógica común, y como ejemplo, parece bastante tentador pensar que el reflujo gastroesofágico no conllevaría un alto deterioro de la movilidad, pero podría darse el caso de que las personas que están inmovilizadas en la cama probablemente desarrollen la patología.

Una forma posible de minimizar el impacto de las comorbilidades, al medir condiciones de salud específicas con un instrumento genérico, sería usar un conjunto de instrucciones que exijan que el paciente evalúe su estado de salud general mientras piensa solo en su enfermedad específica.

Conclusiones

Este estudio nos permite concluir que podemos llevar a cabo un proceso de traslación de puntuaciones de severidad de la patología del reflujo gastroesofágico a instrumento genéricos de calidad de vida relacionada con la salud (EQ-5D-3L y SF-6D).

Así mismo, podemos concluir que hay que tener precaución cuando se utilizan los instrumentos genéricos de CVRS en patologías leves, ya que comorbilidades o dolencias ajenas a la misma pueden distorsionar los valores de utilidad propios de la misma.

Tercer estudio

Objetivo

El objetivo de este estudio es obtener los algoritmos de mapeo necesarios para traducir la medición específica de CVRS obtenida por el FKSI en tres de los instrumentos genéricos basados en preferencias más populares (SF-6D, EQ-5D-3L y HUI-III). Analizaremos dos procedimientos: por un lado, mediante métodos de regresión y, por otro, obteniendo perfiles mediante análisis de conglomeración de datos.

Como beneficio secundario, podremos evaluar cuál de los instrumentos genéricos es más adecuado para capturar el deterioro de la CVRS debido a la condición de ERC.

Método

Diseño del estudio

Se diseñó como un estudio observacional prospectivo de corte transversal. El único criterio objetivo que se estableció en la muestra fue alcanzar un tamaño suficiente para poder llevar a cabo los análisis multivariantes propuestos. Se fijó como objetivo disponer de una muestra mínima de 150 pacientes analizables. Los pacientes fueron reclutados aleatoriamente por los terapeutas colaboradores de la asociación ALCER, sin limitaciones respecto a la procedencia geográfica y cumpliendo los requisitos de ser mayor de edad, estar en tratamiento y con capacidad de responder por sí mismo. Los participantes dieron su consentimiento informado. El protocolo del estudio fue aprobado por el CEIT de la UAM.

Sujetos

La muestra final estuvo compuesta por 161 pacientes, de ambos sexos. La edad media fue de 54,6 años y con una antigüedad media de 2,82 años desde el diagnóstico tanto con patologías congénitas como adquiridas. Los pacientes reciben tratamientos conservadores, diálisis o trasplantes. Se cumplieron las pautas de la declaración de Helsinki

Instrumentos

Se administraron tres cuestionarios para medir la CVRS (EQ-5D-3L, SF-6D y HUI-III) y un instrumento específico de severidad de la sintomatología de enfermedad crónica de riñón (FKSI-9). Además, se administraron el cuestionario de ansiedad/depresión hospitalaria de Hamilton (HADS) y el cuestionario de expresión/rasgo de ira (STAXI).

Análisis estadísticos

La puntuación de severidad de la patología se hizo coincidir con la puntuación factorial obtenida mediante un análisis de componentes principales considerando el instrumento específico como unidimensional, reescalando su valor a una métrica 0-1.

Se realizó una traslación métrica, mediante modelos lineales y no lineales, de los valores de severidad a cada uno de los valores de desutilidad de cada instrumento genérico, obteniendo valores de desutilidad pronosticados para cada valor de severidad.

Se estimaron modelos de regresión lineal, cuadrático y cúbico, así como el modelo Tobit. En la estimación de los diferentes modelos de predicción, se eliminaron aquellos pacientes que puntuaron valores anómalos en al menos dos instrumentos genéricos, ya que su puntuación podía estar reflejando características que no eran propias de la patología de estudio. Se consideraron valores atípicos los pacientes ubicados claramente fuera del intervalo de confianza (95%) para los individuos del modelo lineal (aproximadamente, un residuo tipificado superior a 3).

Junto con la significación estadística de los coeficientes de regresión, se evaluó la bondad de ajuste (GOF) de cada modelo utilizando los estadísticos R^2, error absoluto promedio (MAE) y error absoluto promedio porcentual (MAPE). MAE y MAPE se calcularon en general y por grupos quintiles en función de los puntajes de gravedad, para evaluar el GOF local en los diferentes niveles de gravedad.

No se incluyeron covariables en los modelos de regresión para considerar solo el efecto directo de la enfermedad.

Se llevó a cabo un análisis de perfiles latentes (LCP) para comprobar si la ordenación de los estados de salud resumidos por cada uno de los tres instrumentos genéricos pudiera estar teniendo en cuenta información o niveles diferentes de salud. Se introdujeron en el LCP las desutilidades de los instrumentos genéricos, así como la severidad del instrumento específico (FKSI). A fin de poder describir los conglomerados obtenidos, se introdujeron como covariables inactivas las variables sociodemográficas y las propias de la patología.

Resultados

Las puntuaciones de severidad obtuvieron un valor promedio de 0,261 (DT = 0,219) y con asimetría positiva AS = 0,954 (ET = 0,191).

Las puntuaciones promedio de utilidad obtenidas con los instrumentos genéricos fueron sensiblemente diferentes: MEQ = 0,676 (DTEQ = 0,247), MSF = 0,514 (DTSF = 0,286) y MHU = 0,673 (DTHU = 0,232), siendo el promedio del instrumento SF-6D significativamente menor que las restantes medias ($p < 0,001$). Las correlaciones entre las mismas fueron *r* (EQ, SF) = 0,797, *r* (EQ, HU) = 0,764 y *r* (SF, HU) = 0,763, todas significativas ($p < 0,001$). Las puntuaciones presentaron un claro sesgo negativo en todos los casos, con acumulación de casos en la parte superior de la escala: ASEQ = 0,682 (ET = 0,191); ASSF = 0,904 (ET = 0,195); ASHU = 0,944 (ET = 0,194).

En cuanto al grado de sensibilidad mostrado por los instrumentos, se observó que el EQ-5D-3L fue el menos sensible, obteniéndose solo 36 perfiles, mientras que con el HUI-III se obtuvieron 89 perfiles y con el SF-6D 146.

El modelo de mejor ajuste fue el cúbico para los tres instrumentos genéricos, si bien hay que resaltar que las diferencias de ajuste entre los modelos de distinta forma (lineal, cuadrático y cúbico) fueron mínimas. El patrón cúbico fue elegido además por representar mejor la forma esperada para la evolución de las utilidades

El valor de ajuste de los modelos fue moderado, siendo el de mejor ajuste el correspondiente al SF-6D ($R^2 = 0,619$), mientras que EQ-5D-3L ($R^2 = 0,548$) y HUI-III ($R^2 = 0,565$) fueron inferiores. Sin embargo, el error porcentual del modelo que utiliza el SF-6D como predictor fue mucho mayor (MAPE = 56,9%) que el obtenido por los otros dos modelos, que se mantuvo en torno al 20%.

Como se esperaba, el valor de desajuste por quintiles resultó ser especialmente malo en el quintil correspondiente a los valores altos de utilidad, es decir, en los estados de salud menos graves.

El análisis LCP dio lugar a cuatro conglomerados. La solución alcanzó una $R^2 = 0,87$ con una tasa de error de clasificación del 7%. El análisis LCP permite apreciar que los conglomerados identifican grupos de pacientes con niveles de deterioro progresivo en la enfermedad (FKSI) y también en calidad de vida. Al no existir cruces podemos inferir que otros aspectos de salud estén influyendo sustancialmente en la medición de la CVRS. La progresión de la desutilidad a medida que se avanza entre los estratos de discapacidad dentro de cada instrumento es similar para los tres instrumentos.

Discusión

El paciente tiene una visión global de su estado de salud y resulta difícil aislar el efecto de posibles comorbilidades, efectos adversos o incluso el estado afectivo. Las medidas genéricas reflejan el valor social del estado de salud en el que se encuentra el paciente (comparado con otros estados de salud posibles).

Una estrategia posible para eludir este problema sería diseñar un experimento de elección de preferencias de las condiciones de salud derivadas del instrumento específico. Otra posibilidad podría ser determinar los perfiles de salud genéricos que son realmente prevalentes y significativos en la enfermedad particular, y solo mapear esas condiciones.

Por el momento, el mapeo directo de estados de salud específicos en valores de utilidad genéricos parece ser la opción más aceptada. Pero otra posible forma de determinar el mapeo entre instrumentos genéricos anclados por un instrumento específico podría ser el derivado de generar perfiles a partir de los estados de salud compartidos por grupos de pacientes, utilizando procedimientos de generación de conglomerados. Este procedimiento nos permitirá determinar tantos conglomerados como se considere oportuno y obtener la tabla de correspondencias entre utilidades de distintos instrumentos a partir del valor medio de las utilidades de cada instrumento, representado por el centroide de cada conglomerado. Para un número reducido de conglomerados esta opción es viable, pero el comportamiento con un número elevado de conglomerados puede que no sea tan uniforme y que aparezcan inversiones entre los instrumentos (cruces en los perfiles).

Las puntuaciones de utilidad obtenidas con los instrumentos SF-6D y HUI-III mostraron ser más sensibles a la gravedad de la ECR que las obtenidas con el EQ-5D-3L. La función de distribución acumulada observada de las puntuaciones de desutilidad SF-6D y HUI-III fueron más uniformes, mientras que en el EQ-5D-3L fue más pronunciada (especialmente en los estados de salud más leves).

La estrategia de la utilización de las puntuaciones factoriales para resumir la severidad relacionada con ECR, es técnicamente preferible al uso de la puntuación obtenida directamente de la suma de las respuestas a los ítems del FKSI, ya que pondera cada ítem en función de su

capacidad para ordenar de manera óptima a los pacientes en función de la severidad conjunta captada por todos los ítems que responde el paciente. Además, evita tener que decidir cómo se deben agregar las puntuaciones en la variable criterio (la desutilidad) según los niveles de respuesta en cada uno de los ítems y minimiza el posible impacto de las covariables en determinados niveles particulares de respuesta.

En el caso de predecir la desutilidad de los instrumentos genéricos, el modelo que mejor se adapta es el modelo cúbico. Todos los modelos propuestos presentaron el mismo problema, la gran dispersión de las puntuaciones de utilidad observadas en los estados de salud poco severos del FKSI. Pero este fenómeno no debe ser interpretado como un comportamiento anómalo, sino que obedece a una limitación de los propios instrumentos específicos para captar el efecto de covariables que puedan explicar el nivel de deterioro general y no tanto por la limitación de los instrumentos genéricos para medir los estados de salud más benignos. Lógicamente, se podrían obtener modelos de mejor ajuste incluyendo covariables no específicas de la ERC, pero eso llevaría a modelos de aplicación limitada en otros conjuntos de datos y, por tanto, los modelos de mapeo no serían generalizables.

Conclusiones

Se demostró que es posible mapear las puntuaciones FKSI específicos de ECR, en desutilidades genéricas (SF-6D, HUI-III y EQ-5D-3L), logrando valores adecuados de bondad de ajuste y valores aceptables de cantidad de varianza explicada (entre el 55% y el 62%).

La supremacía del modelo cúbico para el mapeo entre el instrumento específico y los genéricos no fue muy evidente, ya que los valores MAPE de los distintos modelos fueron muy similares. La similitud de los modelos es debida, fundamentalmente, al escaso ajuste obtenido por todos ellos en los valores bajos de desutilidad, mientras que en la parte de desutilidades medias la diferencia de curvatura de los modelos es inapreciable. Este es un problema inherente a los instrumentos genéricos, que han demostrado captar deterioros de salud no atribuibles a la patología específica valorada por el FKSI.

CONCLUSIONES GENERALES

Una vez analizados los estudios presentados podemos establecer las siguientes conclusiones.

En primer lugar, debemos indicar que uno de los principales problemas que encontramos es la forma de definir el índice de severidad del instrumento específico. Como hemos indicado anteriormente, estos no están pensados para la obtención de un índice global de valoración del mismo, pues, o bien no se analiza su estructura interna, o bien esta no nos permite obtener un valor global representativo. Por ello, y como primera propuesta, se considera que la opción más favorable es la de establecer una estructura unidimensional, de manera que cada ítem aporte su peso, y en función de la respuesta dada por el paciente se obtenga una función aditiva que nos permita obtener el valor de la severidad, aun cuando el instrumento tenga su procedimiento de obtención.

Respecto a los instrumentos genéricos, debemos indicar que no miden la calidad de vida teniendo en cuenta las mismas dimensiones, por lo que la traslación (mapping) entre ellos es muy relativa. Al tener un número desigual de dimensiones, el número de perfiles que se pueden obtener es muy distinto, lo que implica que a un perfil determinado de uno de ellos le puede corresponder una amplia variedad de perfiles de otro, y no solo eso, sino que al tener MAUF diferentes, en algunos casos aditivas y en otras multiplicativas, perfiles relativamente severos de un instrumento coinciden por el valor de la utilidad con perfiles relativamente poco severos de otro. Llevar a cabo un proceso de agregación por la media de los valores de un instrumento respecto al otro desvirtúa en gran medida el valor de la utilidad, ya que los valores extremos distorsionan su valor. Es preferible llevar a cabo un análisis que elimine aquellos sujetos que han puntuado de manera atípica o no coherente entre los diferentes instrumentos y no realizar el proceso de agregación.

Con respecto al proceso de mapping entre instrumentos específicos y genéricos debemos indicar que, claramente, estos no son igualmente sensibles a las diferentes patologías. La razón es bien sencilla: hay patologías cuyo nivel de severidad no se ve reflejado en las dimensiones que mide un determinado instrumento genérico y, sin embargo, no por ello debemos considerar que la patología misma sea menos grave que otras cuya sintomatología se ve bien

recogida en el instrumento genérico. Un caso muy frecuente es el deterioro cognitivo, al que se ha dado muy poca importancia tradicionalmente en el ámbito médico, siendo considerado como mucho como "deterioro mental", sin distinción alguna de los trastornos afectivos en la mayoría de instrumentos genéricos. En ocasiones nos podremos encontrar con que, simplemente, el instrumento genérico no recoge el deterioro en la calidad de vida debida a un trastorno particular y carece de sensibilidad frente a ese trastorno.

Otro problema es determinar si es necesario incluir en el mapping el impacto adicional de las covariables del paciente en su calidad de vida. Al realizar el mapping de un instrumento específico sobre otro genérico, nuestro interés fundamental es trasladar una valoración subjetiva del deterioro experimentado con la enfermedad a la métrica de la preferencia poblacional dada al estado de salud del paciente. Normalmente para hacer valoraciones económicas que incluyan la opinión del paciente en la comparación de los tratamientos o alternativas terapéuticas. Para que el modelo sea estimable de manera infalible y fácilmente generalizable, no solemos incluir en el modelo covariables que pueden no haber sido utilizadas en el estudio original o que no han sido registradas en las bases de datos de los estudios clínicos. Sin embargo, sabemos que los modelos de calidad de vida han demostrado ser sensibles a covariables no específicas de la enfermedad como el sexo, la edad, el estado civil, el nivel educativo, el nivel socioeconómico, el hábitat rural o urbano, la presencia de enfermedades concomitantes, etc., y que pueden tener influencia en los valores recogidos por el instrumento genérico. Está claro que la inclusión de las covariables mejorará el ajuste del modelo de mapping, pero deberemos valorar si es realmente relevante y sustancial la cantidad de información que aportan las covariables a la predicción realizada a partir de las opiniones del paciente, que es la información que realmente deseamos incluir en los modelos económicos.

Otra aproximación que se está utilizando en la actualidad es paliar la insensibilidad del instrumento específico a algún aspecto importante de la patología que es capaz mermar de manera sustancial la calidad de vida percibida por el paciente (impactando en su funcionalidad o en sus percepciones, como por ejemplo la presencia de depresión). En esos casos se está proponiendo ampliar el instrumento específico con indicadores sensibles a los ámbitos o áreas funcionales no cubiertas. Pero deberemos ser conscientes de que esta estrategia alterará las propiedades psicométricas del instrumento específico original, siendo una cuestión que quedará pendiente de evaluación.

A pesar de ser atractivo, seguir la misma línea de razonamiento con el cuestionario genérico y ampliarlo con atributos o dimensiones no contempladas inicialmente resulta inviable, ya que eso invalidaría el uso de la función de utilidad multiatributo. Aunque hemos encontrado propuestas de este estilo a la que se denomina ampliación ("bolt-on", atornillado) de dimensiones.

A pesar de haber analizado diferentes métodos para realizar el mapping, son los modelos de regresión lineal, cuadrático o cúbico, sin mucha diferencia de ajuste entre ellos, los más adecuados. Los índices MAE y MAPE deben ser interpretados con cautela, ya que los valores de utilidad muy pequeños pueden inflar notablemente los valores de desajuste, al estar dividiendo por cantidades próximas a cero. El análisis de los índices de ajuste por quintiles resulta mucho más adecuado, ya que nos permite detectar en qué parte de la curva el ajuste es mejor o peor.

Para realizar los procesos de mapping se aconseja modelar valores de desutilidad en lugar de valores de utilidad, en primer lugar, porque la masa de datos generalmente se concentra en torno a estados de salud más indulgentes, y las bajas desutilidades se acercarán más al origen del eje, y en segundo lugar, porque podemos anclar desutilidades de valor 0 (salud perfecta) y por tanto podemos omitir el valor de la constante.

La aplicación de un procedimiento de análisis de perfiles latentes (LCP) nos permite encontrar perfiles comunes dentro de la patología y cada uno de los instrumentos. Hay que considerar que los mismos se han obtenido a partir de valores medios de severidad y desutilidades y están obtenidos mediante el paquete informático Latent Gold, el cual trabaja con probabilidades de pertenencia a uno u otro conglomerado. Una ventaja de este sistema es que podemos tener controladas covariables no introducidas en la predicción de los perfiles y que nos permitirían predecir utilidades de instrumentos genéricos que no han sido cumplimentados por los pacientes. El proceso de predicción puede realizarse mediante interpolación de los valores medios de los perfiles definidos. Un inconveniente de este sistema es que en función de la muestra de estudio pueden producirse cruces de perfiles, por lo que estamos obligados a definir un número de perfiles que impidan que dichas intersecciones se produzcan en demasía y no se permita una interpretación adecuada del resultado.

LIMITACIONES

En los estudios llevados a cabo no se ha incluido el estudio del efecto de ninguna covariable. Sería interesante valorar el efecto parcial de algunas covariables, en especial sexo, rango de edad y presencia/ausencia de comorbilidad, covariables todas ellas que suelen estar consideradas en la mayoría de los protocolos de recogida de datos clínicos. Determinar el conjunto mínimo de covariables puede ser un hallazgo importante en este campo.

El repertorio de modelos sometidos a estimación ha venido impuesto en gran medida por la disponibilidad en los paquetes estadísticos al uso, y en particular el paquete IBM SPSS Statistics. Es nuestro objetivo comenzar a programar modelos mixtos de distribución con lenguajes de programación personalizables.

Los estudios han sido realizados utilizando la versión de tres niveles de respuesta del EQ-5D-3L, que era la disponible cuando se recogieron los datos de cada estudio. Sería interesante validar los resultados con el instrumento EQ-5D-5L, de 5 niveles de respuesta.

REFERENCIAS BIBLIOGRÁFICAS

Abellán, JM., Sánchez, FI., Martínez, JE., & Méndez, I. (2012). Lowering the floor of the SF6D scoring algorithm using a lottery equivalent method. . *Health Economics* (21 (11)), 1271-1285. doi:10.1002/hec.1972

Alandis, S., Ruiz, M. A., Errando, C. Villacampa, F., Arumi, D., Lizarraga, I., & Rejas, J. (2012). Quality of life in patiens with overactive bladder: valdation and psycometric properties of the spanish Overactive Blader Questionnaire-Short Form. *Clin Drug Investigation.*

Awad, A., Voruganti, L., & Heselgrave, R. (1997). A conceptual model of quality of life in schizophrenia: Description and preliminary. *Qual Life Res* (6), 21-26.

Badía, X., & Lizán, L. (2003). Estudios de Calidad de Vida. Atención Primaria. Conceptos, organización y práctica clínica. (C. P. Martín Zurro A, Ed.) *Elsevier España ediciones*, 250-261.

Baron, J., Wu, Z., Brennan, D., Weeks, C., & Ubel, P. (2001). Analog scale, ratio judgment and person trade-off as utility measures: biases and their correction. *Journal of Behavioral Decision Making* (14), 17-34.

Bishop, M., Berven, N., Hermann, B., & Chan, F. (2002). Quality of life among adults with epilepsy, an exploratory model. *Rehabilitation Counseling Bulletin* (45), 87-95.

Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference based single index from UK SF-36 health survey. *J. Clin Epidemiol,* 51, 1115-28.

Bungay, K., Boyer, J., Steinwald, A., & Ware, J. (1996). Health Related Quality of Life: An Overview. En J. Bootman, R. Towsend, & W. McGhan, *Principles of Pharmaeconomics* (2 ed.). Cincinnati, USA: Harvey Whitney Books Company.

Burke, C. (2001). Testing an Asthma quality of life model. *Journal of Theory Construction & Testing,* 5, 38-44.

Campbell, A., Converse, E., & Rodgers, W. (1976). *The Quality of American Life.* Russel Sage Fundation, New York.

Erikson, R. (1993). Descriptions of inequality: The Swedish approach to welfare research. *The Quality of Life (Oxford press)*, 67-83.

Etcheld, M., Van Elderen, T., & Van Der Kamp, L. (2003). Modeling Predictors of quality of life after coronary angioplasty. *Annals of Behavioral Medicine,* 26, 49-60.

Feeny, D., Furlong, W., Boyle, M., & Torrance, GW. (1995). Multi-attribute health status classification Systems. Health Utilities Index. *Pharmacoeconomics* (7), 490-502.

Feinstein, A. (1987). Clinimetrics perspectives. (J. Chron, Ed.) *Pergamon Journals Ltd,* 40 (6), 635-640.

Fernández, J., Hernández, R., & Cueto, A. (31 de octubre de 1994). ¿Qué son los QALYs? *Atención primaria,* 14 (7).

Frey, S., & Stutzer, A. (2002). "What can economists learns from happiness research? *Journal of Economic Literature* (40), 402-435. doi:doi/10.1257/002205102320161320

Hickey, A., O'Boyle, C., McGee, H., & Joyce, C. (1999). The schudle for evaluation of individual quality of life. En C. Joyce, C. O'Boyle, & H. McGee, *Individual Quality of Life. Approaches to conceptualization and assesement.* Amsterdam: Harwood Academic Publisher.

Lawton, M. (1999). Quality of Life in Chronic Illness. *Gerontology* (45), 181-3.

Masters, G. (s.f.). A Rasch model for partial credit scoring. *Psycometrika,* 47 (2), 149-174. doi: 10. 1007/BF02296272.

Meeberg, G. (1993). Quality of life: a concept analysis. *Journal of Advanced Nursing* (8), 18: 32.

Morgenstern, O., & Von Neumann, J. (1953). *Theory of Games and Economic Behavior.* Princeton University Press.

Noll, H. (2002). Towards a European System of Social Indicators: Theoretical Framework and System Architecture. *Social Indicators Research,* 58, 47-87.

Noll, H., & Zapf, W. (1994). Social indicators research: Societal monitoring and social reporting. *Trends and Perspectives in Empirical Social Research*, 1-16.

Ong, L., Visser, M., & Lammes, F. (1999). Doctor–Patient communication and cancer patients' quality of life and satifaction. (E. S. Ltd., Ed.) *ELSEVIER* (41), 145-156.

Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goñi, JM., & Luo, N. (2016). EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. *PharmacoEconomics* (34), 993-1004.

Rapley, M. (2003). *Quality of Life Research. A Critical Introduction.* London: Sage.

Robinson, A, S. A. (2006). Exploring challenges to TTO utilities: valuing states worse than dead. *Health Economic* (15 (4)), 393-402.

Ruiz, M., Rejas, J., Soto, J., Pardo, A., & Rebollo, I. (2003). Adaptación y validación del Health Utilities Index Mark 3 al castellano y baremos de corrección en la población española. *Med Clin* (120 (3)), 89-96.

Sanda, M., Rodney, L., & Dunn, M. (marzo de 2008). Quality of Life and Satisfaction with Outcome among Prostate-Cancer Survivors. *The New England Journal of Medicine*, 1250-1261. doi:DOI: 10.1056/NEJMoa074311

Segurado, A., & Agulló, E. (2002). Calidad de vida laboral: hacia un enfoque integrador desde la Psicología Social. *Psicothema,* 14 (4), 828-836.

Stevens, K., McCabe, C., Brazier, J., & Roberts, J. (2007). Multi-attribute utility function or statistical inference models: A comparison of health state valuations models using the HUI-III2 health state classification system. *Health Economic* (26), 992-1002.

Streiner, D. (2003). Clinimetrics vs. psychometrics: an unnecessary distinction. *Journal of Clinical Epidemiology* (56), 1142-1145.

Tate, D., & Forchheimer, M. (junio de 2002). Quality of life, life satisfaction, and spirituality: Comparing outcomes between rehabilitation and cancer patients. *American Journal of Physical Medicine and Rehabilitation,* 81, 400-410.

Testa, M. (2000). Interpretation of quality of life outcomes: issues that affect magnitude and meaning. *Medical Care* (38: II), 166-174.

Thoits, P., & Angel, R. (1987). The impact of culture on the cognitive structure of illness. *Culture, Medicine and Psychiatry,* 11, 465-94.

Uusitalo, H. (1994). Social statistics and social reporting in the nordic countries. (P. Flora, F. Kraus, H. Noll, & F. Rothenbacher, Edits.) *Social Statistics and Social Reporting in and for Europe*, 99-120.

Villagut, G., Ferre, M., Rajmil, L. Rebollo, P. Permanyer-Miralda, G. & Quintana, J. M. (2005). El cuestionario SF-36 español: Una década de experiencias y nuevos desarrollos. *Gaceta Sanitaria (19, 135-150).*

Vosvick, M., Koopman, C., Gore, F. C., Thoresen, C., Krumboltz, J., & Spiegel, D. (2003). Relationship of functional quality of life to strategies for coping with the stress of living with HIV/AIDS. *Psychosomatics: Journal of Consultation Liaison Psychiatry*, 44-51.

Ware, J., Kosisonski, M., Keller, SD., & Gandek, B. (1993). *SF36 Health Survey Manual and Interpretation Guide.* Boston Massachusetts: The health Instituite, New England Medical Center.

Warrr, P. (2012). How to think about and measure psychological well-being. En Sinclair, M. Wang & L. E. Tetrick (Eds.) Reserach methods in occupational health psychology. Routledge (pp. 100-114).

Weinstein, M., & Siegel, J. (16 de octubre de 1996). Recommendations of the Panel on Cost-Effectiveness in Health and Medicine. *JAMA* (276), 1253-1258. doi:doi:10.1001/jama.1996.03540150055031

WHOQOL Group (1995). The World Health Organization Quality of Life Assessment (WHOQOL): Position Paper from the World Health Organisation. *Social Science & Medicine* (10), 1403- 1409.

Wilson, I., Cleary, P. (1995). Linking clinical variables with health related quality of life. *JAMA* (273), 59-65.

Wu, A. (2000). Quality of life assessment in clinical research: application in diverse populations. *Medical Care,* 38 (III), 30-35.

Yang, Y., Brazier, J., Tsuchiya, A., & Coyne, K. (2009). Estimating a Preference-Based Single Index from the Overactive Bladder Questionnaire. *Value in Health* (12 (1)), 159-166.

ANEXOS

Clin Drug Investig
DOI 10.1007/s40261-016-0377-z

ORIGINAL RESEARCH ARTICLE

Mapping of the OAB-SF Questionnaire onto EQ-5D in Spanish Patients with Overactive Bladder

Miguel A. Ruiz[1] · Laura L. Gutiérrez[1] · Manuel Monroy[1] · Javier Rejas[2]

Abstract

Background and Objective Mapping disease-specific measures onto generic preference-based indexes allows estimating utility values in specific conditions to determine gain of quality-adjusted-life-years when the status of condition varies. The aim of this study was to map a disease specific scale, the Overactive Bladder Questionnaire 5-dimensional health classification system (OAB-5D) derived from the Overactive Bladder questionnaire-Short Form (OABq-SF), onto a preference-based scale, the EuroQol-5D (EQ-5D), in a sample of patients with overactive bladder (OAB) in a Spanish population.

Method A survey addressed to value the health states was conducted among 246 patients at 18 clinics of urology from Spain. A total of 43 out of 243 possible health states have been valued, using VAS (Visual Analog Scale) and TTO (time trade-off) techniques. In addition, ordinary least squares (OLS), generalized linear models (GLM) and Tobit models were estimated. Resulting models were compared and the best one was selected in terms of goodness of fit measures, attribute sign, coefficient magnitude, and statistical significance of regression coefficients. Finally, the internal validity of the best model was calculated by bootstrap resampling.

Results The best model to map the OAB-5D onto EQ-5D could be estimated and the stability of parameter estimations was proved. The mentioned model estimated through OLS regression attained R^2 value of 0.892, with the aggregated data; with GLM (efficient maximum likelihood regression), Pearson χ^2 of 15.3 has been obtained; AIC (Akaike information criterion) = −550.9 and BIC (Bayesian information criterion) = −475.4. OLS model included the following OABq-SF items (and range of weights): A1 (0.102, 0.216); A3 (0.070, 0.171); B3 (0.071, 0.078); B1 (0.076, 0.136); B2 (−0.132, −0.028).

Conclusion It is possible to map the OAB-5D scores onto EQ-5D in the Spanish population, allowing estimating EQ-5D utility scores from OAB specific health conditions.

Key Points

This work was able to map the OABq-SF scores onto EQ-5D in a Spanish population, allowing estimating EQ-5D utility scores from OAB specific health conditions.

Economic evaluations in OAB may be carried out now using specific measurement of patient's health status with overactive bladder.

1 Introduction

Overactive bladder (OAB) is a syndrome characterized by the symptom of urinary urgency accompanied with or without urge incontinence, which is often associated with urinary frequency and nocturia, that appear without a local

✉ Javier Rejas
javier.rejas@pfizer.com

Miguel A. Ruiz
miguel.ruiz@uam.es

[1] Department of Methodology, School of Psychology, Universidad Autónoma de Madrid, Ctra. De Colmenar, km 15, 28049 Madrid, Spain

[2] Health Economics and Outcomes Research Department, Pfizer, S.L.U., Alcobendas, Spain

Published online: 10 February 2016

pathologic or metabolic explanation [1]. Estimates of prevalence for incontinent OAB range from 6 to 35 % in Europe, and up to 16.5 % in the USA [2, 3]. Some authors have reported that only 27 % of OAB patients receive medication to relieve their symptoms [4]. One possible reason could be underreporting of symptoms to the clinician, perhaps due to lack of knowledge about available effective treatments, social stigma associated to bladder control problems, or the extended belief that these symptoms are part of normal aging [5].

Since OAB is defined by patient self-reported symptoms, treatment effectiveness assessment should also be based upon patient perceptions. Patient-reported outcomes (PRO), such as symptom bother and health-related-quality-of-life (HRQOL), are important and clinically relevant outcomes to evaluate, particularly when no clinically objective and reliable means are available, such as in the continent OAB sufferer [6, 7]. OAB symptoms have been shown to be highly bothersome to patients and to negatively impact on overall HRQOL. Importantly, the bother caused by OAB symptoms has been shown to vary among individual patients, and it has been suggested that a decrease in the level of bother may be a better indicator of how satisfied patients are with OAB treatment than relying exclusively on measures of symptom frequency [8]. Additionally, patients with OAB tend to employ coping behaviors that allow them to tolerate their symptoms [5].

On the other hand, the increasing need of a proper allocation of health resources has propelled the development of new techniques directed to compare disease burden across pathologies. There is a growing body of literature on mapping functions of "source" disease-specific HRQOL measures onto "target" generic preference-based measures based on regression models [9]. Utilities are the cornerstone of economic evaluations and their value is needed to compute quality-adjusted life years (QALYs), but patient values are usually gathered using generic instruments such as Medical Outcome Survey- Sort Form-6 Dimensions (SF-6D), Health Utility Index-III (HUI3) or EuroQol-5 Dimensions (EQ-5D). In clinical trials, disease-specific instruments are preferred since they are more sensitive to clinical changes, and when elderly people are being studied it is preferred not to overload patients with additional generic measurements. Hence, the possibility of translating specific PRO outcomes into utilities is a desirable procedure to follow. This methodology is also needed when generic measures have not been included in a closed clinical study and a meta-analysis is needed.

The solution of mapping disease-specific measures onto generic preference-based indexes (such as EQ-5D), allowing to estimate utilities, has been accomplished in cancer [10–12], Parkinson disease [13], insomnia [14], inflammatory bowel disease [15] and osteoarthritis [16], among others. But the mapping of OAB-specific measures onto EQ-5D has not been done in Spanish culture. Mapping is one solution that is gaining popularity as it enables health state utility values to be predicted when no preference-based measure has been included in the study. This approach involves estimating the relationship between a non-preference-based measure and a generic preference-based measure using statistical association (also known as 'cross-walking' [17] or estimating exchange rates between instruments), and requires a degree of overlap between the descriptive systems of the two measures and that the two measures are administered on the same population. Typically, mapping uses two datasets: an estimation dataset that contains respondents' self-reported scores for their own health, using two or more preference- and non-preference-based measures, and a study dataset containing only the non-preference-based measure. Regression techniques are used on the estimation dataset to estimate a statistical relationship between the measures, and the results are then applied to the study dataset to obtain predicted health state utility values [9].

Recent studies have worked out a mapping solution of the Incontinence Quality of Life Questionnaire (I-QOL) onto the EQ-5D for patients with idiopathic or neurogenic OAB [18]. In our study, the OAB-5D was used as source-specific mapping measure [19]. The OAB-5D is a 5-attribute classification system derived from the OABq-SF questionnaire [20] by extracting relevant severity dimensions using a Rasch model [21] and developed as a preference-based direct estimate of utilities for general population with valuations obtained through the time-trade-off (TTO) method. The sample used in the present study did not include neurogenic OAB patients therefore all participants were close to subclinical pathology.

The aim of this study was to map a disease-specific scale, the Overactive Bladder Questionnaire 5-dimensional health classification system (OAB-5D), onto a preference-based scale, the EQ-5D, in a sample of patients with OAB in the Spanish population.

2 Methods

2.1 Study Design

A multicenter prospective observational study, carried out under normal conditions of clinical practice, was designed. All patients were asked to complete the written informed consent in order to be included in the study, which was approved by the Research Ethics Committee of the Universidad Autónoma de Madrid, under the label "Aqua Study A0221077". The study included two visits: (1) baseline visit: at recruitment, the OABq-V8 was

administered to patients, along with the OABq-SF, and a three-day patient diary. Inclusion criteria, socio-demographic data (gender, age, body mass index, ethnicity and study level), and the medical records were also gathered. The EQ-5D was completed at the end of the visit. (2) follow-up visit: after 3 months of starting a new antimuscarinic-based therapy, the OABq-SF was administered again. Treatment effectiveness was assessed twice: the Clinical Global Impression of Improvement scale (CGI-C) [22] was completed by clinicians and the Treatment Benefit Scale (TBS) [23] was responded to by patients. Compliance with treatment was also measured using the Morisky-Green questionnaire [24]. For the purpose of the present work, we only considered data collected during the baseline visit.

2.2 Participants

The study enrolled patients of both genders, above 18 years old, with symptomatic OAB and fulfilling the following inclusion criteria: older than 18 years, with OAB as diagnosed in routine practice, giving informed consent to participate in the study and attaining a score above 8 points in the OAB-V8 questionnaire [5]. All patients included in this study were initiating or intensifying treatment for OAB with an antimuscarinic-based therapy (pharmacological treatment of urinary symptomatology of OAB with drugs that block the muscarinic receptors on the bladder) [25].

A convenience sample was recruited with random selection of patients. Sample size determination was based in the expected change in the domains of the OABq-SF scale, which was expected to vary between 0.99 (HRQoL scale) and 1.14 (symptom bother scale) with a standard deviation of 2.0, in absolute value, corresponding to a small effect-size (half a standard deviation). Furthermore, it was expected to find a minimum correlation of 0.3 between OAB-V8 and both CGI-C and TBS scales. Sample calculation was estimated with a 95 % confidence interval, a type I error <0.05 and a statistical power of 90 %. Finally, sample size was increased by 30 % of the estimated sample size to protect the study against missing data at follow up. Taking into account all the mentioned criteria, a minimum sample size of 164 patients was determined suitable for our research purposes.

2.3 Instruments

2.3.1 Overactive Bladder-V8 (OAB-V8)

The OAB-V8 scale measures the extent of involvement by the major symptoms of the OAB. This scale is adapted from the scale of quality of life in symptoms of OAB-q VH of 33 items that remain 8 items of the OAB-q, but the instructions are modified for filling, leaving as a screening tool. Symptoms are evaluated using a scale of 6-point Likert-type (0 = nothing to 5 = very much) plus a dichotomous question about the sex of the patient (man = 2 additional points). The total score is obtained by adding the individual scores of the items. A score ≥ 8 points indicates that it is likely that the patient suffers from an OAB [5].

2.3.2 Overactive Bladder Questionnaire (OABq-SF)

The OABq-SF is a specific HRQOL questionnaire abbreviated form of the Overactive Bladder Symptom and Health-Related Quality of Life Questionnaire (OAB-q) [26]. The OAB-q was shortened into the OABq-SF to reduce respondent burden while retaining the original reliability, validity, and responsiveness. In the OAB-SF, the symptom-bother scale was reduced to 6 items and the HRQOL was reduced to 13 items [20]. Patients should rate each item on a six-point Likert scale, ranging from "none of the time" to "all of the time" for the HRQOL items, and "not at all" to "a very great deal" for the symptom-bother items. The subscales are summed and transformed into scores ranging from 0 to 100. A 4 weeks recall period was used. Psychometric properties of the Spanish validated version have been reported elsewhere [27].

A preference-based utility index based on five dimensions extracted from the OABq-SF, called the OAB-5D has been developed in UK general population [19]. This instrument classifies patient health states by considering five attributes: Urge, Urine, Sleep, Coping and Concern. Attribute scores are derived from responses to five particular OABq-SF items (Table 1).

OAB-5D scores are obtained by creating a series of dummy variables capturing the impairment levels related to every OAB dimension and, when necessary, by combining two consecutive levels of the scale. Furthermore, a dummy variable (labelled *N2*) is computed to capture the additional severity of OAB symptoms, when the maximum level of severity (score = 6) is attained for two or more of the OAB-5D dimensions. *N2* is included based on the assumption that possible interactions may occur more frequently in severe levels than in mild levels. Since the increment in one unit in the original item Likert scale could not be constant thorough its entire metric, each item is decomposed in five dummy variables, representing presence/absence of each one of the severity levels (1 to 6) on the item considered.

The utility variable (u) has a scale ranging between $u = 0$ (denoting "death") and $u = 1$ (representing "full health"). However, some respondents may value their health state "worse than dead", being possible to attain a negative value as low as $u = -1$ for much deteriorated

Table 1 OAB-5D items for Model 0 and Model 1. All items directly extracted from the OABq-SF

OAB-5D dimensions	Original UK model replicated	Best proposed model
URGE	A1: having an uncomfortable urge to urinate	A1: having an uncomfortable urge to urinate
URINE	A6: having an urine loss associated with a strong desire to urinate	A3: having an accidental loss of small amounts of urine
SLEEP	B3: symptoms interfered with your ability to get a good night's rest	B3: symptoms interfered with your ability to get a good night's rest
COPING	B1: symptoms caused you to plan "escape routes" to restrooms in public places	B1: symptoms caused you to plan "escape routes" to restrooms in public places
CONCERN	B12: symptoms caused you embarrassment	B2: symptoms made you feel like there is something wrong with you

A Symptom Bother Scale, *B* Health Related Quality Of Life Scale, *OAB-5D* overactive bladder questionnaire 5-dimensional health classification system, OABq-SF overactive bladder questionnaire-short form

conditions. It is not expected that negative utility values may be observed for health conditions related to OAB symptomatology and a minimum utility $u = 0$ could be assumed. In such cases, it is easier to model disutility values ($\bar{u} = 1 - u$), anchoring the regression models at the minimum theoretical value and avoiding to include the intercept. Therefore, as health state severity increases a higher disability will be assigned by the model, the sign of regression coefficients for the dummy variables should be positive, and with increasing magnitude as we move up in the level of severity they represent.

2.3.3 EuroQol-5D (EQ-5D)

EQ-5D [28] is a generic measure designed to describe and value quality of life related to health states, including five basic dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression, with three levels of response for each dimension. Each combination of dimension scores is considered a health state, defining a classification system with a total of 243 possible health states. Health states are translated into utility values by a Multi-Attribute Utility Function (MAUF). In the present study, 43 different EQ-5D health states have been observed. For Spanish population, two preference indexes have been estimated from general population valuation of health states, using time trade off (TTO) and visual analogue scale (VAS) methods [29].

The EQ-5D MAUF takes the form:

$$u_i = 1 - \left(k + \sum_{j=1}^{j=5} \sum_{k=1}^{k=3} b_{jk} D_{ijk} + b_{N3} N3_i \right)$$

Where the utility for individual i is obtained by combining the dummy score (0,1) obtained by the patient at level k on each j dimension/attribute, weighted by the corresponding b_{jk} MAUF weight, plus a constant (k), and a weighted interaction term ($N3$). The five EQ-5D dimensions/attributes are: mobility (D_1), self-care (D_2), usual activities (D_3), pain/discomfort (D_4), and anxiety/depression (D_5). *N3* dummy variable is a non-multiplicative interaction term, frequently used in EuroQol models [30], which allows measuring the "extra" disutility when severe problems are reported in at least one EQ-5D dimension. $N3 = 1$ when a person's health state has attained level 3 of severity at least in one attribute. It should be noted that $b_{j1} = 0$ for all dimensions, since the first level ($k = 1$) represents perfect health in that attribute.

2.4 Statistical Methods

The UK model was replicated using ordinary least squares general linear models (OLS) and following the specifications given by the authors [19], including the proposal of merging specific dimension levels (Model 0). During the Spanish cultural adaptation process, some problems were found with the translation of item B12, and it was suspected that this particular item, included in the UK OAB-5D algorithm, could present to some extent conceptual divergence in the Spanish version. Being this so, all possible combinations of items assigned to each one the original OAB-5D dimensions were tested as competing models. Best fitting models were selected while models with non-significant weights were discarded and non-monotonous category weights were proposed for merging. Along with OLS models, generalized linear models (GLM), and Tobit models were also tested.

OLS models were estimated by selecting 5-item OABq tuples, one item for each OAB-5D dimension, and decomposing each item in a set of five dummy variables (recall that a pair of consecutive response levels will be merged) indicating the level of severity reached in that

particular item. An interaction term (*N2*) was also included in the model which would take the value $N2 = 1$ when the maximum level of severity was attained for two or more of the OAB-5D dimensions, and $N2 = 0$ otherwise. No intersection term was included, imposing the model to assign a disutility value 0 to the corresponding OAB-5D perfect health state. The OLS model takes the form:

$$\bar{u}_i = \sum_{j=1}^{j=5} \sum_{k=1}^{k=4} b_{jk} D_{ijk} + b_{N2} N2_i + \varepsilon_i = \hat{\bar{u}}_i + \varepsilon_i$$

Where $D_1 =$ urge, $D_2 =$ urine, $D_3 =$ sleep, $D_4 =$ coping, $D_5 =$ concern; b_{jk} are weights assigned to level k for dimension j, b_{N2} represents the weight for the interaction term *N2* and ε_i represents the prediction error term. Again, all $b_{j1} = 0$ since level 1 for any dimension represents full health at that attribute.

General linear models were obtained using OLS estimation to map OAB-5D items onto the EQ-5D metric, being disutility values $\bar{u}_{\text{VAS}}$ or $\bar{u}_{\text{TTO}}$ the model dependent variable. The transformed disutility was used, instead of utility, since the disutility function $f(\bar{u})$ allows avoiding the estimation of the intercept term. For TTO estimates, negative utility values were coded as 0 in order to make the origin-anchored model meaningful.

GLM estimates were obtained using efficient maximum likelihood (EML) regression. $\bar{u}_{\text{VAS}}$ was selected as dependent variable, given that it does not attain negative values (unlike $\bar{u}_{\text{TTO}}$), which is a restriction for calculating $\text{Log}(\bar{u})$. Besides, negative scores do not have the same linear properties as those displayed by positive scores [31], and it is convenient to work with a distribution containing only positive scores. The value of $\bar{u} = 0$ is one of the possible observed values, for which $\text{Log}(0)$ is undefined. Therefore, an arbitrary small constant ($c = 0.05$) was added to each value. On the other hand, the dummy variable sets derived from the OAB-5D items were used as independent variables, as in the OLS estimation. Taking into account the skewed distribution of the dependent variable, the Gamma distribution and the logarithmic link function were used. In fact, a higher concentration of cases was observed at the lower end of the $\bar{u}_{\text{VAS}}$ scale, most of values (60 %) were located below the mean value of $\bar{u}$ (0.24). In addition to the absence of normality of $\bar{u}_{\text{VAS}}$ distribution, it is not tenable to assume predictor homoscedasticity. In fact, a homogeneous dispersion of $\bar{u}_{\text{VAS}}$ unstandardized residuals along predicted values was not found, indicating variance heteroscedasticity; in this sense, the use of OLS estimation could be problematic and GLM should be preferred [32].

Since lack of linearity and the impossibility of obtaining negative values are prominent issues in mapping OAB-5D values, Tobit regression models were also estimated. Tobit regression obtains parameter estimates from the cumulative distribution of observed values for the dependent variable, instead of using the probability density function, which is less sensitive to departures form linearity. It also allows assuming the existence of bellow censoring in the distribution of the dependent variable, which is the case for disutility values, with 32 % of cases obtaining a disutility of 0 (which is to say a utility of 1). The procedure isolates values at the threshold value and obtains regression estimates as if it could be possible to extend the distribution bellow the censoring threshold, simulating the possibility of obtaining negative values, and predicting the censored value based on a Probit estimate [33].

Data were modeled at individual and aggregated levels. Given that we could assume that health states gathered in our sample would take into account the influence not only of the OAB condition but also the influence of age, gender and other concomitant pathologies (which may not be included in the mapping algorithm), we estimated models by aggregating data using OAB-V8 scores as strata and obtaining stratum average disutilities. We rejected aggregating data based on OAB-5D strata in order to avoid introducing an excessive smoothing. We expected aggregated models to be more stable.

Goodness of fit (GOF) was assessed using adjusted R^2, sign and significance level of regression coefficients, as well as the relative size of coefficients within a given dimension. Model predictive capabilities were also assessed using mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), the latter two not weighting more larger errors. GLM and Tobit competing models were also assessed using χ^2, χ^2/df, and χ^2 likelihood ratio test, Akaike information criterion (AIC) and Swartz Bayesian information criterion (BIC) GOF statistics.

As an additional check of model internal validity, bootstrapping estimates of parameter standard errors were also obtained, in order to avoid problems derived from not meeting asymptotic theory assumptions. As a non-parametric technique that attempts to estimate empirically the sampling distribution of any given statistic [30], the bootstrap distribution was used to obtain robust standard error estimates and confidence intervals. This approach intended to validate a multivariate model by drawing a sufficiently large number of subsamples and estimating the model in each subsample [34]. Estimates from all the subsamples were combined, providing best coefficient estimates observed within replications, observed bias against asymptotic estimates, expected variability and likelihood on differing from zero when asymptotic theory assumptions may not hold. Bootstrap percentile approach was used to build 95 % confidence intervals, generating 1000 bootstrap sub-samples with replacement.

In order to ensure that no sample bias could be influencing model estimation due to the unbalanced presence of more women, measurement invariance of the OABq-SF questionnaire was tested. Using a multi-group confirmatory factor analysis, a common dimension structure for both groups was estimated, assessing the loss in GOF when comparing a model in which the structure is freely estimated in each isolated gender group (configural invariance) with a model in which parameters are restricted to be equal in both groups (strong metric invariance). If the loss in fit is

Table 2 Sociodemographics and health status of patients included in the study

Variable	Count	Percentage
N	246	
Gender: women	188	76.0
Ethnicity: Caucasian	238	96.7
Educational level		
Illiterate	16	6.5
Primary	87	35.4
Secondary	58	17.1
Professional degree	42	17.1
Post graduate	43	17.5
Occupation		
Retired	63	25.6
Housewife	76	36.2
Employee	89	36.2
Long-term disability	8	3.3
Unemployed	10	4.1
Marital status		
Married	179	73.1
Widowed	28	11.4
Single	19	7.8
Separated/divorced	19	7.8

	Mean	SD
Age	57.7	12.78
Height	164	7.36
Weight	71.3	10.84
Body mass index	26.55	4.14

Self-reported OAB-5D health state	Level 0 (%)	Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)	Level 5 (%)
Urge	5 (2.0)	17 (6.9)	59 (24.0)	91 (37.0)	51 (20.7)	23 (9.3)
Urine loss	27 (11.0)	34 (13.9)	67 (27.3)	58 (23.7)	34 (13.9)	25 (10.2)
Sleep	19 (7.7)	39 (15.9)	72 (29.3)	67 (27.2)	29 (11.8)	20 (8.1)
Coping	7 (2.8)	20 (8.1)	107 (43.5)	69 (28.0)	22 (8.9)	21 (8.5)
Concern	7 (2.8)	21 (8.5)	119 (48.4)	64 (26.0)	14 (5.7)	21 (8.5)

	Mean	SD
OAB-V8	23.2	7.01
OABq-SF	22.5	5.58
EuroQol-5D utilities		
Estimated EQ-5D VAS	0.762	0.226
Estimated OAB-5D	0.763	0.074

not statistically significant, both groups are thought to share a common structure. Underlying dimension mean scores were also compared.

Statistical analyses were carried out using IBM SPSS 21.0 for Windows and Mplus 7.11 software.

3 Results

A final sample composed of 246 OAB patients was recruited. Mean age was 57.7 years (SD = 12.79); 76 % were women; 99 % Caucasian; 36 % were employees and 35 % reached primary studies (Table 2). Attending to existing comorbidities, 33.7 % did not experience any comorbidity, 38.6 % experienced one comorbidity, and 17.5 % experience two, being the maximum of four comorbidities. The most frequent comorbidity was high blood pressure (33.7 %), followed by diabetes mellitus (21.1 %), frequent urinary infections (16.7 %) and depression (14.4 %). Mean OABq-SF score was 22.5 (SD = 5.58). Distribution of responses by OAB-5D item and response level covered all the range of possible values (see Table 2). Mean VAS utility was 0.760 (SD = 0.226), ranging from min = 0.014 to max = 1, while mean TTO utility was 0.790 (SD = 0.272), ranging from min = −0.330 to max = 1.0. The observed mean value obtained applying the TTO scoring was slightly higher than the VAS scoring (Table 3). TTO scoring assigned negative values to some patients, and significant differences were observed between scoring method mean values ($t_{239} = -5.608$; $p < 0.001$). In our sample, predominant EQ-5D profiles were related to more lenient health states, while TTO scoring assigned negative tariff values ("worse than dead" valued states) to those EQ-5D profiles of patients experiencing a serious impairment in their health, and also associated with the oldest participants.

OLS estimates for the baseline individual level model using TTO scoring (replicating the original study) were very different from UK estimates. While in the UK model 70 % of weights were statistically significant and all attributes presented at least one significant weight, in the Spanish model only urge presented significant weights at the two highest severity levels (results not shown). At the aggregated level, all attributes presented significant weights in the UK sample (although levels needed to be combined, possibly due to lack of monotonicity) while in the Spanish sample coping was not statistically significant. Using VAS scoring in the Spanish sample, at the individual level, only urge obtained positive and significant weights for the two highest levels, while concern presented one significant weight and all attribute levels obtained negative weights. At the aggregated level urine and sleep were also

Table 3 Health states defined by the sample in the EQ-5D and utility values (tariffs)

E5D profile	Frequency	%	Mean (u_{VAS})	Mean (u_{TTO})
1-1-1-1-1	79	32.1	1.000	1.000
1-1-1-1-2	27	11.0	0.799	0.914
1-1-1-2-1	26	10.6	0.790	0.887
1-1-1-2-2	16	6.5	0.739	0.825
2-2-2-2-2	8	3.3	0.493	0.514
1-1-2-2-2	6	2.4	0.684	0.754
2-1-2-2-1	6	2.4	0.645	0.710
2-1-2-2-2	6	2.4	0.594	0.648
2-1-1-1-1	5	2.0	0.760	0.870
1-1-2-1-1	4	1.6	0.795	0.905
1-1-2-2-3	4	1.6	0.421	0.381
1-1-2-3-2	4	1.6	0.412	0.291
2-1-1-2-2	4	1.6	0.649	0.719
1-1-1-1-3	3	1.2	0.536	0.541
1-1-1-2-3	3	1.2	0.476	0.452
1-1-2-1-2	3	1.2	0.744	0.843
1-1-2-2-1	2	0.8	0.735	0.816
2-1-2-1-1	2	0.8	0.705	0.799
2-1-2-2-3	2	0.8	0.331	0.275
2-1-2-3-3	2	0.8	0.272	0.103
2-2-1-1-1	2	0.8	0.659	0.736
2-2-2-2-1	2	0.8	0.544	0.576
2-2-2-3-2	2	0.8	0.222	0.051
2-2-2-3-3	2	0.8	0.170	−0.031
2-2-3-3-3	2	0.8	0.115	−0.155
1-1-2-3-3	1	0.4	0.361	0.209
1-1-3-3-2	1	0.4	0.357	0.167
1-2-1-1-1	1	0.4	0.749	0.842
1-2-1-2-2	1	0.4	0.638	0.691
1-2-2-1-2	1	0.4	0.642	0.709
1-2-2-2-2	1	0.4	0.583	0.620
1-2-3-3-3	1	0.4	0.205	−0.049
2-1-1-1-2	1	0.4	0.709	0.808
2-1-1-2-1	1	0.4	0.701	0.781
2-1-1-2-3	1	0.4	0.386	0.346
2-1-2-1-2	1	0.4	0.654	0.737
2-1-2-3-1	1	0.4	0.374	0.247
2-1-2-3-2	1	0.4	0.323	0.185
2-2-1-2-1	1	0.4	0.599	0.647
2-2-2-1-1	1	0.4	0.604	0.665
2-2-2-3-1	1	0.4	0.273	0.113
2-2-3-2-3	1	0.4	0.175	0.017
2-3-3-3-3	1	0.4	0.014	-0.330
Total	246	100.0		
Median			0.789	0.887
Mean			0.760	0.790
SD			0.226	0.272
Skewness			−0.885	−1.778
Kurtosis			0.363	2.779

u_{VAS} EQ-5D utility in visual analogue scale, u_{TTO} EQ-5D utility in time trade-off scale, *EQ-5D* EuroQol-5 dimensions, *VAS* visual analogue scale, *TTO* Time Trade Off

Table 4 Estimated regression weights (B) and significance (p) for model 1, using aggregated level data and generalized linear modeling

Item[a]	Dimension levels	Estimates					
		Asymptotic			Bootstrap[b]		
		B	SE	*p*	Bias	SE	*p*
A1	URGE2	**0.102**	0.041	0.013	0.005	0.065	0.043
	URGE3	**0.109**	0.041	0.008	0.005	0.065	0.031
	URGE4	**0.134**	0.041	0.001	0.006	0.064	0.010
	URGE5	**0.216**	0.045	0.000	0.004	0.066	0.001
A3	URINE2	**0.070**	0.019	0.000	−0.003	0.020	0.002
	URINE3	**0.078**	0.021	0.000	−0.003	0.021	0.001
	URINE4	**0.082**	0.023	0.001	−0.002	0.023	0.001
	URINE5	**0.171**	0.028	0.000	−0.005	0.037	0.001
B3	SLEEP2	**0.071**	0.028	0.011	−0.003	0.029	0.007
	SLEEP3	0.049	0.026	0.063	0.000	0.027	0.036
	SLEEP4	**0.065**	0.026	0.013	0.000	0.028	0.009
	SLEEP5	**0.078**	0.034	0.021	−0.004	0.050	0.047
B1	COPING2	0.076	0.051	0.136	0.001	0.090	0.190
	COPING3	0.067	0.047	0.160	0.002	0.087	0.215
	COPING4	**0.094**	0.047	0.049	0.000	0.089	0.125
	COPING5	**0.136**	0.055	0.014	−0.001	0.091	0.045
B2	CONCERN2	**−0.132**	0.056	0.019	0.000	0.090	0.040
	CONCERN3	−0.093	0.054	0.088	−0.003	0.089	0.102
	CONCERN4	−0.105	0.054	0.053	−0.003	0.089	0.076
	CONCERN5	−0.028	0.055	0.611	−0.006	0.091	0.354
	N2	−0.065	0.044	0.140	0.009	0.052	0.098

Bold: $p \leq 0.05$

[a] A = items from Symptom Bother Scale, B = items from health-related quality of life scale

[b] Based on 973 samples

significant but some levels needed to be combined to preserve monotonicity.

After estimating competing models by creating all possible item combinations, extracting items from each conceptual attribute/dimension, the best fitting model for the Spanish sample differed in some of the items used to measure each dimension. Item A-6 had to be replaced with item A-3 for the urine attribute, and item B-12 had to be replaced with item B-2 for the concern attribute. Item levels were collapsed as suggested in the UK model.

At the individual level, urge (levels 4 and 5) and urine (level 5) attributes were significant both using VAS and TTO approaches. Using aggregated data, attributes urge, urine and sleep were statistically significant using both VAS and TTO scorings; coping was significant at levels 4 and 5 (VAS) and concern at level 2 (VAS and TTO), and close to significance at other levels (Table 4). Positive regression coefficients indicated that scores in that level of a given dimension predict higher levels of impairment. Nevertheless, negative sign was obtained for concern regression estimates (as it had happened at the individual level) and a non-monotonic pattern was found for the ordered severity levels of this dimension.

Adjusted R^2 for the replicated baseline model indicate that urge and urine were able to explain 58 % of disutility variance when using VAS scoring at the individual level; urge explained 45 % of the variability of the EQ-5D scores when using TTO. At the aggregated level, urge, urine, sleep and concern explained 88 % of variability of EQ-5D scores using VAS scoring, while they explained 78 % of variance when using TTO scoring. It was not possible to compare adjusted R^2 with the original ones since the author did not report this GOF index. The size of MAE index indicated an adequate fit at both individual and aggregated level (0.061).

Regarding the best-fitting model, at the individual level, urge and urine attributes explained 60 % (VAS) and 48 % (TTO) of the variability when mapping the scales. At the aggregated level, all five OAB-5D dimensions explained 89 % (VAS) and 80 % (TTO) of the variability, showing an important increment with respect to the individual level. MAE values did not improve at the aggregated level (see

Table 5 Goodness of fit statistics for the replicated and the best-fitting models using VAS utility scoring

Fitting model	Best-fitting	
	Individual level	Aggregated level
Ordinary least squares models		
Adjusted R^2	0.605.	0.892
Mean absolute error	0.158	0.158
Mean absolute percentage Error	148.4 %	93.5 %
Root mean residual	0.207	0.207
Generalized linear models		
χ^2	123.823	15.309
χ^2/df	0.571	0.069
Likelihood ratio χ^{2a}	275.779*	743.652*
Akaike information criterion	−124.431	−550.914
Bayesian information criterion	−44.472	−475.381
Mean absolute error	0.163	0.160
Mean absolute percentage error	160.2 %	149.9 %
Root mean residual	0.220	0.212
Tobit models		
Akaike information criterion (AIC)	204.992	−467.559
Bayesian information criterion (BIC)	285.143	−386.936
Mean absolute error	0.224	0.158
Mean absolute percentage error	186.6 %	93.5 %
Root mean residual	0.299	0.207

df degrees of freedom, *VAS* visual analogue scale

* $p \leq 0.001$

[a] Comparing against null model

Table 5). Adding a small constant ($c = 0.05$, equal to the lower limit of the following observed disutility) to 0-value disutilities, to avoid division by 0, MAPE = 148.4 % and RMSE = 0.207, while excluding 0-value disutilities, MAPE = 37.8 % and RMSE = 0.196.

Results obtained using the generalized linear models did not improve estimation notably. Estimating the replicated model with individual data only urgency, at its highest level 5, and incontinence, at the intermediate level 3, attained significant regression coefficients. The signs of coefficients were negative for some levels of the different dimensions, and a non-monotonic increase in coefficients was observed (results not shown). Estimating the replicated model with aggregated data, incontinence and sleep were significant at all levels, urge at the highest level and concern at the three lowest levels. The coefficient signs were negative for all levels of concern (results not shown). GOF statistics were more favorable for the aggregated model: $\chi^2/df = 0.069$, likelihood ratio omnibus test for the set of coefficients was significant ($\chi^2 = 743.6$, $p < 0.001$), MAE = 0.160 and MAPE = 149.9 %.

Bootstrap OLS regression coefficient estimates for the best fitting model (see Table 4) were very close to the asymptotic estimates, with small parameter bias, although bootstrap standard error estimates were slightly higher. Significance levels were slightly higher for bootstrap estimates, except for some Sleep levels and for the interaction term *N2*.

Tobit models obtained worse MAE = 0.224, and MAPE = 187 % errors for individual data, while at the aggregated level errors were similar to those obtained with the OLS model.

Figure 1 represents the relation between OAB-5D rank of mean severity associated to each health state and EQ-5D disutilities using VAS scoring system. Three sets of values are represented: average disutility for each OAB-5D health state on which ranks have been computed at the individual level, average observed disutility aggregated over OAB-V8 scores, and disutility predicted by the estimated model at the aggregated level. Average mean disutility scores associated with different OAB-5D severity levels (computed by the set of items included in the model as predictors) did not follow an overall linear trend, while a curvilinear trend may be observed at the higher severity end. Observed disutilities present a consistent floor effect (censoring) below the value of 0 (corresponding to perfect health) a gap is observed between the values of 0 and 0.20, due to the EQ-5D VAS scoring system. Above that value

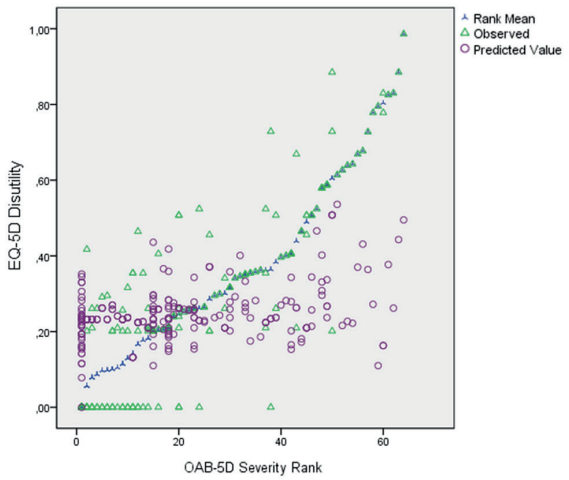


Fig. 1 OAB-5D severity rank of health states versus EQ-5D VAS disutility values computed as: average rank disutility, observed average by OAB-V8 scores, and model predicted scores. *OAB-5D* overactive bladder questionnaire 5-dimensional health classification system, *EQ-5D VAS* EuroQol-5 dimensions visual analogue scale, *OAB-V8* overactive bladder questionnaire-V8

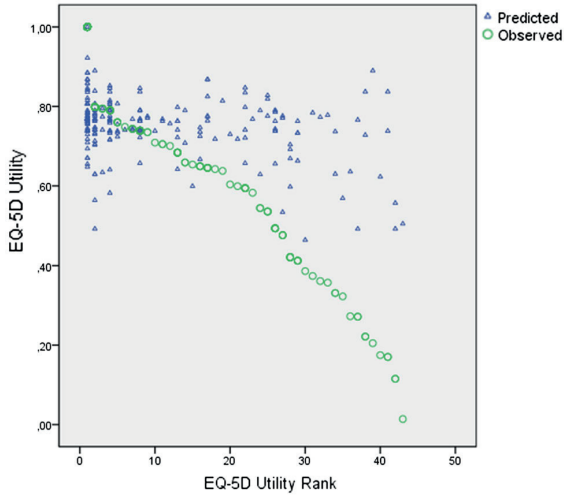


Fig. 2 EQ-5D severity rank of health states versus observed and predicted EQ-5D VAS utility values. *EQ-5D* EuroQol-5 dimensions, *EQ-5D VAS* EuroQol-5 dimensions visual analogue scale

most observed scores match with the OAB-5D aggregated score. Predicted scores present high variability, especially at the lower end of severity, where disutilities are overestimated, while for more severe conditions, disutilities tend to be underestimated (Fig. 2).

Results from Confirmatory Factor Analysis suggested that no differences between genders existed with respect the measurement instrument. Strong measurement invariance could be assumed with no loss of fit with respect the configural invariance ($\Delta\chi^2 = 9.1$, $df = 6$, $p = 0.168$). Underlying factor means were not significantly different ($z < |0.709|$, $p > 0.239$). No significant differences were found between genders with respect to EQ-5D utility mean scores ($t < |0.993|$, $p > 0.324$) nor with respect to OABq-SF scores ($t < |0.581|$, $p > 0.562$).

4 Discussion

In the present study, a series of mapping models have been assessed, aimed towards converting a specific quality-of-life instrument (which evaluates the interference and severity of the OAB Syndrome) into the metric of a generic preference instrument (EQ-5D). The best possible model for estimating utility values, based on health states defined by the OABq-SF, was obtained. The EQ-5D was selected because it is a widely used instrument, which ensures the possibility of comparison with other studies. Mapping OAB-5D onto this generic instrument should preserve the specificity of the disease, and allow computing preference-based utilities and derived measures, such as QALYs.

Three estimation procedures were compared; ordinary least squares with the general linear models, maximum likelihood with the generalized linear and Tobit models, showing similar results in terms of significant dimensions, significant levels and GOF. Models were estimated using individual level data and data aggregated based on mean OAB severity, in order to avoid the influence of outlier health states not related to OAB condition. All possible models produced by combining items from each OAB-5D dimension (urgency, urine loss, sleep, concern and coping) were assessed, and the one with best interpretability and fit was selected.

Replicating the model proposed by the developers of the OAB-5D system, OLS estimates obtained in the Spanish sample were systematically smaller than those obtained in the Anglo-Saxon sample. Only urgency, urine and sleep were statistically significant in our model. Since sample size used in the reference study was not substantially higher ($n = 312$) than the one obtained with our sample ($n = 246$), it could be the case that generic health states valuations presented lower variability for the different specific health states in the reference study. Additionally, the coping dimension did not result significant in the majority of the estimated models in the Spanish culture. This finding could be explained by the fact that the generic system EQ-5D pays little attention to psychological aspects of deterioration (whether it is cognitive or affective) so that it is hard to encounter an anchor for this dimension with the generic instrument. To avoid this problem some authors are proposing nowadays to bolt-on additional dimensions

into generic instruments to increase sensitivity to specific symptomatology [35].

The best fitting model obtained significant coefficients for most severity levels in all dimensions, using aggregated data. In this model, urgency and incontinence obtained significant coefficients for all levels, sleep and coping for the higher levels, and concern for the intermediate level 3. Nevertheless, this model differed from the replicated model in the items used to measure the urine and concern dimensions. Adjusted R^2 model fit was increased approximately by 47 % when data were aggregated by OAB severity, using average of disutility. We should be cautious with the interpretation of adjusted R^2, given that it does not necessarily indicates a better fit, since the number of data points is reduced with aggregation. Mean residuals for both levels of data were equal, and MAE did not improve with data aggregation. MAPE improved from 148 % at individual level to 94 % at the aggregated level.

In the case of generalized linear models, the dimensions urine and sleep resulted significant at all levels, except in one of the proposed models. Concern was significant at the intermediate and high levels in the reference model and at levels 2 and 4 in the proposed model. Urgency was significant at the higher level of severity (5) in the proposed models. Coping was not significant in any of the models proposed, neither with individual level data nor with aggregated data. As it was the case using OLS estimation, the fit of generalized models improves with aggregation; χ^2 diminishes notoriously, as well as the AIC and BIC values. Root mean residual (RMR), MAE and MAPE were worse with GLM estimates.

The best set of items for mapping OABq-SF scores is defined by the following OAB-5D combination (Model 1): D_1 = urgency (Item A1: "Having annoying desire to urinate"); D_2 = urine (Item A3: "Having accidental loss of small amount of urine"); D_3 = sleep (Item B3: "the symptoms interfered with your capability to sleep at night"); D_4 = coping (Item B1: "The symptoms made you plan escape routes to the bathroom in public places"); D_5 = concern (Item B2: "The symptoms made you feel that something was not right"). Utility values should be computed applying the set of coefficients given in Table 4. For instance, a patient with the following set of OABq-SF scores: A1 = 4, A3 = 2, B3 = 5, B1 = 4, B2 = 3, will obtain an estimated utility score of $u = 1 - (0.134 + 0.070 + 0.078 + 0.094 - 0.093 - 0.065) = 0.782$.

It is important to mention that OLS, GLM and Tobit estimations found some negative B coefficients for different levels of severity (2 to 5), and for both data aggregation levels, especially in the concern dimension. This behavior has been persistent even when different items were considered. In the best fitting model, concern presented negative coefficients at all levels, and higher coefficient in absolute value for the less severe level. This pattern suggests that patients may be adapting to their condition and they are little worried by their OAB condition they experience a reduction on the impact over their health state.

In relation with the interaction terms *N2*, we found that it is not an essential term, since it did not reach significance for any of the estimated models. This finding may be explained by the high presence of low disutility values close to 0 (perception of no deterioration in their own health state). Hence, the role of this "added severity" term is not needed and it does not add value to the estimation. This same result has been found in mappings performed in the UK [19] and in the USA [36]; these authors concluded that adding this term did not improve any model and therefore it was not included in their specifications. The fact that the interaction was not significant, would also confirm that we are dealing with an additive model [37, 38].

A somewhat problematic issue has been the observed distribution of utilities, which was markedly skewed, with accumulation of scores in the upper boundary score. Utilities values below 0 may be observed when using TTO scoring, as it has been previously found [19, 29], and we also found 2 % of cases with negative utility. These negative values are usually understood as an aversion to risk preference for health states perceived as worse than dead. But the definition of utility does not allow the presence of values above 1, which would mean "better than full health". The Tobit procedure allows to partial out the presence of a censoring limit in the observed scores. Unfortunately, at the aggregated level, this procedure did not improve predictions and coefficients were of similar sign and size as OLS estimates.

Some authors have proposed to include other explanatory variables like age in the mapping algorithm [18], in order to account for severity of symptoms related to concomitant health conditions. This strategy would improve fit, and it would explain why patients with no OAB symptoms obtain high disutility values, but it would also blur the ability of the OAB-5D system to account for health disability.

Our research was based on two separate motivations. On one hand we needed to develop a scoring algorithm suitable for the Spanish population, especially when one of the items included in the UK scoring system was knows to have cultural problems. On the other hand, we desired to study the benefits of different statistical models, in order to address which one would be more suitable to account for non-linearity and censoring effects. The disease we are interested in has presented the inconvenience to be a non-severe health condition, while patients suffering from it are usually of advanced age. The first issue entails that the specific instrument might not cover all the range of

disutilities covered by a generic instrument, avoiding a better fit. The latter issue may have introduced an excess of variability in our data, especially at the lower end of the OAB-5D scores. Nevertheless, we have been able to obtain the best-fitting model system which also provides a meaningful scoring system for computing utility values. One main limitation in our work is the fact that we are mapping a self-reported quality-of-life instrument onto a general population preference instrument. It is known that differences exist between these two approaches for valuating health, particularly when patients are able to adapt to their condition and under-estimate severity or impairment.

5 Conclusion

The objective of estimating models with which mapping can be performed has been achieved, although it cannot be affirmed that there is total overlap of the functions in the psychological variables of these patients; however, it can be concluded that there is association between the specific instrument of the disease (OABq-SF) and the generic instrument based on preferences (EQ-5D). This finding resembles previous findings in the mapping of specific instruments (OAB and insomnia) to generic instruments. In the same manner, it is relevant to consider what is the purpose of the results of the mapping, taking into account the differences in the application of these in the different scopes (psychological, medical, economic). In addition to consider the degree of precision of a model, the substantive interpretation must be addressed and the reach of the results in a specific sample.

Acknowledgments Authors thank the participants in the AQUA study who were responsible for collection and recording of the data used in this research. We also thank Pfizer, SLU for funding English edition of manuscript.

Compliance with Ethical Standards

Funding No funding has been received for the conduct of this study.

Disclosure of conflict of interests We used the database of an original study funding by Pfizer, SLU, Alcobendas (Madrid), Spain. However, the work included in the present manuscript has not received any funding and it is solely the work of the authors, except for English editing of the manuscript that was performed by NOVA-trasnet and was funded by Pfizer, SLU. Javier Rejas is full employee at Pfizer, SLU. Miguel A. Ruiz, Laura L. Gutiérrez and Manuel Monroy declare that they do not have any conflict of interest as a result of participating in this research.

Authors' contributions All authors had complete access to the data, participated in the analysis and/or interpretation of results, drafted and approved the content of the manuscript. MAR and JR were responsible of the design and idea of this work, and for the interpretation of data and results. MM and LLG performed the statistical analysis, participated in data interpretation and drafted part of the manuscript. MAR and JR were responsible for literature review, extraction of references and drafted the final version of manuscript.

References

1. Abrams P, Cardozo L, Fall M, Griffiths D, Rosier P, Ulmsten U, et al. The standardization of terminology of lower urinary tract function: report from the standardization sub-committee of the International Continence. Neurourol Urodyn. 2002;21:167–78.
2. Stewart WF, van Rooyen JB, Cundiff JW, Abrams P, Herzog AR, Corey R, et al. Prevalence and burden of overactive bladder in the US. World J Urol. 2003;20:327–36.
3. Stewart W, Minassian VA, Hirsch AG, Kolodner K, Fitzgerald M, Burgio K, et al. Predictors of variability in urinary incontinence and overactive bladder symptoms. Neurorol Urodyn. 2010;29:328–35.
4. Milsom I, Abrams P, Cardozo L, Roberts RG, Thuroff J, Wein AJ. How widespread are the symptoms of an overactive bladder and how are they managed? A population-based prevalence study. BJU Int. 2001;87:760–6.
5. Coyne KS, Zyczynski T, Margolis MK, Elinoff V, Roberts RG. Validation of an overactive bladder awareness tool for use in primary care settings. Adv Ther. 2005;22:381–94.
6. Ricci JA, Baggish JS, Hunt TL, Stewart WF, Wein A, Herzog AR, et al. Coping strategies and health care-seeking behavior in a US national sample of adults with symptoms suggestive of overactive bladder. Clin Ther. 2001;23:1245–59.
7. Berges R, Pientka L, Hofner K, Senge T, Jonas U. Male lower urinary tract symptoms and related health care seeking in Germany. Eur Urol. 2001;39:682–7.
8. Coyne KS, Sexton CC, Kopp ZS, Ebel-Bitoun C, Milson I, Chapple C. The impact of overactive bladder on mental health, work productivity and health-related quality of life in the UK and Sweden: results from EpiLUTS. BJU Int. 2011;108:1459–71.
9. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2010;11:215–25.
10. Versteegh MM, Leunis A, Luime JJ, Boggild M, Uyl-de Groot CA, Stolk EA. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. Med Decis Mak. 2012;32:554–68.
11. Kim EJ, Ko SK, Kang HY. Mapping the cancer-specific EORTC QLQ-C30 and EORTC QLQ-BR23 to the generic EQ-5D in metastatic breast cancer patients. Qual Life Res. 2012;21:1193–203.
12. Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. Eur J Health Econ. 2010;11:427–34.
13. Cheung YB, Tan LCS, Lau PN, Au WL, Luo N. Mapping the eight-item Parkinson's Disease Questionnaire (PDQ-8) to the EQ-5D utility index. Qual Life Res. 2008;17:1173–81.
14. Gu N, Botterman M, Ji X, Bell C, Carter J, van Hout B. Mapping of the insomnia severity index and other sleep measures to EuroQol EQ-5D health state utilities. HQLO. 2011;9:1–34.
15. Buxton MJ, Lacey LA, Feagan BG, Oliver R. Mapping from disease-specific measures to utility: an analysis of the relationship between the Inflammatory Bowel Disease Questionnaire and Crohan's Disease Activity Index in Crohn's disease and measures of utility. Value Health. 2007;10:214–20.
16. Grootendorst P, Marshall D, Pericak D, Bellamy N, Feeny D, Torrance GW. A model to estimate Health Utilities Index Mark 3

utility scores form WOMAC Index scores in the patients with osteoarthritis of the knee. J Rheumatol. 2007;34:534–42.
17. Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, et al. The Beaver dam health outcomes survey: initial catalog of health-state quality factors. Med Decis Mak. 1993;13:89–102.
18. Kay S, Tolley K, Colayco D, Khalaf K, Anderson P, Globe D. Mapping EQ-5D utility scores from the incontinence quality of life questionnaire among patients with neurogenic and idiopathic overactive bladder. Value Health. 2013;16:394–402.
19. Yang Y, Brazier J, Tsuchiya A, Coyne K. Estimating a Preference-Based Single Index from the Overactive Bladder Questionnaire. Value Health. 2009;12:159–66.
20. Coyne KS, Lai JS, Zyczynski T, Kopp Z, Avery K, Abrams P. An overactive bladder symptom and quality-of-life short form: development of the overactive bladder questionnaire short form (OAB-q SF). 34th Joint Meeting of the International Continence Society and the International UroGynecological Association, August 23–27, Paris, France; 2004.
21. Young T, Yang Y, Brazier JE, Tsuchiya A, Coyne K. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Qual Life Res. 2009;18:253–65.
22. Studenski S, Hayes RP, Leibowitz RQ, Bode R, Lavery L, Walston J, et al. Clinical global impression of change in physical frailty: development of a measure based on clinical judgment. JAGS. 2004;52:1560–6.
23. Bech P. Rating scales for psychopathology, health status, and quality of life. Berlin: Springer-Verlag; 1993. p. 33–5.
24. Morisky DE, Green LW, Levine DM. Concurrent and predictive validity of a self-reported measure of medication adherence. Med Care. 1986;24:67–74.
25. Andersson KE. Antimuscarinic mechanisms and the overactive detrusor: an update. Eur Urol. 2011;59:377–86.
26. Coyne K, Revicki D, Hunt T, Corey R, Stewart W. Psychometric validation of an overactive bladder symptom and health-related quality of life questionnaire: the OAB-q. Qual Life Res. 2002;11:563–74.
27. Arlandis S, Ruíz M, Errando C, Villacampa F, Arumí D, Lizarraga I, et al. Quality of life in patients with overactive bladder: validation and psychometric properties of the Spanish Overactive Bladder Questionnaire-short Form. Clin Drug Investig. 2012;32:523–32.
28. EuroQol, Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy. 1990;16:199–208.
29. Badia X, Roset M, Montserrat S, Herdman M, Segura A. La versión española del EuroQol: descripción y aplicaciones. Med Clin (Barc). 1999;112(Supl 1):79–86.
30. Stolk E, Oppe M, Scalone L, Krabbe P. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health. 2010;13:1005–13.
31. Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. Soc Sci Med. 1997;44:1519–30.
32. Neal D, Simons J. Inference in regression models of heavily skewed alcohol use data: a comparison of ordinary least squares, generalized linear models and bootstrap resampling. Psychol Addict Behav. 2007;21:441–52.
33. McDonald JF, Moffitt RA. The uses of tobit analysis. Rev Econ Stat. 1980;62:318–21.
34. Fox J. Bootstrapping regression models: appendix to An R and S-PLUS companion to applied regression. Thousand Oaks: Sage; 2002.
35. Yang Y, Rowen D, Brazier J, Tsuchiya A, Young T, Longworth L. An exploratory study to test the impact on three "Bolt-On" items to the EQ-5D. Value Health. 2014;18:52–60.
36. Shaw J, Simon A, Yu S, Chen S, Iannachione V, Johnson J, et al. A median model for predicting United States population-based EQ-5D health state preferences. Value Health. 2010;13:278–88.
37. Fox J. Applied regression analysis and generalized linear models. London: Sage; 2008.
38. Furlong W, Feeney D, Torrance GW, Goldsmith S, De Pauw Z, Zhu Z, et al. Multiplicative multi-attribute utility function for health utilities index mark 3 (HUI3) system: a technical report. McMaster University, Centre for Health Economics and Policy Analysis;1998. p. 98–11.

Monroy et al. Health and Quality of Life Outcomes (2018) 16:177
https://doi.org/10.1186/s12955-018-1003-y

Health and Quality of Life Outcomes

RESEARCH Open Access

Mapping of the Gastrointestinal Short Form Questionnaire (GSF-Q) into EQ-5D-3L and SF-6D in patients with gastroesophageal reflux disease

Manuel Monroy[1], Miguel A. Ruiz[1*], Javier Rejas[2] and Javier Soto[2]

Abstract

Background: The short, self-administered Gastroesophageal Reflux Disease (GERD) Symptom Frequency Questionnaire (GSFQ) is a specific Quality of Life (QoL) instrument which measures the impact of GERD symptoms on QoL. This study aims to map the specific scores in GSFQ into two generic instruments: SF-6D and EQ-5D-3 L, in order to obtain utility estimates derived from the GERD condition.

Method: A national representative sample of GERD patients was selected, stratified by gender, age (< 45, ≥45 years) and GERD severity (0-I, II-IV Savary-Miller score) for validation purposes. Age, gender, BMI, GERD diagnose, GERD severity, associated comorbidities and risk factors were recorded. GSFQ, SF-6D, EQ-5D-3 L, and the HRQoL Visual Analogue Scale (VAS) were answered by patients. Several mapping methods were estimated, regression using dummy variables, and linear, quadratic and cubic regression using optimal factor scores. The use of a GERD aggregated summary severity derived from the GSFQ was dimed the best predictor. Overall Mean Absolute Error (MAE), overall Mean Absolute Percentage Error (MAPE) were used as goodness-of-fit (GOF) indexes to compare models.

Results: A total of 3405 patients were recruited by 490 clinicians. Mean age was 49 (±14.4) years and 49.8% were women. Reported comorbidities were clustered in 6 antecedents and 15 concomitant pathologies. Aggregation of levels for the frequency of symptoms items was found more suitable for estimation. Regression weights were found to follow a monotonous progressive pattern. Overall MAE ranged from 0.092 to 0.094 for SF-6D utility prediction and from 0.008 to 0.08 for EQ-5D-3 L, while MAPE values ranged from 27.9 to 29% for SF-6D and from 36.8 to 38.4% for EQ-5D-3 L. Cubic regression GOF demonstrated a better fit.

Conclusions: It is possible to translate specific GSFQ scores assessing GERD condition into generic SF-6D and EQ-5D-3 L utility values. Although regression using dummy variables is a suitable mapping procedure, other alternative mapping methods convey better fit, in particular cubic regression.

Background

Gastroesophageal Reflux Disease (GERD) appears when stomach contents flux back to the esophagus. It happens when the valve located between the esophagus and the stomach does not close properly. Most frequent disease symptoms are acidity and acid reflux. Other less frequent but associated symptoms are heartburn without clear motive, panting, throat ache and cough, among others [1, 2].

GERD can be classified into four severity levels, ranging from the appearance of edema and erythema, causing some degree of esophagus erosion, up to esophageal ulcers or Barret's esophagus. Consumption of alcohol or carbonated drinks, obesity and smoking are known to be GERD risk factors [3].

According to the DIGEST international study, approximately 7.7% of the population suffers from GERD [4]. Attending to the current consensual definition: "GERD should be used to include all individuals who are

* Correspondence: Miguel.ruiz@uam.es
[1]Faculty of Psychology, Universidad Autónoma de Madrid, C/ Ivan Pavlov 6, 28049 Madrid, Spain
Full list of author information is available at the end of the article

exposed to the risk of physical complications from gastro-esophageal reflux, or who experience clinically significant impairment of health related well-being (quality of life) due to reflux related symptoms, after adequate reassurance of the benign nature of their symptoms" [5]. Furthermore, it is commonly accepted that self-reporting is one of the main sources of diagnosis [6] and patients should report experiencing symptoms at least twice a week [2, 7] for a diagnosis of GERD.

It is important to remark that the impairments caused by GERD symptoms are highly variable and may affect quality of life even when there are no endoscopic findings [2]. Patients tend to adopt eating behaviors in order to prevent or attenuate their clinical situation. The Agency for Healthcare Research and QALY reports that the more frequent treatments are antacids (neutralizing stomach acids) and type 2 histamine receptor antagonists (H_2RA) or proton pump inhibitors (PPI), both reducing the production of stomach acid [8, 9]. The impacts of GERD symptoms on patients' health-related quality of life (HRQoL) is usually ascertained by means of patient-reported-outcomes measurements (PROMs) such as the Gastrointestinal Short Form Questionnaire (GSF-Q) [10].

HRQoL measures are particularly important for GERD sufferers given their diagnostic capabilities, while they also reveal important issues to health service providers for several reasons. First, HRQoL has been shown to have a direct relation with mortality, hospitalization and consumption of clinical resources. Second, it has been shown to have a low to moderate relation with other disease-specific indicators, hence contributing complementary information for assessing clinical impairment [11]. Presently, HRQoL has been identified as a clinical target in itself, both in patients with limited life expectancy and for therapies directed towards disease coping or symptom accommodation, as much as for biological improvement (as is the case for most chronic diseases). Preference-based measures (PBMs) play a central role in these evaluations. They allow patients to describe the impact of ill health and have an associated "utility" score for each health-state description. These utility scores can then be used to calculate quality-adjusted life-years (QALYs), which is an outcome metric used in many economic evaluations of potential health benefits [12].

In the past, clinical studies did not always include a PBM. Often they included one or more of the many PROMs that are not full PBMs because they do not have an associated, preference-based scoring system. On the other hand, PROMs have proved to be very sensitive to variations in patient health conditions, and this is one of the reasons for their extended use in clinical studies. Furthermore, when a major research need is to compare result with those of other pathologies or comorbidities, it will not be possible to use disease-specific PROMs, and generic HRQoL instruments should be preferred. Most popular generic instruments (like SF-6D, EQ-5D and HUI3), offer the possibility of computing the utility score associated to each health condition (as captured by the instrument attribute profile), reflecting the population preference towards each health state in a situation of uncertainty. This peculiarity allows using them in computing QALYs and in health economics in general.

It is usually the case that a disease-specific PROM instrument will be preferred in research about a particular disorder and when the use of generic instruments has been avoided because they do not capture properly the different levels of disease symptomatology on patients' HRQoL. Also, because there is evidence suggesting that generic measurements might have poor sensitivity to change in some health conditions, such as GERD or others non-threatening illnesses, or are incapable of discriminating well between patients using different drugs to treat their health problems [13, 14]. In such cases, the usual strategy is to map the specific measurements into a generic instrument allowing further comparison with other studies in which the specific instruments may not be pertinent or are otherwise unavailable (e.g., retrospective databases) [15, 16].

Aligned with such an approach, since 2008, NICE's preferred measure of health-related quality of life in adults has been EQ-5D, to derive utilities set values for health economic evaluations (see Guide to the methods of technology appraisal 2013, at https://www.nice.org.uk/process/pmg9/chapter/foreword.

The aim of the present study was to obtain the mapping algorithms needed for translating the specific HRQoL measure obtained by the GSF-Q into two of the most popular preference-based generic instruments, the SF-6D and the EQ-5D-3 L. As a secondary benefit, we will be able to assess which one of the generic instruments is more suitable for capturing HRQoL deterioration due to GERD conditions.

Methods

Study design

The present study is a secondary analysis carried out using the data gathered for the cultural validation of the GSF-Q into Spanish [17]. The original study was developed to ensure adequate estimation of psychometric properties, and was designed as an observational study that would provide a rich data set, not only for instrument validation but particularly for mapping studies, beyond what could be obtained in controlled clinical trials. This was a cross-sectional, single time point assessment design. The original sample design was thought to ensure representativeness of three strata: gender, age (< 45, ≥45 years) and symptom severity (Savary-Miller: 0-I, ≥II). Patients

were selected at random by demand of attention and covering each sample stratum. Scales were administered in a single visit. Patients were over 18 years of age, able to read Spanish, and signed an informed consent form. The Ethics Committee of one of the participating centers in the validation study was responsible for approving the study design. Clinicians were recruited at random and proportionally on the geographical extension and service demand in the Spanish Autonomous Communities. The study recruited the participation of 510 gastroenterologists, and they were requested to provide 4 to 8 subjects each. Additional data on the study design may be found elsewhere [17].

Participants

The final sample was composed by 3405 patients, from whom 2251 completed all the questionnaires, sociodemographic and clinical data. Half of the participants were women (49.8%), 63.9% were obese, 40.1% smokers, 42.8% consumed alcohol, and 46.5% consumed carbonated beverages. GERD was diagnosed in 80% of cases, 46.3% were under IBP treatment, 16.5% used H_2RA, and 25.3% used antacids. It should be mentioned that 48.4% were on treatment for at least one other comorbidity (Table 1). All patients had signed informed consent forms, and the Helsinki declaration guidelines were met.

Table 1 Sample sociodemographic and clinical descriptors

Variable	Level	Frequency	Percent
Age (decades)	18–30	147	6.5
	31–40	392	17.4
	41–50	510	22.7
	51–60	529	23.5
	61–70	431	19.1
	71–80	187	8.3
	> 80	55	2.4
Gender	Male	1131	50.2
	Female	1120	49.8
Smoking	Yes	903	40.1
	No	1348	59.9
Alcohol	Yes	963	42.8
	No	1288	57.2
antiH2	Yes	348	15.5
	No	1903	84.5
Treated for comorbidities	Yes	1161	51.6
	No	1090	48.4
GERD Level	0	396	17.6
	1	521	23.1
	2	583	25.9
	3	220	9.8
	4	102	4.5
	Unknown	429	19.1
Body Mass Index	Infra-weight	17	.8
	Normal	796	35.4
	Over-weight	1438	63.9
Carbonated Drinks	Yes	1047	46.5
	No	1204	53.5
IBP	Yes	965	42.9
	No	1286	57.1
Antacid	Yes	556	24.7
	No	1695	75.3

Instruments

Three questionnaires were used to measure HRQoL, the 2 most popular generic ones and a GERD specific instrument.

The *Gastrointestinal Short Form Questionnaire* (GSF-Q) [6, 7], was used to measure GERD symptom impact on HRQoL. The questionnaire is composed of six items, plus 2 filter items. The first four gauge the impact of GERD symptoms during the most recent week (upper abdomen pain, breastbone pain, limited eating, heartburn) using a 5-point Likert scale (0 = Never, 4 = All of the time). The last two inquire about the number of days per week with daytime or nighttime disturbances (0–7 days). The total score is obtained by adding up individual item scores, and it is customary to rescale it into a 0–100 severity scale. A higher score represents a higher impact on HRQoL and scores are usually interpreted by comparison with population norms [17].

EuroQol-5 Dimension-3 Levels (EQ-5D-3L) [18, 19] is a generic, preference-based HRQoL instrument. It gathers the level of deterioration for 5 attributes: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression; using 3-level items (1 = none, 2 = some problems, 3 = a lot of problems). Each combination of levels creates a health profile, with a total of 243 possible health states, although not all of them are equally likely. Profile [11111] corresponds to perfect health and profile [33333] represents the worse possible health state. Based on population preference ranking, health states are translated to a social utility value using a multi-attribute utility function (MAUF). Different MAUFs are used for different countries, mainly using estimates based on Time Trade-Off (TTO) and Visual Analogue Scale (VAS) methods [20]. The basic form of the EQ-5D-3 L MAUF is:

$$u_i = 1 - \left(q + \sum_{j=1}^{j=5} \sum_{k=1}^{k=3} b_{jk} D_{ijk} + b_{N3} N3_i \right)$$

Where the utility/preference value for health state i (u_i) is obtained by subtracting from 1 the health state disutility ($\overline{u}_i$). Disutility is obtained by weighting (b_{jk}) the

deterioration level k attained in dimension D_j, plus an interaction term ($N3_i$), which adds a constant b_{N3} when any of the dimensions reaches its maximum deterioration level, plus a constant (q). It should be noted that $b_{j1} = 0$ for the first level of any dimension ($k = 1$), which represents no deterioration in that dimension [21].

The *Medical Outcomes Survey Sort Form-6 Dimension* (SF-6D) [22, 23] is a generic, preference-based HRQoL instrument derived from the 36-item MOS SF-36 [24]. It gathers the level of deterioration for 6 dimensions: physical functioning, role limitations, social functioning, pain, mental health, and vitality; using a recoding of 11 specific items into 4 to 6 levels. A total of 18,000 health profiles are possible, with the profile [111111] corresponding to perfect health and [645655] representing the worse possible health state. Different MAUFs have been estimated for deriving preference utilities in different countries, with the peculiarity that no severity (interaction) constant is used. As in the previous case, a value of 0 is assigned to the first level for each dimension/attribute.

Statistical analyses

The first step consisted in checking the unidimensionality of GSF-Q items and, if met, deriving an overall severity index due to GERD condition. This severity index will be used to short generic health states (EQ-5D-3 L or SF-6D) when their corresponding profiles differ only in the permutation of one severity level, e.g.: [11112] vs. [11121]. A first approach was to estimate a unidimensional latent variable model assuming the latent variable to be continuous and items/indicators to be ordinal while using the WLSMV estimation method. A second approach was to decompose each k-categories item into a series of k dummy variables (0 = No, 1 = Yes) and coding lower level dummy categories as fulfilled (1) when a particular item-level was reached. A Partial Credit model [25] (an extension of the Rasch model) using ML estimation was obtained. In this way, estimated category thresholds could be compared across items and monotonic distribution of item step thresholds could be checked. Observed EQ-5D and SF-6D utility mean scores were compared using standard t-test and using bootstrap estimates in order to avoid the influence of skewness and extreme utility values.

Once a summary GERD-specific severity index was obtained, this index was mapped onto each of two utility values (separately), and several models were tested (see below) in order to predict the utility value associated to each GERD severity condition.

Disutility values ($d_i = 1\text{-}u_i$) were modeled, instead of utility values, for several reasons. First, the data-mass usually concentrates around more lenient health states, and low disutilities will fall closer to the axis origin. Second, it is always possible to estimate a model without the intercept term, anchoring 0 value disutilities (perfect health) at the 0 GERD severity value. Since GERD is not necessarily a disabling condition, and in order to attenuate the impact of possible comorbidities in the disutility value for each individual, disutilities were aggregated, using the mean value, by GERD severity, before modelling.

The following regression models estimated linear, quadratic and cubic trends, using density function values, and Tobit and Probit, using cumulative distribution values. The following covariates were tested for inclusion: Age (decades), BMI (low, normal and overweight), GERD diagnosis (Yes), smoking, alcohol consumption, carbonated drinks consumption, IBP treatment, H_2RA treatment, antacid treatment, and treatment for comorbidities. In order to anchor the best possible health states in both instruments, the GERD severity factor scores were rescaled into the range 0–1, and regression models were fit through the origin.

Along with the statistical significance of regression coefficients, model goodness-of-fit (GOF) was assessed using R^2, mean absolute error (MAE) and mean absolute percentage error (MAPE). MAE and MAPE were computed overall and by quintile group based on severity scores to assess local GOF at the different levels of severity. Bootstrap estimates for model coefficient standard errors were also obtained to avoid the influence of outlier observations in the assessment of parameter significance levels. General internationally-accepted guidelines proposed for instrument mapping were followed [13].

All analyses were conducted using the SPSS for Windows statistical software, version 22.0 and Mplus 7.

Results

GSF-Q scores ranged between 0 and 30 with a mean value of 10.54 (SD = 5.94). GERD Severity summary scores (factor scores) ranged between − 1.40 and 1.88 with a mean value of 0 (SD = 0.636) with a symmetric distribution (Skewness = 0.021, SE = 0.052).

At the individual level, SF-6D mean utility scores (M_{SF} = 0.656, SD_{SF} = 0.207) were significantly lower than EQ-5D-3 L scores (M_{EQ} = 0.744, SD_{EQ} = 0.206), both under asymptotic assumptions ($t = -27.54$, $p < 0.001$) and using 10.000 bootstrap samples: Difference 95% CI = (− 0.093, − 0.081), suggesting that slightly higher utilities were obtained with the EQ-5D. As expected, both utility scores showed a marked negative skewness, SF-6D: $Skewness_{SF} = -0.784$, $SE_{SF} = 0.052$; EQ-5D-3 L: $Skewness_{EQ} = -1.049$, $SE_{EQ} = 0.052$, with a high correlation between them ($r = 0.733$, $p < 0.001$).

The first eigenvalue of the correlation matrix was λ = 3.55 and all further eigenvalues were below 1. The confirmatory factor analysis for the 1-dimension solution (assuming variables to be ordinal) attained good

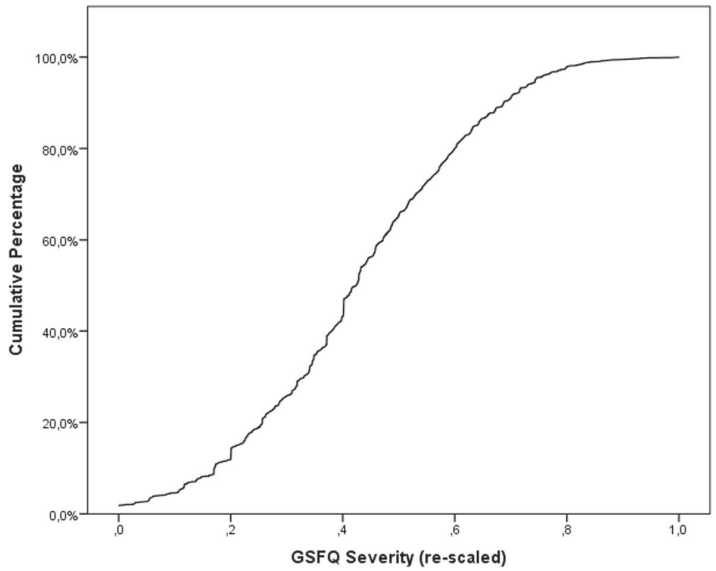


Fig. 1 GSF-Q severity score cumulative distribution

GOF indexes with CFI = 0.951 and TLI = 0.918. Figure 1 shows the cumulative distribution for rescaled factor scores, exhibiting a smooth ogive distribution with no evident changes in curvature. This figure may be used as normative data to obtain percentiles from severity scores. Figure 2 represents the response category thresholds for each item with respect to the latent normal severity score. In this figure, severity scores are expressed in standard deviations from the mean latent severity of 0 and, for each GSF-Q item, partial credit thresholds for each step rating

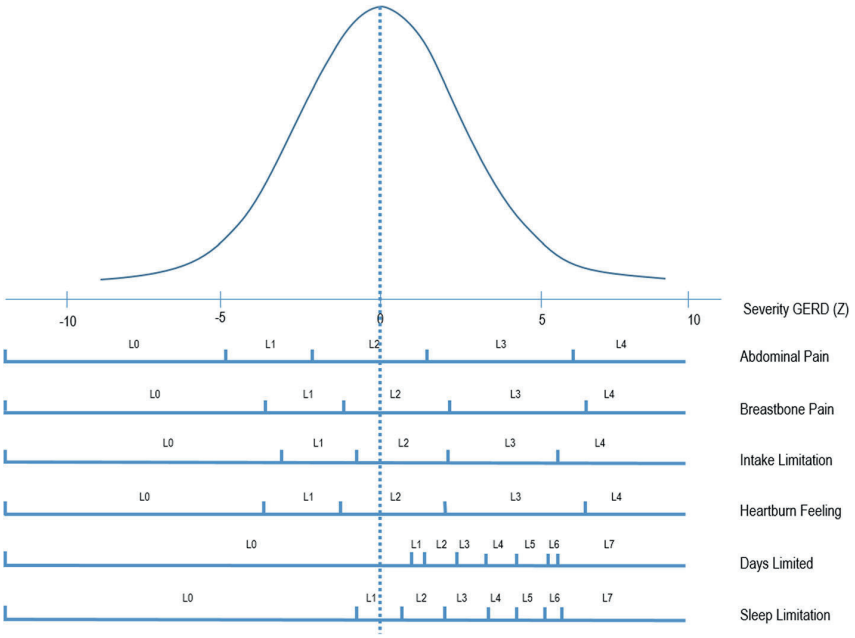


Fig. 2 GSF-Q item thresholds assuming unidimensionality (Partial Credit Model)

response are plotted, showing a rather even spread and separation of rating categories for the first four items, and a displacement of the category thresholds above the mean severity for the last two items of daytime and nighttime limitations. This later result is in accordance with the smaller weight received by the two last items in computing the factor score.

The resulting equation needed for computing re-scaled estimated factor scores from observed GSF-Q items scores may be expressed as follows:

$$\hat{f}_i = (0.183x_{1i} + 0.204x_{2i} + 0.100x_{3i} + 0.174x_{4i} + 0.047x_{5i} + 0.044x_{6i} + 1.4025) \times 0.30479$$

Where x_1 to x_4 are the scores in the first 4 GSF-Q items (0 = Never, 4 = Always), x_5 is the number of days with disability, x_6 is the number of nights with GERD problems, and 1.4025 and 0.30479 are scaling constants moving the factor scores into the 0–1 range.

EQ-5D-3 L showed to be particularly less sensitive to GERD severity. Only 78 (32%) of the 243 possible EQ-5D-3 L profiles were observed and 17 (7%) of them gathered more than 90% of patients. Table 2 shows the most frequent EQ-5D-3 L profiles observed in our sample. In the case for SF-6D utility scores, 975 (5.4%) out of the 18,000 possible health states were observed, 35 (0.2%) profiles presented a prevalence above 5/1000, gathering only 25.5% of cases.

The best fitting model for mapping GSF-Q into SF-6D disutilities was a cubic model including variables GERD severity (linear, quadratic and cubic), age (in decades), gender, BMI group (infra, normal, and over-weight), and being treated for comorbidities (see Table 3). The model GOF was good ($R^2 = 0.888$), with MAE = 0.092 and MAPE = 27.9% (Table 4) Fig. 3.

The best fitting model form mapping GSF-Q onto EQ-5D-3 L disutilities was the cubic model including GERD severity (linear, quadratic and cubic), age (in decades), gender, and being treated for comorbidities. BMI group was not significant and the following GOF statistics were obtained: $R^2 = 0.831$, MAE = 0.086 and MAPE = 37.0%.

Discussion

Specific HRQoL instruments are the preferred choice for measuring patient perceptions on their health condition

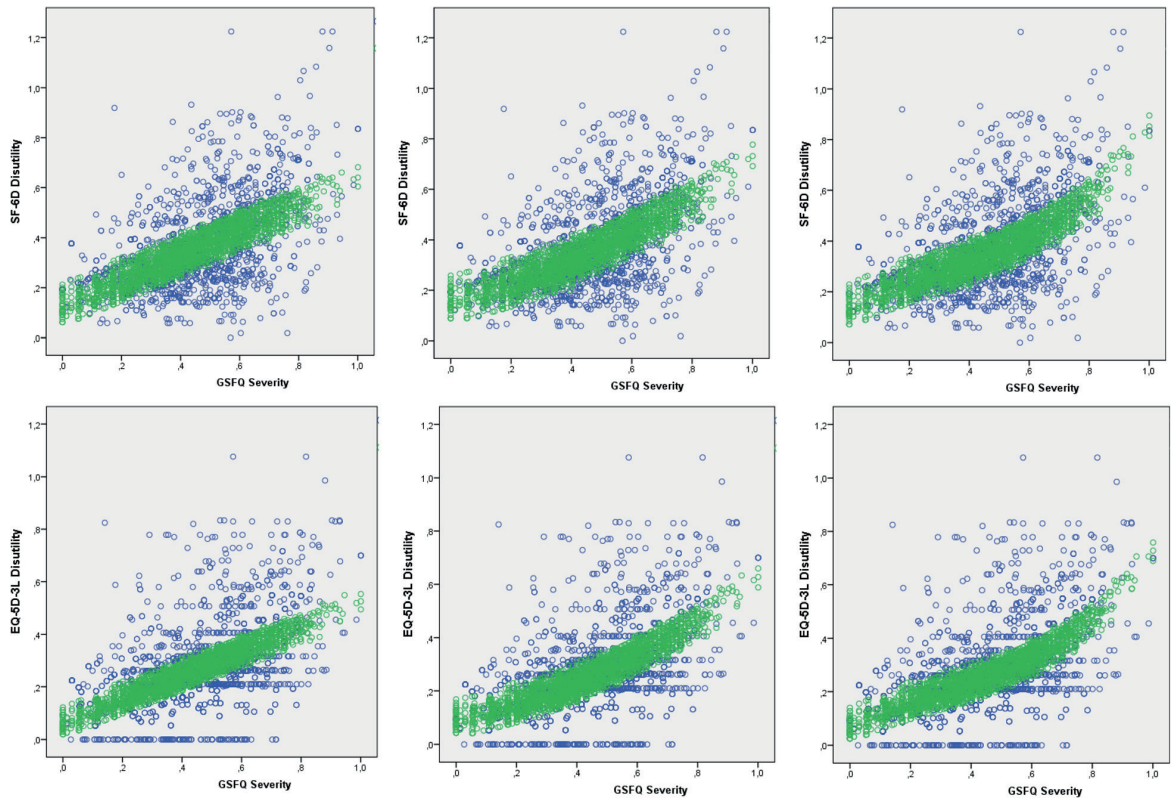


Fig. 3 SF-6D: (up) and EQ-5D-3 L (down) observed (blue) and predicted (green) utility values vs. GERD severity for the linear (left), quadratic (center) and cubic (right) models

Table 2 Most prevalent EQ-5D-3 L and SF-6D health state profiles, associated utilities, and frequencies (cases, percentages and cumulative percentages; partial listing)

Profile	Utility	Freq.	Percent	Cum. %
EQ-5D				
11111	1.00	566	25.1	25.1
11121	.79	376	16.7	41.8
11122	.74	294	13.1	54.9
11222	.68	166	7.4	62.3
21222	.59	118	5.2	67.5
11112	.80	107	4.8	72.3
11221	.74	99	4.4	76.7
22222	.49	52	2.3	79.0
21221	.65	50	2.2	81.2
21121	.70	36	1.6	82.8
21122	.65	31	1.4	84.2
11223	.42	30	1.3	85.5
22232	.22	30	1.3	86.9
22221	.54	22	1.0	87.8
11232	.41	17	.8	88.6
21232	.32	17	.8	89.3
11233	.36	16	.7	90.0
11211	.79	15	.7	90.7
11123	.48	14	.6	91.3
21233	.27	12	.5	91.9
22233	.17	12	.5	92.4
11113	.54	10	.4	92.8
21111	.76	10	.4	93.3
21223	.33	10	.4	93.7
22223	.23	10	.4	94.2
11212	.74	8	.4	94.5
12222	.58	7	.3	94.8
21112	.71	7	.3	95.2
21131	.43	7	.3	95.5
21231	.37	7	.3	95.8
11231	.46	6	.3	96.0
22332	.17	6	.3	96.3
12221	.63	5	.2	96.5
11131	.52	4	.2	96.7
12121	.69	4	.2	96.9
12233	.26	4	.2	97.1
21132	.38	4	.2	97.2
12223	.32	3	.1	97.4
21211	.71	3	.1	97.5

Table 2 Most prevalent EQ-5D-3 L and SF-6D health state profiles, associated utilities, and frequencies (cases, percentages and cumulative percentages; partial listing) *(Continued)*

Profile	Utility	Freq.	Percent	Cum. %
SF-6D				
111222	.86	52	2.3	2.3
111112	.94	30	1.3	3.6
111223	.86	30	1.3	5.0
111122	.88	29	1.3	6.3
111123	.88	22	1.0	7.2
111322	.84	19	.8	8.1
111111	1.00	18	.8	8.9
111212	.92	18	.8	9.7
111224	.79	18	.8	10.5
211224	.78	17	.8	11.2
212324	.73	17	.8	12.0
112324	.74	16	.7	12.7
211222	.84	16	.7	13.4
111323	.84	15	.7	14.1
211323	.83	15	.7	14.7
111225	.76	14	.6	15.4
111324	.78	14	.6	16.0
113424	.55	14	.6	16.6
211324	.76	14	.6	17.2
212323	.79	14	.6	17.9
212325	.69	14	.6	18.5
312323	.77	14	.6	19.1
111121	.93	13	.6	19.7
112322	.80	13	.6	20.3
112323	.80	13	.6	20.8
113324	.72	12	.5	21.4
211223	.84	12	.5	21.9
212322	.79	11	.5	22.4
111221	.92	10	.4	22.8
111325	.74	10	.4	23.3
112222	.82	10	.4	23.7
211322	.83	10	.4	24.2
212224	.74	10	.4	24.6
311324	.75	10	.4	25.1
312324	.71	10	.4	25.5
111124	.81	9	.4	25.9
112122	.84	9	.4	26.3
113323	.78	9	.4	26.7
212223	.81	9	.4	27.1

Table 3 Estimated model coefficients

Model	Predictor	SF-6D disutility				EQ-5D-3 L disutility			
		B	SE	Beta	Sig	B	SE	Beta	Sig
Linear	GSF-Q severity	.481	.014	.592	<.001	.441	.011	.688	<.001
	Age (decade)	.012	.002	.124	<.001	.010	.002	.132	<.001
	Gender (Female)	.041	.005	.077	<.001	.027	.005	.064	<.001
	BMI (Grouped)	.019	.006	.080	<.001	–	–	–	ns
	Comorbidities (Treated)	.043	.003	.132	<.001	.034	.006	.082	<.001
Quadratic	GSF-Q severity	.241	.044	.297	<.001	.195	.042	.304	<.001
	GSF-Q severity (square)	.291	.051	.214	<.001	.309	.049	.288	<.001
	Age (decade)	.013	.002	.138	<.001	.013	.002	.171	<.001
	Gender (Female)	.045	.005	.084	<.001	.032	.005	.076	<.001
	BMI (Grouped)	.031	.004	.217	<.001	.009	.004	.078	.020
	Comorbidities (Treated)	.040	.006	.075	<.001	.031	.006	.074	<.001
Cubic	GSF-Q severity	.610	.091	.751	<.001	.527	.170	.485	<.001
	GSF-Q severity (square)	−.822	.245	−.605	.001	−.665	.207	−.620	.001
	GSF-Q severity (cube)	.891	.192	.444	<.001	.768	.068	.823	<.001
	Age (decade)	.012	.002	.128	<.001	.013	.002	.167	<.001
	Gender (Female)	.043	.005	.080	<.001	.030	.005	.071	<.001
	BMI (Grouped)	.041	.006	.078	<.001	–	–	–	ns
	Comorbidities (Treated)	.023	.004	.163	<.001	.032	.006	.077	<.001

ns not significant

because of their high sensitivity to changes due to disease management and treatment suitability. However, mapping specific HRQoL into generic utility scores can present methodological problems. Albeit the good psychometric properties of instruments like GSF-Q for measuring the impact of GERD on patients' daily lives [10, 17, 26], GERD is a relatively mild health disabling disease, as compared to other possible health states measured by generic instruments. Besides, it is difficult to instruct patients to restrict their thinking to only one specific disease-related disability, isolating their judgments from other comorbidities that might be present, or from the impact of normal disabilities associated with to aging, when responding to generic instruments. The final result is that generic instruments might capture the effects of other disabilities and limitations which are not be directly related to the specific disease being mapped.

One possible strategy for avoiding these problems would be to design a preference-choice experiment with the health conditions vignettes derived from the specific instrument [27]. Unfortunately, it could be expected that marginal disutilities could be oversized if other, very severe health conditions are included as anchoring. Another possibility could be to describe specific health conditions only by the set of generic health profiles that are

Table 4 Estimated model goodness of fit statistics

Model	Overall		MAE						MAPE (%)					
	R^2	Adj. R^2	Overall	Q1	Q2	Q3	Q4	Q5	Overall	Q1	Q2	Q3	Q4	Q5
SF-6D														
Linear	.885	.885	.094	.067	.076	.078	.123	.127	29.0	39.0	26.1	23.0	30.8	26.2
Quadratic	.887	.886	.093	.065	.075	.069	.122	.126	28.1	33.4	27.2	23.6	31.3	25.2
Cubic	.888	.887	.092	.065	.075	.069	.120	.125	27.9	33.6	26.0	23.3	31.4	25.2
EQ-5D-3 L														
Linear	.827	.826	.008	.065	.076	.073	.101	.124	38.4	62.3	37.0	28.7	32.6	32.0
Quadratic	.830	.829	.086	.065	.076	.069	.099	.121	36.8	52.2	40.0	29.0	32.8	30.2
Cubic	.831	.831	.086	.065	.076	.076	.098	.119	37.0	33.6	38.2	28.9	33.1	29.8

MAE Mean Absolute Error, *MAPE* Mean absolute percentage Error, *Q1-Q5* quintile groups

prevalent and meaningful in the particular disease, and only mapping those conditions. This approach could be used when observed distributions like the one obtained for the EQ-5D-3 L are found (see Table 2), and a reduced number of health states gather the majority of patients. But, very large samples would need to be used, if the intent is to obtain representative results, and it could be cumbersome when the number of possible health states is very large, as has happened with the SF-6D (Table 2).

In the time being, directly mapping specific health states onto generic utility values seems to remain the best option, and special care should be taken, by aggregating generic utility values over specific severity scores, in order to smooth out the impact of non-specific effects on the mapping estimates. The present paper reports the first study mapping GSF-Q onto two of the most widely used generic HRQoL instruments. In fact, our study could be considered to have high ecological validity due to the large sample used and its ample representativeness.

In our study, GERD was found to be a quite lenient pathology, with mean utility values of 0.656 (SF-6D) and 0.744 (EQ-5D-3 L). In fact, the most prevalent health-attribute level reported was the first (no deterioration), in both generic instruments, except for the attributes/dimensions of pain and Mental Health (see Table 5). Even the scaling of the response levels of one's own GSF-Q suggests that the third response level (L2 in Fig. 2) had been selected by patients in order to be located above the mean in the latent (error-free) severity score for all items, except for the number of days with problems. These results are in agreement with regular GERD diagnosis, which states that stomach problems should be present more than 2 days a week in order to be consistent with GERD [7].

Obtained SF-6D utility scores were shown to be more sensitive to GERD-severity than those obtained from EQ-5D-3 L.The distribution of the former was more spread out, with less likelihood of ceiling effects, and did not exhibit a gap between perfect health, $u(11111) = 1$, and the following larger value, as it was the case with the later, $u(11121) = 0.79$. The observed cumulative distribution function of SF-6D disutility scores was more uniform; the distribution function of EQ-5D-3 L disutilities was steeper (especially in the milder health states) and the distributions did not cross over within their ranges.

GSF-Q scores showed good unidimensional behavior which allowed summarization of GERD-related severity in a single score using factor analysis weights. Unidimensionality analyses endorsed the possibility of summarizing the different GERD symptoms in an aggregated overall score, also obtaining an adequate scaling of response levels. In our case, this strategy should be preferred against one using item-response dummy coding in the regression models, since it avoids deciding how to aggregate item response levels [28] and minimizes the possible impact of covariates in particular response levels.

For each of the generic instruments, the best-fitting model was selected. In both cases, the model including GSF-Q severity (observed, squared and cubed), age, gender, and being treated for comorbidities attained the best fit, and the SF-6D model additionally included BMI. The sign of the regression coefficients were in accordance with predicting a higher disutility as GSF-Q severity scores increase. The inclusion of significant covariates by all models suggests that the loss in HRQoL may be influenced not only by GERD symptoms but also by personal comorbidities present. This is to say that GERD symptoms may be not very prominent when assessing HRQoL using a generic instrument if other health conditions might be present, such as aging, being treated for comorbidities and overweight.

R^2 values were within the range 0.885–0.888 for model SF-6D, and within 0.827–0.831 for model EQ-5D-3 L. Overall MAPE = 27.9% for predicting SF-6D and MAPE = 37.0% for predicting EQ-5D-3 L when using predictions derived from the cubic model. Computing predicted SF-6D disutility MAPE by GSF-Q severity quintile groups, MAPE ranged between 33.6% for Q1 and 23.3% for Q3 while for predicted EQ-5D-3 L disutility, MAPE ranged between 33.6% for Q1 and 28.9% for Q3 (see Table 4). As expected, the error magnitude was smaller near the

Table 5 Percentage of responses by dimension level for each dimension/attribute of the EQ-5D-3 L and SF-6D generic instruments

Dimension	EQ-5D-3 L			Dimension	SF-6D					
	Dimension Level				Dimension Level					
	1	2	3		1	2	3	4	5	6
Mobility	78.4%	21.2%	0.4%	Physical Function	35.6%	27.8%	22.2%	2.5%.	10.5%	1.4%
Self-care	91.5%	8.1%	0.4%	Role Limitation	54.9%	12.2%	18.3%	19.6%	*	*
Daily activities	66.0%	32.7%	1.3%	Social Function	36.4%	28.7%	27.1%	7.0%	0.8%	*
Pain	32.8%	59.9%	7.3%	Pain	12.3%	19.5%	34.8%	18.7%	13.2%	1.4%
Anxiety/Depression	54.6%	39.1%	6.3%	Mental Health	10.8%	61.8%	17.6%	7.8%	2%	*
				Vitality	5.3%	19.3%	22.8%	28.9%	16.9%	6.8%

* Unused dimension level

location of the centroid; while it was particularly high when predicting EQ-5D-3 L disutilities using the linear model (up to 62.3% in Q1).

Some additional covariates, like smoking and drinking carbonated beverages or alcohol, approached statistical significance, but all models were kept as parsimonious as possible, and only statically-significant predictors were included ($p < 0.05$). Bootstrap estimates were generated, based on 1000 samples with replacement, obtaining parameter estimate bias smaller than |0.002| and significance levels$\hat{p} \leq 0.002$.

Mapping disease-specific instruments onto generic health related measures is a common methodological strategy due to the high sensitivity of specific instruments and the wide generalizability of generic measures. Mapping the GERD-specific GSF-Q scores onto generic utilities (SF-6D and EQ-5D-3 L) was shown to be possible, attaining adequate goodness-of-fit values. In both cases, the best-fitting model was the more complex one; the model based on GSF-Q severity, raised to the cubic power, and including generic covariates: age, gender, BMI and treatment for comorbidities. However, the model for predicting EQ-5D-3 L disutilities did not include BMI as a statistically significant covariate.

The use of cubic prediction models needs special care, since small variations in the cubed predictors can entail excessively large predicted values, including those for predictors out of the range of the observed data used for prediction, that can produce unreasonable predictions. In our case this prevention is needless, given that all GERD severity values are scaled within the 0–1 range (any value will have to be inside the range of values used for estimation), and possible covariate values are limited to the observed repertoire.

In our study, we found that utility values associated with GERD-specific conditions were rather high, suggesting that this disease is not very disabling (in general). Nevertheless, patients with utility values as low as SF-6D = – 0.3150 and EQ-5D-3 L = – 0.0757 were observed, although they were not always associated with the worst GSF-Q severity scores. Given the reduced number of prevalent health states obtained for the generic instruments (especially for EQ-5D-3 L) the question arises whether some characteristic or "natural" disease-related health states could be identified for each generic instrument, discarding other comorbidity-influenced health states. From a nosological point of view, it looks quite tempting to think that GERD would not entail a high deterioration in mobility, but it could be the case that bed-ridden people might very likely develop GERD. One possible way to minimize the impact of comorbidities, when measuring specific health conditions with a generic instrument, would be to use a set of instructions demanding that the patient assess his or her overall health condition while thinking only of his or her specific disease.

Limitations

The present study has been carried out with a Spanish population, and we cannot ensure that other cultural or eating habits would not distort our results.

Conclusions

In the present study two methods are presented allowing the mapping of specific GERD-severity scores obtained by use of the GSF-Q, onto generic HRQoL values, as measured by the SF-36 and EQ-5D-3 L instruments. In both cases, the cubic model attains best adjustment.

Mapping is an approach that enables utilities to be predicted for the calculation of quality-adjusted life-years when no preference-based information has been elicited what will allow to elaborate health economic evaluations in a simpler way, since it is not necessary to have data of no preference-based instruments. The results of this study will allow to carry out economic evaluations in the world of gastroesophageal reflux disease which will help in the future when it is necessary to make decisions with new alternatives that arrive at the market.

Abbreviations

BMI: Body Mass Index; CFI: Comparative fit index; CIED: Cardiac implantable device; EQ-5D-3 L: EuroQol 5 Dimensions 3 Levels; GERD: Gastroesophageal Reflux Disease; GOF: Goodness-of-fit; GSF-Q: Gastrointestinal Short Form Questionnaire; H2RA: H_2 Receptor Antagonist; HRQoL: Health Related Quality of Life; M: Mean; MAE: Mean absolute error; MAPE: Mean percentage absolute error; MAUF: multi-attribute utility function; MOS SF-36: Medical Outcome Survey Sort Form – 36 items; p: significance level; PBM: Preference-based measure; PPI: Proton pump inhibitors; PROM: patient-reported-outcome measurement; Q1: 1st Quartile; Q3: 3rd Quartile; R^2: R-square GOF statistic; RMSEA: Root mean square error of approximation; SD: Standard Deviation; SF-6D: Medical Outcome Survey Sort Form 6 dimensions; TLI: Tucker-Lewis fit index; VAS: Visual Analogue Scale

Ethical approval and consent to participate

This is a secondary analysis. The original study obtained the approval by the Ethical Research Committee from the Hospital Universitario La Paz, Madrid (Spain). Signed informed consent and permission to use personal health information were obtained from all participating patients.

Availability of data and materials

The datasets generated and analyzed during the current study are not publicly available due to the fact that Ethics Committee approvals were not obtained for sharing of datasets outside of the research team, but are available from the corresponding author on reasonable request.

Authors' contributions

All authors have contributed substantially in the manuscript preparation, interpretation of results or study design and management. The principal authors take full responsibility for the data presented in this study, analysis of the data, conclusions, and conduct of the research, and had full access to those data and has maintained the right to publish any and all data independent of any third party. All authors read and approved the final manuscript.

Consent for publication

Not applicable. Al results are reported as aggregated data.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Faculty of Psychology, Universidad Autónoma de Madrid, C/ Ivan Pavlov 6, 28049 Madrid, Spain. [2]Faculty of Economics, Universidad Calos III de Madrid, C/ Madrid, 126, 28903 Getafe, Madrid, Spain.

Received: 29 September 2017 Accepted: 29 August 2018
Published online: 10 September 2018

References

1. Winkelstein A. Peptic esophagitis: a new clinical entity. JAMA. 1935;104:906–9.
2. Ing AJ, Ngy MC, Breslin ABX. The pathogenesis of chronic persistent cough associated with gastroesophageal reflux. Am J Rtspir Crit Care Med. 1994; 149:160–7.
3. Marzo M, Alonso P, Bonfill X, Fernández M, Fernández J, Martínez G, Mearín F, Mascort JJ, Piqué JM, Ponce J, Sáez M. Guía de práctica clínica sobre el manejo del paciente con enfermedad por reflujo gastroesofágico (ERGE). Gastroenterol Hepatol. 2002;25:85–110.
4. Stranghellini V. Three month prevalence rates of gastrointestinal symptoms and the influence of demographic factors: results from the domestic international Gastroenterology Surveillance Study (DIGEST). Scand J Gastroenterol Suppl. 1999:20–8.
5. Dent J, Brun J, Fendrick AM, Fennerty MB, Jansens J, Kahrilas PJ, et al. An evidence-based appraisal of reflux disease management. The Genval Workshop Report. Gut. 1999;44(Supl 2):S1–S16.
6. Carlsson R, Dent J, Bolling-Sternevald E, et al. The usefulness of a structured questionnaire in the assessment of symptomatic gastroesophageal reflux disease. Scand J Gastroenterol. 1998;33:1023–9.
7. Arín A, Iglesias MR. Enfermedad por reflujo gastroesofágico. Anales del Sistema Sanitario de Navarra (2003).
8. Hardin SM, Rchter JE, Guzzo MR, Schan CA, Alexander RW, Bradley LA. Asthma and gastroesophageal reflux: acid suppressive therapy improves asthma outcome. Am J Med. 1996;100:395–405.
9. Kahrilas PJ. Gastroesophageal Reflux Disease. JAMA. 1996;276:983–8.
10. Pare P, Meyer F, Armstrong D, Pyzyk M, Pericak D, Goeree R. Validation of the GSFQ, a self-administered symptom frequency questionnaire for patients with gastroesophageal reflux disease. Can J Gastroenterol. 2003;17:307–12.
11. Alonso J. La medida de la calidad de vida relacionada con la salud en la investigación y en la práctica clínica Unidad de Investigación en Servicios Sanitarios. Institut Municipal d' investigación Mèdica (IMIM) 1999.
12. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. N Engl J Med. 1977;296:716–21.
13. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2010;11(2):215–25.
14. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability? Health Econ. 2010;19(2):125–9.
15. Kearns B, Ara R, Wailoo A, et al. Good practice guidelines for the use of statistical regression models in economic evaluations. Pharmacoeconomics. 2013;31:643–52.
16. Wailoo AJ, Hernandez-Alava M, Manca A, et al. Mapping to Estimate Health-State Utility from Non-Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. Value Health. 2017; 20:18–27.
17. Ruiz MA, Suárez JM, Pardo A, García-Vargas M, Pascual V. Cultural adaptation to Spanish and validation of the Gastrointestinal Short Form Questionnaire. Gastroenterol Hepatol. 2009;32(1):9–21.
18. EuroQoL Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy. 1990;6(3):199–208.
19. Badia X, et al. La versión española del EuroQol: descripción y aplicaciones. Med Clin (Barc). 1999;112(Supl. 1):79–86.
20. Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. Soc Sci Med. 1997;44:1519–30.
21. Szende A, Oppe M, Devlin N (Eds.). EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Springer; 2007.
22. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271–92.
23. Abellan Perpinan JM. Utilidades SF-6D para España. Guía de uso 2012/8. Sevilla: Catedra de Economía de la Salud. Universidad Pablo de Olavide. Consejería de Salud de la Junta de Andalucía; 2012.
24. Vilagut G. El cuestionario de salud SF36 español: una década de experiencia y nuevos desarrollos. Unidad de Investigación en Servicios Sanitarios. Institut Municipal d'Investigació Mèdica (IMIM-IMAS). Barcelona. España. Gac Sanit. 2005;19(2):135–50.
25. Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982; 47(2):149–74.
26. Teruel C, Faro V, Muriel A, Mañas N. Sensitivity and specificity of the Gastrointestinal Short Form Questionnaire in diagnosis of gastroesophageal reflux disease. Rev Esp Enferm Dig. 2016;108(4):174–80.
27. Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ. 2009;18(11):1261–76.
28. Coyne K, Revicki D, Hunt T, Corey R, Stewart W. Psychometric validation of an overactive bladder symptom and health-related quality of life questionnaire: The OAB-q. Qual Life Res. 2002;11(6):563–74.

Mapping of the FKSI Specific Kidney Disease Measure onto Three Generic Preference-Based Measures to Generate Utility Values

Authors: Manuel Monroy, Engr; Miguel A Ruiz, Ph.D.; Julio Bogueat, BA; Isabel García, Md; Juan C Julián, BA

Keywords: Social utility; health related quality of life; mapping; regression models; patient reported outcome measurement (PROM); FKSI; MOS SF-6D; HUI-3; EuroQol-5D.

Abstract

AIM

To obtain the mapping algorithms for translating the specific severity measures obtained by FKSI upon three generic utility instruments (SF-6D, EQ5D-3L y HUI-3) for chronic kidney disease (CKD) patients.

METHODS

A cross-sectional prospective observational study design, including CKD patients recruited at random on demand for attention, was used. The final sample was composed by 161 patients suffering CKD. Women were 42%, mean age 54.6 years (SD=15.5) and average disease seniority of 2.82 years (SD=1.61). Three questionnaires measuring HRQoL (EQ5D-3L, SF-6D and HUI-3) were administered along with CKD symptom severity specific (FKSI-9). Different regression models (linear, quadratic, cubic) were fitted for each separate generic instrument estimating preference based utility values from FKSI symptom severity, and compared using several goodness-of-fit statistics. Empirical grouping of patients based on utilities and severity was explored using Latent Profile Analysis.

RESULTS

Observed FKSI scores distributed between 0 and 29 points with an average M=7.5 (SD=6.39). Utility mean values for the three instruments were noticeably different: M_{EQ}=0.676 (SD_{EQ}=0.247), M_{SF}=0.514 (SD_{SF}=0.286) y M_{HU}=0.673 (SD_{HU}=0.232). In the three cases, the best fitting model was the cubic one, with the best fit attained by SF-6D (R^2=0.619) while EQ-5D (R^2=0.548) and HUI-3 (R^2=0.565) were lower. Latent profile analysis distinguished four clusters, with R^2=0.87 and 7% classification-error-rate.

CONCLUSIONS

Obtained results allow to transfer CKD deterioration values into social utility values for those health states, as measured by the three most widely used HRQoL instruments.

Background

Chronic kidney disease (CKD) is a progressive and irreversible loss of renal function, due to different causes (diabetic nephropathy, hypertension, glomerulonephritis, hereditary renal failure, pharmacological poisoning, etc.) that lead to the kidneys losing their ability to eliminate waste, concentrate urine and preserve electrolytes in the blood, progressing towards the total loss of kidney function. At advanced stages, usual treatments are kidney transplantation, hemodialysis and peritoneal dialysis, all of which have a notable impact on daily life and the quality of life of patients.

The Spanish society of nephrology (SEN)[1] reported a CKD prevalence of 1,234 patients per million population (pmp) in 2016, varying between 1,752 pmp in the age group between 45 and 64 years and 2,888 pmp in people over 75 years. Etiology is known to vary with age, being vascular causes more frequent in patients older than 75 years, those caused by diabetes in the 65-74 age group, polycystic disease in the 45-65 age group and hereditary origin in ages less than 45 years. The prevalence of renal replacement therapy is 521 pmp in hemodialysis, 67 pmp in peritoneal dialysis and 647 pmp in renal transplant. The percentage of mortality in 2016 was 8%, being the most frequent causes cardiovascular problems and associated infections. Median survival is 6.3 years, with a 5 years' survival percentage of 57%.[2]

Given the great impact of CKD on patient's wellbeing, it is important to bring forward indicators capable of quantifying the patient's vital state, towards an adequate therapeutic follow-up and, in particular, those reported by the patient. One of the patient reported outcome measures (PROM) most widely used as an indicator of the patient health status is Health Related Quality of Life (HRQoL), a measure that reflects the patient's subjective perception (without intervention of the clinical staff) in a repertoire of dimensions such as: emotional state, level of pain, physical functioning, social functioning and general perception of one's own health.[3]

HRQoL is a particularly important output due to its diagnostic capabilities, since it has been shown to be directly related to mortality, hospitalization and consumption of clinical resources[4]. HRQoL has also shown relationship with other specific disease indicators, adding complementary information for assessing clinical deterioration. Nowadays, HRQOL is accepted as a clinical goal by itself in patients with limited life expectancy or in therapies seeking to cope with the disease or to accommodate to symptoms, typical aspects of chronic diseases. PROMs have proven to be very sensitive when we study variations in health status for a particular pathology, being this sensitivity one of the reasons for usual inclusion in clinical studies.[5,6,7,]

Researchers commonly prefer to use pathology-specific questionnaires in patient follow-up, due to their greater sensitivity to health changes and better targeting to the pathology under study. But when the aim is to compare results with those of other pathologies or to perform economic evaluations, it is preferred to use generic (non disease-specific) HRQOL instruments. This is not without limitations, since generic instruments may capture information on patient characteristics (such as age, comorbidities or unwanted treatment effects) which may not be relevant or might be insensitive to mild health conditions.

The most popular generic instruments (such as SF6D, EQ-5D and HUI3) offer the possibility to calculate the utility value associated with each health condition (according to the profile given by the attributes measured by the instrument), which reflects the population preference towards each state of health, in a choice situation of uncertainty. This feature allows the use of utilities in the calculation of quality of life adjusted for years of life (QALY) and in any health-economy study in general.

In real life research, it is usually the case that we prefer to use a disease specific PROM instrument instead of a generic one, and not to include one of the later, so as not to overload the patient with self-reported measures. In such cases, the usual strategy is to perform a metric translation (mapping) from the specific measurement over the generic instrument.[8] The mapping is also of interest when we want to compare our results with those obtained using a different generic instrument or even when there was no generic instrument available (as in the studies of retrospective databases or in meta-analyses).

OBJECTIVE

The objective of this study is to obtain the mapping algorithms necessary to translate the specific HRQoL measurement obtained using the FKSI specific CKD health index into three of the most popular preference-based generic instruments (SF6D, EQ-5D-3L and HUI-3). We will compare two procedures, on based on regression methods and another obtaining profiles by means of cluster analysis.

As a secondary benefit, we will be able to assess which one of the generic instruments is more adequate for capturing deterioration in HRQoL due to CKD condition.

METHODS

Study design

The present study was designed as a cross-sectional prospective observational study. The sample was designed with the aim of reaching a large enough size to carry out the proposed multivariate analyses. Patients were included randomly on demand for treatment in the participating centers. A minimum target sample size of 150 patients with complete responses was determined. Patients were recruited by the collaborating therapists from the ALCER association, without limitations regarding the geographical origin, and including them as they gave their informed consent. The study protocol was approved by the Universidad Autómoma de Madrid (Spain) Research Ethic Committee. The Helsinki Declaration guidelines were met.

Participants

The following inclusion criteria were applied: both genders, age above 18 years old, being in treatment of a chronic kidney disease, no cognitive impairment, being able to answer the questionnaires on their own, and having given their informed consent.

The final sample consisted of 161 patients, 41.6% of the participants being women. Mean age was 54.6 years (SD = 15.5) and with an average time from diagnosis of 2.82 years (SD = 1.61). A total of 18.6% were obese, 37.3% suffered from congenital pathology (14.9% Polycystic, 16.1% Glomerulonephritis, 2.5% Pyelonephritis) and

44.7% from acquired pathology (28.1% due to diabetes, hypertension or cardiovascular accident, 2.5% due to food or drug intoxication and 1.2% due to trauma). The most common treatments were: conservative treatment (13.7%), dialysis (62.7%) and transplantation (23.1%). Form them, 34.8% were above the cut-off score for clinical anxiety, while 23.6% could be classified with a clinical level of depression. Average anger expression score was 32.6 (SD = 9.45). It is worth mentioning that 88.2% were also in treatment, at least, for another comorbidity (Table 1).

Instruments

An ad-hoc data collection form was designed including the four questionnaires to be administered: three to measure generic HRQoL (EQ5D-3L, SF6D and HUI3) and a specific instrument measuring severity of CKD symptomatology (FKSI-9). In addition, Hamilton's Hospital Anxiety-Depression Scale (HADS) and the State-Trait Anger Expression Inventory (STAXI) were also administered. Three data collection forms were created, so that each one of the HRQoL questionnaires was presented first in turn, with the aim to control for any possible carryover effect among the quality of life measurements.

Questionnaire EQ5D-3L

EuroQol-5D-3L (EQ5D-3L)[9,10] is a generic instrument of HRQOL based on population preferences. It assesses the level of deterioration in 5 attributes: mobility, self-care, daily activities, pain/discomfort and anxiety/depression; using items with 3 response levels (1=none, 2=some problems, 3=many problems). Each combination of levels creates a health profile, with a total of 243 possible health states, although not all are equally likely. The profile [11111] corresponds to perfect health and the profile [33333] represents the worst possible state of health (pits). Based on the sorting of health profiles according to social preferences, each health state is translated into a social utility value, which may be computed from the 5 attribute levels using multi-attribute utility function (MAUF). Different MAUFs are used in different countries, mainly using estimates based on standard gamble, time trade-off and visual analog scale (VAS) procedures. The basic MAUF equation is additive:

$$u_i = 1 - \left(q + \sum_{j=1}^{j=5} \sum_{k=1}^{k=3} b_{jk} D_{ijk} + b_{N3} N3_i \right)$$

where the utility/preference value for health status *i* is obtained by subtracting from 1 the disutility of the health status. Disutility is obtained by weighting by b_{jk} the level of deterioration *k* reached in dimension *j* (dummy variable D_{ijk}) plus an interaction term (*N3*), which adds a constant when any of the dimensions reaches its maximum level of deterioration, plus a constant (*q*). It should be noted that the first level on any dimension ($k = 1$), represents that there is no deterioration in that dimension, $D_{ijk}=0$, and the perfect health profile is anchored at a utility value of 1.

SF6D questionnaire

The medical results survey (MOS), in its 6 dimension utility form (SF6D) [11], is a generic HRQoL instrument based on preferences derived from the MOS SF-36 (36 items). It summarizes the level of deterioration in 6 dimensions: physical functioning, role limitations, social functioning, pain, mental health and vitality; using a coding in 4 to 6 levels of 11 items. It is possible to obtain a total of 18,000 health profiles, with profile [111111] corresponding to perfect health, and [645655] representing the worst possible state of health. Different MAUFs have been estimated to derive utilities in different countries, with the particularity that no constant of severity (interaction) is used. A value of 0 is assigned to the first level for each dimension/attribute.,[12],[13]

HUI3 questionnaire

The Health Utilities Index Mark 3 (HUI-3)[14] survey covers several aspects of health, intentionally restricted to skills (physical and emotional), and excludes role performance and social interaction. It covers eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain, using five to six response levels. Each combination of levels indicates a unique health status. The MAUF of this instrument is multiplicative and different functions have been estimated for different countries; the state of perfect health [11111111] has a utility of 1, while the utility for the lowest level in the eight attributes [66566565] is -0.36, which is considered a health situation equivalent to worse than being dead. The MAUF for deriving the utility from a profile in the Spanish population, is given by:

$$u_i = (1{,}0078 \times b1 \times b2 \times b3 \times b4 \times b5 \times b6 \times b7 \times b8) - 0{,}0078$$

Where the values $b1$-$b8$, in the pits-Full Health metric, are the coefficients calculated in the Spanish population[15], which correspond to the response level attained in each one of the dimensions.

FKSI-DRS questionnaire

The Functional Assessment of Cancer Therapy - Kidney Symptom Index - Disease Related Symptoms (FKSI-DRS-9)[16,17,18] is a self-reported questionnaire composed by the first 9 items of the instrument FKSI-15, which are answered on a 0 to 4 points Likert scale assessing the level of limitation due to the symptoms of kidney diseases. This instrument was used as the disease-specific measure of deterioration. Two dimensions may be distinguished: physical and psychological. Items used to assess perceive individual change were discarded from the FKSI-15. The scoring ranges from 0 to 36 points. A higher score reflects greater deterioration. [19]

Statistical analyses

The criterion score on the specific kidney disease health status was obtained computing the factor score for FKSI items, assuming one overall dimension (Principal Components extraction, factor score regression method), which produces a summary score with 0 mean and standard deviation proportional to the eigenvalue of the dimension.

To interpretation easier, it was re-scaled to a 0-1 metric, since the attainable minimum and maximum scores are known. The score obtained was considered the specific kidney symptomatology indicator of reported severity.

Once the specific severity indicator was obtained, a metric translation of the indicator values was performed on each of the three generic measures of HRQoL used in this study, each one separately. In this way, the predicted utility values were obtained for each generic instrument given a level of kidney symptoms severity. Several regression models, linear and non-linear, were tested and compared using various goodness-of-fit statistics.

In all regression models, the values of disutility ($d_i = 1 - u_i$) were used, instead of the values of utility, for the following reasons. First, the data mass is usually concentrated around the most favorable health states with least disutilities, so that the points of greatest mass are close to the origin of the coordinate axes, the independent and dependent variables (severity and disutility) are measured in the same direction and the slope of the model is always positive. Secondly, it is always possible to estimate a model without the intersection term, anchoring the 0 value of disutility (perfect health) at the origin, and making it match with the minimum severity value of the FKSI (which will also be 0).

Subsequently, it suffices to subtract from 1 the predicted disutility to obtain the model utility predicted value.

The following regression models were estimated: linear, quadratic and cubic, using the density function values; and Tobit, using cumulative values of the distribution function. To anchor the best possible health states in both instruments, symptom severity scores were scaled within the 0-1 range.

Before estimation of the different prediction models, those patients with evident outlier values in two or more of the generic instruments were discarded, since their score could be reflecting peculiarities that were not typical of the pathology under study. Outlier values were identified as those clearly falling outside the 95% individual confidence interval for the linear model (departing in more than 3 standardized residuals).

Along with the statistical significance for the regression coefficient estimates, goodness of fit (GOF) of each model was assessed using R^2 statistic, average absolute error (MAE) and percent average absolute error (MAPE). MAE and MAPE were computed overall and by quintile groups according to the severity scores, in order to assess the local GOF at the different levels of severity. MAE and MAPE indices should be studied with caution since very small utility values can inflate the mismatch values substantially, when dividing by quantities close to 0.

Covariates were not included in the regression models (age, disease seniority, number of treatments, comorbidities, depression level, etc.) with the aim to consider only the direct effect of the disease. In addition, the inclusion of covariates would limit the use of the models in retrospective studies in which the possible covariates could have not been gathered.

As an additional procedure, a latent profile analysis (LCP) was carried out exploring how health states summarized by the three generic instruments rank patients. It could be the case that patients are sorted differently by each generic instrument or that utility measures might show different sensitivity at different levels of severity. The disutilities of the generic instruments (HUI-III, SF-6D and EQ-5D) as well as the severity of the specific instrument (FKSI) were included as active variables in the LCP. Sociodemographic variables and disease descriptors were also included as inactive covariables in order to describe the profiles obtained.

All analyzes were carried out using IBM SPSS v23 software and LatentGold V.5.0.

RESULTS

Observed direct scores on the FKSI renal symptoms severity scale were distributed between 0 and 29 points, with a mean M = 7.5 (SD = 6.39), while transformed factor scores varied between 0 and 1, with an average value of 0.261 (SD = 0.219), and with positive skewness $g_1 = 0.954$ (SE = 0.191). Only 6.2% of patients scored at the minimum scale value.

Average utility scores obtained with the generic instruments were significantly different: $M_{EQ} = 0.676$ ($SD_{EQ} = 0.247$), $M_{SF} = 0.514$ ($SD_{SF} = 0.286$) and $M_{HU} = 0.663$ ($SD_{HU} = 0.232$), being SF6D mean significantly lower than the other two ($p < 0.001$). Correlations between them were all significant ($p < 0.001$), r(EQ, SF) = 0.797, r(EQ, HU) = 0.764 and r(SF, HU) = 0.763. Scores presented a clear negative bias in all cases, with accumulation of cases at the top of the scale: $g_{EQ} = 0.682$ (SE = 0.191), $g_{SF} = 0.904$ (SE = 0.195), $g_{HU} = 0.944$ (SE = 0.194).

Figure 1 shows the cumulative distribution of re-scaled factor scores. The equation needed to compute the corrected factor scores out from the scores on the individual items of the FKSI is given by

$$\hat{f}_i = 0,226 * \begin{bmatrix} (0,155*(X_1\text{-}1,593))+(0,121*(X_2\text{-}1,185))+(0,029*(X_3\text{-}0,673))+ \\ (0,128*(X_4\text{-}1,216))+(0,282*(X_5\text{-}1,525))+(0,139*(X_6\text{-}0,772))+ \\ (0,090*(X_7\text{-}0,512))+(0,089*(X_8\text{-}0,148))+(0,071*(X_9\text{-}0,062))+1,17 \end{bmatrix}.$$

Where X_1-X_9 are the scores on the FKSI items, and the values 0.266 and 1.17 are scale constants needed to translate the values into the 0-1 range.

Regarding the degree of sensitivity shown by the instruments, it was observed that the EQ-5D was the least sensitive, obtaining only 36 profiles of the possible ones and accumulating 55.9% of the patients in 4 of them (11111, 11112, 11121, 11122), while the 89 profiles were obtained using the HUI-3 and 146 using the SF-6D (Table 2).

Table 3 shows the percentage of patient accumulated at the different response levels of the attributes and for each one of the instruments. It can be observed that the

patients tend to be located at the less severe health levels, although patients can be found in the higher levels of severity of most attributes.

The cubic model was the best fitting one for the mapping functions three, although it should be noted that the differences in fit were minimal between the models of different shape (linear, quadratic and cubic). The cubic pattern was chosen due to better represent the expected evolution of the utilities, starting at a floor value corresponding to the perfect health state (disutility = 0) and growing towards an asymptotic value at the ceiling of the scale (disutility = 1). Table 4 shows the coefficients needed to estimate the disutilities for the three instruments. Predicted utilities are obtained by subtracting from 1 the value of predicted disutility.

Moderate fit was attained by all models, with the SF-6D reaching the best fit (R^2 = 0.619), while EQ-5D (R^2 = 0.548) and HUI-3 (R^2 = 0.565) were lower. However, the relative error obtained with the SF-6D model was much higher (MAPE = 56.9%) than the 20% obtained by the two other models. As expected, the size of residuals stratified by quintiles turned out to be especially bad at the quintile corresponding to high utility values, that is, in less serious health conditions.

Although determining the number of clusters for this validation test is not crucial, the LCP analysis identified 4 clusters with centroids shown in Table 5. The solution reached good fit R^2 = 0.87 with an error classification rate of 7%. Cluster profiles (Figure 2) show that averages are arranged in parallel (without crossings) implying that clusters are collecting groups of patients with levels of progressive deterioration in the disease (FKSI) and also in the three generic instruments of HRQOL. In the absence of crosses, we can infer that there are no other aspects of health not being considered, which might be influencing substantially the measurement of HRQoL, but those corresponding to the CKD itself. It is also true that if we would increase enough the number of conglomerates, profiles would end up showing crossings between clusters. Inspection of profiles also shows that the SF-6D tends to assign slightly higher disutility values, and the EQ-5D usually assigns lower disutility values. Progression of disutility when moving between disability strata within each instrument is rather similar for all three instruments.

DISCUSSION

Disease specific HRQoL are the preferred choice for measuring health given their high sensitivity to changes in the patient health state (treatment effectiveness, disease progression, coping with symptoms, etc.). Therefore, using generic instruments instead implies loosing sensitivity and also involves other measurement problems since it is difficult to make the patient isolate the health aspects related only to the pathology that is being assessed. Naturally, patients have an overall view on their health state and it is difficult to filter out the effect of possible comorbidities, adverse events or the affective state. However, even if it is unadvisable to use generic instruments for an accurate assessment of the health state and, therefore, for patients follow-up, there are research situations where obtaining generic measures is crucial. We must remember that the generic measures reflect the social value of the patient health state (compared to other possible health states) and not really their vital situation. Which is the reason why they are the measures of choice in pharmaco-economic valuations.

A possible strategy to avoid these problems would be to design preference elicitation choice experiments using vignettes based on the health conditions derived from the specific instrument, but this would not prevent from the inflation of marginal utilities due to other serious comorbidities being present. Another possibility would be to determine the generic health profiles that are really prevalent and meaningful in the particular disease, and only to mapping those conditions. This approach could be used when observed distributions are found such as that obtained for EQ-5D-3L, where a small number of health states gather together the majority of patients. However, if we intend to obtain representative results, very large samples should be used, and it could be cumbersome when the number of possible health states is very large, as has happened empirically with the SF-6D (with 127 states) or the HUI- 3 (with 64 states, see Table 3).

For the time being, the direct mapping of specific health states into generic utility values seems to be the most accepted option[20]. Nevertheless, another possible way to determine the mapping between generic instruments anchored by a specific instrument could be to identify empirical profiles of health states shared by groups of patients, using cluster generation procedures such as the LPA. This procedure would allow to determine as many clusters as considered appropriate, and to obtain the table of correspondences between utility values of different instruments based on the average utility value on each

instrument, represented by the centroid of each cluster. We have seen that for a small number of clusters this option is possible (see Table 5), but the behavior with a high number of strata (clusters) might not be as uniform as in our case, and inversions between the instruments (profile crossings) may appear, which could be difficult to understand. In fact, in our findings, the LPA solution may have been particularly insensitive to comorbidities due to the removal of extreme cases (Figure 3).

In our study, CKD has shown to be a quite disabling pathology, with low average utility values: $M_{EQ} = 0.676$, $M_{SF} = 0.514$ and $M_{HU} = 0.673$. However, we have observed a large number of patients whose scores are at the lowest level (without deterioration or with mild deterioration) in most of the attributes of the generic instruments (Table 4). It is also true that our sample, even being representative of patients with CKD in touch with patient organizations, is not a sample with a high level of deterioration since 50% of the subjects obtain scores between 0 and 5 points (from a possible maximum of 36 points).

Utility scores obtained using SF-6D and HUI-3 instruments showed to be more sensitive to CKD severity than those obtained using the EQ-5D-3L. This behavior is known and currently a new version of the EQ-5D is being developed with five levels per attribute[21,22]. In addition, the distribution of scores of the first two instruments was more disperse and they did not show a gap between perfect health and the following health profile. The observed cumulative distribution functions for SF-6D and HUI-3 disutility scores were more uniform, while the EQ-5D-3L showed a steeper function, especially at the mild health states.

In the regression models, the strategy of using factor scores to summarize CKD severity is technically preferable to the use of the score obtained directly from the algebraic sum of FKSI item scores, since each item is weighted according to its individual reliability for optimally sorting patients according to CKD severity. Furthermore, it avoids having to decide on how to sum-up the scores when building the criterion variable (disutilities) based on the response levels in each item, and minimizes a possible impact of the covariates over particular levels of response.

Although we have not considered any covariates in the prediction of disutilities, we did check for the influence of other variables in the mapping functions. Variables able to contribute in explaining additional variability present in utility scores where "number of concomitant diseases", "anxiety" and "frequency of anger situations", and also "years

since diagnosis" in the case of predicting SF-6D disutilities, results departing from the inclusion of obesity, age and hypertension in the EPIRCE study[2].

The model with best fit for predicting disutility values was the cubic model. All proposed models presented the same problem, the great dispersion of the utility scores observed at the non-severe health states of the FKSI (Figure 4). But this phenomenon should not be understood as an anomalous behavior, rather it reflects the limitation of specific instruments themselves to capture the effect of covariates (that may explain the overall level of deterioration), and not so much due to the limitation of generic instruments for measuring benign health states. In fact, a not irrelevant group of patients obtained very high disutility values (probably due to other aspects of their health deterioration) but with a very low specific CKD deterioration level. Studying these cases with large residuals and low FKSI scores, we found that they were subjects with notable high levels of anxiety and depression, among other possible confounding factors. Better fitting models could have been obtained including covariates not specific to CKD (such as age, psychological health, comorbidities, type of treatment, etc.), but this would lead to a limited applicability of models to other data sets and, subsequently, the mapping models would not be generalizable.

Our study on the behavior of utilities in subpopulations of cases produced the stratification of the sample by levels of severity. The clusters corresponded to strata of patients with progressive levels of deterioration, in which all instruments showed a similar progression, both generic and specific. Although the technique used is very sensitive to the presence of atypical cases, the solution obtained discriminated levels of deterioration but not the presence of this type of cases (perhaps due to the previous filtering of outliers).

CONCLUSIONS

The mapping of disease-specific instruments into health related generic measures is a common methodological strategy which takes advantage of the high sensitivity of specific instruments and the broad generalizability of generic measures. It was shown that it is possible to map CKD specific FKSI scores into generic disutilities (SF-6D, HUI-3 and EQ-5D-3L), achieving adequate goodness of fit values and an acceptable amount explained variance (between 55 % and 62%).

The supremacy of the cubic model was not very evident, since the MAPE values of the different models were very similar. The similarity of the models is due to the lack of fit obtained by all of them at low values of disutility (best health states). This is an inherent problem for generic instruments, which have shown to capture health impairments not attributable to the specific deterioration measured by the FKSI.

Our results allow to transfer the values of CKD impairment onto the utility attributed by society to those health states, as they are appraised by the three HRQoL instruments most frequently used in research.

LIMITATIONS

The present study has been carried out in the Spanish population and it is possible that cultural biases might be present.

Availability of data and materials

All anonymized data may be shared on reasonable demand to the authors.

Competing interests

The authors declare that they have no competing interests

Funding

This research was carried out with no external funding

Author's contributions

MR was responsible for the design of the study; MR, JB and JCJ decided on clinical variables and psychological measurements to be gathered from patients, MM & MR analyzed and interpreted the patient data regarding the estimation mapping procedures. MM and JCJ performed the bibliographic search on epidemiological data, JB and IG were responsible for data collection; MM and MR were major contributors in writing the manuscript. All authors read and approved the final manuscript.

TABLES AND FIGURES

Table 1. Sociodemographic and clinical descriptors

Variable		Freq.	%	Variable	Level	Freq.	%
Place of treatment	Dialysis Unit	30	18.6	Etiology	Hereditary	60	37.3
	Hospital	39	24.2		Acquired	72	44.7
	Home	21	13.0		Unknown	29	18.0
	Patient Association	65	40.4	Number of concomitant diseases	0	19	11.8
	Unknown	6	3.7		1	45	28.0
Age (years)	18-29	8	5.0		2	29	18.0
	30-39	25	15.5		3	20	12.4
	40-49	27	16.8		4	22	13.7
	50-59	38	23.6		5	9	5.6
	60-69	33	20.5		6	7	4.3
	70-79	24	14.9		7	1	0.6
	≥ 80	4	2.5		8	3	1.9
	Unknown	2	1.2		9	4	2.5
Gender	Female	67	41.6		10	2	1.2
	Male	93	57.8	Treatment	Conservatory	22	13.7
Body Mass Index	Underweight	5	3.1		Dialysis	101	62.7
	Normal	56	34.8		Transplantation	37	23.0
	Overweight	65	40.4		Not known	3	0.6
	Obesity	30	18.6	Anxiety	Subclinical	85	52.8
	Unknown	5	3.1		Uncertain	20	12.4
Education	No studies	12	7.5		Present	56	34.8
	Primary	44	27.3	Depression	Subclinical	104	64.6
	Secondary	33	20.5		Uncertain	19	11.8
	Vocational Training	37	23.0		Present	38	23.6
	Postgraduate	29	18.0	Disease seniority (years)	≤5	49	30.4
Civil State	Single	34	21.1		6-10	27	16.8
	Married	87	54.0		11-15	20	12.4
	Divorced	22	13.7		16-20	17	10.6
	Widower	13	8.1		> 20	40	24.8
	Other	2	1.9		Unknown	8	5.0

Table 2. Generic HRQoL prevalent profiles, frequency, percentage and cumulative percentage. Partial listing.

Instrument											
EQ5D				SF6D				HUI			
Profile	n	%	Cum %	Profile	n	%	Cum %	Profile	n	%	Cum %
11111	44	27.3	27.3	211224	3	1.9	1.9	21111212	10	6.2	6.2
11121	20	12.4	39.8	211123	2	1.2	3.1	11111111	9	5.6	11.8
11112	14	8.7	48.4	212123	2	1.2	4.3	21111111	9	5.6	17.4
11122	12	7.5	55.9	212125	2	1.2	5.6	21111112	8	5.0	22.4
11222	6	3.7	59.6	212224	2	1.2	6.8	21111211	8	5.0	27.3
21111	5	3.1	62.7	212225	2	1.2	8.1	11111211	6	3.7	31.1
21222	5	3.1	65.8	212324	2	1.2	9.3	11111112	5	3.1	34.2
11123	4	2.5	68.3	213323	2	1.2	10.6	21111232	5	3.1	37.3
11132	4	2.5	70.8	223114	2	1.2	11.8	21111213	4	2.5	39.8
21121	4	2.5	73.3	233224	2	1.2	13.0	11111113	3	1.9	41.6
21232	4	2.5	75.8	312325	2	1.2	14.3	21111311	3	1.9	43.5
12222	3	1.9	77.6	343445	2	1.2	15.5	21111313	3	1.9	45.3
21223	3	1.9	79.5	344556	2	1.2	16.8	11111122	2	1.2	46.6
22222	3	1.9	81.4	443434	2	1.2	18.0	21111132	2	1.2	47.8
22233	3	1.9	83.2	111121	1	.6	18.6	21111221	2	1.2	49.1
11221	2	1.2	84.5	111122	1	.6	19.3	21111222	2	1.2	50.3
11233	2	1.2	85.7	111123	1	.6	19.9	21111333	2	1.2	51.6
21122	2	1.2	87.0	111124	1	.6	20.5	21111334	2	1.2	52.8
21233	2	1.2	88.2	111133	1	.6	21.1	21111412	2	1.2	54.0
22223	2	1.2	89.4	112223	1	.6	21.7	21111434	2	1.2	55.3
22333	2	1.2	90.7	112323	1	.6	22.4	21112323	2	1.2	56.5
11113	1	.6	91.3	112324	1	.6	23.0	23111212	2	1.2	57.8
11211	1	.6	91.9	112326	1	.6	23.6	41111111	2	1.2	59.0
11212	1	.6	92.5	121222	1	.6	24.2	11111132	1	.6	59.6
11223	1	.6	93.2	131133	1	.6	24.8	11111222	1	.6	60.2
11231	1	.6	93.8	133334	1	.6	25.5	11111232	1	.6	60.9
11232	1	.6	94.4	211111	1	.6	26.1	11111311	1	.6	61.5
12312	1	.6	95.0	211113	1	.6	26.7	11111312	1	.6	62.1
21131	1	.6	95.7	211122	1	.6	27.3	11111313	1	.6	62.7
21211	1	.6	96.3	211211	1	.6	28.0	11111324	1	.6	63.4
21212	1	.6	96.9	211214	1	.6	28.6	11111413	1	.6	64.0
21221	1	.6	97.5	211221	1	.6	29.2	11112222	1	.6	64.6
22221	1	.6	98.1	211222	1	.6	29.8	11121112	1	.6	65.2
22312	1	.6	98.8	211223	1	.6	30.4	11121213	1	.6	65.8
22323	1	.6	99.4	211322	1	.6	31.1	11121333	1	.6	66.5
32233	1	.6	**100.0**	211323	1	.6	**31.7**	11131122	1	.6	**67.1**

Tabla 3. Generic HRQoL percentage of responses by attribute/dimension response level.

Table 3A.

	EQ5D-3L Levels		
Dimension	1	2	3
Mobility	73.3	26.1	0.6
Personal Care	88.8	11.2	0
Daily Activities	68.9	28.0	3.1
Pain	43.5	43.5	13.0
Anxiety/Depression	50.3	36.0	13.7

Table 3B.

	SF-6D Levels					
Dimension	1	2	3	4	5	6
Physical Function	7.5	44.7	29.8	6.8	5.6	5.6
Role Limitations	42.9	15.5	8.7	32.9	*	*
Social Function	23.6	25.5	31.7	11.8	7.5	*
Pain	21.1	18.6	24.2	13.0	14.9	8.1
Mental Health	6.8	52.8	16.8	12.4	11.2	
Vitality	5.0	9.9	16.1	24.8	21.7	22.4

Table 3C.

	HUI-3 Level					
Dimension	1	2	3	4	5	6
Vision	26.7	63.4	5.0	3.1	1.9	0
Hearing	92.5	1.9	3.7	1.2	0.6	0
Speech	95.0	3.7	0.6	0	0.6	*
Mobility	82.6	8.7	4.3	3.1	0.6	0.6
Dexterity	36.6	33.5	17.4	9.9	2.5	0
Cognition	63.4	12.4	20.5	3.1	0.6	*
Pain	31.7	37.9	19.3	9.9	1.9	*

Table 3D.

	FKSI Level				
Dimension	0	1	2	3	4
Lack of Energy	25.5	28.6	19.9	14.9	11.2
General Pain	42.2	21.1	21.1	8.7	6.8
Weight Loss	64.6	13.0	14.9	7.5	0
Bone Pain	43.5	23.0	13.7	9.9	9.9
Exhaustion	29.8	23.0	24.2	12.4	10.6
Breathing	64.0	11.8	13.7	6.2	1.2
Cough	72.0	15.5	5.0	6.2	1.2
Fever	91.9	5.6	0.6	1.9	0
Hematuria	98.1	1.2	0	0	0.6

* Dimension level not used

Table 4. Estimated model coefficients, and goodness of fit statistics, overall (top) and by quintiles (bottom)

	Instrument		
Coefficients	EQ-5D	SF-6D	HUI-3
b_0	.103	.248	.092
b_1	.234	-.061	-.082
b_2	1.462	2.460	2.107
b_3	-1.053	-1.648	-1.455
Fit			
R^2	.548	.619	.565
$F_{3,157}$	61.384*	80.228*	64.441*
MAE	.128	.121	.108
MAPE	22.4%	56.9%	17.6%

* p<0,001. : MAE= Mean Absolute Error, MAPE, Mean absolute percentage Error

	MAE (MAPE)					
	Overall	Q1	Q2	Q3	Q4	Q5
EQ-5D-3L	.128 (22.4)	.118 (14.6)	.139 (16.6)	.122 (17.1)	.133 (27.2)	.132 (36.9)
SF-6D	.121 (56.9)	.093 (24.8)	.101 (40.0)	.134 (72.6)	.137 (59.9)	.153 (82.2)
HUI-3	.108 (17.6)	.073 (8.6)	.074 (8.2)	.114 (15.2)	.142 (19.7)	.152 (38.0)

MAE= Mean Absolute Error, MAPE, Mean absolute percentage Error, Q1-Q5=quintile groups.

Table 5. Cluster centroids (means disutilities) and standard deviations by class latent profiles

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
FKSI Severity	0,1413 (0,098)	0,2754 (0,171)	0,4419 (0,234)	0,7611 (0,203)
EQ5D Disutility	0,0563 (0,095)	0,2541 (0,054)	0,4227 (0,166)	0,6809 (0,133)
SF6D Disutility	0,2283 (0,094)	0,3626 (0,182)	0,5507 (0,221)	0,9493 (0,136)
HUI3 Disutility	0,0808 (0,084)	0,1504 (0,101)	0,3835 (0,192)	0,5990 (0,209)
N (%)	46 (32%)	40 (28%)	36 (25%)	21 (15%)

Figure 1. Cumulative distribution for FKSI observed values.

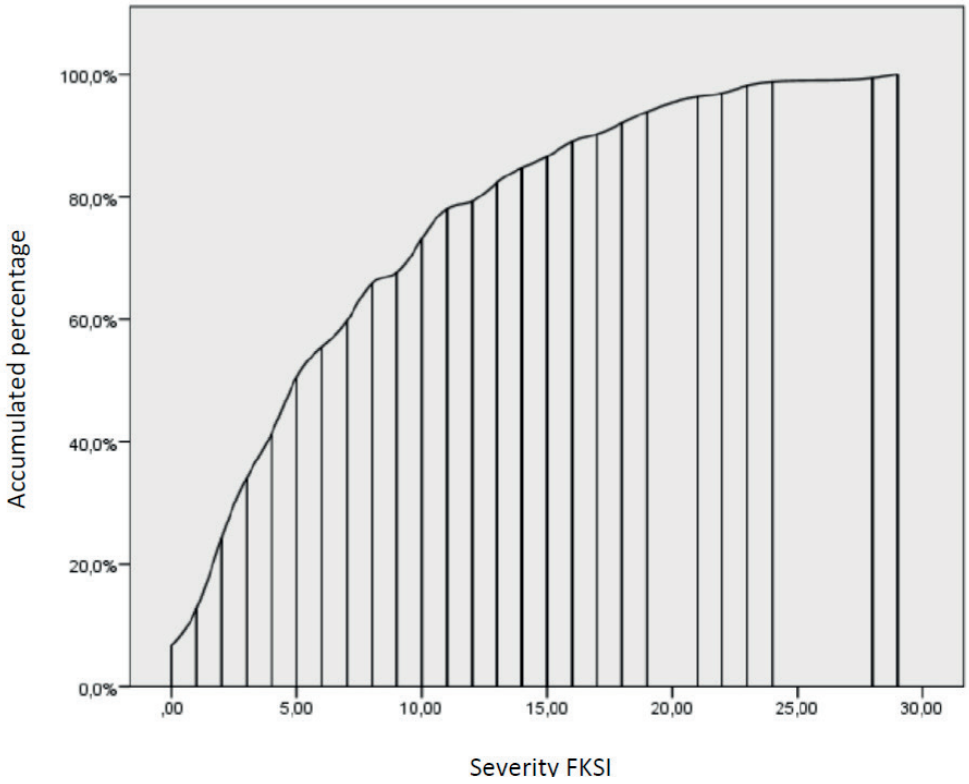


Figure 2. Cluster profiles showing average generic disutilities (EQ-5D, SF-6D, HUI-3) and symptom severity (FKSI).

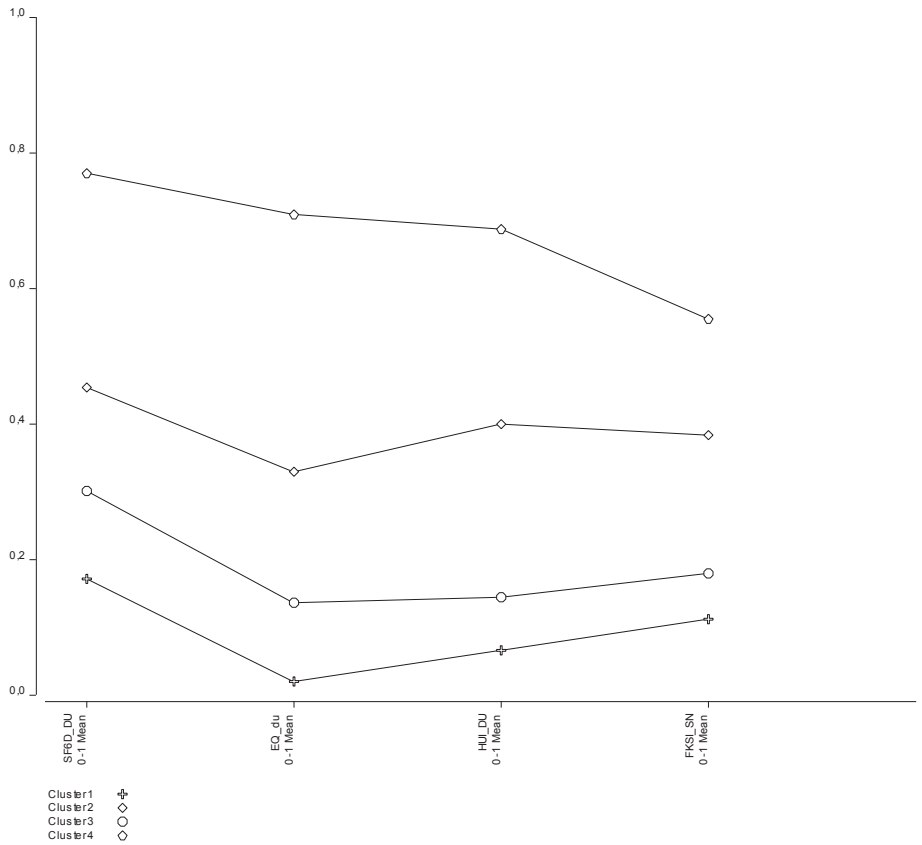


Figure 3. Outlier identification. EQ-5D, SF-6D and HUI-3 vs FKSI scatter-plots showing the fitted linear model with 95% individual confidence interval.

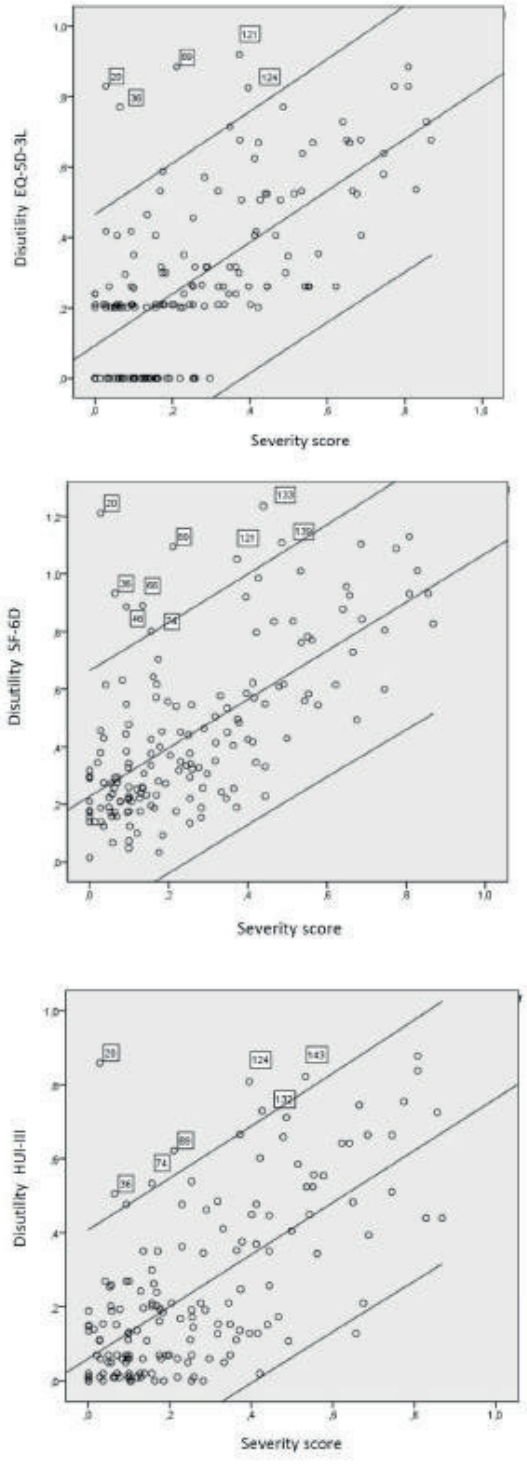
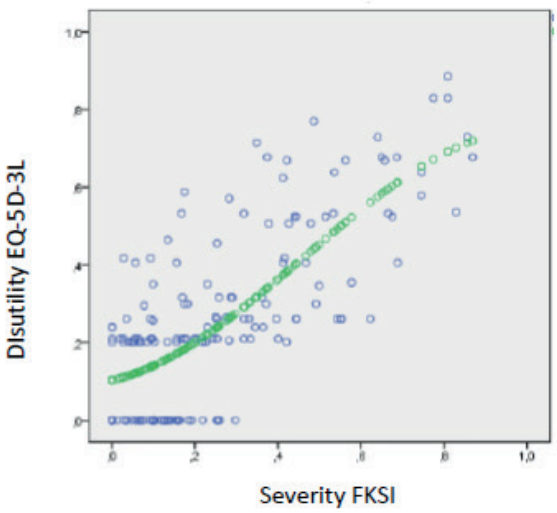
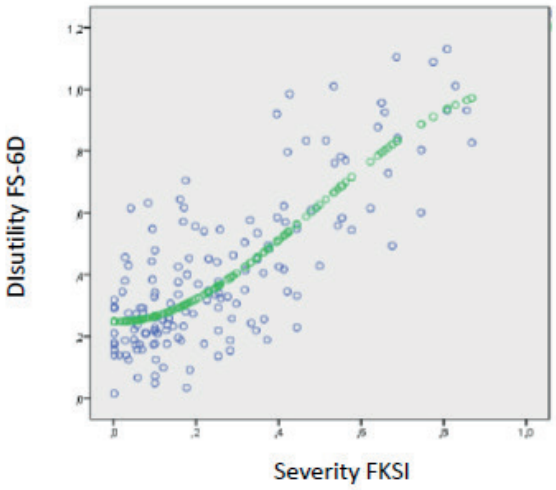
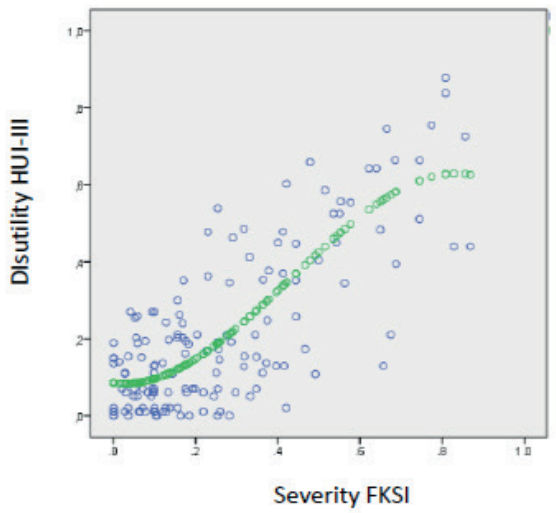


Figure 4. EQ-5D, SF-6D and HUI-3 observed (blue) and predicted (green) disutilities vs FKSI observed values.







REFFERENCES

[1] XLVII Spanish Nephrology Society abstracts [Resúmenes XLVII congreso de la Sociedad Española de Nefrología]. Nefrología 2017; 37(Supl. 1).

[2] Otero A, de Francisco ALM, Gayoso P, García F. Prevalence of chronic renal disease in Spain: Results of the EPIRCE study. Nefrología 2010; 30(1):78-86

[3] Ruiz MA, Pardo A (2005). Calidad de vida relacionada con la salud: definición y utilización en la práctica médica. *Pharmacoeconomics. Spanish research articles, 2 (1):* 31-43.

[4] Perales Montilla C, Duschek S y Reyes del Paso G. Calidad de vida relacionada con la salud en la enfermedad renal crónica: relevancia predictiva del estado de ánimo y la sintomatología somática. Nefrología 2016; 36(3):275–82.

[5] Álvarez-Ude, F. Factores asociados al estado de salud percibido (calidad de vida relacionada con la salud) de los pacientes en hemodiálisis crónica. *Revista de la Sociedad Española de Enfermería Nefrológica* 2001; 14: 64-8.

[6] Souza D, Bernal M. Incidencia, prevalencia y mortalidad del cáncer renal en España: estimaciones y proyecciones para el período 1998-2022. Actas Urológicas Españolas. 2012; 36(9):521-6.

[7] Rebollo-Rubio A, Morales-Asencio J, Pons-Raventos E, Mansilla-Francisco J. Revisión de estudios sobre calidad de vida relacionada con la salud en la enfermedad renal crónica avanzada en España. Nefrología 2015; 35(1):92-109.

[8] Wailoo, A. J., Hernandez-Alava, M., Manca, A., Mejia, A., Ray, J., Crawford, B., ... & Busschbach, J. (2017). Mapping to estimate health-state utility from non–preference-based outcome measures: an ISPOR good practices for outcomes research task force report. *Value in Health* 2017; 20(1):18-27.

[9] EuroQoL Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy 1990; 6(3):199-208.

[10] Badia X, et al. La versión española del EuroQol: descripción y aplicaciones. Med Clin (Barc) 1999; 112 (Supl. 1):79–86.

[11] Brazier J, Roberts J, & Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21(2):271-292.

[12] Abellan Perpinan J.M. Utilidades SF-6D para España. Guía de uso 2012/8. Sevilla: Catedra de Economía de la Salud. Universidad Pablo de Olavide. Consejería de Salud de la Junta de Andalucía 2012.

[13] Vilagut G. El cuestionario de salud SF36 español: una década de experiencia y nuevos desarrollos. Unidad de Investigación en Servicios Sanitarios. Institut Municipal d'Investigació Mèdica (IMIM-IMAS). Barcelona. España. *Gac Sanit.* 2005;19(2):135-50.

[14] Furlong W, Feeny D, Torrance GW, Goldsmith CH, DePauw S, Zhu Z, Denton M, Boyle M. Multiplicative Multi-Attribute Utilitiy Function for the Health Utilites Index Mark 3 (HUI3) System: A Technical Report. Centre for Health Economics and Policy Analysis. McMaster University. Paper 98-11 December 1988.

[15] Ruiz M, Rejas J, Soto J, Pardo A, Rebollo I. Adaptación y validación del Health Utilities Index Mark 3 al castellano y baremos de corrección en la población española. *Med Clin (Barc)* 2003;120(3):89-96.

[16] Cella, D., Yount, S., Brucker, P.S. et al. Development and validation of a scale to measure disease-related symptoms of kidney cancer. *Value Health*. 2007; 10: 285–93.

[17] C Cella, D., Yount, S., Du, H., Dhanda, R., Gondek, K., Langefeld, K., et al. Development and validation of the functional assessment of cancer therapy-kidney symptom index (FKSI). *J Support Oncol* 2006; *4*(4): 191-99.

[18] Butt, Z., Peipert, J., Webster, K., Chen, C., & Cella, D. General population norms for the functional assessment of cancer therapy–Kidney Symptom Index (FKSI). *Cancer* 2013; 119(2), 429-37.

[19] Ortega F, Rebollo P, Bobes J, González M, Saiz P. Interpretación de los resultados de la calidad de vida relacionada con la salud de pacientes en terapia sustitutiva de la insuficiencia renal terminal. Nefrología 2000; 20(5):431-9.

[20] Mukuria, C., Rowen, D., Harnan, S., Rawdin, A., Wong, R., Ara, R., & Brazier, J. An Updated Systematic Review of Studies Mapping (or Cross-Walking) Measures of Health-Related Quality of Life to Generic Preference-Based Measures to Generate Utility Values. *Applied health economics and health policy*, 2019; 1-19.

[21] Herdman, M., Gudex, C., Lloyd, A., Janssen, M. F., Kind, P., Parkin, D., et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research* 2011; 20(10): 1727-36.

[22] Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research* 2013; *22*(7), 1717-27.