Rev. salud pública. 17 (5): 789-800, 2015

DOI: http://dx.doi.org/10.15446/rsap.v17n5.3961

Modelo de simulación y minería de datos para identificar y predecir cambios presupuestales en la atención de pacientes con hipertensión arterial

Simulation and data mining model for identifying and prediction budget changes in the care of patients with hypertension

Luis Joyanes-Aguilar[1], Néstor J. Castaño[2] y José H. Osorio[3,4]

1 Escuela Superior de Ingeniería y Arquitectura, Universidad Pontificia de Salamanca. Madrid, España. joyanes@gmail.com
2 Facultad de Ciencias e Ingeniería, Universidad de Manizales. Colombia. njcastanop@gmail.com
3 Facultad de Ciencias para la Salud, Universidad de Caldas. Manizales, Colombia. jose.osorio_o@ucaldas.edu.co
4 Facultad de Ciencias para la Salud, Universidad de Manizales. Manizales, Colombia.

Recibido 23 Agosto 2013/Enviado para Modificación 4 Febrero 2014/Aceptado 9 Mayo 2015

RESUMEN

Objetivo Presentar un modelo de simulación en el cual se establece el impacto económico que, para el sistema de seguridad social, produce la evolución diagnóstica de pacientes asociados con la hipertensión arterial.
Métodos La información utilizada corresponde a la contenida en los Registros Individuales de Salud (RIPs). Se realizó una caracterización estadística y se planteó un modelo de almacenamiento matricial en Matlab. Se utilizó minería de datos para la elaboración de predictores y finalmente, se construyó un entorno de simulación para determinar el costo económico de la evolución diagnóstica.
Resultados La población que evoluciona desde el diagnóstico corresponde a un 5,7 % y el sobrecosto de producirlo es de 43,2 %.
Conclusiones Se abre la posibilidad para realizar investigaciones orientadas a establecer las relaciones diagnósticas dentro de toda la información reportada en los RIPs, con el fin de establecer indicadores econométricos que determinen cuáles son las evoluciones diagnósticas con mayor relevancia en el impacto presupuestal.

Palabras Clave: Simulación por computador, minería de datos, predicción, salud pública (*fuente: DeCS, BIREME*).

ABSTRACT

Objective To present a simulation model that establishes the economic impact to the health care system produced by the diagnostic evolution of patients suffering from arterial hypertension.

Methodology The information used corresponds to that available in Individual Health Records (RIPs, in Spanish). A statistical characterization was carried out and a model for matrix storage in MATLAB was proposed. Data mining was used to create predictors. Finally, a simulation environment was built to determine the economic cost of diagnostic evolution.
Results 5.7 % of the population progresses from the diagnosis, and the cost overrun associated with it is 43.2 %.
Conclusions Results shows the applicability and possibility of focussing research on establishing diagnosis relationships using all the information reported in the RIPS in order to create econometric indicators that can determine which diagnostic evolutions are most relevant to budget allocation.

Key Words: Computer simulation, data mining, forecasting, public health (*source: MeSH, NLM*).

Los países desarrollados han afrontado en las últimas décadas un incremento en el costo de operación de sus sistemas de atención en salud, por encima de su crecimiento económico (1). Esta situación ha llevado a plantear múltiples reformas al sistema de salud en diferentes países (2). En Colombia, la última gran reforma realizada en 1993, conocida como la Ley 100, destaca como su principal logro un incremento en el nivel de cobertura (3); sin embargo la calidad del servicio presenta un deterioro significativo (4). Adicionalmente, el sistema no se encuentra completamente financiado (5) y presenta un déficit acumulativo año a año (6). En el presente trabajo se propone y se prueba un modelo para realizar un estudio detallado del comportamiento estadístico y económico de enfermedades reportadas y tratadas dentro del Plan Obligatorio de Salud (POS). El modelo permite determinar dos elementos fundamentales: la relación diagnóstica y el costo incremental positivo o negativo producido por esta relación como lógica de su evolución. Desde el punto de vista económico, establecer el costo de la relación de evolución diagnóstica es de utilidad para el análisis de costo beneficio de las políticas de gestión administrativa o de promoción y prevención de la salud. La propuesta planteada está basada en el modelado y la simulación de eventos discretos, conceptos que han sido utilizados en el sector de la salud en áreas como sistemas de espera en salas de cirugía (7,8), análisis de costo en tratamientos de pacientes con diabetes (9), análisis de impacto de las políticas públicas en salud (10). El del presente estudio radica en la posibilidad de elaborar presupuestos económicos predictivos en el área de la salud en Colombia, basados en la unión de métodos actualmente empleados en el manejo de la información de salud pública como la minería de datos (11-13) y la simulación estadística.

METODOLOGÍA

Se modeló el comportamiento de pacientes de escasos recursos económicos cuyos costos clínicos son cubiertos por la División Territorial de Salud de Caldas. Se determinaron los intervalos de tiempo en que los pacientes requirieron de los servicios (cirugías, consultas generales, consultas especializadas, exámenes diagnósticos entre otros), su diagnóstico, cuántos ingresos al año se presentaron en las Instituciones Prestadoras de salud (IPS), edad, sexo, lugar de residencia y los costos de atención. La población analizada comprendió 7360 pacientes que solicitaron servicios clínicos en el periodo 2007-2008; sólo se incluyeron los pacientes que presentaban información, dentro de los 22 diagnósticos expuestos en la Tabla 1.

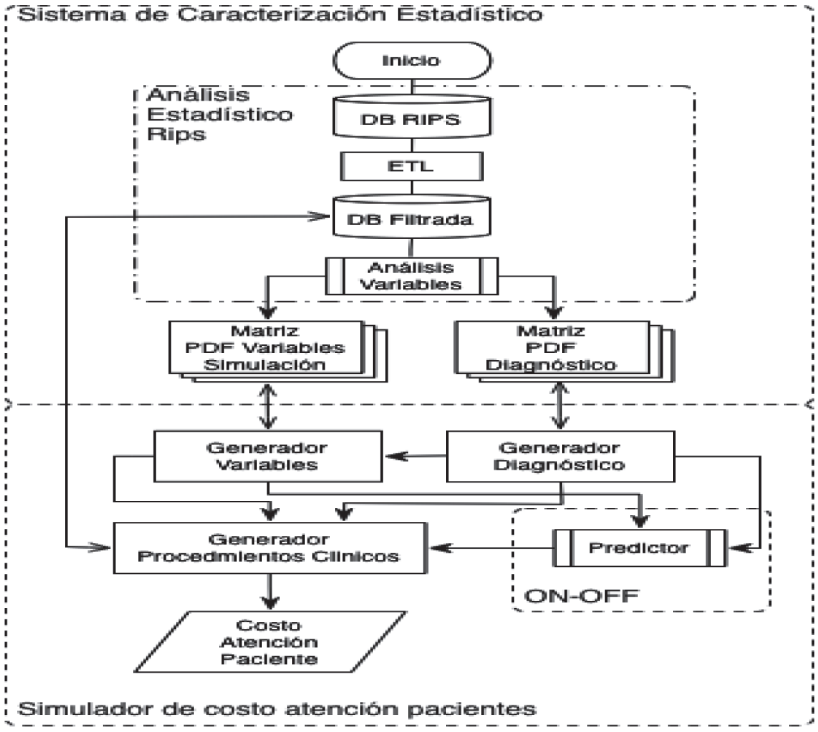
Tabla 1. Codificación CIE10 de los diagnósticos analizados

CIE10	Código real diagnóstico	CIE10	Código real diagnóstico
J459	Asma, no especificada	R55X	Síncope y colapso
R072	Dolor precordial	I255	Cardiomiopatía isquémica
R060	Disnea	I200	Angina inestable
I48X	Fibrilación y aleteo auricular	J440	Enfermedad pulmonar obstructiva
I10X	Hipertensión arterial primaria	N40X	Hiperplasia de la próstata
M069	Artritis reumatoide, no especificada	I500	Protocolo de insuficiencia cardíaca congestiva
E039	Hipotiroidismo, no especificado	I279	Enfermedad pulmonar del corazón, no especificada
I209	Angina de pecho, no especificada	I519	Enfermedad cardiaca, no especificada
J448	Otras enfermedades pulmonares obstructivas crónicas especificadas	J449	Enfermedad pulmonar obstructiva crónica no especificada
I250	Enfermedad cardiovascular aterosclerótica- así descrita	E119	Diabetes mellitus no insulinodependiente
E116	Diabetes mellitus no insulinodependiente con otras complicaciones	J441	Enfermedad pulmonar obstructiva crónica no especificada con exacerbación aguda

La selección de estos diagnósticos se basa en la relación estadística encontrada con la utilización de WEKA-Waikato Environment for Knowledge Analysis-, software de carácter académico utilizado en la minería de datos. La codificación diagnóstica en los RIPs es la CIE10 -Clasificación internacional de enfermedades, décima versión-, que es la codificación de enfermedades producida por la Organización Mundial de la Salud. Para la realización del entorno de simulación se plantearon dos fases, una de caracterización de la información y otra de implementación de los entornos de simulación.

La descripción general de la estructura del modelo se muestra en la Figura 1.

Figura 1. Modelo para la generación de presupuestos basados en la evolución diagnóstica



Análisis estadístico: Los RIPs (Registros Individuales de Salud Pública) son el subsistema de información prioritario para la evaluación y monitoreo del funcionamiento del Sistema General de Seguridad Social en Salud en Colombia), están conformados por 11 tablas (CT: Control, AF: Transacción, US: Usuarios, AD: Descripción Agrupada, AC: Consulta, AP: Procedimientos, AH: Hospitalización, AU: Urgencias, AN: Recién nacidos, AM: Medicamentos, AT: Otros servicios) de las cuales se extrajeron la siguientes variables:

Cédula de Ciudadanía, Sexo, Estrato Socioeconómico, Edad, Fecha de Autorización, Tipo de Autorización, Funcionario que Autorizó, Diagnóstico, Procedimientos Autorizados, Valor de los Procedimientos y Ciudad de Residencia.

La etapa inicial del proyecto consistió en la pre-validación de la

información existente en los archivos planos suministrados por la División Territorial de Salud de Caldas; después de esta etapa se procedió a enlazar Msql con Matlab, para cargar los datos y hacer la caracterización estadística. Una vez se establecieron las funciones de densidad de probabilidad de cada una de las variables analizadas se procedió al almacenamiento en forma matricial de las funciones de densidad que poseían un tipo de distribución probabilística multinomial (Ecuación 1).

Ecuación 1. Función de densidad de probabilidad multinomial

$$F(x_1 x_2 \ldots . x_n, p_1 p_2 \ldots . . p_3) = \binom{n}{x_1 x_2 \ldots . x_n} p_1^{x_1} p_2^{x_2} \ldots p_3^{x_3}$$

Las funciones de densidad de probabilidad de tipo normal, como la edad de los pacientes se generaron con funciones pre-establecidas en Matlab (Ecuación 2).

Ecuación 2. Función de densidad de probabilidad normal

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

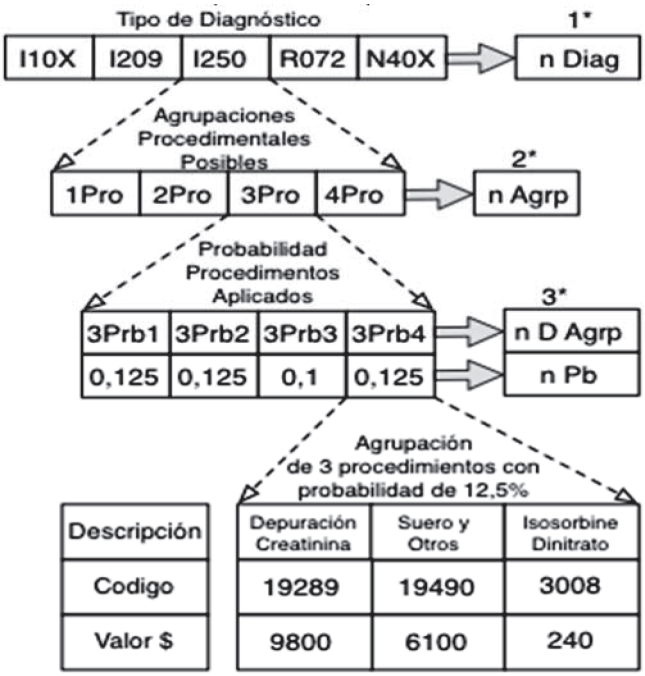
Matriz de Probabilidad de los Diagnósticos Analizados: En esta matriz se encuentra la probabilidad de los 22 diagnósticos utilizados en este estudio, todos codificados utilizando el CIE10. Esta matriz se utilizó para alimentar la función multinomial que genera el diagnóstico para cada paciente simulado.

Matriz de probabilidad de las variables asociadas a cada diagnóstico: Se planteó la utilización de una matriz jerárquica basada en agrupaciones utilizadas por otros autores (14). Las agrupaciones se realizaron por diagnóstico, número de procedimientos y tipo de procedimientos clínicos aplicados de acuerdo al número. En la primera dimensión cada columna representa un diagnóstico específico y las filas representan la variable que se desea generar dentro del diagnóstico.

En la Figura 2, se realiza una descripción detallada de la forma como se estructuró el almacenamiento de los vectores de probabilidad de los procedimientos clínicos aplicados.

En el primer nivel de la estructura de almacenamiento se forma una matriz de orden A{1xN}, donde N corresponde al número de diagnósticos analizados. En el segundo nivel A{1x1..n3..N}{1x1..k3..K} en donde K corresponde al máximo número de procedimientos aplicados a un paciente en un ingreso a una institución prestadora de salud – IPS -, dentro de un diagnóstico que corresponde a la posición A{1xn3} que corresponde a I 250 (enfermedad cardiovascular aterosclerótica). Dentro del elemento A{1xn3} {1xk3}{1x1..j4..J} se encuentra el vector de densidad de probabilidad correspondiente a las diferentes agrupaciones posibles de tres procedimientos clínicos, donde J corresponde al número máximo de diferentes agrupaciones de n3 procedimientos.

Figura 2. Estructura de almacenamiento matricial de los procedimientos clínicos aplicados a los pacientes



Simulador de costo/atención paciente: La simulación de eventos discretos utilizada en este proyecto se trabajó utilizando los modelos planteados por otros autores (15,16). El espacio temporal en que se realizaron todas las mediciones y la simulación es de un año. Las variables

de entrada requeridas para iniciar la simulación son el número de pacientes que se van a simular y la segunda una entrada tipo ON-OFF que activa o no la utilización del predictor. Los procesos que se realizaron durante la simulación se describen a continuación.

Generador diagnóstico: Se genera un número aleatorio entero con distribución multinomial correspondiente a uno de los 22 diagnósticos caracterizados, consultando el vector de probabilidad almacenado.

Generador de variables: En este proceso se generan las variables que se muestran en la tabla edad, sexo, estrato social, número de ingresos a las IPS al año, lugar de nacimiento, lugar de residencia.

Generador de procedimientos clínicos: Se genera el conjunto de procedimientos que se le aplican a un paciente cuando ingresa a una IPS, por ejemplo: en un solo ingreso se le puede realizar una consulta general, en otro una consulta especializada y al tiempo ser hospitalizado, medicado, examinado o remitido a otra IPS. Estos son procedimientos que se le aplican a un paciente con un diagnóstico asignado con anterioridad. Para esta generación aleatoria se aprovecha el tipo de almacenamiento utilizado en la matriz de probabilidad de las variables asociadas a cada diagnóstico; las variables que intervienen son: el tipo de diagnóstico, el número de procedimientos y la agrupación de procedimientos típicos de cada diagnóstico.

Predictor: Con el propósito de analizar la relación existente entre las variables se realizaron diversos experimentos en WEKA (Waikato Environment For Knowledge Analysis). Para establecer la evolución diagnostica se determinó cuántos y cuáles pacientes tenían el diagnóstico I200 (angina inestable), I 500 (Protocolo de insuficiencia cardiaca congestiva), J 449 (Enfermedad pulmonar obstructiva crónica no especificada). En el año 2008, se estableció cuántos habían cambiado de diagnóstico con respecto al año inmediatamente anterior y se determinó qué tipo de diagnóstico presentaban antes de la evolución. Se encontró una población de 247 pacientes de los cuales se disponían 1 692 registros. Experimentado con clasificadores de tipo Bayesiano (17), meta-clasificadores tipo Bagging y diferentes árboles de decisión, se seleccionó un predictor tipo RepTree tipo árbol de decisión, con un porcentaje de clasificaciones correctas de 84,2 %. Los resultados de la clasificación se describen en la Tabla 2.

Tabla 2. Indicadores Estadísticos de la Clasificación REPtree

Indicador	Valor
Correctamente Clasificados	84,2 %
Incorrectamente Clasificados	15,8 %
Kappa	0,729
Error Medio Absoluto	0,14
Error Cuadrático Medio	0,27
Error Absoluto Relativo	34,73 %

Costo de atención paciente: En este proceso se realizó la acumulación discriminada por diagnóstico de los costos de atención de cada paciente simulado, por último se produjo un informe para establecer comparativos.

Validación del simulador: El procedimiento que se utilizó para validar los datos generados por el simulador fue el siguiente:

Se generó un número de 7 360 pacientes igual al máximo disponible en la base de datos durante los años 2007 y 2008 para los 22 diagnósticos analizados.

Se utilizaron los datos no simulados para obtener los histogramas de las diferentes variables, a los que se les denominó valor real. Para la variable costos se totalizaron los costos reales y se compararon los simulados.

Se aplicó la prueba de bondad de ajuste Chi-cuadrado para validar los datos simulados.

Simulación: se simuló un crecimiento en el número de pacientes que evolucionan de diagnóstico, se inició en 100 y se terminó en 1 000, incremento que se utilizó para visualizar el grado de influencia individual de cada uno de los diagnósticos en los costos. La simulación calcula la diferencia económica existente entre la evolución y la permanencia en el diagnóstico simulado.

RESULTADOS

En la Tabla 3 se muestran los resultados de validación de la hipótesis nula para las variables edad y número de ingresos para el diagnóstico I 10x de hipertensión básica. La prueba se aplicó con éxito a todas las variables simuladas. Por último, se corrió la simulación 10 veces y se promedió el costo económico para cuatro diagnósticos, generando aleatoriamente un total de 7 360 pacientes. Posteriormente se compararon estos datos con

lo existentes en la base de datos suministrada y se obtuvo un intervalo de confianza superior al 95 %.

Tabla 3. Prueba Chi-Cuadrado para la validación de las variables edad y números de ingresos anuales simulados, población con diagnostico I10x

Indicador	Edad	No. Ingresos
H	0	0
P	0,3850	NAN
X2	75,9055	0,1797
DF	73	1

En la Tabla 4, se comparan los costos de atención producidos en dos escenarios de simulación, en el primero se asume que los pacientes no evolucionan y permanecen en su diagnóstico y en el segundo se simula la evolución a los diagnósticos I200, I500, J449. Se puede apreciar que la evolución de los pacientes produjo aumento presupuestal de un 43,2 %; este cálculo sirve para realizar apropiaciones presupuestales más confiables y definir una política de prevención con sus respectivos niveles de control.

Tabla 4. Estimación de costos de atención ($) registrados en la Tabla AP de RIPs para 420 pacientes sin evolución y con evolución diagnóstica

Nro. de SIM	Costo atención sin evolución	Costo atención con evolución
10	28 220 168	37 549 279
20	27 284 549	38 394 405
100	27 .202 559	38 337 141
200	27 059 004	38 763 005
500	26 993 425	38 219 738

DISCUSIÓN

En el presente estudio, se estableció un modelo que permite calcular el costo de la evolución diagnóstica. Se planteó un caso de estudio realizado con las enfermedades asociadas a la hipertensión arterial, en el que se tasó el costo de la evolución diagnóstica en 43,2 %. Los resultados abren la posibilidad de realizar investigaciones orientadas a establecer las relaciones diagnósticas dentro de toda la información reportada en los RIPs, con el fin de establecer indicadores econométricos que determinen cuales son las evoluciones diagnósticas con mayor relevancia en el impacto presupuestal. Algunos autores (18), plantean que el principal problema del sector de la salud no son los sistemas de aseguramiento o las políticas, sino la medición errónea de las cosas. Inmersos en el análisis de costos se pueden encontrar en la literatura diferentes estudios analíticos (19), donde existe un grupo al que se le aplica un tratamiento clínico o se le suministra un medicamento y otro grupo de control el cual no se interviene. Este tipo de investigaciones

son importantes para validar la influencia de una variable en los costos de atención, pero su debilidad radica en que no se presenta un modelo en el cual se puedan realizar predicciones. En otro tipo de análisis de costos (20) el enfoque principal radica en cuál es la participación porcentual de diferentes variables o programas en los costos de presupuestos en salud. Algunos autores (21) presentan una investigación que posee relación con el estudio acá planteado. Se utiliza la minería de datos con el método de razonamiento basado en casos, para determinar los sobrecostos generados por diagnósticos elaborados por médicos practicantes. Conocer si un diagnóstico influye en los costos define una prioridad; muchos estudios terminan en este punto, pero conocer cuáles variables son las relevantes en la producción del diagnóstico, cómo interactúan, cuál es su gestión y evolución, mejoran el plan de intervención y la minería de datos es crucial en esta tarea. La posibilidad de predecir el comportamiento presupuestal que implica la evolución diagnostica de un paciente, muestra que la simulación y la minería de datos son herramientas adecuadas para la realización de presupuestos en salud. Las diferencias entre otros autores (21) y el trabajo presentado en éste artículo, son los tipos de variables utilizadas; ya en uno se analiza la calidad del diagnóstico y en el otro se acepta el diagnostico como una verdad y se analizan los costos en los procedimientos practicados.

Por otro lado en uno se utilizan algoritmos de razonamiento basado en casos y en otro se utilizan árboles de decisión y finalmente, en ambos estudios se utiliza la minería de datos para establecer clasificadores; en el primero para determinar si las variables encontradas corresponden al diagnóstico, y en el otro para determinar los diagnósticos que evolucionan y a qué evolucionan, pero la gran diferencia radica en que en el último, se construye un entorno de simulación para la toma de decisiones, y éste hecho constituye un avance valioso. El entorno de simulación permite calcular el costo de la evolución diagnostica, pero podría ser utilizado para determinar qué sucedería si se aumenta o se disminuye la media de edad de la población; así como se manipula la variable edad, podría manipularse cualquier otra existente en el modelo y establecer la influencia de esta variación en los costos.

El almacenamiento multidimensional, matricial y jerárquico que se puede realizar en Matlab, es de gran utilidad para la asociación entre la generación numérica de variables aleatorias y su correspondencia dentro de una posición matricial, que puede contener información de tipo numérico,

o de carácter gráfico. El cambio de diagnósticos asociados a I 500, I 200, J 449 que se producen año a año, sí produce un aumento en el costo de los servicios de salud. Esta investigación produjo como resultado el diseño, desarrollo de software y puesta en marcha de procesos que permiten realizar apropiaciones presupuestales más precisas, basadas en análisis estadísticos, modelos de clasificación y predicción ♠

Agradecimientos: Los autores agradecen a la División Territorial de Salud de Caldas por aportar los datos para la investigación y al grupo de investigación en Informática y Telecomunicaciones de la Universidad de Manizales, por proporcionar los recursos técnicos necesarios.

REFERENCIAS

1. Tavakoli M, Davies H. Cost and Efficiency within Health Care Systems. Health Care Management Science. 2004; 7: 5-6.
2. Cavagnero E. Health sector reforms in Argentina and the performance of the health financing system. Health Policy. 2008; 88: 88-9.
3. Morales L. Financiamiento del Sistema de Seguridad Social en Salud en Colombia. Proyecto Cepal. Santiago/Chile: Naciones Unidas. 1997; 55: 60-71.
4. Echeverry E. La Salud en Colombia: Abriendo el Siglo 21 y la Brecha de las Inequidades. Gerencia y Políticas de Salud. 2002; 3: 76-4.
5. Clavijo S, Peña M. Reformas al régimen de salud: entre la emergencia social y el déficit estructural. Rev de Asuntos Publicos. 2010; 5: 4-8.
6. Clavijo S, Vera A. Los Desafios Fiscales de Colombia (2010-2014). ANIF. Bogota: ANIF; 2010. pp 1-10.
7. Morton A, Bevan G. What's in a wait? Contrasting Management Science and Economic Perspectives on Waiting for Emergency Care. Health Policy. 2008; 85: 207-17.
8. Lamiri, M., Grimaud, F., Xie, X.: Optimization Methods for a Stochastic Surgery Planning Problem. International Journal of Production Economics, Special Issue on Introduction to Design and Analysis of Production Systems 2009; 120(2), 400–410.
9. Girod I, Valensi P, Laforêt C, Moreau-Defarges T, Guillon P, Baron F.An Economic Evaluation of the Cost of Diabetic Foot Ulcers: Results of a Retrospective Study on 239 Patients. Diabetes & Metabolism. 2003; 29: 1262-636.
10. Johansson P. Economic Evaluation of Public Health Programmes - Constrains and Opportunities. Sthockholm :Deparment Public Health Sciences. 2009; 1: 22-33.
11. Gosain A, Kumar A. Analysis of Health Care Data Using Different Data Mining Techniques. Intelligent Agent & Multi-Agent Systems. 2009: 1-6.
12. Sherafat K. Interoperability of Data and Knowledge in Distributed Health care Systems. IEEE International Workshop on Software Technology and Engineering Practice; 2005. pp. 1-10.
13. Mcshea M, Holl R, Badawi O. A Collaboration Between Industry, Health-Care Providers, and Academia. IEEE Engineering in Medicine And Biology Magazine. 2010; 18-5.

14. Codrington A, Chaussalet T, Millard P, Whittlestone P Kelly J.A System for Patient Management Based Discrete-Event Simulation and Hierarchical Clustering. IEEE Computer Society, Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems; 2006. pp. 1-6.
15. Anderson G. Evaluation in health informatics: computer simulation. Computers in Biology and Medicine. 2002; 32: 151-64.
16. Rodríguez BJM, Serrano D, Monleón T, Caro J. Discrete-event simulation models in the economic evaluation of health technologies and health products. Gaceta Sanitaria. 2007; 22: 151-61.
17. Ueno K, Hayashi T, Iwata K, Honda N, Kitahara Y, Paul TK. Prioritizing Health Promotion Plans with k-Bayesian Network Classifier. Seventh International Conference on Machine Learning and Applications; 2008. p. 10-15.
18. Jerrell J, McIntyre R. Health-Care Costs of Pediatric Clients Developing Adverse Events during Treatment with Antipsychoticsv. Value In Health. 2009; 12: 716-22.
19. Guerrero I. Assessing the Economic Value of Public Health Programs Based on Risk: The Case of the Cancer Plan in France. Value In Healt. 2010; 13: 552-56.
20. Clarke P, Leal J, Kelman C. Estimating the Cost of Complications of Diabetes in Australia Using Administrative Health-Care Data. Value In Health. 2008; 199-06.
21. Zhuang Z, Churilov L, Burstein F. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research. 2007; 662-75.