Ediciones pensando el futuro
Editions thinking the future
Informática 2013
XV CONVENCIÓN Y FERIA INTERNATIONAL
15TH INTERNATIONAL CONVENTION AND FAIR
www.informaticahabana.com
www.informaticahabana.cu
ISBN: 978-959-7213-02-4

SLD 161 CARACTERIZACIÓN Y ANÁLISIS DE LA BASE DE DATOS DE CÁNCER DE MAMA SEER-DB

SLD 161 CHARACTERIZATION AND ANALYSIS OF THE DATABASE OF BREAST CANCER SEER-DB

Guillermo Gilberto Molero-Castillo[1], Yaimara Céspedes González[2], María Elena Meda Campaña[3]

1 Doctorado en Tecnologías de Información, guillemolero@comunidad.unam.mx, Universidad de Guadalajara, México

2 Desoft S.A., Ministerio de la Informática y las Comunicaciones, MIC, yaimara@mic.cu, Cuba

3 Centro de Investigación en Sistemas y Gestión de la Información, Universidad de Guadalajara, emeda@cucea.udg.mx, México

RESUMEN: *En el presente trabajo se describe la caracterización de la fuente de datos relacionada al cáncer de mama en pacientes mujeres de origen hispano. Así como el proceso de análisis empleado con estas bases de datos para determinar la calidad de la serie de datos y las variables significativas a emplearse en el proceso de predicción del caso de estudio, esto a través de técnicas de Minería de Datos. La principal consideración fue determinar cuántas y cuáles son las variables oncológicas apropiadas para el estudio. Asimismo, se analizó la variabilidad y distribución de las principales variables oncológicas registradas en las bases de datos disponibles.*

Palabras Clave: Base de Datos, Cáncer de mama, Minería de Datos, Series de tiempo.

ABSTRACT: *This paper describes the characterization of data source related to breast cancer in women patients of Hispanic origin. As well as the process of analysis used with these databases to determine the quality of the data series and the significant variables used in the prediction process of the case of study, this through Data Mining techniques. The main consideration was to determine how many and which are the appropriate clinical-oncological variables for the study. Also the variability and distribution of the main clinical-oncological variables recorded in the databases available were analyzed.*

KeyWords: Database, Breast Cancer, Data Mining, Time Series.

1. INTRODUCCIÓN

En la última década se ha observado un incremento considerable en la aplicación de Minería de Datos a problemas relacionados con series de tiempo [1]. Dichos trabajos han sido orientados principalmente al *agrupamiento* (*ej.* análisis de la causa de muerte de pacientes), *clasificación* (*ej.* predicción del consumo de fármacos), *detección de anomalías* (*ej.* análisis de historias clínicas para la identificación de enfermedades), *síntesis y descubrimiento de reglas* (*ej.* identificación de patologías).

Las series de datos temporales son un caso particular de patrones secuenciales, su análisis ofrece una valoración de la estacionalidad de la serie, describiendo las oscilaciones de los datos con relación a un valor promedio e identificando la presencia de posibles tendencias [2].

En [3] se define como una serie de datos temporales al "*conjunto de valores ordenados cronológicamente que permiten predecir y describir el comportamiento de una o más variables en un determinado*

periodo". Algunas veces estas series pueden ser muy extensas, conteniendo billones de observaciones [4]. En las series temporales se identifican cuatro tipos de patrones [3]: tendencia, variación estacional, accidental y cíclica.

La *tendencia* (T) refleja la evolución de la serie durante un determinado periodo. Este periodo varía según la naturaleza de la serie, el cual puede ser estacionario o constante, lineal, exponencial u otras. La *variación estacional* (S) es el comportamiento que agrupa las oscilaciones repetitivas en periodos de tiempo. Estos periodos pueden ser estaciones del año, días, meses, bimestres, trimestres, semestres, años, entre otros. Mientras que la *variación accidental* (A) es un patrón que corresponde a las fluctuaciones accidentales que se dan por la ocurrencia de fenómenos imprevisibles, como la presencia de huracanes, que afectan a la variable en estudio de manera esporádica y no permanente. También es conocido como variación irregular. Por su parte, la *variación cíclica* (C) se presenta cuando los datos reflejan oscilaciones periódicas no regulares, ocasionadas por asumir periodos no establecidos. Generalmente aparecen en series de datos climatológicos, por ejemplo en ciclos de sequía.

En este trabajo se describe la caracterización y el análisis de la fuente de datos relacionada al cáncer de mama en pacientes mujeres de origen hispano. Los datos analizados corresponden a series de datos provenientes de la Base de Datos del Programa de Vigilancia, Epidemiología y Resultados Finales (SEER) del Instituto Nacional del Cáncer (NCI) de los Estados Unidos. La principal consideración fue determinar cuántas y cuáles son las variables oncológicas apropiadas para el estudio. Asimismo, se analizó la variabilidad y distribución de las principales variables establecidas en la Base de Datos SEER, como: Origen del paciente, Edad del paciente, Año de diagnóstico, Estado civil del paciente, Tipo de la enfermedad y Confirmación del diagnóstico. Por último, se dan a conocer algunas consideraciones y conclusiones finales.

2. FUENTE DE DATOS

La fuente de datos a partir de la cual se realizó el proceso de análisis del cáncer de mama en pacientes mujeres de origen hispano, fueron datos provenientes de la Base de Datos del Programa de Vigilancia, Epidemiología y Resultados Finales (Surveillance, Epidemiology and End Results; SEER por sus siglas en inglés) del Instituto Nacional del Cáncer (NCI) de los Estados Unidos.

El Programa de Vigilancia, Epidemiología y Resultados Finales (SEER) es el responsable del registro nacional del cáncer y la principal fuente de información autorizada para esta enfermedad en los Estados Unidos. Se encarga de la recopilación de la información sobre casos de cáncer diagnosticados (incidencia), sobre las muertes atribuidas a esta enfermedad (mortalidad) y la supervivencia de pacientes con cáncer. Esto con el fin de comprender y abordar el cáncer en la población de los Estados Unidos. En la actualidad, son diversas las investigaciones que se realizan a través del uso de los registros del cáncer, los cuales están a disposición de investigadores, médicos, funcionarios de salud pública, legisladores, políticos, grupos de investigación y público en general; esto con el fin de [5]:

- Monitorear las tendencias del cáncer con el paso del tiempo.
- Mostrar patrones del cáncer en distintas poblaciones.
- Apoyo para establecer prioridades en la asignación de recursos.
- Guiar la planeación y evaluación de programas para el control del cáncer.
- Promover actividades de investigación en el área médica y de epidemiología.

Así, la información sobre casos de cáncer y muertes por esta enfermedad es crucial para elaborar informes sobre las tendencias del cáncer, determinar si los esfuerzos de prevención y control son eficaces, propiciar la participación en investigaciones y se emprendan acciones cuando se reporten posibles aumentos en la incidencia del cáncer.

2.1 Base de Datos SEER

En 1973, el Programa SEER comenzó a reunir y registrar datos sobre diversos casos de cáncer en los estados de Connecticut, Iowa, Nuevo Mexico, Utah, Hawai y áreas metropolitanas de Detroit, San Francisco y Oakland [6]. En los últimos 30 años, SEER ha añadido más poblaciones a la lista de vigilancia y ahora existen millones de casos registrados en la base de datos. Abarcando, en la actualidad, aproximadamente el 28 % de la población de los Estados Unidos [7].

En los registros de la Base de Datos SEER se recopilan datos demográficos del paciente, localización del tumor primario, morfología del tumor, etapa del cáncer al momento del diagnóstico, tratamiento, seguimiento de la enfermedad, entre otros. La obtención y registro de los datos se da a través de establecimientos médicos, como hospitales, consultorios y laboratorios de patología, que envían información sobre los casos evaluados a sus respectivos registros estatales de cáncer [8]. Por lo general, la mayor parte de la información proviene de hospitales, donde empleados autorizados, llamados registradores, transfieren la información de las histo-

rias clínicas de los pacientes a bases de datos locales, para posteriormente ser enviado al registro central del cáncer [9].

2.2 Acceso a la Base de Datos SEER

La Base de Datos SEER fue adquirida a través de un acuerdo, firmado y enviado, de confidencialidad para el acceso a la versión actual de la fuente de datos. La cual fue proporcionada por el Programa SEER a través de dos vías: el *primero* mediante el envío del disco DVD-SEER, que contiene información relacionada al cáncer de mama y otros tipos de cáncer; y la *segunda* mediante la descarga de archivos comprimidos, disponibles en uno de los servidores del Instituto Nacional del Cáncer de los Estados Unidos. Para descargar estos archivos fue necesario un nombre de usuario y contraseña proporcionados por el SEER.

El acceso a estas fuentes de datos permitió comprobar que los archivos comprimidos en formato ZIP presentan el mismo contenido de registros que se incluyen en el DVD-SEER. Así, el objetivo de tener estas dos fuentes de datos se basa fundamentalmente en el propósito de realizar un exhaustivo análisis de los datos de casos de cáncer de mama diagnosticados en pacientes mujeres de origen hispano que ofrece el SEER.

Se observa que en la actualidad el total de campos o variables establecidas en la base de datos SEER es de 124. Las cuales no sólo se utilizan para registrar datos sobre casos de cáncer de mama, sino también para almacenar otros tipos de cáncer, como: pulmón, estómago, esófago, ovario, próstata, hígado, páncreas, colon, entre otros.

3. ANÁLISIS DE LA FUENTE DE DATOS

El análisis de datos es una de las actividades fundamentales en el proceso de Minería de Datos, mediante el cual se establece el contacto directo con el problema a resolver. El análisis de las bases de datos disponibles se realizó en dos etapas.

La *primera* consistió en una revisión y análisis preliminar del total de variables listadas en la Base de Datos SEER, esto con el fin de establecer aquellas relevantes en función del periodo de sus registros; descartando las variables que presentan una escasa o nula cantidad de registros disponibles. La Base de Datos SEER tiene registros a partir de 1973 por lo que se hizo el análisis a partir de esa fecha.

En la *segunda* etapa se determinó la calidad de la serie de datos para establecer las variables significativas y el periodo de años que se empleará en el proceso de predicción de la supervivencia y mortalidad del cáncer de mama en pacientes mujeres de origen hispano.

3.1 Análisis preliminar de la Base de Datos SEER

Para el análisis preliminar del total de las variables establecidas en la base de datos SEER, se consideró aquellas variables con suficientes registros disponibles, esto es, que tuvieran por lo menos más del 50% de datos registrados a lo largo de 1973-2008, que es el periodo de registro establecido en las bases de datos proporcionadas por el SEER. Esto con la finalidad de tener una amplia representación de variables con periodos similares, descartando las que presentan alta cantidad de registros faltantes.

Así, se analizó la frecuencia de registros en las 124 variables. Para esto, se organizaron los archivos de la fuente de datos de texto plano (bases de datos textuales independientes) en un archivo único (BreastCancer.txt). Esto con el propósito de importar la serie de datos a una tabla específica de una base de datos que fue caracterizada en SQL Server. Con el fin de concentrar todas las variables y registros de casos de cáncer de mama en una sola fuente de datos.

En general, el total de archivos de texto plano que se integraron (compilaron) fueron cuatro, todos llamados "BREAST.TXT", los cuales fueron hallados en carpetas definidas como: yr1973_2008.seer9, yr1992_2008.sj_la_rg_ak, yr2000_2008.ca_ky_lo_nj y yr2005.lo_2nd_half, con un total de 630,218; 140,829; 269,286 y 1403 registros de casos de cáncer de mama, respectivamente.

Así, el proceso para importar los registros del archivo único "BreastCancer.txt", a una base de datos, fue realizado a través de la siguiente función definida en SQL Server:

```
BULK INSERT SEER..BREASTCANCER FROM
'c:\DataSEER\BreastCancer.txt'
WITH (ROWTERMINATOR = '\n')
```

Una particularidad que se observó en las series de la Base de Datos SEER fue que los campos no estaban separados por espacios, tabuladores, comas o cualquier otro carácter que permitiera identificar y organizar cada una de las variables. Esto se debe fundamentalmente a la necesidad, por parte del SEER, de reducir el tamaño de los archivos que contienen grandes volúmenes de datos, con el fin de hacerlos portables y de fácil transferencia, como es el caso de los registros del cáncer de mama.

Ante esta situación, una vez importadas las series de datos y con base en la información proporcionada por el SEER sobre el *nombre de la variable*, la *posición* y la *longitud* de cada una de las variables dentro de la fuente de datos, se elaboró una función definida en SQL Server para la separación y organización de las variables dentro de una sola tabla, quedando conformada por 124 campos y un total de 1'041,736 registros.

Los 1'041,736 registros corresponden a todos los casos de cáncer de mama que fueron registrados desde 1973 a 2008 para todos los tipos de raza y origen clasificados por el Instituto Nacional del Cáncer y Programa de Vigilancia, Epidemiología y Resultados Finales (SEER) de los Estados Unidos. Por lo que, en función del objetivo de este trabajo, se filtraron sólo los datos de pacientes de origen hispano, quedando en total 67156 registros.

En general, como producto del análisis preliminar de la cantidad de registros de las 124 variables, en función del total disponible (67156), se pudo observar que 73 cuentan con suficientes registros disponibles, esto es, tienen porcentajes por encima del 50 % de datos registrados a lo largo de 1973-2008. Mientras que las otras variables, en total 51, registran una alta cantidad porcentual de registros faltantes o no disponen de datos.

3.2 Análisis de la calidad del conjunto de datos

La segunda etapa de análisis consistió en el estudio de la calidad de la serie de datos, asociado fundamentalmente a la cantidad de registros válidos continuos en un determinado periodo, con la finalidad de determinar las variables significativas y el periodo de datos que se emplearán en el proceso de predicción de la supervivencia y mortalidad del cáncer de mama en pacientes de origen hispano.

Para este proceso, y con base en los resultados obtenidos del análisis preliminar, se examinaron los registros disponibles de cada una de las variables, en total 73. El análisis consistió en seleccionar aquellas variables significativas, que tienen relación directa con el cáncer de mama, con registros suficientes en periodos consecutivos y bajo la opinión de oncólogos especialistas; esto con el fin de reforzar la investigación sobre la supervivencia y mortalidad del cáncer de mama en pacientes de origen hispano. Se logró la participación de estos especialistas a través de un compromiso de colaboración académica y la realización de una estancia de investigación, ambas realizadas durante el periodo 2011 y 2012.

Así, para el análisis de los datos y la selección de las variables, se establecieron las siguientes consideraciones:

a. La variable debe tener relación directa con el Cáncer de Mama y no con otros tipos de cáncer, que también son registrados por el Programa SEER.
b. Cada variable debe tener por lo menos 4 años de datos consecutivos, a partir de 1973, hacia adelante. Por lo que se tomó como base el año más cercano o próximo con datos. Por ejemplo: 1973-1977, 1974-1978, 2004-2008.
c. La variable analizada, además de tener los 4 o más años consecutivos de datos, debe presentar por lo menos el 90 % de registros válidos consecutivos, esto es, para cada variable se acepta la existencia de hasta un 10 % de registros nulos y/o faltantes.
d. Aunado a las consideraciones anteriores, y con el fin de reforzar la investigación, la selección de las variables estuvo sujeta a la opinión y consideraciones de médicos especialistas, basado en sus experiencias, sobre la importancia y contribución de las variables en procesos de diagnósticos clínicos y predicción de escenarios médicos asociados con la enfermedad.

Al aplicar las consideraciones anteriores, se observó que 35 de las 73 variables cumplen con los criterios establecidos (Tabla I).

Tabla I: Variables seleccionadas consideradas significativas

No.	Variable	Año de Inicio	Año Final	Años Fav.	Válidos	% Acept.
1	CASENUM	1973	2008	35	67156	100
2	REG	1973	2008	35	67156	100
3	MAR_STAT	1973	2008	35	67156	100
4	ORIGIN	1973	2008	35	67156	100
5	NHIA	1973	2008	35	67156	100
6	AGE_DX	1973	2008	35	67156	100
7	YR_BRTH	1973	2008	35	67152	99.9
8	PLC_BRTH	1973	2008	35	67156	100
9	SEQ_NUM	1973	2008	35	67156	100
10	DATE_mo	1973	2008	35	67156	100
11	DATE_yr	1973	2008	35	67156	100
12	LATERAL	1973	2008	35	67156	100
13	BEHO3V	1973	2008	35	67156	100
14	GRADE	1973	2008	35	67156	100
15	DX_CONF	1973	2008	35	67156	100
16	REPT_SRC	1973	2008	35	67156	100
17	EOD10_SZ	1988	2003	15	35698	100
18	EOD10_PN	1988	2008	20	63400	100
19	EOD10_NE	1988	2008	20	63400	100
20	CS_SIZE	2004	2008	4	27702	100
21	D_AJCC_S	2004	2008	4	27702	100
22	SURGPRIM	1998	2008	10	50672	100
23	NO_SURG	1973	2008	35	67156	100
24	RADIATN	1973	2008	35	67156	100
25	RAD_SURG	1973	2008	35	67156	100
26	REC_NO	1973	2008	35	67156	100
27	TYPEFUP	1973	2008	35	67156	100
28	AGE_REC	1973	2008	35	67156	100
29	AJ_3SEER	1988	2003	15	35698	100
30	NUMPRIMS	1973	2008	35	67156	100
31	STCOUNTY	1973	2008	35	67156	100
32	SURV_TM	1973	2008	35	67156	100
33	STAT_REC	1973	2008	35	67156	100
34	DTH_CL	1973	2008	35	67156	100
35	O_DTH_CL	1973	2008	35	67156	100

El número de variables seleccionadas representa el 47.9 % del total de variables que fueron elegidas a través del análisis preliminar (73 variables), y el 28.2 % del total general establecido en la Base de Datos SEER (124 variables). Lo que muestra que para el proceso de predicción del caso de estudio se emplearon casi la tercera parte del total de variables registras en la base de datos SEER.

Las variables seleccionadas, por lo general, presentan un alto porcentaje de aceptación (100%). Mientras que las variables descartadas (38), aun teniendo un alto porcentaje de aceptación, alcanzando hasta un 100%, no fueron seleccionadas debido a que fundamentalmente no tenían relación directa con el cáncer de mama, presentaban redundancia o duplicidad de información con otras variables y por la opinión y sugerencia de médicos especialistas que colaboraron en esta etapa del trabajo de investigación.

La variable SEX, que registra el sexo del paciente al momento del diagnóstico, que tiene relación directa con el cáncer de mama, también fue descartada. Esto debido a uno de los alcances de este trabajo, que es trabajar con registros de casos de cáncer de mama en pacientes mujeres de origen hispano. Así, de los 67156 registros, 312 corresponden a pacientes hispanos de sexo masculino, los cuales fueron descartados, quedando en total 66844 registros (Fig. 1).

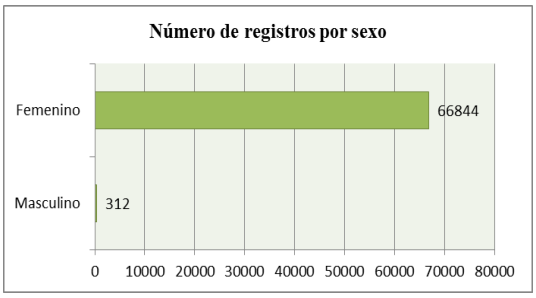

Fig. 1: Casos de cáncer de mama por sexo del paciente

Además, se observó que la mayor cantidad de las variables seleccionadas alcanzan 35 años favorables de registros válidos consecutivos, otros como: EOD10_PN y EOD10_NE presentan 20 años favorables, mientras que EOD10_SZ y AJ_3SEER alcanzan los 15 años. En el caso de SURGPRIM, ésta cuenta con 10 años favorables y finalmente CS_SIZE y D_AJCC_S poseen 4 años favorables. Aun cuando las variables CS_SIZE y D_AJCC_S presentan el mínimo de años con datos consecutivos requeridos (2004-2008), éstas se consideraron significativas por presentar series consecutivas y por registrar información relacionada con el tamaño del tumor y la etapa de la enfermedad, respectivamente.

4. ANÁLISIS DE LA VARIABILIDAD Y DISTRIBUCIÓN DE LAS PRINCIPALES VARIABLES ONCOLÓGICAS

Una vez establecidas las variables significativas, se realizó el análisis de variabilidad y distribución de las principales variables oncológicas disponibles en la fuente de datos, tales como: origen del paciente (ORIGIN), año de diagnóstico del cáncer (DATE_yr), estado civil del paciente (MAR_STAT), edad del paciente al momento del diagnóstico (AGE_DX), tipo de la enfermedad (BEHO3V) y confirmación del diagnóstico del cáncer (DX_CONF).

4.1 Origen del paciente

Para el análisis de variabilidad y distribución de la variable *Spanish/Hispanic Origin* (ORIGIN), que permite identificar pacientes de origen hispano, se estimó la cantidad de registros de acuerdo al origen del paciente, que comprende las siguientes categorías: México (incluye chicano); Puerto Rico; Cuba; Sur o Centro América (Excepto Brasil); Otro origen hispano (incluye Europa, excluye República Dominicana); Español, hispano, latino (no asignado en ninguna de las categorías anteriores); Sólo apellido español; y República Dominicana.

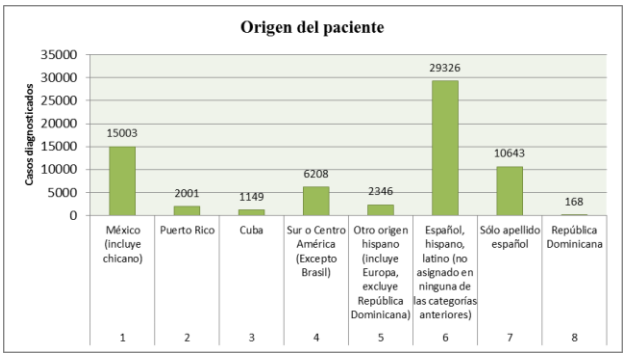

Fig. 2: Casos de cáncer de mama de acuerdo al origen del paciente

Se observa, Fig. 2, que la mayor cantidad de casos registrados de cáncer de mama se concentra en las categorías: *México* (incluyendo chicano); *Español, hispano o latino* (no asignado en otras categorías); y *Sólo apellido español*, con 15003, 29326 y 10643 registros, respectivamente. Otro grupo importante de casos diagnosticados se presenta en la categoría *Sur o Centro América* (Excepto Brasil), con 6208 registros. Un grupo menor de registros se concentra en las categorías: *Puerto Rico* (2001), *Cuba* (1149), *Otro origen hispano* (incluyendo Europa, excluyendo República Dominicana) -(2346)- y *República Dominicana* (168).

Por lo anterior, con base en la información presentada, se puede inferir que las variadas diferencias del número de casos registrados de cáncer de mama, por categoría, son básicamente proporcionales al tamaño de la población de dichas categorías (México, Puerto Rico, Cuba; Sur o Centro América, República Dominicana, entre otros) que viven en territorio Estadounidense; así como por la fecha de inicio en la que se empezó a registrar información sobre cada una de las categorías; originando esta variabilidad de una categoría a otra.

4.2 Edad del paciente al momento del diagnóstico

La Fig. 3 muestra la distribución y variabilidad, por grupo de edades, de los casos de cáncer de mama registrados al momento del diagnóstico. La variable *Age at diagnosis* (AGE_DX) es la responsable de concentrar esta información.

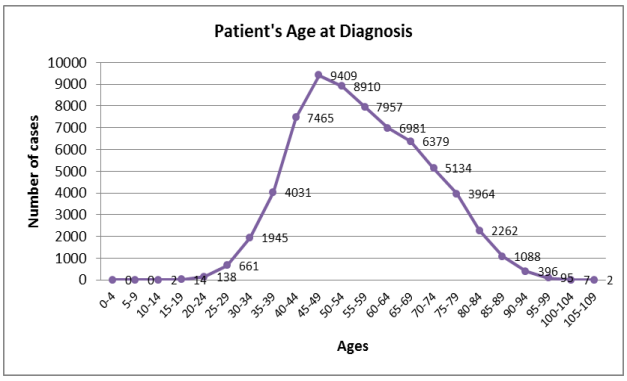

Fig. 3: Distribución por edades de casos de cáncer de mama diagnosticados

Se observa, Fig.3, la presencia de 2 casos diagnosticados de cáncer de mama en pacientes mujeres de 14 años. Asimismo, en otros grupos de edades, menores a 30 años, también se observa la presencia de esta patología: 15-19 con 14 casos, 20-24 con 138 casos y 25-29 con 661 casos. Esto indica que el cáncer de mama no sólo está apareciendo a edades tempranas, sino que es una las primeras causas de enfermedad y muerte en la mujer menor de 30 años.

Además, se observa un incremento progresivo de casos diagnosticados de esta enfermedad a parir de los 30 años, alcanzando el mayor número de registros entre los 45 y 49 años (9409 registros), donde las edades: 48 y 49 años representan los picos más altos de casos diagnosticados con 1939 y 1936 registros, respectivamente. De 50 a 90 años también se tiene una importante presencia de casos diagnosticados, siendo más evidente entre los 50 y 75 años. A partir de los 75 años se observa un significativo descenso del número de casos diagnosticados, hasta alcanzar 2 registros de pacientes mujeres que están entre 105 y 109 años, específicamente, una diagnostica a los 105 y otra a los 107 años.

4.3 Año de diagnóstico del cáncer

Esta variable, *Year of diagnosis* (DATE_yr), representa el año en que el tumor fue diagnosticado por primera vez, ya sea clínicamente o microscópicamente confirmado. Así, con base en la información de la variable, se observa (Fig. 4) que el número de casos de cáncer de mama diagnosticados a lo largo de 1973 y 2008 presenta un incremento exponencial. Por lo que se puede inferir que esta patología es una afección con tendencia no uniforme, puesto que con el tiempo el número de casos detectados se han incrementado significativamente.

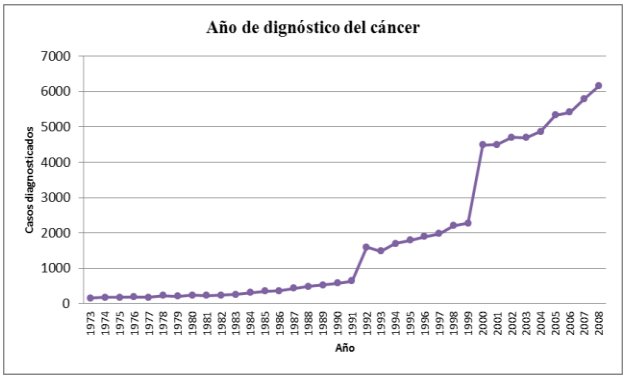

Fig. 4: Variabilidad de casos de cáncer de mama diagnosticados por año. Periodo 1973-2008

Además, se observa que a partir de 2000 a 2008 el incremento es mucho más evidente, esto comparado con años anteriores (1973 a 1999). Esto es, tan sólo en los últimos 9 años (2000-2008) el registro del número de casos diagnosticados fue de 45953, representando el 68.75 % del total de registros disponibles (66844); mientras que para el periodo 1973-1999 (27 años) el total de casos registrados fue de 20891 (31.25 %).

Otro detalle que salta a la vista es que en tan sólo un año, 1999-2000, el número de casos registrados se incrementó en casi el 100 %, esto es, de 2280 registros en 1999 a 4490 en el 2000. Esta situación puede ser a consecuencia del incremento de pacientes con esta patología y/o porque se amplió el área de cobertura de vigilancia y seguimiento de la enfermedad.

En consecuencia, es notorio que el cáncer de mama se está convirtiendo en un importante problema de salud pública que adquiere cada vez mayores dimensiones, constituyéndose estos hechos en el estímulo fundamental para la realización de este trabajo.

4.4 Estado civil del paciente

Para la variable *Marital Status at DX* (MAR_STAT), que permite identificar el estado civil del paciente al momento del diagnóstico del tumor, se estimó el número de registros por situación marital de los casos de cáncer de mama disponibles. Las categorías o valores que comprende esta variable que fueron analizadas son: soltera, casada, separada, divorciada, viuda y desconocida.

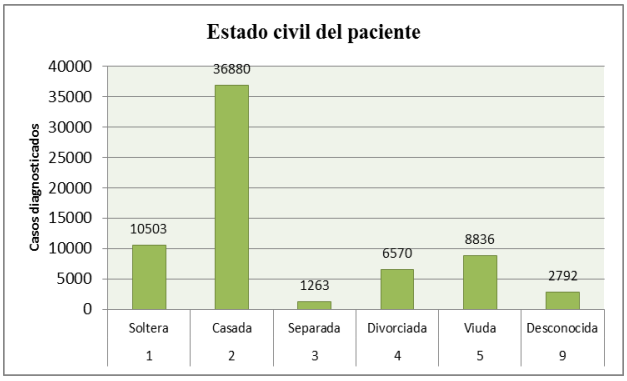

Fig. 5: Distribución de casos de cáncer de mama de acuerdo al estado civil del paciente

La Fig. 5 muestra la distribución y variabilidad del número de casos registrados de cáncer de mama de acuerdo al estado civil del paciente. Básicamente, se aprecia que el número de pacientes en condición de casadas (36880 registros) supera ampliamente al resto de categorías, lo que indica que el mayor número de pacientes, al momento del diagnóstico, se encontraba casada. Mientras que las condiciones: soltera, separada, divorciada, viuda y situación desconocida, presentan valores de 10503, 8836, 6570, 2792 y 1263 registros, respectivamente. Lo que indica que la enfermedad puede aparecer en todas las condiciones o estado civil de la mujer, representando una grave enfermedad que puede cobrar numerosas vidas en la población femenina a nivel mundial e indudablemente en las mujeres de origen hispano, esto si no se previene, detecta y controla a tiempo la enfermedad.

4.5 Tipo de la enfermedad

La Fig. 6 muestra la distribución y variabilidad del tipo de cáncer de mama registrados en la variable *Behavior code ICD-O-3* (BEHO3V), la cual se conforma por los siguientes cuatro tipos: Benigno, Potencial maligno o benigno, Carcinoma in situ (no invasivo) y Carcinoma maligno (invasivo).

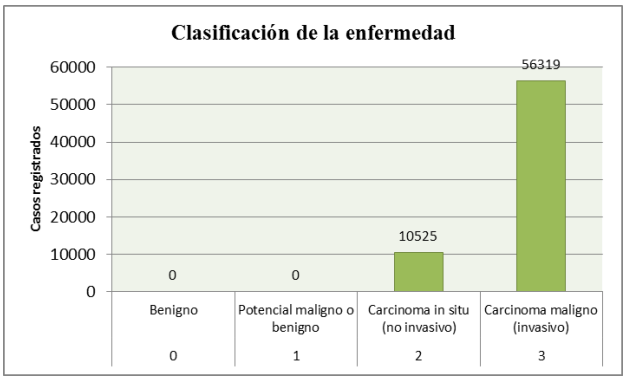

Fig. 6: Distribución de casos de cáncer de mama por tipo de la enfermedad

En general, se aprecia (Fig. 6) que la mayor cantidad de casos registrados se concentra en el tipo *Carcinoma maligno (invasivo)*, con 56319 registros, representando 84.25 % del total de casos disponibles. Mientras que el resto de casos (10525 registros) se encuentran distribuidos en el tipo *Carcinoma in situ (no invasivo)*, representando el 15.75 %. Asimismo, se observa que los tipos *Benigno* y *Potencial maligno o benigno* no presentan registros. Por lo que, con base en la información presentada, esta variable es considerada importante para el seguimiento y análisis del comportamiento de la neoplasia.

4.6 Confirmación del diagnóstico del cáncer

La Fig. 7 muestra la distribución y variabilidad del mejor método utilizado para la confirmación de la presencia del cáncer de mama, *Diagnostic Confirmation* (DX_CONF). Los métodos utilizados son: *Confirmación microscópica* (Histología positiva, Citología positiva, Método no especificado de confirmación microscópica positiva); *Confirmación no microscópica* **(**Pruebas de laboratorio positiva, Visualización directa sin confirmación microscópica, Radiología y otras técnicas de imagen sin confirmación microscópica, Sólo diagnóstico clínico); y *Confirmación desconocida* (Se desconoce confirmación microscópica, Certificado de defunción).

Fig. 7: Distribución de los métodos utilizados para la confirmación de la enfermedad

Se observa (Fig. 7) que la mayor cantidad de casos diagnosticados fue a través de *Confirmación microscópica*, fundamentalmente a través de exámenes de Histología, que dieron positivas en total 65812 casos, el cual representa el 98.46 % del total de casos diagnosticados. Otros casos, en menor cantidad, también fueron confirmados a través de Citología positiva (498 casos) y Método no especificado (18 casos). Asimismo, se distingue que un grupo menor de casos diagnosticados fueron confirmados a través de *evaluaciones no microscópicas*, como: Pruebas de laboratorio (1), Visualización directa (12), Radiología y otras técnicas de imágenes (116) y Sólo diagnóstico clínico (108). Además, se observa la presencia de 279 casos de los que se desconoce la forma de confirmación.

5. CONCLUSIONES

El análisis de datos es una de las actividades fundamentales en el proceso de Minería de Datos. Los datos analizados corresponden a series de datos provenientes de la Base de Datos del Programa de Vigilancia, Epidemiología y Resultados Finales (SEER) del Instituto Nacional del Cáncer (NCI) de los Estados Unidos.

Este análisis se realizó en dos etapas, en la primera se hizo una evaluación preliminar de la disponibilidad de datos de todas las variables listadas en la Base de Datos SEER, esto con el fin de establecer aquellas relevantes en función del periodo de sus registros, descartando las que presentan una escasa o nula cantidad de registros disponibles; y en la segunda etapa se determinó la calidad de la serie de datos para establecer las variables significativas y el periodo de años que se empleará en el proceso de predicción del caso de estudio.

De la evaluación quedaron 35 variables consideradas significativas, que representa el 28.2 % del total de variables registradas en la base de datos SEER (124 variables) y el 47.9 % del total de variables que fueron elegidas a través del análisis preliminar (73 variables).

Posterior a la determinación de las variables significativas, se analizó la variabilidad y distribución de las principales variables establecidas en la Base de Datos SEER, como: Origen del paciente, Edad del paciente, Año de diagnóstico, Estado civil del paciente, Tipo de la enfermedad y Confirmación del diagnóstico; este análisis proporcionó la identificación de tendencias y comportamientos del conjunto de datos disponibles.

El trabajo realizado implicó retos importantes, como el análisis de un amplio conjunto de datos clínicos, así como el manejo de técnicas de Minería de Datos para analizar los casos clínicos de Cáncer de Mama; permitiendo extender la visión de la minería de datos y su aplicación a problemas de diverso índole, en este caso aplicado a medicina.

6. REFERENCIAS BIBLIOGRÁFICAS

[1] Keogh E., Lin J. y Truppel W. (2003). *Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research*. Proceedings of the Third IEEE International Conference on Data Mining, pp. 115, ISBN: 0-7695-1978-4, Florida, Estados Unidos

[2] Kessler M. (2003). *Apuntes de métodos estadísticos de la Ingeniería y Apuntes de estadística industrial*. Departamento de Matemática Aplicada y Estadística, Universidad de Cartagena, pp. 73, España

[3] Puerto J. y Paz M. (2001). *Análisis descriptivo de series temporales aplicadas al precio medio de la vivienda en España*. Management Mathematics for European Schools, pp. 41, España

[4] Chiu B., Keogh E. y Lonardi S. (2003). *Probabilistic Discovery of Time Series Motifs*. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 493-498, ISBN: 1-58113-737-0, Washington, Estados Unidos

[5] NCI (2012). *SEER Training Modules*. Surveillance, Epidemiology and End Results Program. < http://training.seer.cancer.gov/modules_reg_surv.html>. Último acceso 12 de marzo de 2012

[6] SEER (2012a). *About the SEER Program*. Surveillance, Epidemiology and End Results Program. < http://seer.cancer.gov/about/> Último acceso 15 de febrero de 2012

[7] SEER (2012b). *Home SEER*. Surveillance, Epidemiology and End Results Program. < http://seer.cancer.gov/> Último acceso 16 de febrero de 2012

[8] CDC (2012). *Los registros del cáncer proporcionan datos fiables sobre el cáncer*. Centros

para el Control y la Prevención de Enfermedades. <www.cdc.gov/spanish/especialesCDC/CancerRegistros/>. Último acceso 18 de febrero de 2012

[9] ACS (2012). *Cancer registries.* American Cancer Society. <www.cancer.org/Cancer/CancerBasics/cancer-surveillance-programs-and-registries-in-the-united-states>. Último acceso 20 de febrero de 2012

7. SINTESIS CURRICULAR DEL AUTOR

Guillermo Gilberto Molero-Castillo es Maestro en Ciencia e Ingeniería de la Computación por la Universidad Nacional Autónoma de México (UNAM), actualmente es candidato a Doctor (PhD) en Tecnologías de Información por la Universidad de Guadalajara (UDG), México. Es graduado en Ingeniería de Sistemas y Computación (1999-2003), donde obtuvo el primer puesto en el cuadro de méritos de su Generación. Ha cursado diferentes cursos y diplomados. Sus líneas de investigación son Inteligencia Artificial, Reconocimiento de Patrones, Minería de Datos, Inteligencia Computacional e Inteligencia de Negocios. Ha laborado en instituciones, como: SaitoSoft S.A. de C.V. como Líder de Proyectos en los desarrollos para la Comisión Nacional de Derechos Humanos (CNDH), Gas Metropolitano y Suprema Corte de Justicia de la Nación (SCJN), en PetroSoft S.A. de C.V. como analista en el desarrollo del Sistema Experto para el Bombeo Neumático Continuo del Activo Cantarell (PEMEX), así como en Consultoría de Crews, S.A. de C.V. como Analista Científico. Ha participado en eventos nacionales e internacionales, es autor de varias publicaciones científicas.