Diseño de un sistema para la gestión y procesamiento estadístico de bases de datos genéticas en Cuba

Design of a system for the management and statistical processing of genetic databases in Cuba

Daniel Medero García,[I] Pedro Quintero Rojas,[II] Beatriz Marcheco Teruel.[III]

Resumen

Las investigaciones realizadas por el Centro Nacional de Genética Médica de Cuba en los últimos diez años han generado registros con información sobre decenas de miles de personas y familias en el país. Los individuos incluidos en los registros médicos presentan enfermedades genéticas, defectos congénitos, o participan en investigaciones aún sin padecer una enfermedad determinada, como es el caso de las parejas de gemelos. La dispersión de los datos registrados y la falta de herramientas informáticas para gestionarlos han condicionado la no obtención de un máximo provecho del procesamiento de esa información, que debe ser utilizada en beneficio del propio sistema de salud. En el presente trabajo se propone el diseño y la implementación de un sistema que integra y permite gestionar, filtrar, analizar y graficar el conjunto de datos expuesto anteriormente. Finalmente se llevan a cabo pruebas que verifican su eficiencia y funcionamiento y se proponen recomendaciones para aumentar su extensibilidad.

Palabras clave: Bases de datos, registros genéticos, gestión de datos, registro gemelos, Cuba

Abstract

The research carried out by the National Center for Medical Genetics of Cuba in the last ten years has generated records with information on tens of thousands of people and families all over the country. The individuals included in the medical records have genetic diseases, congenital defects, or participate in research even without suffering a specific disease, as is the case of twin pairs. The dispersion of the registered data and the lack of computer tools to manage them have conditioned the lack of obtaining maximum benefit from the processing of this information, which must be used to benefit the health system itself. In the present work we propose the design and implementation of a system that integrates and allows to manage, filter, analyze and graph the set of data previously exposed. Finally, tests are carried out to verify their efficiency and functioning and recommendations are proposed to increase their extensibility.

Keywords: Databases, genetic records, data management, registry of twins, Cuba.

[I] Lic. en Ciencias de la Computación. Universidad de La Habana.

[II] Lic. en Ciencias de la Computación. Facultad de Matemática y Computación. Universidad de La Habana.

[III] Dra. en Ciencias Médicas. Especialista de primer y segundo grado en genética clínica. Investigadora Titular y Profesora Titular. Centro Nacional de Genética Médica. E-mail: beatriz@infomed.sld.cu

Introducción

En el mundo de la información digitalizada crece cada vez más el volumen de datos almacenados, y ello puede traer como consecuencia que sea compleja su utilización y procesamiento. Una de las necesidades que surgen cuando el conjunto de datos es muy grande, es la de filtrado; con la que se puede obtener un grupo más específico y un resultado más rápido. Por otra parte, realizar análisis estadísticos puede ayudar a determinar correlaciones entre variables de un sistema, lo que produce información que pueden ser de gran utilidad.

El Centro Nacional de Genética Médica (CNGM) de Cuba, realiza estudios para la comprensión del rol de los factores genéticos en el origen de enfermedades que se transmiten de generación en generación. Se especializa en obtener información sobre qué ocurre exactamente en el ciclo celular de los seres vivos y cómo puede ser que, por ejemplo, entre seres humanos se transmitan características biológicas, de apariencia física y psicológica. Este tiene la misión de coordinar la red de centros y servicios de esta especialidad en el país.

Ubicado en la Universidad de Ciencias Médicas de La Habana, la institución se ocupa de la formación de los médicos especialistas de toda Cuba que tienen como función la atención clínica y el diagnóstico de malformaciones congénitas y enfermedades genéticas, su investigación y prevención. Asimismo, es sede de las dos maestrías que existen en el país para la especialidad. Dirige la asignatura de genética médica en el pregrado para las carreras de la salud.

Este centro tiene un conjunto de bases de datos en diversos formatos que contienen una extensa información sobre decenas de miles de individuos y familias. Una de ellas es la del Registro Nacional de Gemelos que muestra datos sociales y del estado de salud de 55 mil parejas de gemelos cubanos idénticos y no idénticos.

Dispone además de la base de datos de la investigación sobre mestizaje étnico de la población cubana con más de 1000 individuos de todo el país. De estos, por ejemplo, se conservan datos de un cuestionario aplicado que recoge información sobre su situación socioeconómica, su estado de salud y mediciones antropométricas.[1]

Se cuenta con otra base de datos con poco más de 1400 individuos centenarios. Estos forman parte de un estudio realizado por un equipo multidisciplinario integrado por especialistas de varias instituciones. Como apoyo a esta investigación, se han elaborado estudios de marcadores genéticos y de estrés oxidativo a los centenarios participantes. Actualmente se explora la posible relación entre las características genéticas que se han encontrado en esos análisis y cómo ellas 'predisponen' favorablemente a los individuos a vivir más años. Además, existen otras bases de datos digitalizadas y por digitalizar que se encuentran en formatos diferentes (Excel, MySQL, Access).

La posibilidad de realizar en Cuba grandes estudios poblacionales dada la organización del Sistema Nacional de Salud, si bien por una parte constituye una fortaleza, tiene a su vez la debilidad de no contar con herramientas informáticas que permitan obtener el máximo provecho del procesamiento de esa información. Esta puede ser utilizada en beneficio del propio sistema de salud y de la población en general con la realización de análisis que consideren desde el comportamiento global a nivel de país hasta las situaciones propias de una provincia o municipio que la diferencien del resto.

Un sistema centralizado que permitiese hacer consultas dinámicas a los datos contenidos, que a su vez sea capaz de analizarlos y generar reportes sería de gran utilidad. Este sistema pudiese incluso generalizarse de tal forma que fuese capaz de gestionar todo un conjunto de conceptos, no solamente desde el punto de vista biomédico. Si además se le incluyese la posibilidad de análisis estadístico y realizar gráficos, al abarcarse diversos campos de la ciencia el sistema se utilizaría en un ámbito más amplio.

Ante esta necesidad, la dirección del Centro Nacional de Genética Médica se dirigió a la Facultad de Ciencias de la Computación de la Universidad de La Habana para solicitarle la elaboración de un sistema informático, con el requisito de que este pudiera ser utilizado para interactuar dinámicamente con las bases de datos existentes y realizar análisis estadísticos.

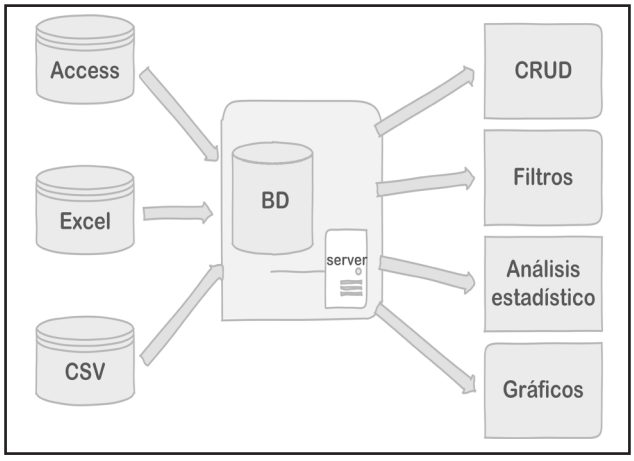
La presente investigación se planteó como problema científico la necesidad de unificar las bases de datos disponibles en diferentes formatos e incorporarle herramientas de gestión y soporte estadístico para facilitar su análisis y encontrar correlaciones. Para ello se buscó generar un sistema en el que se puedan realizar filtrados, que permita estimar, a través de distintas funciones, el riesgo de una persona de padecer ciertas enfermedades: asma, hipertensión, diabetes, accidente cerebrovascular, diferentes tipos de cáncer, entre otras, dado que se expone a diversos factores de riesgo y tiene o no antecedentes familiares de la misma. Se solicitó el diseño e implementación de un instrumento informático que permitiese elaborar filtros, gráficos y análisis entre variables. Ello facilitaría la identificación de factores genéticos que podrían o no ser predictores de un incremento en la esperanza de vida en las personas y diseñar estrategias dirigidas a promover la salud desde esas perspectivas.

Con el propósito de ofrecer solución a la problemática anteriormente expuesta, el presente trabajo se trazó como objetivos el diseño y la implementación de un sistema capaz de: concentrar en una sola herramienta las informaciones de diferentes bases de datos y permitir la inserción, lectura, actualización y eliminación de registros, crear filtros avanzados sobre esos datos, realizar diferentes tipos de gráficos para mostrar la información de una forma más visual y hacer análisis estadístico sobre las diferentes variables del nuevo sistema.

Método

Se realizó un estudio de las distintas herramientas disponibles para la gestión de datos y realización de análisis estadísticos. Se compararon las mismas atendiendo a las ventajas y desventajas que ofrecen y capacidad de solución para el problema de investigación a resolver. Se determinó que era posible construir una aplicación web que unificara las bases de datos antes mencionadas. La misma debía ser capaz de realizar operaciones de creación, lectura, actualización y eliminación de datos, construir filtros, hacer análisis estadístico y generar gráficos tal como se muestra en la figura 1. Ello permitiría lograr que la información estuviera de forma homogénea, y que también se pudiera acceder a la misma desde un navegador web una vez se realice el despliegue de la aplicación en un servidor.

Figura 1. Diseño general de la aplicación web.

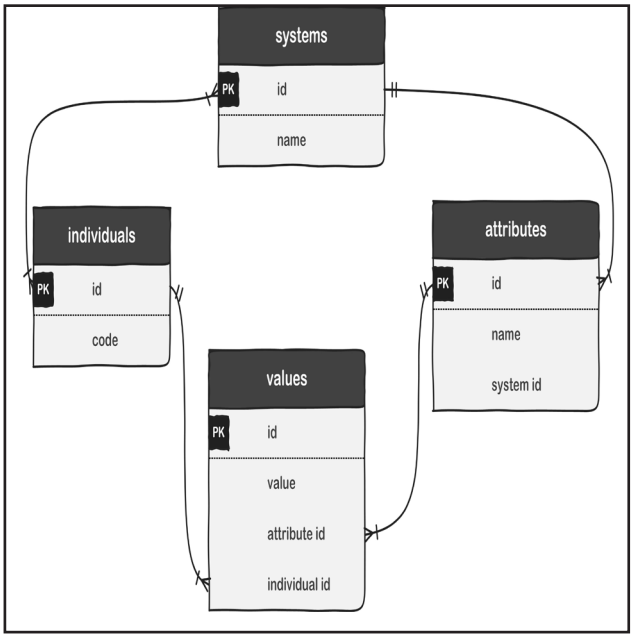


CRUD=crear, leer, actualizar y borrar
CSV=valores separados por comas.
* Las iniciales proceden del inglés en ambos casos.

La base de datos del sistema estuvo compuesta por cinco conjuntos de entidades cuya estructura se muestra en la figura 2 y su funcionamiento se argumenta a continuación:

- ***systems*:** en ella se guardan el nombre y la descripción de las bases de datos existentes (e.g. *Centenarios*, *Gemelos*, *Mestizaje*). Cada uno de estos registros se relaciona con un conjunto de *individuals* y de *attributes*. De esta forma se podría leer: un *sistema* está compuesto por un conjunto de *individuos* y un conjunto de *atributos*.
- ***individuals*:** almacena el *code* (código) de cada uno de los individuos. Tiene una relación muchos a muchos con *systems* puesto que una persona puede pertenecer a varios sistemas y un sistema puede contener a varias personas. Está también relacionada con *values* que es donde se guardarán todos sus datos.
- ***attributes*:** se registran todos los atributos correspondientes de los sistemas, razón por la cual está relacionada con *systems*. Un sistema contiene un conjunto de atributos; pero un atributo pertenece a un único sistema. A su vez está enlazada con *values* de tal forma que un atributo contiene muchos valores; pero a un valor le corresponde un único atributo.
- ***values*:** contiene todos los valores de todos los individuos del sistema. Para conocer el valor del atributo de una persona se requiere de las llaves foráneas de su relación con *attributes* e *individuals*.

Figura 2. Modelo relacional de la base de datos



A continuación, se describe mediante un ejemplo esta estructura de la base de datos:
Supongamos que tenemos los siguientes datos de individuos centenarios en el formato de la tabla 1.

Tabla 1. Registro de individuos centenarios. Cuba.

Código	Edad	Provincia
10199012	100	Pinar del Río
13982378	107	Matanzas
10289087	101	Granma

Para este ejemplo los conjuntos de entidades *systems, individuals, attributes* y *values* contendrían los datos de las tablas 2, 3, 4 y 5 respectivamente.

Tabla 2. Ejemplo de systems para el registro de individuos centenarios de Cuba.

Id	Name
1	Centenarios

Tabla 3. Ejemplo de *individuals* para el registro de individuos centenarios de Cuba.

Id	Code
1	10199012
2	13982378
3	10289087

Tabla 4. Ejemplo de *attributes* para el registro de individuos centenarios de Cuba.

Id	Name	System id
1	edad	1
2	provincia	1

Tabla 5. Ejemplo de *values* para el registro de individuos centenarios de Cuba.

Id	Value	Attribute id	Individual id
1	100	1	1
2	107	1	2
3	101	1	3
4	Pinar del Río	2	1
5	Matanzas	2	2
6	Granma	2	3

Proceso de extracción y transformación de datos
en este se construyó un módulo con distintas funcionalidades para cada formato de las bases de datos a extraer. Dicho módulo es el encargado de la unificación de los datos dispersos en la base de datos anteriormente definida. Una vez integrados todos los datos en el servidor se implementaron módulos que permiten al usuario la gestión de los mismos, la posibilidad de consultarlos, realizar análisis estadísticos y gráficos.

Para la implementación del sistema, se utilizó el *framework* de aplicaciones web *Ruby on Rails*, también conocido como *Rails*. Fue escrito en el lenguaje de programación Ruby, siguiendo el paradigma de la arquitectura Modelo Vista Controlador (MVC). Se combinó la simplicidad con la posibilidad de desarrollar aplicaciones del mundo real escribiendo la menor cantidad de código posible y con un mínimo de configuración. El lenguaje de programación Ruby permite la metaprogramación, de la cual *Rails* hace uso, lo que resulta en una sintaxis muy legible. *Rails* se distribuye a través de RubyGems[2], que es el formato oficial de paquete y canal de distribución de bibliotecas y aplicaciones. Los principios de la filosofía de *Ruby on Rails* incluyen *DRY*[1]que significa que las definiciones deben hacerse una sola vez y *Convención sobre configuración* en la que el programador sólo necesita definir aquella configuración que no es convencional.

Diseño de la base de datos
El sistema de gestión de base de datos utilizado fue MySQL. Se usó una de las gemas disponibles en RubyGems[2] llamada "mysql2" que posee la capacidad de servir de intermediario entre *Rails* y MySQL para su conexión y la consulta e iteración de los datos. La estructura de la base de datos fue modificada ligeramente en función de aspectos convencionales de *Rails*.

Gestión de datos
Para la administración de datos se utilizaron las facilidades de *Rails* usando los generadores de *scaffolds* como punto de partida. En programación, el *scaffolding* es un método para construir aplicaciones basadas en bases de datos. En este el programador escribe una especificación que describe cómo debe ser usada la base de datos. Luego el compilador utiliza esa especificación para generar el código que la aplicación usará para crear, leer, actualizar y eliminar registros de la base de datos. Una vez generados los modelos, controladores y vistas se terminan de implementar las funciones básicas de creación, muestreo, edición y eliminación. En el caso de la creación y edición de un atributo, se incluyó una entrada en la que el usuario seleccione entre todos los nomencladores del sistema.

Extracción de bases de datos
Dada la heterogeneidad de las bases existentes se decidió utilizar las tareas *rake* (*rake tasks*) de *Rails*. *Rake* es una herramienta de automatización y gestión de tareas de Rails. Este permitió especificar tareas y describir dependencias, así como grupos de tareas en

un *namespace* (espacio de nombres). Se implementó una tarea por cada base de datos que se necesita exportar. Cada uno de ellos extrae y transforma los datos según el formato y las especificidades de cada una.

Filtros
Para la realización de filtros se crearon tres ficheros: la vista filters_view.erb, el *javascript* filters_javascript.js y el controlador filters_controller.rb.
En el fichero filters_view.erb se declaró la interfaz con que el usuario va a interactuar. Finalmente se colocó una sección donde al usuario se muestran los resultados de la consulta en forma de tabla.
El fichero filters_javascript.js es el encargado de la funcionalidad de la vista. Una vez se reciben los datos en el servidor este se encarga de procesar cada uno de los objetos enviados, realizar las consultas correspondientes y devolver el resultado a la vista. El encargado de esta tarea es el fichero filters_controller.rb. Se realizan dos consultas SQL. La primera es para extraer los id de los individuos que cumplen con las condiciones definidas en los filtros. La segunda es para extraer los atributos especificados por el usuario de estos individuos. Finalmente, el resultado es llevado a formato JSON y enviado al cliente. Este a su vez genera una tabla al usuario con los registros y los atributos seleccionados.

Análisis estadístico
Para el análisis estadístico de los datos se decidió utilizar la herramienta R.
Para la comunicación de *Ruby on Rails* y R se decidió utilizar RServe, por ser la más actualizada además de ofrecer varias funcionalidades.

Gráficos
Para la generación de gráficos se utilizó la biblioteca de *javascript* Highcharts.[3,4] Este permite la creación de una variedad de gráficos interactivos de barras, líneas, puntos, área, entre otros. Además, es de código abierto y gratis para uso no comercial. Recibe un parámetro de categorías y otro de sus respectivos valores. Se persigue que el usuario tenga la posibilidad de poder graficar las variables (atributos) introducidos en el sistema. Se ofrece la opción de escoger una variable a categorizar en el eje *x* y múltiples variables en el eje *y,* y de cada una seleccionar una función agregada y un gráfico en específico.

Aspectos éticos
El presente trabajo se realizó mediante una colaboración acordada entre el Centro Nacional de Genética Médica y la Facultad de Matemática y Computación de la Universidad de La Habana. Para garantizar la confidencialidad de la información almacenada en las bases de datos de los registros genéticos no se utilizaron los mismos en la implementación del sistema, sino que se utilizaron los nombres de las variables y se generaron datos no reales. La investigación se realizó observando los preceptos de la Declaración de Helsinki revisada, de la Asociación Médica Mundial.[4]

Resultados
A continuación, se muestran algunos ejemplos de los resultados de la aplicación. Estos pertenecen a la simulación realizada mediante generación de una base de datos de Excel denominada 'Centenarios' en la cual se simuló información de un total de 1488 individuos con 1404 atributos (variables) a semejanza del registro real.

La tabla 6 muestra la cantidad de elementos luego del proceso de extracción. En el caso de *attvalues* no se insertaron los valores *null* o cadenas vacías.

Tabla 6. Cantidad de registros luego del proceso de extracción.

Tabla	Cant. registros
systems	1
patients	1,488
attribs	1,404
attvalues	2,087,664
nomenclatures	1,153
items	11,518

Los resultados de la demostración de la funcionalidad del análisis estadístico mediante la comparación entre el análisis realizado con la aplicación implementada y el obtenido al utilizar el paquete SPSS, muestra que no hay diferencias en los resultados (tablas 7, 8 y 9). En la tabla 7 aparecen los resultados del análisis de correlación de Spearman entre las variables peso y hemoglobina realizado con la nueva aplicación y con la misma funcionalidad del paquete estadístico SPSS, obteniéndose resultados similares.

Tabla 7. Resultados del análisis de la correlación entre peso y hemoglobina (método de Spearman) mediante SPSS y la aplicación diseñada.

	SPSS	Aplicación
p value	0.059685	0.058956
rho	0.072	0.072194
peso (N)	1047	1047
hemoglobina (N)	685	685

En la tabla 8 aparece el análisis del test chi-cuadrado para las variables color de piel y hemoglobina realizado con la nueva aplicación y comparado con la misma funcionalidad del paquete estadístico SPSS, no se observan diferencias.

Tabla 8. Resultados del *test* chi-cuadrado entre color de piel y hemoglobina realizado mediante SPSS y la Aplicación diseñada.

	SPSS	Aplicación
statistic value	816.995	816.995378
p value	0.984	0.984134
df (degrees of freedom)	906	906
valid cases (N)	971	971

En la tabla 9 aparece el análisis T-test entre las variables presión arterial sistólica y presión arterial diastólica realizado con la nueva aplicación y comparado con la misma funcionalidad del paquete estadístico SPSS, obteniéndose valores semejantes.

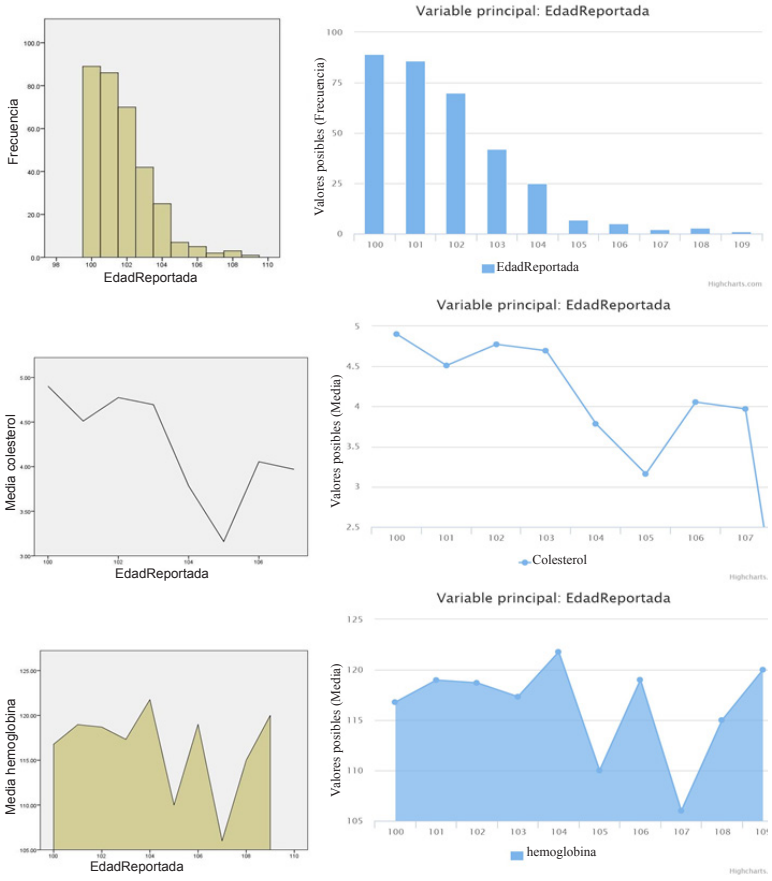
Tabla 9. Resultados del T-test entre presión arterial sistólica y presión arterial diastólica realizado mediante SPSS y la Aplicación diseñada.

	SPSS	Aplicación
statistic value	105.028	105.027884
df (degrees of freedom)	1322	1322
conf. interval	47.100 - 48.893	47.099 - 48.892
estimated mean of the differences	47.996	47.996220
PAS* (N)	1323	1323
PAD** (N)	1323	1323

* PAS: presión arterial sistólica
** PAD: presión arterial diastólica

La figura 3 muestra ejemplos de los gráficos generados con el uso de la aplicación diseñada y su comparación con los de SPSS.

Figura 3. Ejemplos de gráficos generados mediante el uso de SPSS y por la aplicación desarrollada



Izquierda: generados por SPSS.

Derecha: generados por la aplicación desarrollada.

Se verificaron también las funcionalidades con que el usuario añade, actualiza, modifica y elimina elementos del sistema por medio de las facilidades ofrecidas por *Ruby on Rails*.

Discusión

Con el presente trabajo se logró diseñar e implementar un software que permite al Centro Nacional de Genética Médica gestionar las bases de datos de investigaciones y registros de pacientes con la funcionalidad de tener incorporada la posibilidad de realizar análisis estadísticos y producir gráficos.

Como puede apreciarse en las tablas 7,8 y 9 no existe diferencia entre los resultados obtenidos en los análisis estadísticos realizados con la aplicación diseñada, al compararlos con los obtenidos mediante el uso de la herramienta SPSS, que es una de las más empleadas por los investigadores para el análisis de los resultados. Consideramos que estos resultados fueron posibles gracias al empleo de R, ya que es un lenguaje de programación orientado exclusivamente al análisis estadístico. Se aprovechó además su integración a Rails y su conjunto de funcionalidades para la realización de *tests* y su extensibilidad.

Al experimentar el funcionamiento de la aplicación para la realización de gráficos se observan resultados similares a los obtenidos en la comprobación de la funcionabilidad de la realización de análisis estadístico.

Se puede observar la similitud de los gráficos generados mediante la aplicación diseñada y SPSS al utilizar las mismas variables. No obstante, es importante señalar que los generados por SPSS son estáticos, mientras que los del sistema son interactivos. En este caso la utilización de la biblioteca de *javascript* Highcharts,[4] es la que permite la creación de gráficos interactivos. Una de las principales ventajas de esta aplicación es que es capaz de realizar la extracción de sus bases de datos originales y fusionarlos en un formato común. Luego ofrece la posibilidad de realizar filtros, análisis estadístico y graficación.

La comprobación de los filtros realizada por medio de observaciones y consultas SQL directas a la base de datos ha demostrado su correcto funcionamiento.

Conclusiones

El diseño e implementación de esta herramienta web logró la integración de un conjunto de sistemas de datos en un formato común. El uso de la herramienta R garantizó el desarrollo de la funcionalidad de análisis estadístico.

En las pruebas realizadas de las funciones de la aplicación se pudo verificar que, en comparación con herramientas como SPSS, la precisión de salida es muy similar.

La herramienta implementada ofrece las funcionalidades solicitadas por el Centro Nacional de Genética Médica para lograr la homogeneidad de los datos obtenidos en sus investigaciones. Además, posee la flexibilidad para integrar nuevos sistemas y aumentar su extensibilidad en un futuro.

Referencias bibliográficas

1. Marcheco-Teruel, Beatriz, Esteban J Parra, Evelyn Fuentes-Smith, Antonio Salas, Henriette N Buttenschøn, Ditte Demontis, María TorresEspañol, Lilia C Marín-Padrón, Enrique J Gómez-Cabezas, Vanesa Álvarez-Iglesias y cols.: Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. PLoS Genet, 10(7):e1004488, 2014.
2. RubyGems. [en línea]. Seattle: RubyGems; abril 2009 [fecha de acceso: 15/01/2017]. Disponible en: https://rubygems.org
3. Kuan Joe. Learning Highcharts 4. 2nd Ed. Birmingham: Packt Publishing Ltd; 2015.
4. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. World Medical Association. Clinical Review & Education [en línea]. 2013 [citado 1 de septiembre de 2017]. Disponible en: http://www.up.ac.za/media/shared/Legacy/sitefiles/file/45/2875/declarationofhelsinki_fortaleza_brazil2013.pdf
5. Highsoft. Highcharts [en línea]. Vik i Sogn: Highsoft; 2009 [fecha de acceso: 15/01/2017]. Disponible en: http://www.highcharts.com.