Universidad de las Ciencias Informáticas

FACULTAD 6

Título: *SEEGEN-R: Sistema Estadístico de Epidemiología Genética – basado en R. Extensión para los estudios de Epidemiología Genética.*

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Autor(es): Omar Dixán Puig Pupo

Maidelyn Padrón Rodríguez

Tutor(es): Msc. Elvismary Molina de Armas

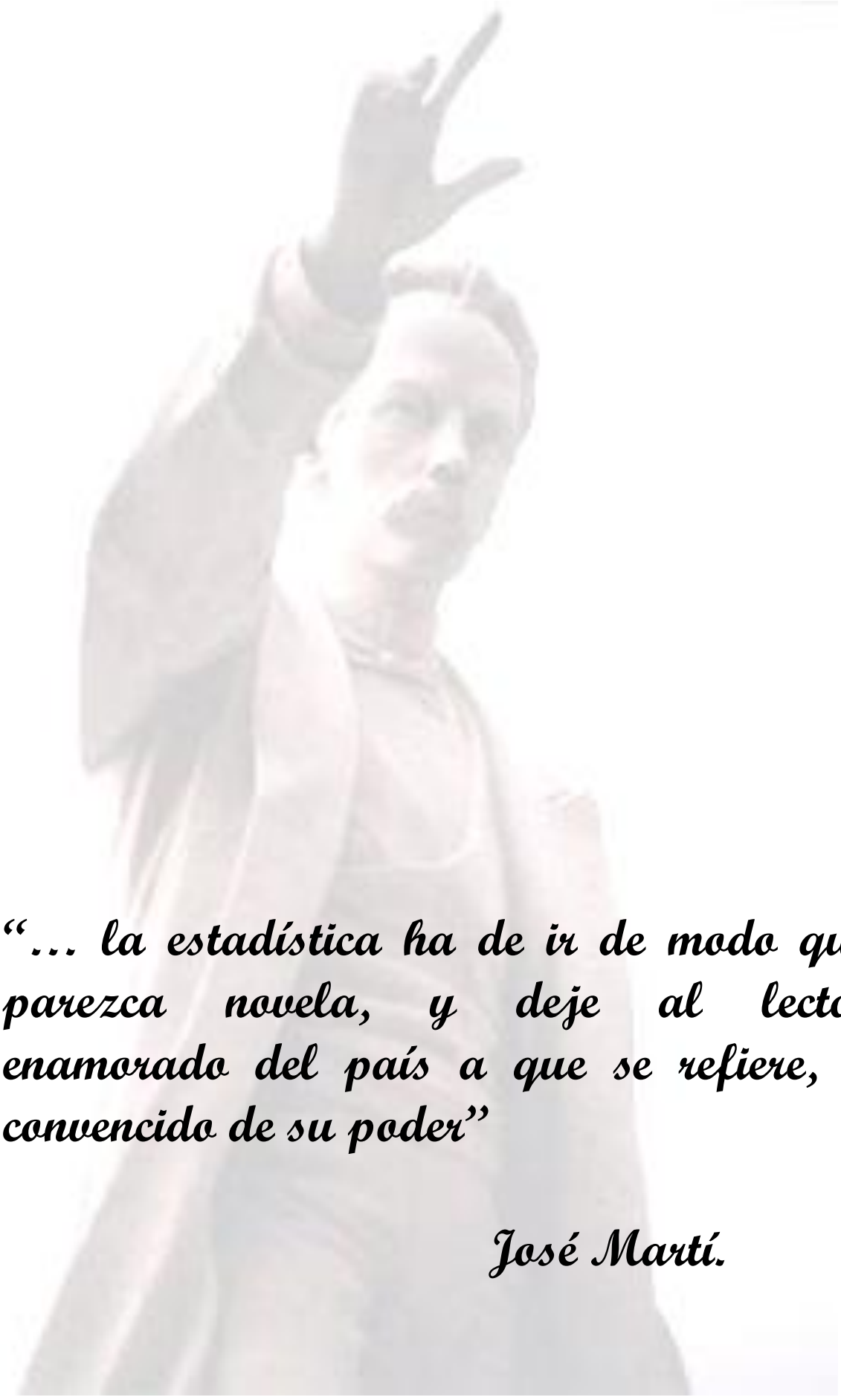
Msc. Yunier Emilio Tejeda Rodríguez

Ing. Yudiel la Rosa González

Consultante: Dr. C. Roberto Lardoeyt Ferrer.

La Habana, Junio de 2013

"Año 55 de la Revolución"



"... la estadística ha de ir de modo que parezca novela, y deje al lector enamorado del país a que se refiere, y convencido de su poder"

José Martí.

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año _____.

Omar Dixán Puig Pupo

_____

Firma del Autor

Maidelyn Padrón Rodríguez

_____

Firma del Autor

Ing. Yudiel La Rosa González

_____

Firma del Tutor

Msc. Elvismary Molina de Armas

_____

Firma del Tutor

Msc. Yunier Emilio Tejeda Rodríguez

_____

Firma del Tutor

DATOS DE CONTACTOS

Tutores:

Msc. Elvismary Molina de Armas
Universidad de las Ciencias Informáticas, La Habana, Cuba
Email: emolina@uci.cu

Msc. Yunier Emilio Tejeda Rodríguez
Universidad de las Ciencias Informáticas, La Habana, Cuba
Email: yuniere@uci.cu

Ing. Yudiel la Rosa González
Universidad de las Ciencias Informáticas, La Habana, Cuba
Email: ylarosag@uci.cu

Consultante:

Dr. Roberto Lardoeyt Ferrer
Centro Nacional de Genética Médica, La Habana, Cuba
Email: lardgen@infomed.sld.cu

AGRADECIMIENTOS

A nuestro comandante en jefe Fidel Castro y la Revolución Cubana por permitir que nuestro sueño de ser profesionales se haga realidad.

A todos los que ayudaron en nuestra formación, a nuestros familiares y amigos por su apoyo y confianza, por su comprensión en los momentos difíciles y por estar ahí cuando los necesitamos.

Al Dr. Roberto y a los tutores, el ingeniero Yudiel La Rosa y a los máster en ciencias Elvismary Molina y Yunier E. Tejeda, por su ayuda para el desarrollo de la tesis y en especial a la ingeniera Dayana Joseph, que a pesar de no ser tutora, nos proporcionó su apoyo incondicional, su paciencia, su confianza y colaboración para poder lograr la realización de este trabajo.

A todas aquellas personas que de una u otra forma, colaboraron o participaron en nuestra formación como persona y profesional, hacemos extensivo nuestro más sincero agradecimiento.

DEDICATORIA

Dedicamos el esfuerzo y sacrificio, todos nuestros éxitos y vicisitudes de tantos años de estudios y vida, esperando ser un orgullo para ustedes:

De Maidelyn:

Quiero dedicarle este trabajo ***a mi mamá Mabel,*** *por estar junto a mí en cada paso que he dado en la vida, por ser la principal fuente de inspiración para seguir adelante sin rendirme. Porque sin escatimar esfuerzo alguno, has sacrificado gran parte de su vida para formarme en la profesional que me he convertido. A quien la ilusión de su vida ha sido convertirme en una persona de bien. A quien nunca podré pagar su sacrificio ni aún con el tesoro más grande del mundo. A ti debo este logro, y contigo lo comparto.*

A mi hermana Yairys*, por su cariño, por su ejemplo como mujer luchadora y porque de ti aprendí a vivir feliz con el fruto de mi propio esfuerzo. En especial* ***a mi hermanita Mayalis****, por llegar a este mundo para alegrarnos la vida, por darme todo su cariño, por llenar de felicidad mis días con las carticas que me ha dedicado, por sus abrazos y sus besos, porque por ella trato de ser una mejor persona cada día para ser un ejemplo para ti, a ambas las adoro y las amo con la vida.*

A mis primas Aisa y Yanet*, por el gran significado que tienen en mi corazón, por ayudarme y apoyarme cuando lo he necesitado, por lo bien que me hacen sentir cuando estamos juntas, por su cariño y ternura, por su preocupación, para ustedes mi triunfo.*

A mi padrastro Rafael (el Nene), *por ayudar a mi mamá a criarme, por quererme y cuidarme como su propia hija, por velar que fuera siempre por el mejor camino, por contribuir con mi formación de ser una mejor persona.*

A mi novio Rubier, *por cuidarme, por valorarme, por apoyarme, por darme fuerzas, por su comprensión, por su paciencia, por soportar mis malos ratos, por todo este tiempo juntos sin importar cuan lejos estamos, por sobre todas las cosas por quererme, y* ***a su familia*** *por acogerme como suya, por darme su cariño y mostrar su preocupación.*

A mi abuela Elsa, *por su cariño y por estar presente en esta etapa de mi vida tan importante.*

A Mailé*, por preocuparse por mí, por correr conmigo cada vez que necesitaba algo, por su cariño y por*

el tiempo que me ha dedicado.

A mi compañero de tesis Omar, *por pasar juntos malas noches, por ser tan comprensivo y darme la oportunidad de desarrollar la tesis a su lado.*

A todos mis amigos de la UCI *y compañeros de estudios del 6107, del 6201, del 6309 hasta el 6509. Pero no quisiera dejar de mencionar a mis amigos* ***Liuva, Karel, Erick, Yissel, Martha, Felix, Eduardo y Daldis,*** *con los cuales he pasado uno de los mejores momentos de mi vida, la vida universitaria y en especial a dos personas que quiero mucho, una es* ***a Yudeily (Yudy),*** *por ser mi mejor amiga en estos 5 años, por soportarme, por ayudarme siempre que lo necesité, por compartir conmigo en las buenas y en las malas, por todos los buenos consejos que me has dado, gracias Yudy por ser como eres, y* ***a tu familia****, por dejarme entrar en ella,* ***a mimi (Maribel)*** *gracias por todo tu cariño. Y la otra persona es* ***Fernando (el Ferna),*** *por estar ahí siempre que lo necesité, por ser mi amigo y por ser para mí como el hermano varón que siempre he querido.*

A mis compañeros del deporte*. En especial* ***a Henry, a Miguel Socorro, a Miguel Albuerne y a los profes Antonio Larrude y Jose Luis,*** *por enseñarme las técnicas que aprendí y por hacerme reír en cada entrenamiento.*

A Yamisleidys, Osleidys, Yasiel, Adriana, Tomás, Jezabel y a la China, *por ser mis amigos desde antes de entrar a la universidad y por seguir ocupando un pedacito de mí.*

A la FEU *y a todo el piquete que compartió conmigo todas esas noches de reuniones, de trabajo para algún evento, esos días largos de sueño y cansancio para que la facultad saliera bien y los muchachos disfrutaran con nuestro esfuerzo, a la FEU por ser mi segunda universidad.*

A todos los profesores *que han contribuido con mi formación profesional desde pequeña, en especial* ***a la Nena, a Jorgito, a Lidia, a Mercedes y a María Isabel,*** *por guiarme por el buen camino, por confiar en mí como persona y en mis capacidades y por construir las bases de lo que hoy he logrado.*

A todos los amigos que han formado parte de mi historia, la historia que de aquí en adelante podré contar. Y a todo aquel que luego del abrazo y del saludo, preguntó ¿Cómo va la Tesis?, a todos ellos muchísimas gracias….

De Omar:

Aunque no se puede resumir tantos sentimientos en unas pocas líneas es meritorio nombrar aquellos que siempre estuvieron conmigo durante este arduo camino.

A ti, esposa mía, *por soportar la lejanía y esperarme con paciencia, por pasar el dolor de tener que despedirme cada vez que volvía a tus brazos, por darme una nueva luz en mi vida como lo es nuestra hija.*

A ti, madre querida, *por darme vida y enseñarme a andar, por preocuparte por mí en todo momento y estar presente en las buenas y en las malas.*

A mi hermano, *por enseñarme el color de la paciencia pero sobre todo por el cariño que tenemos.*

A mi padre y abuelos, *que me enseñaron qué es "sacrificio" y a no tenerle miedo a la desventura.*

A mis suegros, *por prestarme de su tiempo y aconsejarme en esta investigación, a los cuales les debo el regalo de la mejor mujer del mundo.*

A mis maestros, *que moldearon mi forma de pensar y me abrieron los ojos en este mundo nuevo.*

A todos aquellos que fueron partícipes de mi vida profesional y personal, por los cuales he llegado hasta aquí, hoy. Este trabajo como producto de sus enseñanzas está dedicado a ustedes.

RESUMEN

En el Centro Nacional de Genética Médica (CNGM), los cálculos estadísticos de los estudios de Epidemiología Genética se hacen utilizando herramientas propietarias, cuyas licencias son muy costosas. Estas están diseñadas para hacer análisis estadísticos y ninguna de estas herramientas reúne las funcionalidades necesarias para realizar estos estudios, siendo necesario utilizar varias de ellas para la obtención de los resultados finales, lo cual implica un gasto de tiempo considerable en la realización del análisis de la información.

Debido a esto, el CNGM en colaboración con la Universidad de las Ciencias Informáticas concertó la construcción de la aplicación SEEGEN-R (Sistema Estadístico de Epidemiología Genética- basado en R). Esta aplicación realiza estudios de Genética Poblacional, que al igual que los estudios de Epidemiología Genética y de Epidemiología Tradicional, están comprendidos dentro de la disciplina que lleva por nombre este software. Por lo antes expresado se decide incluirle una extensión para los estudios de Epidemiología Genética para obtener una herramienta más potente que centralice los análisis estadísticos.

Esta extensión informática permite realizar estudios específicos para la genética: estudios de asociación genética, estudios de interacción, estudios clásicos en gemelos, estudios de agregación familiar y sus variantes, realizando diferentes cálculos estadísticos utilizando el paquete de R como motor estadístico, dando la opción de escoger el tipo de estudio que se desea realizar. En esta herramienta se utilizó RUP como metodología de desarrollo, Netbeans v7.2 como entorno de desarrollo, RCaller y RUniversal como librerías para la conexión de los lenguajes de programación Java y R.

PALABRAS CLAVES: Epidemiología Genética, análisis estadístico, Centro Nacional de Genética Médica, estudios epidemiológicos.

Tabla de Contenidos

TABLA DE CONTENIDOS

Tabla de Contenidos

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

A mediados de los años 80 emergió una nueva disciplina como resultado de la interacción de muchos años entre las ciencias Genética y la Epidemiología. Este nuevo campo se llamó Epidemiología Genética, su trabajo se centra en la interacción que existe entre los factores genéticos y los factores medioambientales en la ocurrencia de las enfermedades humanas. La expansión de esta nueva ciencia se aceleró aún más con los extraordinarios avances científicos en el campo de la biología molecular que han ampliado nuestro entendimiento de enfermedades genéticas a nivel molecular y tienen aplicaciones para la clasificación, diagnosis y tratamiento para muchos desórdenes Mendelianos. [1]

En Cuba la principal fortaleza de la genética es la genética comunitaria, no solo por la población que posee ni por la accesibilidad de ésta a los servicios médicos, sino por la organización de un sistema nacional de salud que va desde el nivel primario, por ejemplo consultorios o policlínicos, hasta el terciario que son las instituciones, como el Centro Nacional de Genética Médica (CNGM). Dado el drástico descenso en el coste de "genotipado[1]", los estudios de Epidemiología Genética suelen disponer de una gran cantidad de información y para realizar el análisis estadístico de estos es preciso utilizar un software que integre la mayor cantidad de funcionalidades y que permita realizar dichos análisis de forma rápida y factible. [2]

El Centro Nacional de Genética Médica (CNGM) tiene dentro de sus funciones principales las investigaciones básicas y aplicadas en el campo de la Genética Médica, la Inmunología, la Bioquímica, Epidemiología Genética y otras disciplinas afines dirigidas a la obtención de nuevos conocimientos, evaluación y desarrollo de nuevas tecnologías, productos y procedimientos de trabajo, con el fin de mejorar los niveles de salud de nuestro pueblo y disminuir el impacto de las enfermedades con implicación genética en el cuadro de la morbimortalidad[2] del país y realizar aportes al desarrollo de estas ramas de las ciencias, teniendo en cuenta las potencialidades que se derivan de su integración.[3]

Con el fin de cumplir estas metas el CNGM en colaboración con la Universidad de las Ciencias Informáticas (UCI) concertaron la construcción del software SEEGEN-R. Esta aplicación realiza los estudios de Genética Poblacional, que al igual que los estudios de Epidemiología Genética y de Epidemiología Tradicional, están comprendidos dentro de la disciplina mencionada anteriormente. Este sistema está basado en extensiones independientes, con lo que se ahorra mucho trabajo en términos

[1] **Genotipado:** es la técnica de laboratorio que se utiliza para determinar la información genética de un organismo, o genotipo, y poder diferenciarlo del resto.
[2] **Morbimortalidad:** Son las enfermedades causantes de la muerte en determinados poblaciones, espacios y tiempo.

de mantenimiento y reutilización de código, facilitando el rápido crecimiento del software. Una característica que tiene a su favor es la usabilidad y además utiliza R como lenguaje de programación estadístico, aprovechando gran parte de sus funcionalidades y facilidad de acceso a una amplia variedad de paquetes estadísticos y gráficas de alta calidad.

Los cálculos estadísticos de los estudios de Epidemiología Genética se hacen a través de herramientas propietarias, por lo que deben pagar las licencias correspondientes, estas están diseñadas específicamente para el análisis estadístico por tanto se necesita una capacitación al personal para dominar su entorno de trabajo. Además con ninguna de estas herramientas el especialista puede realizar una investigación que agrupe los estudios de casos y controles teniendo en cuenta los familiares enfermos de los casos, los estudios de casos particulares como el de gemelos monocigóticos frente a dicigóticos, ni de asociación, solo por mencionar algunos de ellos. Al no existir un software que agrupe estas funcionalidades, se hace necesario utilizar varios de ellos para la obtención de los resultados finales, lo cual implica un gasto de tiempo considerable en la realización del análisis de la información. Por lo antes expresado, es necesario incluir una extensión para los estudios de Epidemiología Genética en la aplicación informática SEEGEN-R para obtener una herramienta más potente que realice y centralice los análisis estadísticos.

Por lo anteriormente expuesto se define como **problema a resolver:** *¿Cómo realizar análisis estadísticos en estudios de Epidemiología Genética en el CNGM?*

Para darle solución al problema planteado se necesita realizar una investigación que abarque el **objeto de estudio**: *Análisis estadísticos en la disciplina "Epidemiología Genética".*

Implicado en **el campo de acción**: *Proceso de gestión de análisis estadísticos en estudios de Epidemiología Genética.*

Para satisfacer las necesidades del CNGM, el trabajo persigue como **objetivo general** de la investigación: *Desarrollar una extensión para el análisis estadístico de Epidemiología Genética para la aplicación informática SEEGEN-R v1.0.*

Este objetivo general se desglosa en los siguientes **objetivos específicos**:

- Identificar las funcionalidades de la extensión informática "Epidemiología Genética".
- Diseñar las funcionalidades de la extensión informática "Epidemiología Genética".
- Implementar las funcionalidades de la extensión informática "Epidemiología Genética".
- Validar el funcionamiento de la extensión informática "Epidemiología Genética".

Para dar cumplimiento a los objetivos específicos se definen las siguientes **tareas de la investigación**:

- Revisión bibliográfica sobre los siguientes temas:
  - ❖ Estudios de asociación genética.
  - ❖ Estudios de agregación familiar.
  - ❖ Estudios clásicos en gemelos.
  - ❖ Estudios de interacción.
- Análisis de las aplicaciones existentes que posibilitan realizar análisis estadísticos en el campo de la epidemiología genética.
- Realizar el estudio de la arquitectura del sistema SEEGEN-R.
- Identificación de los requerimientos de la extensión informática “Epidemiología Genética”.
- Realizar diagramas de clases de diseño y la descripción de las clases de la extensión informática “Epidemiología Genética”.
- Realizar la descripción del diagrama de componentes de la extensión informática “Epidemiología Genética”.
- Implementación de las funcionalidades de la extensión informática “Epidemiología Genética”.
- Diseño de los casos de prueba correspondientes a las funcionalidades de la extensión informática “Epidemiología Genética”.
- Ejecución de pruebas a las funcionalidades de la extensión informática “Epidemiología Genética”.

Estructura del documento

Capítulo 1. Fundamento teórico: En este capítulo se presenta una breve descripción del proceso de gestión de análisis estadísticos en estudios de Epidemiología Genética en Cuba. Además, se describen las herramientas, metodologías y tecnologías posibles a utilizar para desarrollar la extensión informática.

Capítulo 2. Características del sistema: En este capítulo se define lo que debe hacer la aplicación a través de la caracterización del negocio. Se describen los procesos que son objeto de automatización. Se dan a conocer los requerimientos funcionales y no funcionales, los actores y casos de uso del sistema a desarrollar.

Capítulo 3. Análisis y diseño del sistema: En este capítulo se definen las clases del diseño con las que va a contar la extensión. Se exponen los diagramas de clases del diseño, los diagramas de secuencia realizados en el diseño, así como los patrones de diseño y de arquitectura aplicados.

Capítulo 4. Implementación y prueba: En este capítulo realiza la representación de los diagramas de componentes de la extensión. Además se brinda una descripción de los principales métodos implementados, se muestran imágenes de la interfaz desarrollada y se hace referencia a su validación.

CAPÍTULO 1. FUNDAMENTO TEÓRICO

Introducción

En el presente capítulo se presenta una descripción del proceso de gestión de análisis estadísticos en estudios de Epidemiología Genética en Cuba. Se explican brevemente los estudios de asociación genética, clásicos en gemelos, así como los de interacción y de agregación familiar. Además, se describen las herramientas, se fundamentan las metodologías y tecnologías posibles a utilizar para desarrollar la extensión informática.

1.1. Estudios de Epidemiología Genética

La epidemiología es, en la acepción más común, el "estudio de las epidemias", es decir, de las “enfermedades que afectan transitoriamente a muchas personas en un sitio determinado". Su significado deriva del griego Epi (sobre), Demos (Pueblo) y Logos (ciencia). Una definición técnica es la que propone que la epidemiología es "el estudio de la distribución y determinantes de las enfermedades en poblaciones humanas". De esta manera podemos definir como Epidemiología Genética "el estudio de la distribución y determinantes genéticos de las enfermedades en poblaciones humanas". [4]

La disciplina “Epidemiología Genética”, contempla tres grandes grupos de estudios, los Estudios de Genética Poblacional, Estudios de Epidemiología Tradicional y uno que lleva su mismo nombre, Estudios de Epidemiología Genética. Su medio es la identificación de la función que cumplen los factores genéticos, en interacción con factores ambientales, en el origen de las enfermedades en los seres humanos [5]. Este último grupo a su vez, agrupa varios estudios, a continuación se describen algunos de ellos:

1.1.1. Estudios de agregación familiar

Un aspecto fundamental de la epidemiología genética es el estudio de la agregación (o recurrencia) de ciertas enfermedades en determinadas familias. King MC, Lee GM, Spinner NB, Thomson G y Wrensch MR, propusieron tres preguntas que permiten identificar los alcances de los estudios sobre recurrencia familiar:

1. ¿Hay enfermedades que afectan a varios miembros de una misma familia?

2. ¿Se relaciona dicha agregación familiar con una exposición ambiental común, con una susceptibilidad heredada, o con una herencia cultural de factores de riesgo?

3. De existir, ¿cómo se hereda la susceptibilidad genética?

La observación de la prevalencia de cierta enfermedad en familiares de los casos índices[3] y de los controles[4], permite determinar la existencia de una agregación familiar. Dicha agregación existe cuando los familiares de los individuos afectados corren un mayor riesgo de padecer la enfermedad que los familiares de individuos no afectados. Este método es eficiente y poco costoso, pero una de sus limitaciones es que la información sobre las características de los familiares de los casos y controles puede dar lugar a sesgos. Por ejemplo, si el investigador sabe de la presencia de la enfermedad en la familia del participante, existe la posibilidad de un diagnóstico desacertado. La capacidad de recordación de los familiares y su conocimiento de las características del trastorno también pueden ser mayores cuando existe un pariente afectado. [5]

Los estudios de agregación familiar son los primeros que se realizan para identificar el papel de los genes en la aparición de las distintas enfermedades. Responde a la primera pregunta de la estrategia familiar de la epidemiología genética: ¿Se agrega la enfermedad objeto de estudio en las familias? Existen tres tipos de estudios de agregación:

Estudios casos y controles

Lo primero en este tipo de estudio es detectar a las personas con la enfermedad de interés (casos) y luego seleccionar cierta cantidad de personas libres de la enfermedad (controles). Los casos y los controles son estudiados para investigar cuáles factores difieren entre ellos. En estos estudios los participantes se seleccionan según su estado de enfermedad y se mira en retrospectiva que pudo haberla causado. Algunas de sus ventajas son:

- Son más económicos y más rápidos especialmente en el estudio de enfermedades cuyo tiempo de desarrollo es muy largo (por ejemplo enfermedades crónicas).
- Muchos factores de riesgo pueden estudiarse simultáneamente.
- Son adecuados para estudiar enfermedades raras (muy poco frecuentes).

Para este tipo de estudio se utiliza una entrada de datos que contiene los datos de los casos y los controles como se expone en la tabla 1.

Identificación del individuo	Estado de salud	Total de familiares enfermos	Total de familiares
1	Caso	9	18
2	Control	3	10
3	Control	2	16

[3] **Caso índice:** Es el individuo afectado, por medio del cual se incorpora su familia al estudio.
[4] **Controles:** Individuos no afectados.

4	Caso	5	12
…	Caso / Control	…	….

Tabla 1: Entrada de datos: Estudios de agregación familiar para casos y controles

El campo total de familiares puede ser específico para un tipo de familiar, por ejemplo: madre, padre, hermano carnal, abuela, abuelo, tíos, etc.… o puede tratarse como se ve en el ejemplo del total de familiares estudiados para el individuo.

A partir de este conjunto de datos se obtiene una tabla de contingencia[5] que recoge los valores contabilizados por grupos (ver tabla 2). Por ejemplo para este caso sería:

	Total de familiares enfermos	**Total de familiares**
Casos	14	30
Controles	5	26

Tabla 2: Tabla de contingencia: Estudios de agregación familiar para casos y controles

Teniendo como entrada la tabla de contingencia, se aplican en este estudio los métodos de Chi Cuadrado de Pearson, Chi Cuadrado de Pearson con corrección de Yates y la Prueba exacta de Fisher como métodos estadísticos para realizar el análisis de agregación familiar para casos y controles. Obteniendo los valores de estos cálculos el especialista puede interpretar la información del estudio realizado. Comparando familiares de casos y familiares de controles se puede estimar que en la medida que aumenta la proporción de genes en común, mayor es la frecuencia o probabilidad de aparición de la enfermedad.

Estudio de casos y población

El estudio de casos y población tiene como objetivo medir la recurrencia o frecuencia de ciertas enfermedades en determinadas familias. Para este tipo de estudio se utiliza la siguiente entrada de datos.

Identificación del individuo	**Tipo de Familiar**	**Total de Familiares**	**Familiares enfermos**
1	Tíos	4	2
2	Primos	3	0

Tabla 3: Entrada de datos: Estudios de agregación familiar para casos y población

Como salida del estudio se muestra para cada tipo de familiar la probabilidad de expresión de la

[5] **Tablas de contingencia:** En estadística se emplean para registrar y analizar la relación entre dos o más variables, cada una con un número determinado de resultados posibles, habitualmente de naturaleza cualitativa. Una de las características determina las filas de la tabla y la otra las columnas.

enfermedad (calculada mediante la razón de casos favorables sobre casos posibles), el resultado de las Dócimas de λ[6] mide la razón establecida entre la proporción de expresión de una enfermedad en los familiares sobre la proporción de igual enfermedad en la población, esta última se tiene como dato preliminar. Este estudio también se puede realizar a través de las dócimas de Chi-Cuadrado. Los resultados arrojados permiten demostrar que la enfermedad resulta ser más frecuente en las familias que en la población general.

Estudio de casos particulares

El estudio de casos particulares es muy similar al de casos y controles, solo que los casos y controles van a ser familiares particulares desde el punto de vista de la información genética que los vincula. Tiene varias opciones, entre las que se encuentran estudios para gemelos monocigóticos frente a dicigóticos, estudios para gemelos dicigóticos frente a un hermano carnal, estudios para primos hermanos y para cónyuges. Por ejemplo una de sus variantes es:

Los estudios de gemelos monocigóticos frente a dicigóticos, donde los casos van a ser los gemelos enfermos y los controles van a ser sus cógemelos, y la variable a medir va a ser la cigocidad, el objetivo es identificar si existe dependencia entre el cógemelo enfermo y la cigocidad.

Identificador del gemelo	Cigocidad	Estado del cógemelo
1	MC	Enfermo
2	DC	No enfermo

Tabla 4: Entrada de datos: Estudios de agregación familiar para casos particulares

A partir de estos datos se calcula la tabla de contingencia y se realizan los cálculos de métodos de Chi Cuadrado de Pearson, Chi Cuadrado de Pearson con corrección de Yates y la Prueba exacta de Fisher. Los resultados arrojados por estos estudios permiten demostrar la causa genética de la agregación familiar, ya que puede obedecer a múltiples causas: herencia cultural, factores ambientales y genéticos.

1.1.2. Estudios de asociación genética

Los estudios de asociación genética buscan establecer la relación estadística entre variables genéticas poblacionales y un fenotipo determinado (ejemplo: rasgo, riesgo de enfermedad, etc.). Estos están siendo utilizados para descubrir el componente genético que subyace a las enfermedades comunes de alta prevalencia como la diabetes mellitus[7], la enfermedad coronaria o la insuficiencia cardíaca. Se

[6] **Lambda (Λ o λ):** La letra lambda es del alfabeto griego y es usada como símbolo en diversas ramas.
[7] **Diabetes mellitus:** Es una enfermedad que se caracteriza por un aumento de los niveles de glucosa (azúcar) en la sangre.

trata generalmente de estudios de cohortes prospectivos o de tipo casos-controles en los cuales se establece el peso relativo del componente genómico con respecto a otros factores como el ambiente, en el riesgo de desarrollar la enfermedad [6]. Los estudios de asociación buscan relacionar un marcador genético particular con una enfermedad (o un rasgo complejo) a través de una población, más que dentro de familias.

Cuando se planea un estudio de asociación genética, se consideran cuatro componentes mayores: la enfermedad o el rasgo a ser estudiado (punto final de estudio), el grupo de individuos en el cual el rasgo o enfermedad va a ser medido (diseño propiamente dicho), los marcadores genéticos que van a ser genotipificados y, por último, el método analítico para testear la asociación entre el genotipo y el fenotipo (plan estadístico).

Los estudios de asociación genética de rasgos complejos presentan hasta el momento algunos desafíos adicionales a los tecnológicos y logísticos, ya que habitualmente son mal interpretados y por lo tanto parecen ser pobremente reproducibles. Habitualmente se interpreta como significativa la asociación entre una variable genética y un fenotipo cuando el valor de $p < 0,05$. Sin embargo, la mayoría de tales asociaciones tienen dificultades para reproducirse consistentemente. [6]

Para este tipo de estudio se tiene en cuenta el estado de salud y el polimorfismo de riesgo del individuo como se expone en la tabla 5.

Identificador del individuo	**Estado de salud**	**Polimorfismo genético**
1	Enfermo	GG
2	No enfermo	AG

Tabla 5: Entrada de datos: Estudios de asociación genética

A partir de estos datos se calcula las tablas de contingencias siguiendo cuatro modelos.

Tabla de contingencia y cálculo del Odds Ratio en un estudio de asociación genética según el modelo de codominancia.

	GG	**GA**	**AA**	**Total**	**OR (AA vs GG)**	**OR (GA vs GG)**
Casos	A	B	C	N1	$\frac{CxD}{FxA}$	$\frac{BxD}{ExA}$
Controles	D	E	F	N2		
Total	M1	M2	M3	T		

Tabla 6: Tabla de contingencia: Estudio de asociación genética-Modelo de codominancia.

Tabla de contingencia y cálculo del Odds Ratio en un estudio de asociación genética según el modelo

de dominancia.

	GG	GA+AA	Total	OR
Casos	A	(B+C)	N1	$\frac{[(B+C)x\ D]}{[(E+F)x\ A]}$
Controles	D	(E+F)	N2	
Total	M1	M2+M3	T	

Tabla 7: Tabla de contingencia: Estudio de asociación genética-Modelo de dominancia

Tabla de contingencia y cálculo del Odds Ratio en un estudio de asociación genética según el modelo de recesividad.

	GG+GA	AA	Total	OR
Casos	(A+B)	C	N1	$\frac{[(D+E)x\ C]}{[(A+B)x\ F]}$
Controles	(D+E)	F	N2	
Total	M1+M2	M3	T	

Tabla 8: Tabla de contingencia: Estudio de asociación genética -Modelo de recesividad

Tabla de contingencia y cálculo del Odds Ratio en un estudio de asociación genética según el modelo aditivo.

	G	A	Total	OR
Casos	(2A+B)	(2C+B)	2N1	$\frac{[(2D+E)x(2\ C+B)]}{[(2A+B)x\ (2F+E)]}$
Controles	(2D+E)	(2F+E)	2N2	
Total	2M1+M2	2M3+M2	2T	

Tabla 9: Tabla de contingencia: Estudio de asociación genética--Modelo aditivo

Este estudio utiliza los métodos de Chi Cuadrado de Pearson, Chi Cuadrado de Pearson con corrección de Yates, Prueba exacta de Fisher y Odd Ratio para realizar los análisis estadísticos.

Los resultados obtenidos arrojan información sobre la localización de secuencias polimórficas del ADN relacionados con la aparición de las enfermedades en el hombre. Los estudios de asociación permiten identificar cuán frecuente es una determinada enfermedad en el hombre si presenta un determinado polimorfismo génico cuya presencia incrementa el riesgo de padecerla.

1.1.3. Estudios clásicos en gemelos

Los estudios de gemelos establecen las bases de un trastorno examinando la probabilidad de un niño de padecer un trastorno si sus gemelos tienen ese trastorno, conocida también como la tasa de concordancia [7]. Este tipo de estudio se aprovecha del hecho de que hay dos tipos de gemelos, que

difieren en la proporción de material genético compartido. "Fraternal", o dicigóticos, son los gemelos que tienen la mitad de sus genes en común, mientras que "idénticos" o monocigóticos, son los gemelos que comparten todos sus genes. Los investigadores estudian los gemelos idénticos criados por separado para examinar las influencias ambientales. Como alternativa, los gemelos fraternos criados en el mismo hogar pueden ser comparados con los gemelos idénticos criados juntos para identificar influencias genéticas [8]. Este tipo de estudio presenta algunos beneficios:

Los estudios de gemelos son importantes porque permiten la investigación de cuestiones relacionadas con:

- Estabilidad del desarrollo y el cambio.
- Las contribuciones diferenciales de factores genéticos y ambientales a través de diferentes comportamientos y estados relacionados (por ejemplo, la iniciación, la dependencia).

Para este tipo de estudio se tiene en cuenta la cigocidad y el estado de salud de los gemelos como se expone en la tabla 10.

Identificador del gemelo	Cigocidad	Estado de salud
1	MC	Enfermo
2	DC	No enfermo

Tabla 10: Entrada de datos: Estudios de clásicos en gemelos

A partir de estos datos se calcula la tabla de contingencia, la cual contiene la frecuencia absoluta de los gemelos monocigóticos enfermos, gemelos monocigóticos sanos, gemelos dicigóticos enfermos y gemelos dicigóticos sanos. Para realizar este estudio se aplica la Estadística Descriptiva a la tabla de contingencia obtenida. Los resultados permiten demostrar el papel de los factores genéticos en la aparición de las enfermedades en el hombre. Además brinda el riesgo que tiene un gemelo de padecer una enfermedad específica dado que su cogemelo presenta dicha entidad clínica, a través de la proporción de concordancia para el caso índice. Al calcular la heredabilidad permite conocer la contribución de los factores genéticos en la aparición de las enfermedades.

1.1.4. Estudios de interacción

Los estudios de interacción son uno de los más importantes dentro los estudios de Epidemiología Genética y a su vez contiene tres tipos de interacciones más (Interacción Genoma-Ambiente, Interacción Gen-Gen e Interacción Gen-Ambiente), estos estudios comprenden la predisposición de una persona hacia una enfermedad determinada dado su acerbo genético y los factores del entorno donde persiste. Para realizar los estudios de Interacción los especialistas deben analizar

estadísticamente los datos obtenidos de una investigación previa sobre los estudios de asociación de factores de riesgo y ambientales y estudios de agregación familiar.

La existencia de una interacción entre el gen y el medio ambiente es, como tal, una asociación estadística, y esta no es necesariamente causal. Sin embargo, es importante recalcar que estos métodos constituyen herramientas útiles en el análisis de interacciones genético-ambientales, ya que identifican factores que podrían resultar importantes para la prevención del trastorno en estudio [5]. La presencia de una combinación de determinados alelos puede producir que se establezcan interacciones gen*gen que influyan en la aparición o no de una enfermedad. El efecto de la interacción depende de genotipos específicos presentes en distintos genes o regiones génicas. Las interacciones gen-gen pueden ser simplemente aditivas o mucho más complicadas. Sin embargo, la dificultad en la detección de interacciones gen-gen es un problema común en estudios epidemiológicos. [9]

Para este tipo de estudio se toman todas las variables genéticas y ambientales que son significativas para la dolencia en cuestión y se construye una tabla de contingencia por par, teniendo en cuenta el diseño (Casos-casos, Casos-control, Cohorte) escogido por el especialista.

Caso o Control	Variable genética 1	Variable genética 2	Variable genética 3	Variable ambiental 1	Variable ambiental 2	Variable ambiental 3
CASO	SI	SI	NO	SI	NO	SI
CASO	NO	NO	NO	SI	NO	SI
CONTROL	SI	SI	NO	SI	NO	NO
CONTROL	SI	NO	NO	SI	SI	NO
CONTROL	SI	SI	NO	SI	SI	SI
CASO	NO	NO	NO	SI	NO	SI
CASO	NO	SI	SI	NO	NO	NO

Tabla 11: Entrada de datos: Estudios de interacción

Tabla de contingencia y cálculo del Odds Ratio en un estudio de interacción para el diseño de casos y controles.

	Genotipo presente		**Genotipo ausente**	
	Presente	**Ausente**	**Presente**	**Ausente**
Caso	a	b	e	f
Control	c	d	g	h
Odds Ratio	**OR_{11} = ah/cf**	**OR_{01} = bh/df**	**OR_{10} = eh/gf**	**OR_{00} = 1.0 (ref.)**

Tabla 12: Tabla de contingencia: Estudio de interacción - Diseño de casos y controles

Tabla de contingencia y cálculo del Riesgo Relativo en un estudio de interacción para el diseño de

cohorte.

Estatus de la enfermedad	Genotipo presente		Genotipo ausente	
	Presente	Ausente	Presente	Ausente
Caso	a	b	e	f
Control	c	d	g	h
Riesgo	$r_{11} = a/(a+c)$	$r_{01} = b/(b+d)$	$r_{10} = e/(e+g)$	$r_{00} = f/(f+h)$
Riesgo Relativo	$RR_{11} = r_{11}/r_{00}$	$RR_{01} = r_{01}/r_{00}$	$RR_{10} = r_{10}/r_{00}$	$RR_{00} = 1.0(ref.)$

Tabla 13: Tabla de contingencia: Estudio de interacción - Diseño de cohorte

Este estudio utiliza los métodos de Chi Cuadrado de Pearson, Chi Cuadrado de Pearson con corrección de Yates, Prueba exacta de Fisher, Odd Ration (OR) y Risk Relative (RR) para realizar los análisis estadísticos. Los diseños anteriores, permiten demostrar como la coexistencia de factores de riesgo de naturaleza ambiental y la presencia de un genoma predisponente, potencializan el riesgo de padecer determinadas enfermedades. Los resultados obtenidos de los cálculos realizados demuestran el verdadero papel modulador del ambiente sobre la predisposición genética. Desde el punto de vista práctico permite demostrar cuánto se puede reducir el riesgo de padecer una enfermedad si se controlan los factores de riesgo de naturaleza ambiental.

1.2. Software que permiten realizar cálculos estadísticos

En la actualidad existen diferentes aplicaciones que se utilizan para realizar análisis estadísticos, como parte de la investigación fueron analizadas varias de ellas, en busca de posibles soluciones que pudieran ayudar a resolver el problema planteado. A continuación se detallan algunas características de esos programas:

1.2.1. IBM SPSS Statistics

SPSS [10] es un programa estadístico informático muy usado en las ciencias sociales y las empresas de investigación de mercado. Es uno de los programas estadísticos más conocidos teniendo en cuenta su capacidad para trabajar con grandes bases de datos y una interfaz sencilla para la mayoría de los análisis. El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos. Cada uno de estos módulos se compra por separado. Actualmente, compite no sólo con sistemas licenciados como lo son SAS, MATLAB, Statistica, Stata, sino también con software de código abierto y libre, de los cuales el más destacado es el Lenguaje R.

Este software presenta muchas funcionalidades desde el punto de vista estadístico pero es un software

propietario, altamente costoso de adquirir su licencia, lo que significaría un gasto sustancial para la economía del país y desde el punto de vista de la Epidemiología Genética, realiza pruebas no paramétricas, entre ellas el Chi Cuadrado, además permite realizar cálculos básicos de estadística descriptiva y calcula el coeficiente de correlación de Pearson, pero SPSS no tiene una interfaz amigable para los genetistas al no estar enfocado para ellos, el especialista tendría que ser estadístico para poder hacer un estudio con este software.

1.2.2. PSPP

PSPP es una herramienta para el análisis estadístico de muestras de datos. PSPP lee un fichero de sintaxis y un fichero de datos, analiza los datos, y escribe los resultados en una lista de ficheros o por la salida estándar. El lenguaje aceptado por PSPP es parecido a los aceptados por los productos estadísticos de SPSS. Los detalles del lenguaje de PSPP se proporcionan en el manual de la página web del proyecto. PSPP produce salidas de dos maneras: tablas y diagramas. [11]

Una de las características ampliamente destacables de esta aplicación es que se integra a la perfección con OpenCalc y Gnumeric, lo que nos permite manejar los datos estadísticos directamente desde estas hojas de cálculo, una funcionalidad que ayudará a ahorrar tiempo y a trabajar de manera mucho más ordenada para evitar errores de tipeo y crear resultados confiables. La capacidad de cálculo es astronómica, puede trabajar hasta con un billón de variables y casos de manera simultánea. También se puede destacar la velocidad de procesamiento y la estabilidad de la aplicación, por lo que se puede decir, que es una de las mejores herramientas estadísticas libres que se puede encontrar. [12]

El proyecto PSPP es libre, una alternativa al software propietario SPSS, de hecho pretende ser su sustitución en versión libre, aunque por el momento dispone de menos funcionalidades que el SPSS. Actualmente tiene la mayor parte de sus módulos en desarrollo.

1.2.3. InfoStat

InfoStat es un software para análisis estadístico de aplicación general. Cubre tanto las necesidades elementales para la obtención de estadísticas descriptivas y gráficos para el análisis exploratorio, como métodos avanzados de modelación estadística y análisis multivariado. Una de sus fortalezas es la sencillez de su interfaz combinada con capacidades profesionales para el cálculo y el manejo de datos. Debido al origen universitario, el programa tiene muchas facilidades para la enseñanza de la estadística que no son fáciles en encontrar en otros programas similares.

Una propiedad casi única del software estadístico es la habilidad de conectarse con R, una plataforma

de desarrollo de algoritmos estadísticos de dominio público de gran crecimiento. InfoStat se conecta con R de dos maneras: mediante un intérprete integrado que permite ejecutar script de R sin salir del ambiente de trabajo de InfoStat y mediante el desarrollo de aplicaciones utilizando el motor de cálculo de R pero con la interfaz amigable que los usuarios esperan. [13]

Este software es propietario y para trabajar con él y mantenerlo actualizado es necesario adquirir la licencia por una suma considerable de dinero, lo cual sería un gasto más para el país. Desde el punto de vista de la Epidemiología Genética se pueden crear tablas de contingencia, calcular el Odds Ratio (OR) y el Risk Relative (RR) utilizados en los estudios de interacción, pero esto no es suficiente para realizar un estudio completo de Epidemiología Genética.

1.2.4. Arlequín

El programa Arlequín (versión actual 2001), liberado en 1997, todavía es muy popular. Consiste en un ambiente de programa informático tipo explorador en genética de poblaciones, capaz de manejar grandes muestras de datos moleculares, mientras que conserva la capacidad de analizar datos genéticos convencionales. Arlequín puede usar muchos tipos diferentes de datos, tales como datos moleculares y frecuencias genotípicas o haplotípicas, incluyendo datos codominantes o recesivos, pero aún no, datos dominantes. El formato de datos se especifica en un archivo de entrada. El usuario puede crear un archivo de datos desde el principio, utilizando un editor de texto y palabras claves apropiadas, o utilizando el “Project Outline Wizard”. Los datos pueden importarse de los archivos creados por otros programas, incluyendo MEGA, BIOSYS, GENEPOP y PHYLIP.

Pueden existir datos ausentes o ambiguos. Existe un manual de usuario muy detallado, que incluye una gran cantidad de información teórica, fórmulas y referencias. Este programa tiene la capacidad de analizar una cantidad importante de datos y hay una opción “Batch Files” (archivo por tandas) disponible. [14]

Ventajas:

- Apoyo adecuado a través de un manual detallado, que incluye suficiente información teórica, fórmulas y referencias; tiene, además, un sitio bien organizado en la Web con secciones especializadas como ‘Preguntas más frecuentes’.
- La interfaz gráfica es muy fácil de usar.

Desventajas:

- Hay que aprenderse muchas características y opciones.

- Organizar el archivo de datos puede ser una labor compleja y debe ser estructurado correctamente.

Este software está disponible en versiones para Windows, Mac y Linux, todas gratuitas. Desde el punto de vista de la disciplina Epidemiología Genética, es muy usado, pero es un programa de propósito general para análisis de datos en genética de poblaciones, no para los estudios epidemiológicos tratados al inicio del capítulo, por lo tanto no ayuda a resolver el problema planteado.

1.2.5. Epidat

Epidat es un programa de epidemiología y estadística de libre distribución. Es una de las herramientas más útiles y actualizadas que constituyen la base del análisis epidemiológico y su aplicación en entornos técnicos tan diversos como la vigilancia de la salud, el análisis de situación, la epidemiología clínica, el análisis demográfico y el examen de riesgos, entre muchos otros. [15]

En el desarrollo de Epidat 4.0 participan estadísticos, epidemiólogos e informáticos de Galicia, Cuba y OPS. Esta nueva versión se está programando en Java, debido a la versatilidad de este lenguaje, que permite el funcionamiento de la aplicación en distintos sistemas operativos, tales como Windows, Linux y Macintosh. El contenido de Epidat ha ido creciendo con las sucesivas versiones hasta llegar a 12 módulos en la versión 3, ahora ampliados a 19 en su versión 4.0, como resultado de las discusiones del equipo de trabajo y de las numerosas sugerencias aportadas por los usuarios. Estos módulos abarcan una amplia variedad de técnicas estadísticas y epidemiológicas, que cubren las necesidades más generalizadas de los epidemiólogos y técnicos de salud, así como las carencias presentes en los paquetes estadísticos más frecuentemente empleados por los profesionales de la Epidemiología. [16]

La nueva aplicación está actualmente en fase de desarrollo, se han liberado 9 paquetes de los 19 propuestos. Para el manejo del programa no es necesario tener amplios conocimientos epidemiológicos, ni estadísticos. Desde el punto de vista de la Epidemiología Genética es uno de los más utilizados para realizar estudios de esta disciplina. Permite crear tablas de contingencia y calcular el Odds Ratio (OR) y el Riesgo Relativo (RR), hacer análisis para estudios de cohortes y de casos y controles, además puedes hacer distribuciones de probabilidades y cálculos de Chi Cuadrado, pero no tiene estadística descriptiva, por lo que los especialistas tienen que hacer combinaciones con otros programas para completar los estudios.

1.2.6. SEEGEN-R

El Sistema Estadístico de Epidemiología Genética- basado en R (SEEGEN-R), es el resultado de un proyecto cubano que está actualmente en desarrollo. Tiene como objetivo principal construir un programa informático que agrupe los diferentes estudios del campo de la Epidemiología Genética y

permitir una mayor fluidez en la realización de sus estudios. Su arquitectura está basada en extensiones y cada una de estas puede ser programada de forma independiente y luego integrada al sistema, siendo esta característica una de las más importantes de la arquitectura de SEEGEN-R**.**

El software actualmente tiene un módulo para los estudios de Genética Poblacional, pero no cuenta con funcionalidades para realizar estudios de Epidemiología Genética. Es importante señalar que esta aplicación utiliza la biblioteca de clases *SagesAPIPluginStudy* para integrar los estudios de cada extensión que se desarrolle, al menú principal del Módulo Base.

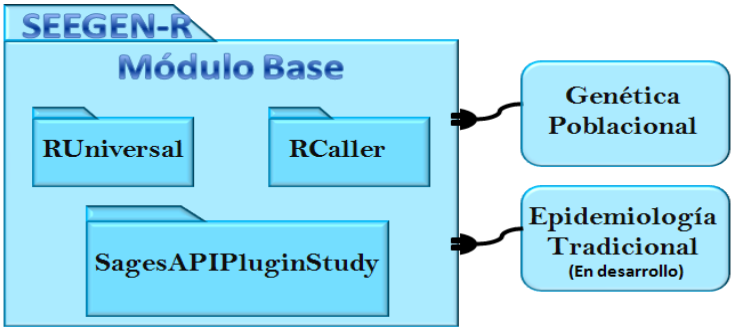


Figura 1: Arquitectura de SEEGEN-R

1.2.7. Conclusiones del estudio de las herramientas

Se ha podido evidenciar que ninguno de estas aplicaciones de forma independiente, es suficiente para realizar todos los análisis estadísticos de Epidemiología Genética. Teniendo en cuenta que en nuestro país no todos los genetistas tienen acceso a la Red Nacional de Salud, el Centro Nacional de Genética Médica ha decidido desarrollar en conjunto con la Universidad de Ciencias Informáticas la aplicación de escritorio SEEGEN-R (Sistema Estadístico de Epidemiología Genética- basado en R), para que esta aplicación sea distribuida por todos los Centros Municipales de Genética Médica. Con el objetivo de contribuir al desarrollo de esta herramienta se determina realizar una extensión a este sistema, que abarque los tipos de estudios genéticos enunciados anteriormente en el epígrafe 1.1.

1.3. Metodologías de desarrollo de software

Las metodologías imponen un proceso disciplinado sobre el desarrollo de software con el fin de hacerlo más predecible y eficiente. En cualquier proceso de desarrollo de software esta define: “Qué” debe hacer el software, “Quién” debe realizar cada actividad, “Cuándo” hacerla y “Cómo” se debe hacer. El objetivo fundamental del proceso de desarrollo es elevar la calidad del software en cada una de las fases por las que transita mediante una mayor transparencia y control sobre el mismo. No existe una metodología de software universal, sino que cada equipo es responsable de elegirla según las características de su proyecto. [17]

1.3.1. Rational Unified Process (RUP)

Rational Unified Process, como su nombre lo dice, es un proceso de ingeniería de software. Este proceso proporciona un enfoque disciplinado para asignar tareas y responsabilidades dentro de una organización de desarrollo. Su objetivo es asegurar la producción de alta calidad de software que satisfaga las necesidades de sus usuarios finales, dentro de un horario y presupuesto predecible. RUP en su ciclo de desarrollo cuenta con cuatro fases y nueve flujos para sus actividades, las cuales se desarrollarán siguiendo un modelo en cascada. A los 6 primeros flujos se les conoce como “Flujos Ingenieriles” o de “Trabajo Básico” y a los 3 restantes de “Apoyo” o “Soporte”.

Esta metodología crea como base los casos de uso, los cuales describen los requerimientos de la aplicación desde el punto de vista del usuario. Además define en cada momento del ciclo de vida del proyecto, qué artefactos, con qué nivel de detalle, y por cuál rol, se deben crear. Con RUP se presentarán al cliente los artefactos al final de una fase y se valorarán las precondiciones para la siguiente. Lo más importante de RUP es que es una metodología muy organizativa y el objetivo de cada actividad es la creación de los artefactos. [18]

Por todo lo anteriormente expuesto se ha decidido utilizar RUP como metodología de desarrollo de software, ya que está pensado para adaptarse a cualquier proyecto. Es una guía efectiva de cómo utilizar de manera efectiva UML (Lenguaje Unificado de Modelado) y le proporciona a cada miembro de un equipo fácil acceso a una base de conocimientos con guías, plantillas y herramientas para todas las actividades críticas de desarrollo. Toda la documentación que genera, posibilita un buen entendimiento de todo lo que se realizó, a futuras generaciones que vayan a utilizar este trabajo.

1.4. Herramientas y Tecnologías

Las herramientas son un conjunto de programas que son usados por los diseñadores y los programadores para construir sistemas. La utilización de tecnologías permite aplicar conocimientos y habilidades con el objetivo de conseguir una solución que permita resolver un problema determinado y lograr satisfacer las necesidades del proyecto.

1.4.1. Lenguaje Unificado de Modelado (UML)

Para entender que es UML (Lenguaje Unificado de Modelado), primero se debe comprender que un lenguaje de modelado es una herramienta del mundo actual de desarrollo de software, que tiene como objetivo representar gráficamente los procesos involucrados en el sistema. UML ofrece un estándar simple para la representación del sistema, incluye aspectos conceptuales tales como procesos de negocio, funciones del sistema y aspectos concretos como expresiones de lenguajes de programación,

esquemas de bases de datos y compuestos reciclados. UML visualiza, especifica, construye y documenta los artefactos del sistema en desarrollo. Es independiente, pero para utilizarlo óptimamente se debería usar en un proceso dirigido por los casos de uso, centrado en la arquitectura, iterativo e incremental.

UML prescribe una notación estándar y semánticas esenciales para el modelado de un sistema orientado a objetos. Previamente, un diseño orientado a objetos podría haber sido modelado con cualquiera de la docena de metodologías populares, causando a los revisores tener que aprender las semánticas y notaciones de la metodología empleada antes que intentar entender el diseño en sí. Ahora con UML, diseñadores diferentes, modelando sistemas diferentes, pueden sobradamente entender cada uno los diseños de los otros. [19]

Algunas ventajas de UML:

- Mayor rigor en la especificación.
- Permite realizar una verificación y validación del modelo realizado.
- Se pueden automatizar determinados procesos y permite generar código a partir de los modelos y a la inversa (a partir del código fuente generar los modelos). Esto permite que el modelo y el código estén actualizados, con lo que siempre se puede mantener la visión en el diseño, de más alto nivel, de la estructura de un proyecto. [2]

1.4.2. Visual Paradigm v8.0

Visual Paradigm es una herramienta CASE (Ingeniería de Software Asistida por Computación). La misma propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación. Esta herramienta ha sido concebida para soportar el ciclo de vida completo del proceso de desarrollo del software a través de la representación de todo tipo de diagramas.

Se caracteriza por:

- Disponibilidad en múltiples plataformas (Windows, Linux).
- Soporte de UML versión 2.1.
- Generación de código - Modelo a código, diagrama a código.
- Ingeniería inversa.

- Modelado colaborativo con CVS y Subversión (control de versiones).
- Distribución automática de diagramas - Reorganización de las figuras y conectores de los diagramas UML.
- Editor de detalles de casos de uso - Entorno todo en uno para la especificación de los detalles de los casos de uso, incluyendo la especificación del modelo general y de las descripciones de los casos de uso.

Esta herramienta permite aumentar la calidad del software, a través de la mejora de la productividad en el desarrollo y mantenimiento del software. Aumenta el conocimiento informático de una empresa ayudando así a la búsqueda de soluciones para los requisitos. También permite la reutilización del software, portabilidad y estandarización de la documentación, además del uso de las distintas metodologías propias de la Ingeniería del Software. [20]

1.4.3. Lenguaje de programación R

R es uno de los entornos que más se está desarrollando en la actualidad. Tiene alrededor de 13 librerías estadísticas definidas en su paquete base y ofrece un buen número de paquetes de rutinas especializadas, muy actuales dentro de los paquetes recomendados. Muchas otras se pueden descargar e instalar de la página de paquetes de los colaboradores [21]. Además ofrece una gran variedad de estadísticas (modelos lineales y no lineales, una serie de análisis, clasificación...), técnicas gráficas, y es altamente extensible. R es un conjunto integrado de servicios de software para la manipulación de datos, cálculo y representación gráfica. [22]

Otra característica importante y atractiva de R está dada por el hecho de que la salida que proporciona cualquier función se puede manipular convenientemente, pues R guarda estos resultados como objetos. Lo anterior significa que usted puede decidir, de toda la información que genera la ejecución de una función, qué es lo que realmente desea mostrar; si es que quiere mostrar algo o puede tomar una parte de esta salida para ser incorporada a la entrada de otra función. Esta característica de R facilita la elaboración de los informes finales a los investigadores encargados, pues la salida del procesamiento estadístico puede ser presentada de una forma muy accesible y atractiva para los investigadores biomédicos. [21]

Este trabajo utiliza R por todo lo antes mencionado y ha estado basado fundamentalmente en:

- La robustez del lenguaje.
- La constante actualización y la amplia literatura disponible.

- Amplias facilidades de manipulación de bases de datos.
- La obtención de informes con un formato predeterminado y con la información que se desea.
- Las facilidades gráficas.
- Facilidades para la documentación de todo el proceso de manipulación de los datos y procesamiento estadístico.

1.4.4. Lenguaje de programación Java

Java, el lenguaje de programación creado por la empresa Sun Microsystems, se ha consolidado firmemente como uno de los más utilizados en la actualidad y ha demostrado ser un lenguaje muy efectivo en programación general. Examinando la arquitectura Java se puede decir que sus principales características son:

Orientado a objetos: Java fomenta los diseños que conlleven a componentes reutilizables, extensibles y sostenibles. Estos componentes son lo bastante flexibles como para controlar los cambios que se puedan producir con el tiempo. Soporta las características propias de la programación orientada a objetos: clase, objeto, herencia, encapsulamiento y polimorfismo.

Interpretado: Los programas de Java en lugar de ser compilados en ejecutables nativos, su código es interpretado en una Máquina Virtual de Java (MVJ) y de este modo pueden ejecutarse sin tener que volver a compilarlos.

Robusto: Java es un lenguaje basado en tipos lo que evita las diferencias implícitas entre tipos, hay referencias en lugar de punteros por lo que no se puede hacer referencia a un puntero en memoria corrompiendo accidentalmente la memoria.

Seguro: Garantiza la seguridad del código que se está ejecutando y evita que el código no seguro realice operaciones seguras.

De arquitectura neutral: Si una empresa desarrolla un nuevo sistema operativo o un hardware completamente nuevo, no tiene que empezar desde cero sin ningún software. Con tan solo agregar la MVJ en la plataforma recién diseñada, se pueden ejecutar todos los programas de Java existentes. [23]

1.4.5. Entorno de Desarrollo Integrado

Un entorno de desarrollo integrado, llamado también IDE (siglas en inglés de Integrated Development Environment), es un programa informático compuesto por un conjunto de herramientas de

programación. Puede dedicarse en exclusiva a un solo lenguaje de programación o bien puede utilizarse para varios. Un IDE es un entorno de programación y consiste en un editor de código, un compilador, un depurador y un constructor de interfaz gráfica (GUI). Los IDEs pueden ser aplicaciones por sí solas o pueden ser parte de aplicaciones existentes.

Netbeans v7.2

El IDE NetBeans es un entorno galardonado de desarrollo integrado disponible para Windows, Mac, Linux y Solaris. El proyecto NetBeans consiste en un IDE de código abierto y una plataforma de aplicaciones que permiten a los desarrolladores crear rápidamente web, empresa, escritorio y aplicaciones móviles utilizando la plataforma Java, así como PHP, JavaScript, Ajax, Groovy y Grails, y C / C + +. Este proyecto está apoyado por una vibrante comunidad de desarrolladores y ofrece una amplia documentación y capacitación de recursos, así como una variada selección de terceras extensiones.

NetBeans IDE 7.2 ofrece un rendimiento significativamente mejorado y la experiencia de codificación, con las nuevas capacidades de análisis de código estático en el Editor de Java y el más inteligente proyecto de exploración [24]. La versión 7.2 de esta herramienta tiene un mejor soporte para las últimas tecnologías Java. Provee de primera clase un soporte integral para las más nuevas tecnologías y últimas mejoras del lenguaje Java. Es el primer IDE en dar soporte al JDK v7, Java EE v6 y JavaFX v2.0. Algunas de sus características son:

- Rápido y con un editado de código inteligente.

Un IDE es mucho más que un editor de texto, NetBeans le hace honor al nombre; su editor de texto es capaz de endentar las líneas, comparar las palabras y signos como paréntesis, puntos y punto y comas. Además ofrece un sistema de avisos para errores de código de tipo sintáctico y semántico.

- Fácil y eficiente gestión de proyecto.

NetBeans IDE proporciona diferentes vistas de los datos, desde las ventanas de proyectos múltiples hasta las herramientas útiles para la creación de sus aplicaciones y su gestión eficiente.

Por lo antes expuesto se seleccionó el IDE Netbeans en su versión 7.2 para el desarrollo de la herramienta, es el primer IDE en dar soporte al JDK v7 y permite el modelado UML.

1.4.6. Bibliotecas de clases

Las bibliotecas de clases o más conocidas en las ciencias de la computación como *library* (en inglés), son un conjunto de subprogramas utilizados para desarrollar software. Las bibliotecas contienen

código y datos, que proporcionan servicios a programas independientes, es decir, pasan a formar parte de estos. A continuación se muestran las bibliotecas de clases utilizadas en el desarrollo de la extensión informática, para enlazar java con el lenguaje R.

RCaller v2.1.1.1

RCaller es una librería para llamar funciones de R desde Java, la cual utiliza otra biblioteca de clases llamada RUniversal para el trabajo con gráficas y un archivo ejecutable llamado Rscrits que se encarga de gestionar las peticiones a R [25]. El funcionamiento de esta biblioteca consta del ejecutable Rscripts y un archivo XML temporal que contiene el resultado del estudio realizado. En SEEGEN-R, esta librería se utiliza para realizar el trabajo estadístico en los diferentes estudios genéticos.

RUniversal v1.0.2.1

Este paquete contiene algunas funciones para convertir objetos R al estilo de variables Java y XML. El código Java generado es interpretado por bibliotecas dinámicas como Beanshell[8]. [26]

Para el desarrollo de esta herramienta se utilizarán las librerías RCaller para enlazar Java y R, y esta a su vez utiliza RUniversal para el trabajo con gráficas y para devolver los resultados.

1.5. Patrón arquitectónico

Un patrón de arquitectura de software describe un problema particular y recurrente del diseño, que surge en un contexto específico, y presenta un esquema genérico y probado de su solución. Los patrones de arquitectura expresan el esquema fundamental de organización para sistemas de software. Proveen un conjunto de subsistemas predefinidos; especifican sus responsabilidades e incluyen reglas y guías para organizar las relaciones entre ellos [27]. Existen diversos patrones de arquitectura, entre los que se destacan: la arquitectura en capas y la arquitectura basada en extensiones. A continuación se verán algunas de las principales características de estas arquitecturas.

1.5.1. Arquitectura en N capas

La arquitectura basada en capas se enfoca en la distribución de roles y responsabilidades de forma jerárquica proveyendo una forma muy efectiva de separación de responsabilidades. El rol indica el modo y tipo de interacción con otras capas, y la responsabilidad indica la funcionalidad que está siendo desarrollada. En toda arquitectura de capas los elementos agrupados en una misma capa pueden comunicarse entre sí. Existen varias propuestas de arquitecturas de capas para sistemas informáticos, donde las variantes más usadas son:

[8] **BeanShell**: es un intérprete pequeño y libre de código fuente de java, orientado a objeto y con características de script. (http://www.beanshell.org/intro.html)

❖ **Arquitectura en dos capas**

La arquitectura de dos capas fue la primera en utilizar la estructura cliente-servidor. Presenta dos capas o niveles: el Nivel de Aplicación que es donde se dispone de todas las interfaces y donde el usuario puede realizar las actividades y el Nivel de Base de Datos o Repositorio de Datos, la cual es útil pues es utilizada para guardar toda la información del sistema. [28]

Figura 2: Arquitectura en dos capas

❖ **Arquitectura en tres capas**

Esta arquitectura está representada por tres niveles: el Nivel de aplicación, que a diferencia de la arquitectura de dos capas, esta solo trabaja con la arquitectura semántica de la aplicación sin importar su implementación ni su estructura física; el Nivel de dominio de la aplicación, el cual es el encargado de la estructura física y el dominio de la aplicación, una de las ventajas de esta aplicación es que los cambios se realizan únicamente en el servidor; y el Nivel de repositorio donde se almacenan todos los datos. [28]

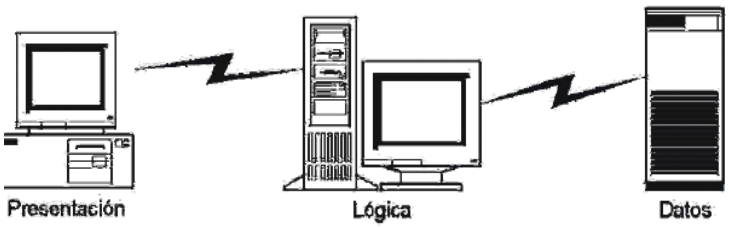


Figura 3: Arquitectura en tres capas

- **Capa de presentación:** Es la que ve el usuario (también se le denomina "capa de usuario"), presenta el sistema al usuario, le comunica la información y captura la información del usuario en un mínimo de proceso (realiza un filtrado previo para comprobar que no hay errores de formato). También es conocida como interfaz gráfica y debe tener la característica de ser "amigable" (entendible y fácil de usar) para el usuario.
- **Capa de negocio:** Es donde residen los programas que se ejecutan, se reciben las peticiones del usuario y se envían las respuestas tras el proceso. Se denomina capa de negocio (e incluso de lógica del negocio) porque es aquí donde se establecen todas las reglas que deben cumplirse. Esta

capa se comunica con la capa de presentación, para recibir las solicitudes y presentar los resultados, y con la capa de datos, para solicitar al gestor de base de datos almacenar o recuperar datos de él. También se consideran aquí los programas de aplicación.

- **Capa de datos:** Es donde residen los datos y es la encargada de acceder a los mismos. Está formada por uno o más gestores de bases de datos que realizan todo el almacenamiento de datos, reciben solicitudes de almacenamiento o recuperación de información desde la capa de negocio.

1.5.2 Arquitectura basada en extensiones

La arquitectura basada en extensiones permite que varios componentes sean desarrollados por distintos equipos de forma independiente, para una vez implementados integrarlos en la aplicación principal, para la cual fueron desarrollados, de esta forma los desarrolladores pueden integrar funcionalidades y nuevos servicios por la flexibilidad que posee [29]. Con esta característica se ahorra mucho trabajo en términos de mantenimiento y reutilización de código, facilitando el rápido crecimiento del software.

1.6. Conclusiones

En este capítulo se han descrito las herramientas y el lenguaje a utilizar, así como la metodología de desarrollo de software para llevar a cabo la aplicación. Dado el estudio de las herramientas estadísticas SPSS, PSPP, InfoStat, Epidat, Arlequín y SEEGEN-R, se determinó para dar solución al problema de investigación realizar una extensión para SEEGEN-R por las características que posee. Además como resultado del estudio de las tecnologías, metodologías y herramientas a utilizar, se ha llegado a la conclusión de que la extensión se construirá usando los lenguajes de programación Java y R y por consiguiente como tecnologías para su integración, las bibliotecas RCaller v2.1.1.1 y RUniversal v1.0.2.1. Además se determinó hacer uso de UML por su estrecha integración con la metodología adoptada, RUP. Como herramienta de modelado se utiliza Visual Paradigma v8.0 y como herramienta IDE el NetBeans IDE v7.2.

CAPÍTULO 2. CARACTERÍSTICAS DEL SISTEMA

Introducción

En este capítulo se define lo que debe hacer la aplicación a través de la caracterización del negocio, se identifican los actores, casos de uso del mismo y su descripción. Además se describen los procesos que son objeto de automatización. También se dan a conocer los requerimientos funcionales y no funcionales, los actores y casos de uso del sistema a desarrollar, así como sus descripciones.

2.1. Modelos de negocio

El modelo de negocio es el estudio de la organización, es fundamental para la comprensión y evolución de una empresa. Caracteriza los elementos más significativos.

2.1.1. Actores del negocio

Actor	Descripción
Consejo Científico	Indica el inicio de la investigación y recibe los resultados de la misma.

Tabla 14: Actores del negocio

2.1.2. Trabajadores del negocio

Trabajador	Descripción
Genetista	Especialista en Genética encargado de realizar estudios de epidemiología.

Tabla 15: Trabajadores del negocio

2.1.3. Diagrama de casos de uso del negocio

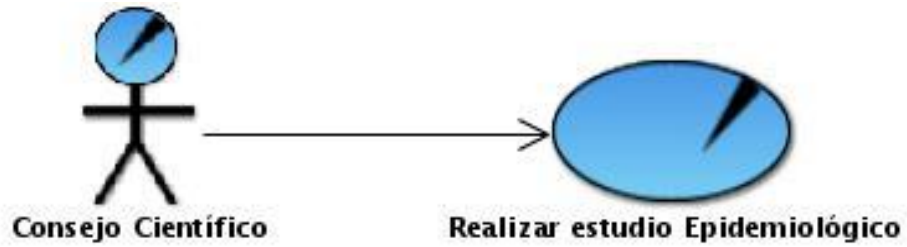


Figura 4: Diagrama de casos de uso del negocio

2.1.4. Descripción textual del caso de uso del negocio

Caso de Uso del negocio	Realizar estudio Epidemiológico.
Actores	Consejo Científico (inicia)
Resumen	El caso de uso se inicia cuando el Consejo Científico indica el inicio de la investigación. El Genetista realiza los estudios correspondientes, comunica los resultados al Consejo Científico

	terminando así el caso de uso.
Acción del actor	**Respuesta del proceso de negocio**
1- El Consejo Científico indica el inicio de la investigación.	2- El genetista aplica instrumento de recolección de datos.
	3- El genetista realiza los estudios epidemiológicos.
	4- El genetista obtiene los resultados del estudio.
	5- El genetista comunica al Consejo Científico los resultados.
6- El Consejo Científico recibe los resultados sobre el estudio epidemiológico.	
Precondiciones	Queda creado el estudio epidemiológico.

Tabla 16: Descripción textual del caso de uso del negocio

2.1.5. Diagrama de actividades

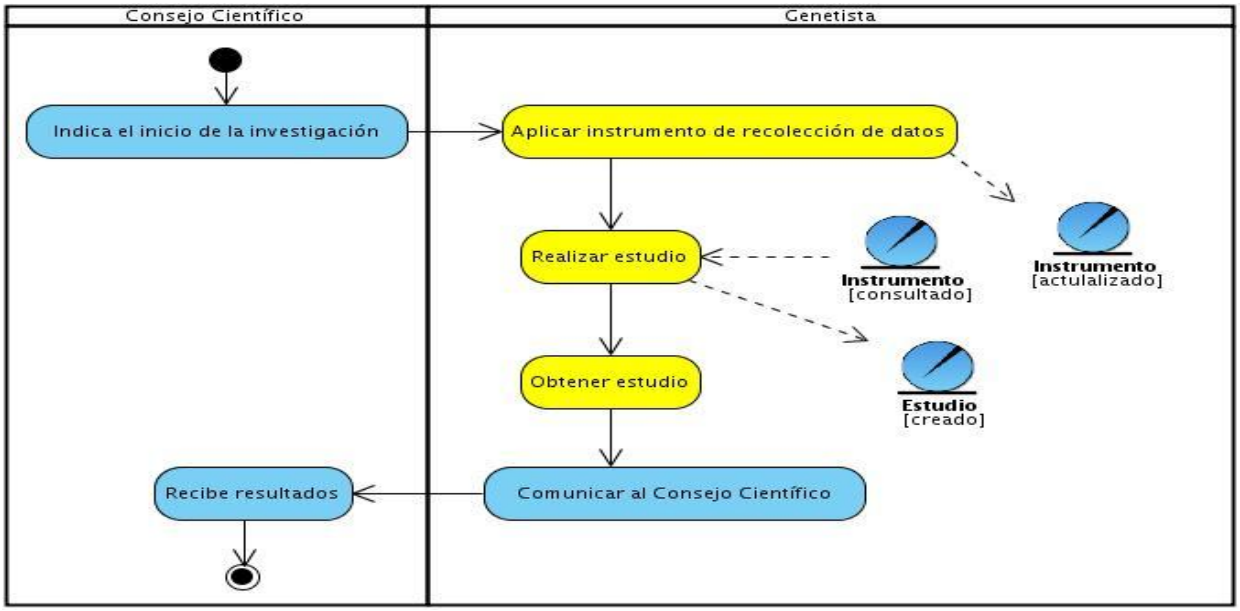


Figura 5: Diagrama de actividades del negocio

2.1.6. Modelo de objetos del negocio

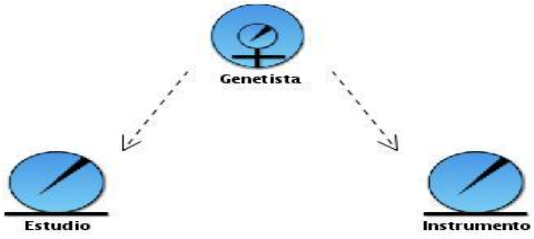


Figura 6: Modelo de objetos del negocio

2.1.7. Reglas del negocio

1. Solo el consejo científico puede indicar el inicio de una investigación.
2. El instrumento es una planilla que recoge los datos necesarios para realizar un estudio epidemiológico.
3. La probabilidad nunca debe ser cero, se debe especificar el último valor en caso de que sea muy pequeña, esto es para el caso de las dócimas de Chi Cuadrado.
4. Se deben realizar dos tablas de contingencia, una para la frecuencia esperada y otra para la observada. Se deben de cumplir supuestos, el primero es que las frecuencias esperadas es igual a uno y que no menos del veinte por ciento de las frecuencias esperadas es igual a cero.
5. Para realizar los estudios el total de familiares afectados no puede ser mayor que el total de familiares.
6. Es significativa estadísticamente la interacción entre dos variables si el OR observado es mayor que el OR esperado, siendo el OR esperado el producto del OR de ambas variables bajo el modelo multiplicativo.
7. Es significativa estadísticamente la interacción entre dos variables si el OR observado es mayor que el OR esperado, siendo el OR esperado la suma del OR de ambas variables menos uno bajo el modelo aditivo.

2.2. Especificación de los requisitos de la aplicación informática

Para la especificación de los requerimientos el cliente debe tener en cuenta además de las funcionalidades o requerimientos funcionales del sistema, los requerimientos no funcionales del mismo.

2.2.1. Requerimientos Funcionales

Los requerimientos funcionales son condiciones o capacidades que deben ser alcanzadas o poseídas por un sistema o componente de un sistema para satisfacer un contrato, estándar u otro documento impuesto formalmente. Con ellos se pretende determinar de manera clara y concisa lo que debe hacer el sistema siguiendo un enfoque funcional [18]. Todas las ideas que los clientes, usuarios y miembros del equipo de proyecto tengan acerca de lo que debe hacer el sistema, deben ser analizadas como candidatas a requisitos. Los requerimientos funcionales agrupados por casos de uso son los siguientes:

RF1: Crear estudio de interacción.

RF 1.1. Cargar juego de datos del Módulo Base para el estudio de interacción.

RF 1.2. Obtener resultados del estudio de interacción.

RF 2: Crear estudio de asociación genética.

RF 2.1. Cargar juego de datos del Módulo Base para el estudio de asociación genética.

RF 2.2. Obtener resultados de estudio de asociación genética.

RF 3: Crear estudio de clásicos en gemelos.

RF 3.1. Cargar juego de datos del Módulo Base para el estudio de clásicos en gemelos.

RF 3.2. Obtener resultados de estudio de clásicos en gemelos.

RF 4: Crear estudio de agregación familiar para casos y controles.

RF 4.1. Insertar datos preanalizados para el estudio de agregación familiar para casos y controles (padre, madre, hijos, hermanastros, tíos, primos, abuelo, abuela, nietos, total de familiares).

RF 4.2. Cargar juego de datos del Módulo Base para el estudio de agregación familiar para casos y controles (padre, madre, hijos, hermanastros, tíos, primos, abuelo, abuela, nietos, total de familiares).

RF 4.3. Obtener resultados del estudio de agregación familiar para casos y controles (padre, madre, hijos, hermanastros, tíos, primos, abuelo, abuela, nietos, total de familiares).

RF 5: Crear estudio de agregación familiar para casos y población.

RF 5.1. Insertar datos preanalizados para el estudio de agregación familiar para casos y población.

RF 5.2. Cargar juego de datos del Módulo Base para el estudio de agregación familiar para casos y población.

RF 5.3. Obtener resultados del estudio de agregación familiar para casos y población.

RF 6: Crear estudio de agregación familiar para casos particulares.

RF 6.1. Insertar datos preanalizados para el estudio de agregación familiar para casos particulares (gemelos monocigóticos frente a dicigóticos, gemelos dicigóticos frente a un hermano carnal, cónyuge y primos hermanos).

RF 6.2. Cargar juego de datos del Módulo Base para el estudio de agregación familiar para casos particulares (gemelos monocigóticos frente a dicigóticos, gemelos dicigóticos frente a un hermano carnal, cónyuge y primos hermanos).

RF 6.3. Obtener resultados del estudio de agregación familiar para casos particulares (gemelos monocigóticos frente a dicigóticos, gemelos dicigóticos frente a un hermano carnal, cónyuge y primos hermanos).

2.2.2. Requisitos No Funcionales

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener. Debe pensarse en estas propiedades como las características que hacen al producto atractivo, usable, rápido o confiable. Normalmente están vinculados a requerimientos funcionales, es decir una vez que se conozca lo que el sistema debe hacer se puede determinar cómo ha de comportarse, qué cualidades debe tener o cuán rápido o grande debe ser. [18]

La extensión se desarrolla siguiendo los estándares y los requerimientos no funcionales de la aplicación SEEGEN-R, para la cual es creada. Se definen a continuación las propiedades de esta aplicación de escritorio, que se tienen en cuenta para el desarrollo del componente.

FNF 1: Apariencia o interfaz externa

La interfaz externa debe estar diseñada para verse en la resolución igual a 1024x760. Los colores predominantes en la misma deben ser el azul y gris. Los resultados de los estudios los muestra siempre en la misma posición del escritorio, posibilitando mayor rapidez en la ubicación de los resultados.

FNF 2: Usabilidad

La aplicación informática debe garantizar un acceso fácil y rápido, contando con un menú que satisfaga las necesidades de los usuarios. Este consta de tres niveles que le facilita al especialista ir directamente al tipo de estudio que desea realizar al dar un clic sobre una de las opciones del menú. La herramienta podrá ser usada sólo por genetistas que posean conocimientos avanzados en el dominio de la especialidad de la epidemiología en la genética.

FNF 3: Soporte

Se debe asegurar el soporte para los usuarios, de manera que se puedan satisfacer sus necesidades a partir de mejoras, una vez puesta en marcha la aplicación.

FNF 4: Rendimiento

Los tiempos de respuestas deben ser como máximo entre 30 y 40 segundos, la velocidad de procesamiento de la información debe ser rápida ya que es una aplicación de escritorio y no posee ningún tipo de intercambio de información a través de la red.

FNF 5. Software.

Se requiere para el funcionamiento de la aplicación disponer cualquier sistema operativo (Windows y Linux); Máquina Virtual de Java versión 1.6 o Superior y el lenguaje de R v1.2 o superior.

FNF 6. Hardware.

Para el desarrollo y ejecución de la aplicación se necesitará:

256 Megabytes (MB) de memoria RAM o más y 50 MB de capacidad en el disco duro como mínimo.

FNF 7. Requisitos Legales.

Las herramientas y las tecnologías en que estará basada la aplicación informática deberán cumplir con las licencias de software libre.

2.3. Definición de los casos de uso del sistema

Un caso de uso es una descripción de los pasos o las actividades que deberán realizarse para llevar a cabo algún proceso. Los personajes o entidades que participarán en un caso de uso se denominan actores. Es además una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento que inicia un actor principal sobre el propio sistema. Muestra la relación entre los actores y los casos de uso en un sistema.

2.3.1. Actores del sistema

Actor	Descripción
Genetista	Especialista en Genética que interactúa con el sistema, es el encargado de realizar los estudios y cálculos estadísticos sobre epidemiología.

Tabla 17: Actores del sistema

2.3.2. Paquetes del sistema

Para la modelación de los casos de uso del sistema se decidió dividirlos en paquetes de acuerdo al siguiente criterio de empaquetamiento:

Los CU requeridos para dar soporte a un determinado proceso de negocio. El paquete Epidemiología Genética responde a este criterio, agrupando todos los casos de uso asociados a ese proceso de negocio y este a su vez utiliza del paquete Módulo Base la funcionalidad Importar datos y Mostrar

resultados para realizar la mayor parte de los casos de uso mencionados anteriormente. De manera que quedaron integrados los 6 casos de uso del sistema y los 15 requisitos funcionales identificados.

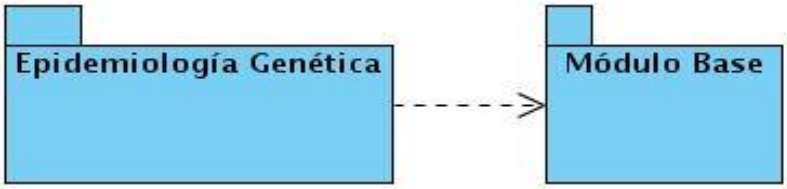


Figura 7: Diagrama de paquetes del sistema

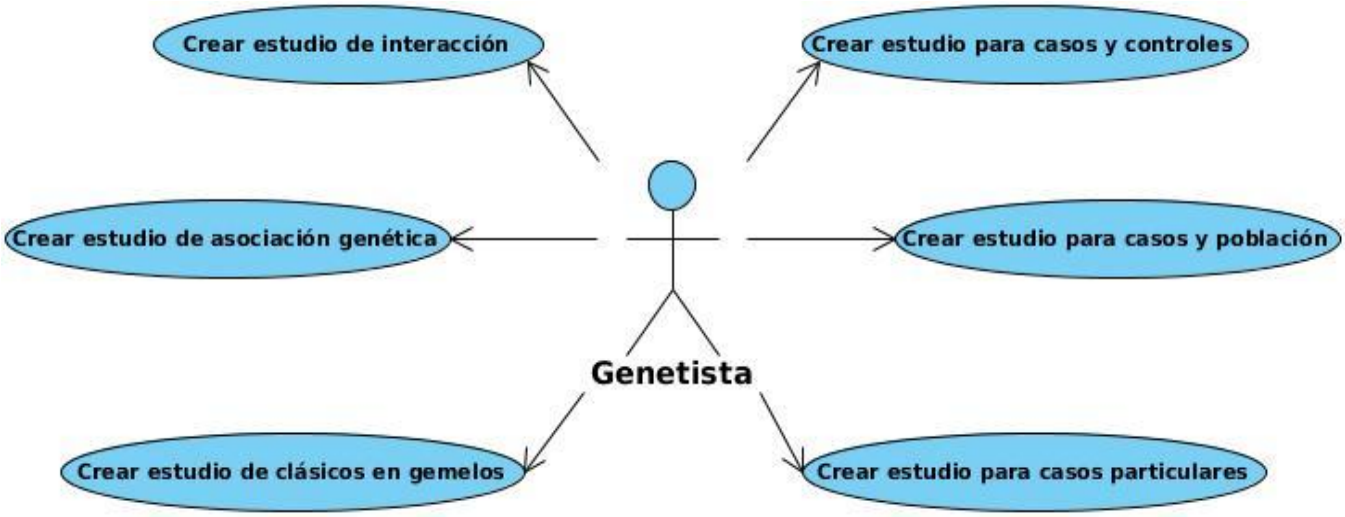
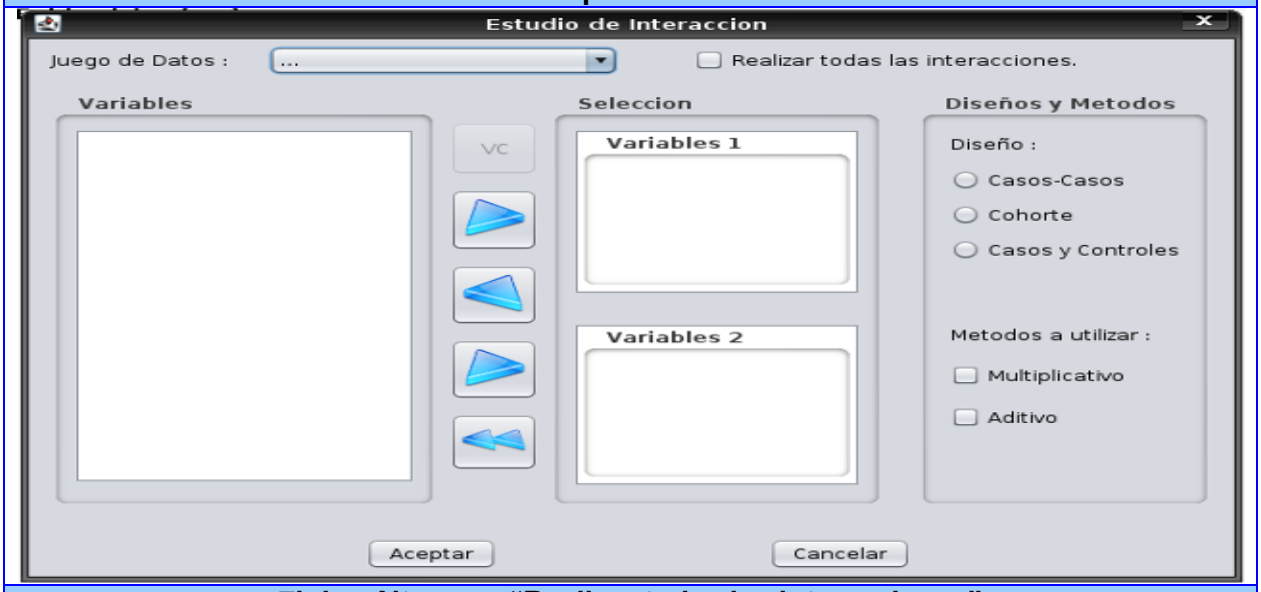


Figura 8: Diagrama de CUS: Paquete de Epidemiología Genética

2.3.3. Descripción de un caso de uso del sistema

Para consultar la descripción detallada de los casos de uso del sistema remitirse al expediente de proyecto, específicamente la plantilla SEEGEN-R_GE_ Modelo de Casos de uso del sistema.

Caso de Uso 1 (CU1)	**Crear estudio de interacción.**
Actor	Genetista
Descripción	El caso de uso inicia cuando el genetista selecciona Crear Nuevo Estudio de Interacción. Se importan los datos desde el Módulo Base y se seleccionan especificaciones del estudio. El caso de uso finaliza cuando se presiona el botón Aceptar.
Referencia	**RF1, RF2**
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1- El genetista selecciona la opción Crear Nuevo Estudio de Interacción.	2- El sistema muestra la interfaz con los siguientes campos a seleccionar : • Crear nueva variable. • Realizar todas la interacciones • Las variables a interactuar

	• El diseño del estudio: 1. Casos-Casos 2. Casos-Controles 3. Cohorte • El o los métodos a utilizar: 1. Multiplicativo 2. Aditivo
3- El genetista selecciona el juego de datos a importar desde el Módulo Base de la aplicación.	4- El sistema muestra las variables del juego de datos seleccionado.
5- El genetista selecciona una de las opciones disponibles.	6- • Si el genetista selecciona Realizar todas las interacciones (ver sección "Realizar interacciones entre variables"). • Si el genetista da clic en el botón Crear nueva variable (ver sección "Crear nueva variable").
Sección: "Realizar interacciones entre variables"	
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1- El genetista selecciona la opción Realizar todas las Interacciones o selecciona las variables a interactuar y presiona la opción Aceptar.	2- El sistema realiza la interacción de las variables seleccionadas utilizando los diseños (Casos-Casos, Casos-Controles y Cohorte) y métodos (Multiplicativo y Aditivo) especificados. El sistema realiza el estudio y envía los resultados al Módulo Base.
"Prototipo de Interfaz"	
	
Flujos Alternos: "Realizar todas las interacciones"	
Acción del Actor	**Respuesta del Sistema**
1.1- El genetista da clic en el botón Cancelar.	2.1- El sistema finaliza el caso de uso.
Sección: "Crear nueva variable"	

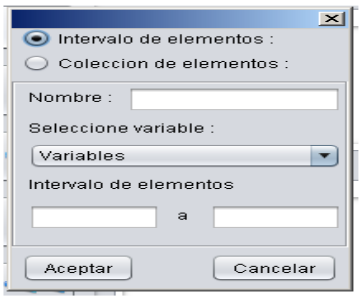
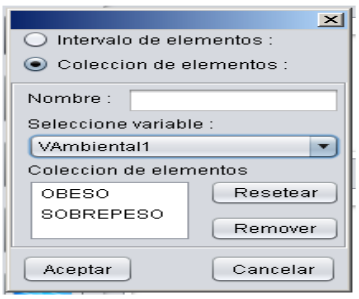
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1- El genetista selecciona la opción crear Variable Criterio	2- El sistema muestra la interfaz donde el genetista selecciona que tipo de variable desea crear; Intervalo de elementos o Colección de elementos (Por defecto – Intervalo de elementos).
3- El genetista introduce los datos y da clic en el botón Aceptar.	4- El sistema crea la variable y la visualiza en la lista de variables.
¨Prototipo de Interfaz¨	
	
Flujos Alternos: “Crear nueva variable”	
Acción del Actor	**Respuesta del Sistema**
1.1- El genetista da clic en el botón Cancelar.	2.1- El sistema finaliza el caso de uso.
Poscondiciones:	Queda creado un estudio de Interacción.
Prioridad:	Crítico.

Tabla 18: Descripción del caso de uso del sistema: Crear estudio de interacción

2.4. Conclusiones

En este capítulo, al realizar el análisis del negocio, se concluye que solo puede iniciar una investigación el Consejo Científico como actor del negocio, y que los trabajadores del negocio son los genetistas. Además con la correcta identificación de 15 requisitos funcionales, se obtuvo un total de 6 casos de usos del sistema, lo cual permitió identificar adecuadamente las entradas para la fase de diseño e implementación. Al concluir con la caracterización de los requisitos no funcionales se pudo identificar las propiedades que se deben tener en cuenta a la hora de desplegar la herramienta. Además se identificó que los requisitos de hardware necesarios para levantar el sistema, no son críticos y son asequibles para los recursos del CNGM, así como los tiempos de respuesta, que son relativamente rápidos.

CAPÍTULO 3. ANÁLISIS Y DISEÑO DEL SISTEMA

Introducción

El diseño es el centro de atención al final de la fase de elaboración y el comienzo de las iteraciones de construcción. Esto contribuye a una arquitectura estable, sólida y a crear un plano del modelo de implementación. En este capítulo se modelan las clases del diseño en correspondencia con la realización de los casos de uso descritos en el capítulo anterior. Se hace énfasis en la arquitectura y patrones de diseño utilizados, además de modelar los aspectos dinámicos de la aplicación a través de los diagramas de secuencia. Se define el diagrama de despliegue de la aplicación.

3.1. Patrón de arquitectura aplicado

Un patrón de arquitectura expresa un esquema de organización estructural esencial para un sistema de software, que consta de subsistemas, sus responsabilidades e interrelaciones. El patrón de arquitectura utilizado por la aplicación SEEGEN-R es una arquitectura basada en extensiones. Esto hace que sea una de las características más importantes de la aplicación. Cada extensión puede ser programada de forma independiente y luego integrada al sistema. Las extensiones más importantes son las de tipo estudio, éstas responden a un grupo de estudios específicos dentro de un campo y son las encargadas de realizar las tareas específicas de estos estudios.

La extensión desarrollada presenta internamente una Arquitectura de N Capas. Según las características que tiene la aplicación se definieron dos capas solamente, la capa de presentación y la capa de negocio. No se identificó para la aplicación la capa de acceso a datos porque no se requiere la persistencia de datos en una base de datos**.**

3.1.1. Vista lógica del sistema

La vista lógica abarca un comportamiento arquitectónicamente significativo. Describe el diseño más importante de las clases, su organización en paquetes y subsistemas, y la organización de estos en capas. Esta vista muestra los requisitos funcionales diseñados en el interior del sistema, o sea, los servicios que el sistema debe proporcionar, que forman el vocabulario del problema y de la solución. [30]

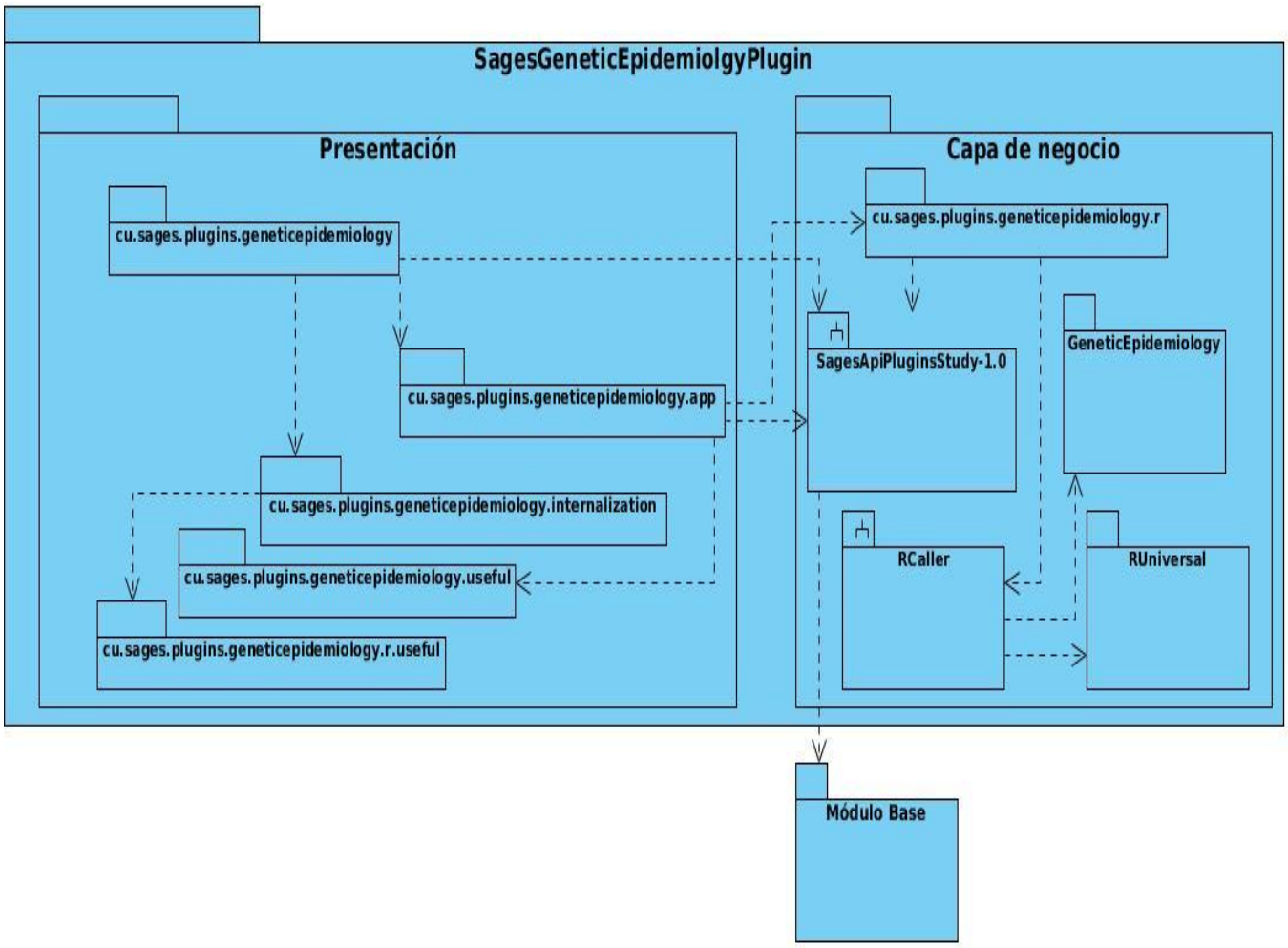


Figura 9: Vista lógica del sistema

En el diagrama de la vista lógica del sistema se representan los dos componentes del patrón de la arquitectura utilizado para el desarrollo de la herramienta. Estos componentes son:

Presentación: Contiene el código que representa la parte que será visualizada en pantalla por el genetista. La capa presentación fue separada en varios paquetes:

- *cu.sages.plugins.geneticepidemiology:* contiene la clase que crea el menú principal de la herramienta.
- *cu.sages.plugins.geneticepidemiology.app:* contiene las clases interfaces que serán visualizadas por el genetista.
- *cu.sages.plugins.geneticepidemiology.r.useful:* es un paquete auxiliar utilizado para la internacionalización del idioma, igual que *cu.sages.plugins.geneticepidemiology.internalization.*

- *cu.sages.plugins.geneticepidemiology.useful:* es un paquete que contiene clases auxiliares para el desarrollo de los estudios.

Capa del negocio: Es el punto de entrada de la aplicación, se mantiene a la escucha de todas las peticiones y ejecuta la lógica de la aplicación. La capa del negocio fue separada en varios paquetes:

- *cu.sages.plugins.geneticepidemiology.r:* contiene las clases controladoras para cada estudio y estas actúan de intermediarias entre la interfaz que le corresponde y el paquete de *GeneticEpidemiology*.

- *SagesAPIPluginStudy:* esta biblioteca de clases se encarga de comunicar la extensión con el Módulo Base, tiene influencia en la capa de negocio a la hora de cargar los datos para cada estudio y se relaciona con el paquete *cu.sages.plugins.geneticepidemiology* para crear el menú con las opciones de la herramienta.

- *RCaller y RUniversal:* RCaller es una biblioteca de clases que permite conectar R con Java, Esta utiliza el paquete *RUniversal* para ejecutarse.

- *GeneticEpidemiology:* es el paquete de R desarrollado para la extensión, tiene implementadas las funciones estadísticas utilizadas por los estudios. Para poder ejecutar la aplicación debe estar instalado el paquete estadístico *GeneticEpidemiology*.

Módulo Base: Es la aplicación sobre la cual se va a agregar la extensión Epidemiología Genética. Brinda varias opciones, entre las que se encuentran: cargar un juego de dados ya elaborado, permite crear uno nuevo si es necesario, da la opción de imprimir, guardar, crear estudios epidemiológicos. Además muestra los resultados de los estudios realizados de forma organizada para el especialista. De este paquete lo más importante para la extensión desarrollada son las funcionalidades Importar juego de datos y Mostrar resultados.

3.2. Aplicación de patrones de diseño

Los patrones de diseño son la base para la búsqueda de soluciones a problemas comunes en el desarrollo de software y otros ámbitos referentes al diseño de interacción o interfaces. Un patrón de diseño identifica: clases, instancias, roles, colaboraciones y la distribución de responsabilidades, además que ayuda a construir clases y a estructurar sistemas de clases.

Con la aplicación de los patrones de diseño se reutilizan las experiencias de otros desarrolladores ya que estos patrones están basados en la recopilación del conocimiento de los expertos en desarrollo de software.

3.2.1. Patrones GRASP aplicados

Creador: Guía la asignación de responsabilidades relacionadas con la creación de objetos, tarea muy frecuente en los sistemas orientados a objetos. Se asigna la responsabilidad a una clase de crear cuando contiene, agrega, compone, almacena o usa otra clase, lo que brinda una alta posibilidad de reutilizar la clase creadora.

Bajo Acoplamiento: Guía la asignación de responsabilidades de forma tal que las clases se comuniquen con el menor número de clases que sea posible, de manera que de producirse una modificación en algunas de ellas se tenga la mínima repercusión posible en el resto de las clases, potenciando la reutilización, y disminuyendo la dependencia entre las mismas.

Alta Cohesión: Mantiene la complejidad dentro de límites manejables y garantiza que las clases con responsabilidades estrechamente relacionadas no realicen un trabajo enorme.

Controlador: Patrón que establece el uso de una clase controladora, que funciona como intermediaria entre cada una de las clases interfaces y el algoritmo que la implementa, de tal forma que es la que recibe los datos del usuario y manda a ejecutar las acciones correspondientes implementadas en el paquete de R.

3.2.2. Patrones GOF aplicados

Facade (Fachada): Patrón que provee una interfaz unificada, simple, que haga de intermediaria entre un cliente y una interfaz o grupo de interfaces de un subsistema. En este caso la clase GeneticEpidemiologyPlugin es la que cumple con estas condiciones, es la encargada de crear el menú principal de la herramienta para el acceso a todas las interfaces de los diferentes estudios.

3.3. Diagrama de clases del diseño

Los diagramas de clases del diseño especifican que clases intervienen en el desarrollo del sistema y como se relacionan. También se puede definir, como una técnica gráfica que se utiliza para modelar la parte estática de la aplicación. A continuación se muestra la clase del diseño para el caso de uso: Crear estudio de interacción.

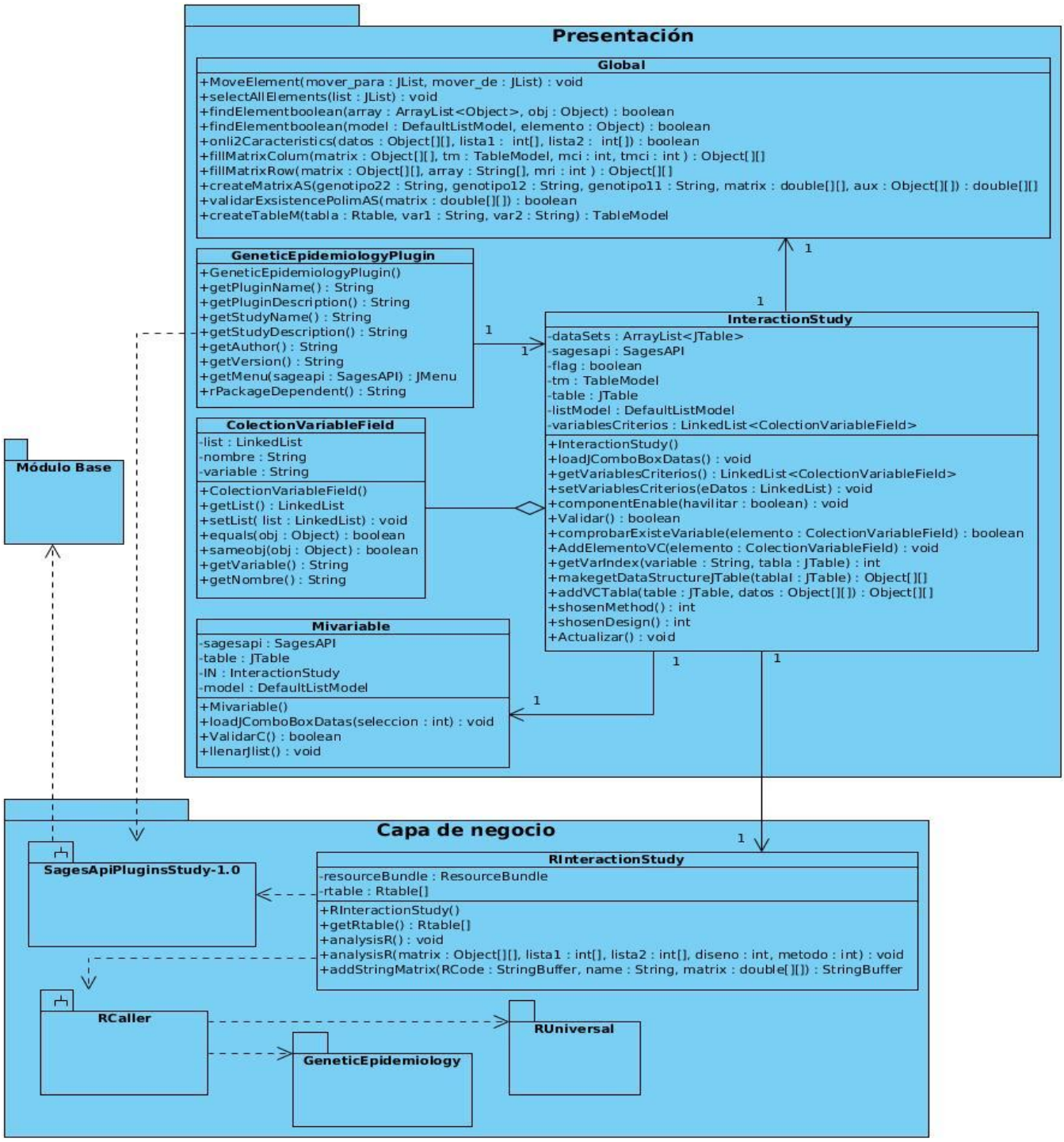


Figura 10: Diagrama de Clases de Diseño: CU1. Crear estudio de interacción

La biblioteca *SagesAPIPluginStudy* es la encargada de comunicar la extensión con el Módulo Base a través de la clase *GeneticEpidemiologyPlugin*, la cual crea el menú principal con todas las funcionalidades de la herramienta y las muestra en la aplicación. *GeneticEpidemiologyPlugin* crea la clase interfaz *InteractionStudy*, ésta contiene el formulario que desarrolla el flujo de trabajo del Estudio

de interacción. La clase *Global* tiene métodos que son utilizados por varias clases, de ella, *InteractionStudy* utiliza el método *selectAllElements*. Además tiene una lista de elementos de tipo *ColectionVariableField* y la clase *Mivariable* se activa cuando el genetista presiona la opción crear nueva variable en la interfaz principal.

La clase *RInteractionStudy* tiene la función de controladora, es la intermediaria entre *InteractionStudy* y la biblioteca *Rcaller*, la cual tiene la tarea de enlazar Java con el lenguaje estadístico R y para su ejecución necesita del paquete RUniversal y llama además a *GeneticEpidemiology*, que es el paquete de R implementado por los desarrolladores y contiene varios de los métodos estadísticos utilizados, implementados en forma de funciones independientes.

Para consultar los restantes diagramas de clases del diseño remitirse al expediente de proyecto, específicamente a la plantilla SEEGEN-R_GE_ Modelo de Diseño_v1.0.

3.4. Descripción de las clases del diseño

La descripción de las clases del diseño es una abstracción lógica de un conjunto de objetos que comparten los mismos atributos, operaciones, relaciones y semánticas, o sea, te permiten tener una representación de cómo funciona internamente y que métodos realiza. A continuación se muestra descripción de la clase InteractionStudy.

Nombre: InteractionStudy	
Tipo de clase: Interfaz	
Responsabilidades: Interfaz del estudio de Interacción.	
Nombre:	**getVariablesCriterios()**
Descripción:	Devuelve la lista de las variables criterio creadas por el genetista.
Nombre:	**setVariablesCriterios(LinkedList eDatos)**
Descripción:	Modifica la lista de Variables Criterios
Nombre:	**componentEnable(boolean havilitar)**
Descripción:	Habilita o deshabilita los componentes del visual.
Nombre:	**comprobarExisteVariable(VariableField elemento)**
Descripción:	Comprueba si el elemento pasado por parámetros existe en la lista de variables criterios existentes.
Nombre:	**AddElementoVC(VariableField elemento)**
Descripción:	Adiciona el elemento pasado por parámetro a la lista de variables criterios.
Nombre:	**addVCTabla()**

Descripción:	Añade la variable criterio creada a la matriz de datos final que recibe R.
Nombre:	**Validar()**
Descripción:	Se encarga de validar que todos los requisitos del estudio estén llenos.
Nombre:	**Actualizar()**
Descripción:	Se encarga de mostrar los juegos de datos que están cargados en la aplicación en el momento que se creó el estudio.
Nombre:	**getVarIndex(String variable, JTable tabla)**
Descripción:	Devuelve la posición de una variable dentro de una tabla.
Nombre:	**makegetDataStructureJTable(JTable tablal)**
Descripción:	Crea la estructura de una matriz de datos en dependencia de las columnas escogidas.
Nombre:	**shosenMethod()**
Descripción:	Devuelve un valor de tipo entero que indica que método fue seleccionado.
Nombre:	**shosenDesign()**
Descripción:	Devuelve un valor de tipo entero que indica que diseño fue seleccionado.

Tabla 19: Descripción de las clases del diseño: Clase InteractionStudy

Para consultar las restantes descripciones de las clases del diseño remitirse al expediente de proyecto, específicamente a la plantilla SEEGEN-R_GE_ Modelo de Diseño_v1.0.

3.5. Diagrama de secuencia

Los diagramas de secuencia muestran las interacciones entre un conjunto de objetos en una aplicación, ordenadas según el tiempo en que tienen lugar [31]. Este contiene detalles de implementación del escenario, incluyendo los objetos y clases que se usan para implementar el escenario y mensajes intercambiados entre ellos. Estos diagramas destacan el orden temporal de los mensajes [32]. A continuación se muestran los diagrama de secuencia para el caso de uso Crear estudio de interacción, uno por cada sección de la descripción del caso de uso.

Caso de uso: Crear estudios de interacción para el escenario “Realizar interacciones entre variables”.

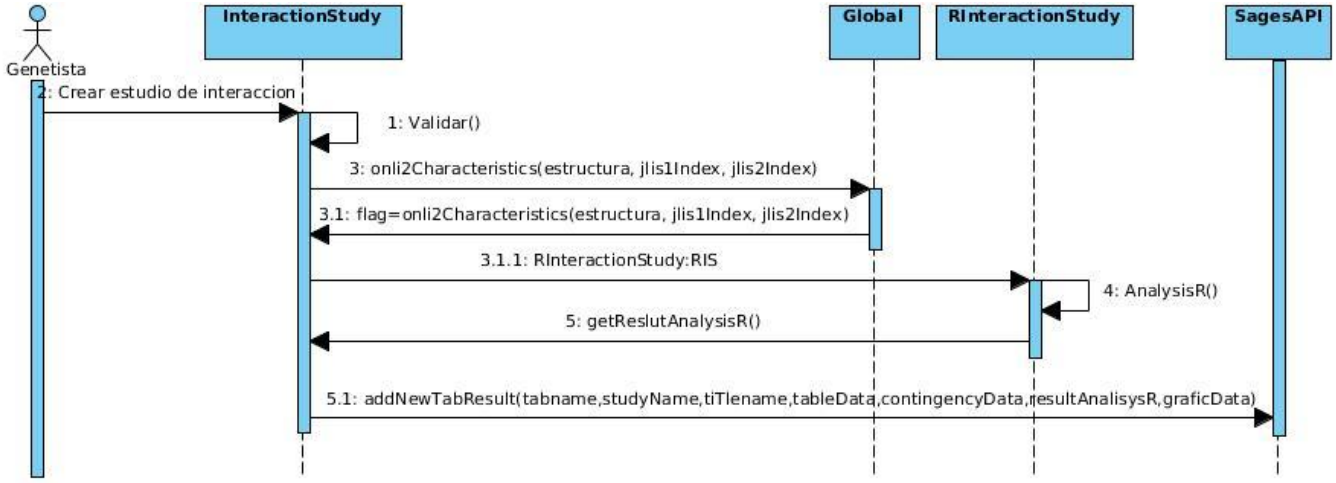


Figura 11: Diagrama de secuencia: CU1-Escenario “Realizar interacciones entre variables”

Caso de uso: Crear estudios de interacción para el escenario “Crear nueva variable”.

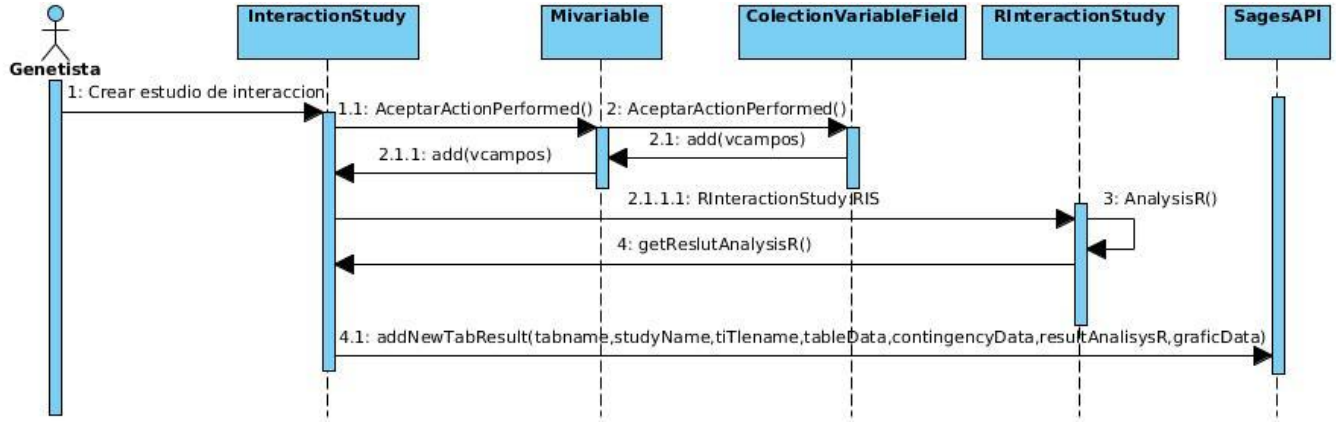


Figura 12: Diagrama de secuencia: CU1-Escenario “Crear nueva variable”

Para consultar las restantes diagramas de secuencia remitirse al expediente de proyecto, específicamente a la plantilla SEEGEN-R_GE_ Modelo de Diseño_v1.0.

3.6. Modelo de despliegue

Es un modelo de objetos que describe la distribución física del sistema en términos de cómo se distribuye la funcionalidad entre los nodos de cómputo. Cada nodo representa un recurso de cómputo, normalmente un procesador o un dispositivo hardware similar. Los nodos poseen relaciones que representan medios de comunicación entre ellos [33]. El modelo de despliegue se utiliza como entrada fundamental en las actividades de diseño e implementación debido a que la distribución del sistema tiene una influencia principal en su diseño. Los genetistas necesitan utilizar esta aplicación de escritorio en cualquier estación de trabajo sin importar el lugar donde se encuentren y no debe estar conectada a la red. Debido a esto solo intervienen, como nodo principal la computadora cliente donde está alojada

la aplicación y una impresora que se utilizará solo en el caso de que se desee imprimir un informe de algún estudio epidemiológico.

Figura 13: Diagrama de despliegue de la aplicación

3.7. Conclusiones

En este capítulo se abordaron los principales elementos del diseño para el desarrollo de la aplicación. Se explicó el patrón de arquitectura en capas, aplicándose el patrón en su versión de dos capas dado los requerimientos del sistema. Además se procedió a la realización de los diagramas de clases del diseño aplicando patrones bien conocidos como GRASP y GOF, lo cual permitió definir una mejor estructura y relación entre las clases así como alcanzar una mejor comprensión del sistema para su posterior implementación. Además quedaron representados los diagramas de secuencia correspondientes para cada caso de uso y se describieron las funciones principales de cada clase. Con la identificación del diagrama de despliegue se concluyó que solo se necesita de una computadora para desplegar la aplicación sin necesidad de estar conectada a la red.

CAPÍTULO 4. IMPLEMENTACIÓN Y PRUEBA

Introducción

En el actual capítulo se desarrolla el modelo de implementación. Se representan los diagramas de componentes elaborados. Además se brinda una descripción de los principales métodos implementados, se muestran imágenes de la interfaz de la extensión. Se hace referencia a la validación del componente y se muestra además el modelo de prueba con la descripción de los casos de prueba basados en casos de uso.

4.1. Diagrama de componentes

El diagrama de componentes muestra las organizaciones y dependencias lógicas entre componentes de software, sean estos ficheros de código fuente, binarios o ejecutables. Los elementos de modelado que lo conforman son los componentes y paquetes y muestran la estructura del sistema en términos de implementación a un alto nivel [31]. No es necesario que un diagrama incluya todos los componentes del sistema, normalmente se realizan por partes. En la figura 14 se muestra el diagrama de componentes que describe el caso de uso Crear estudio de interacción y para un mayor entendimiento, a continuación se relaciona cada componente con las clases que representa:

- *cu.sages.plugins.geneticepidemiology:* Este paquete de componentes solo contiene al componente *GeneticEpedemiologyPlugin.java* encargado de crear el menú principal de la extensión.

- *cu.sages.plugins.geneticepidemiology.app:* Este paquete de componentes contiene varios componentes, pero en el ejemplo a que se hace referencia, se muestra el componente *InteractionStudy.java* y este a su vez solamente tiene la clase *InteractionStudy*, que es la encargada de mostrar la interfaz para que el especialista realice el estudio de interacción.

- *cu.sages.plugins.geneticepidemiology.r.useful:* Este paquete de componentes contiene el componente *Useful.java* y este a su vez tiene solamente la clase *Useful* y es una clase auxiliar para la internacionalización del idioma.

- *cu.sages.plugins.geneticepidemiology.internalization:* Este paquete de componentes contiene a los componentes *lenguage_en_US.properties* y *lenguage_en_US.properties*, estos a su vez tienen solamente las clases *lenguage_en_US* y *lenguage_en_ES* respectivamente y se encargan de la internacionalización del idioma, hasta el momento del español y el inglés.

- *cu.sages.plugins.geneticepidemiology.useful:* Este paquete de componentes contiene los siguientes componentes *Mivariable.java*, *ColectionVariableField.java y Global.java,* estos a su vez tienen

solamente las clases *Mivariable, ColectionVariableField y Global* respectivamente, y son utilizadas por la clase *InteractionStudy* como clases auxiliares para el trabajo con las variables para el desarrollo del estudio de interacción.

- *cu.sages.plugins.geneticepidemiology.r:* Este paquete de componentes contiene varios, pero en el ejemplo a que se hace referencia, se muestra el componente *RInteractionStudy.java* y este a su vez solamente tiene la clase *RInteractionStudy*, que es la que se encarga de pedirle a R las funciones estadísticas necesarias para realizar el estudio de interacción.

- *SagesAPIPluginStudy:* Este componente es de tipo biblioteca de clases, contiene otros 4 elementos más y tiene como función principal integrar la extensión al Módulo base, mostrando todos los estudios de la nueva herramienta como opciones en el menú principal.

- *RCaller y RUniversal: RCaller* es un componente de tipo biblioteca de clases, que para su ejecución necesita del paquete *RUniversal* para crear un archivo XML con los resultados y además convoca el paquete *GeneticEpidemiology*.

- *GeneticEpidemiology:* Este paquete de componentes contiene los componentes: *aditive.model.R, casos.control.R, cohorte.R, PCCI.R, try.catch.R, odds.ratio.R, risk.relative.R, casos.casos.R, PCP.R, clasic.twins.test.R, falconer.test.R, multiply.model.R, OR.R* y *RR.R* dentro de la carpeta R. Cada uno de estos componentes son una clase independiente en lenguaje de programación R, que fueron implementados porque eran necesarios en algún momento de determinado estudio. El diagrama de componentes para el estudio de interacción es el ejemplo que se muestra en la figura 14, en él intervienen las dos últimas funciones de las mencionadas anteriormente. Además de otras como *chisq.test* (para el Chi Cuadrado de Pearson) y *chisq.tes (matrix, correct=T/F)* (para el Chi Cuadrado de Pearson con la corrección de Yates), las cuales si aparecen en el entorno de R, por lo que no fue necesario implementarlas.

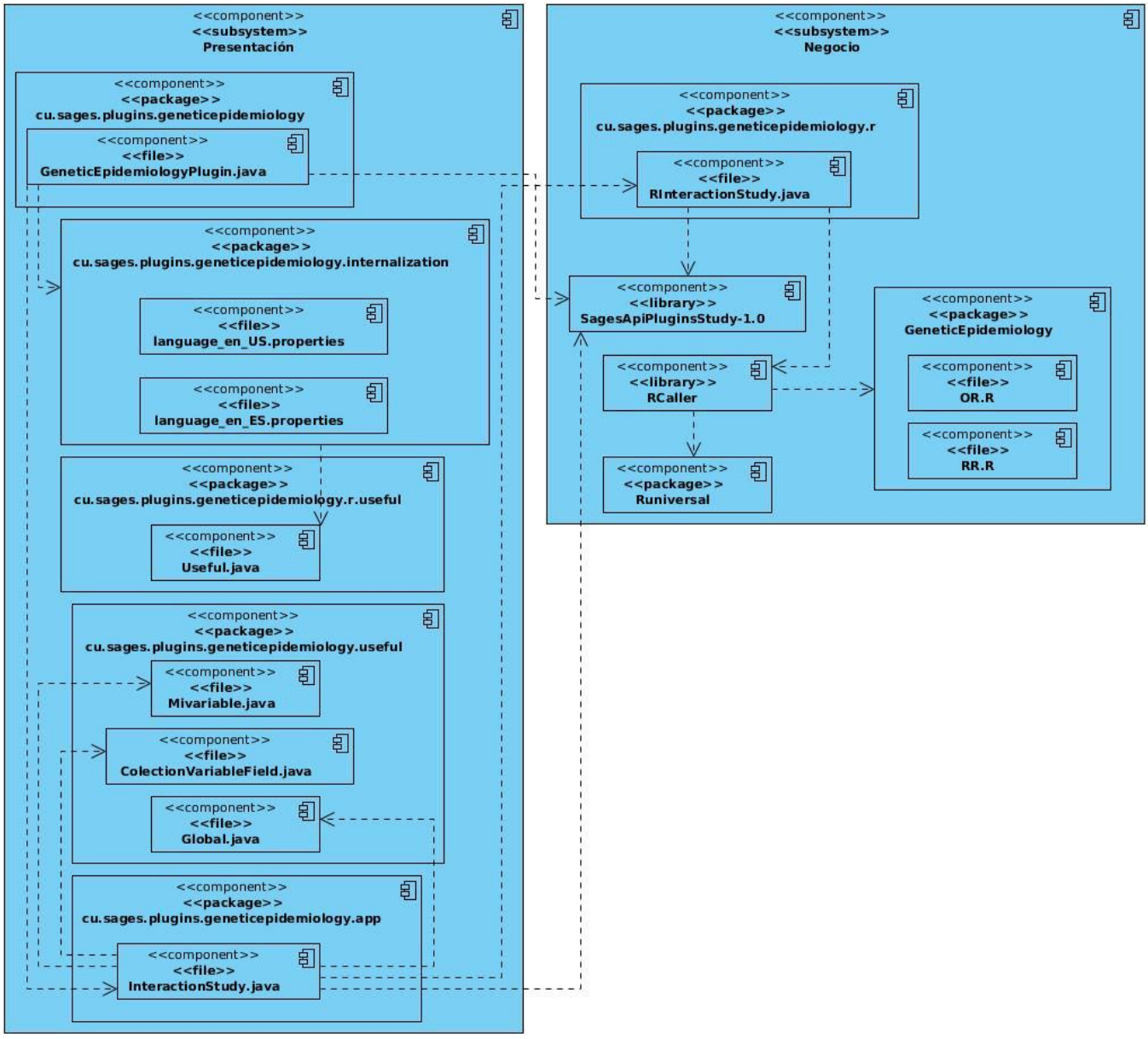


Figura 14: Diagrama de componentes: CUS: Crear estudio de interacción

4.2. Fragmentos de código fuente

En el presente acápite se muestran fragmentos del código fuente en R, pertenecientes a varios componentes del paquete *GeneticEpidemiology* por el impacto y la importancia que se le atribuye dentro de la extensión, ya que sobre él recae gran parte del negocio del sistema. Además se muestra un fragmento del código en Java el cual realiza la función de interactuar con el paquete *GeneticEpidemiology* para realizar el estudio de interacción, capturando el resultado del estudio.

4.2.1. Fragmentos de código en lenguaje en R

La función ***OR***, es el Odds Ratio, se le pasan los 4 valores a que se le va calcular el OR, se efectúa el cálculo y luego se construye el intervalo de confianza. A continuación se muestra un fragmento de código del paquete *GeneticEpidemiology* de la función OR.

```
OR <- function(n00, n01, n10, n11, alpha = 0.05){
  #
  #  Compute the odds ratio between two binary variables, x and y,
  #  as defined by the four numbers nij:
  #
  #    n00 = number of cases where x = 0 and y = 0
  #    n01 = number of cases where x = 0 and y = 1
  #    n10 = number of cases where x = 1 and y = 0
  #    n11 = number of cases where x = 1 and y = 1
  #
  OR <- (n00 * n11)/(n01 * n10)
  #
  #  Compute the Wald confidence intervals:
  #
  siglog <- sqrt((1/n00) + (1/n01) + (1/n10) + (1/n11))
  zalph <- qnorm(1 - alpha/2)
  logOR <- log(OR)
  loglo <- logOR - zalph * siglog
  loghi <- logOR + zalph * siglog
  #
  ORlo <- exp(loglo)
  ORhi <- exp(loghi)
  #
  oframe <- data.frame(LowerCI = ORlo, OR = OR, UpperCI = ORhi, alpha = alpha)
  oframe
}
```

Figura 15: Fragmento de código: Función OR

La función ***casos.controles*** se le pasa una matriz de datos y dos listas con las variables a interactuar y con el uso de la función *xtabs* se crean las tablas de contingencias de cada interacción. Luego con la función *cbind* se concatenan estas tablas y se forman las tablas de contingencias finales (este número de tablas depende de la cantidad de variables de la lista1 multiplicado por la cantidad de variables de la lista2) y finalmente con la función *rbind* se le adiciona una fila con el resultado del cálculo del Odds Ratio (OR). A continuación se muestra un fragmento de código del paquete *GeneticEpidemiology* de la función casos y controles.

```
casos.control <- function(matrix,lista1,lista2){
result <- NULL;
for(i in 1:length(lista1)){
  for(j in 1:length(lista2)){
        ma1 <- xtabs(~matrix[,1]+matrix[,colnames(matrix)==lista1[i]])
        ma2 <- xtabs(~matrix[,1]+matrix[,colnames(matrix)==lista2[j]])
        m <- cbind(ma1,ma2)
        m <- rbind(m,odds.ratio(m))
        result <- c(result,list(m))
        }
   }
   return(result)
}
```

Figura 16: Fragmento de código: Función casos y controles

La función ***PCP*** calcula la proporción de concordancia para el caso par.

La función ***PCCI*** calcula la proporción de concordancia para el caso índice.

La función ***falconer.test*** calcula un estimado de la heredabilidad.

La función ***clasic.twins.tes*** es para el estudio de clásicos en gemelos y en esta se calcula la proporción de concordancia para el caso par del gemelo monocigótico, la proporción de concordancia para el caso par del gemelo dicigótico, la proporción de concordancia para el caso índice del gemelo monocigótico y la proporción de concordancia para el caso índice del gemelo dicigótico. Además se calcula el estimando de heredabilidad por el método de Falconer y se obtiene un vector con los resultados de todas estas funciones. A continuación se muestra un fragmento de código del paquete *GeneticEpidemiology* de la función para clásicos en gemelos.

```
PCP <- function(Ep,En){
result <-Ep/(Ep+En)
return result
}

PCCI <- function(Ep,En){
result <- 2*Ep/(2*Ep+En)
return result
}

falconer.test <- function(PCCIMz,PCCIDz){
result <- 2*(PCCIMz-PCCIDz)
return result
}

clasic.twins.test <- function(EpMZ,EnMZ,EpDZ,EnDZ){
PCPMz <- PCP(EpMZ,EnMZ)
PCPDz <- PCP(EpDZ,EnDZ)
PCCIMz <- PCCI(EpMZ,EnMZ)
PCCIDz <- PCP(EpDZ,EnDZ)
h2 <- falconer.test(PCCIMz,PCCIDz)
vect <- c(PCPMz,PCPDz,PCCIMz,PCCIDz,h2)
return(vect)
}
```

Figura 17: Fragmento de código: Función para clásicos en gemelos

4.2.2. Fragmento de código en lenguaje Java

En la figura 18 se muestra el método ***analysisR***, el cual se encuentra implementado en la clase *RFamilialAgregationStudyCasoControl.java*. Este método recibe como parámetros un nombre y una matriz de datos y se encarga de conectar el entorno de R con la interfaz de Java. A través de *code.addRCod (…)* manda a ejecutar la sentencia con las funciones de R.

```
public void analysisR(String name, double[][] matrix){

    RCaller caller = new RCaller();
    Globals.detect_current_rscript();
    caller.setRscriptExecutable(Globals.Rscript_current);
    caller.setRExecutable(Globals.R_current);
    RCode code = new RCode();

    code.addRCode("library(GeneticEpidemiology)");
    code.setCode(addDoubleMatrix(code.getCode(), "m", matrix, false));
    try{
    code.addRCode("rpta <-try.catch(chisq.test(m, correct = F))");
    caller.setRCode(code);
    caller.runAndReturnResult("rpta");
    statistic = caller.getParser().getAsStringArray("statistic");
    parameter = caller.getParser().getAsStringArray("parameter");
    method = caller.getParser().getAsStringArray("method");
    p_value = caller.getParser().getAsStringArray("p_value");
    resultAnalysisR ="Method :"+method[0]+"\n"+"X-squeare ="+statistic[0]+"\t"
            +"df ="+parameter[0]+"\t"+"p-value ="+p_value[0]+"\n";
    resultAnalysisR +="\n";

    code.addRCode("rpta <- try.catch(chisq.test(m, correct = T))");
    caller.setRCode(code);
    caller.runAndReturnResultOnline("rpta");
    statistic = caller.getParser().getAsStringArray("statistic");
    parameter = caller.getParser().getAsStringArray("parameter");
    method = caller.getParser().getAsStringArray("method");
    resultAnalysisR +="Method :"+method[0]+"\n"+"X-squeare ="+statistic[0]+"\t"
            +"df ="+parameter[0]+"\t"+"p-value ="+p_value[0]+"\n";
    resultAnalysisR +="\n";

    code.addRCode("  rpta <-try.catch(fisher.test(m))");
    caller.setRCode(code);
    caller.runAndReturnResultOnline("rpta");
    statistic = caller.getParser().getAsStringArray("estimate");
    parameter = caller.getParser().getAsStringArray("null_value");
    method = caller.getParser().getAsStringArray("method");
    p_value = caller.getParser().getAsStringArray("p_value");
    resultAnalysisR +="Method :"+method[0]+"\n"+"Odds Ratio ="+statistic[0]+"\t"
            +"Null value ="+parameter[0]+"\t"+"p-value ="+p_value[0]+"\n";
    resultAnalysisR +="\n";

    code.addRCode("rpta <- try.catch(OR(m[1,1],m[1,2],m[2,1],m[2,2]))");
    caller.setRCode(code);
    caller.runAndReturnResultOnline("rpta");
    statistic = caller.getParser().getAsStringArray("OR");
    parameter = caller.getParser().getAsStringArray("LowerCI");
    method = caller.getParser().getAsStringArray("UpperCI");
    p_value = caller.getParser().getAsStringArray("alpha");
    resultAnalysisR +="Method : Odds Ratio \n"+"Odds Ratio ="+statistic[0]+"\t"
            +" Confidence Interval = ("+parameter[0]+" : "+method[0]+")"+"\t"
            +" alpha = "+ p_value[0]+"\n";
    System.out.print(resultAnalysisR);
    }catch(Exception e){
    resultAnalysisR += caller.getParser().getAsStringArray("rpta")[0];
    }
```

Figura 18: Fragmento de código de la clase RFamilialAgregationStudyCasoControl.java

4.3. Interfaces de la aplicación

En este acápite se muestran algunas de las interfaces de la aplicación, ejecutándose con un juego de datos sencillo.

Esta es la interfaz del estudio de agregación familiar para casos y controles, al cual se le insertan los datos preanalizados directamente en la tabla de contingencia.

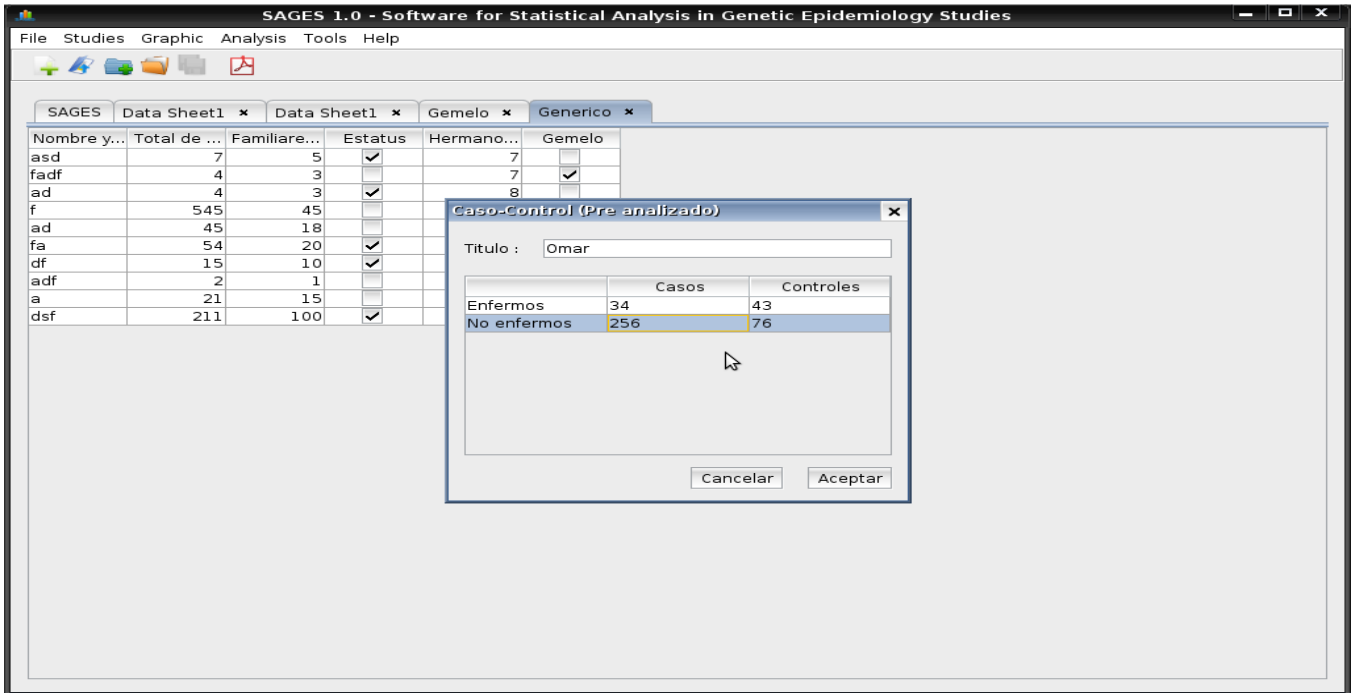


Figura 19: Interfaz del estudio de agregación familiar para casos y controles

Esta interfaz muestra los resultados del estudio de agregación familiar para casos y controles.

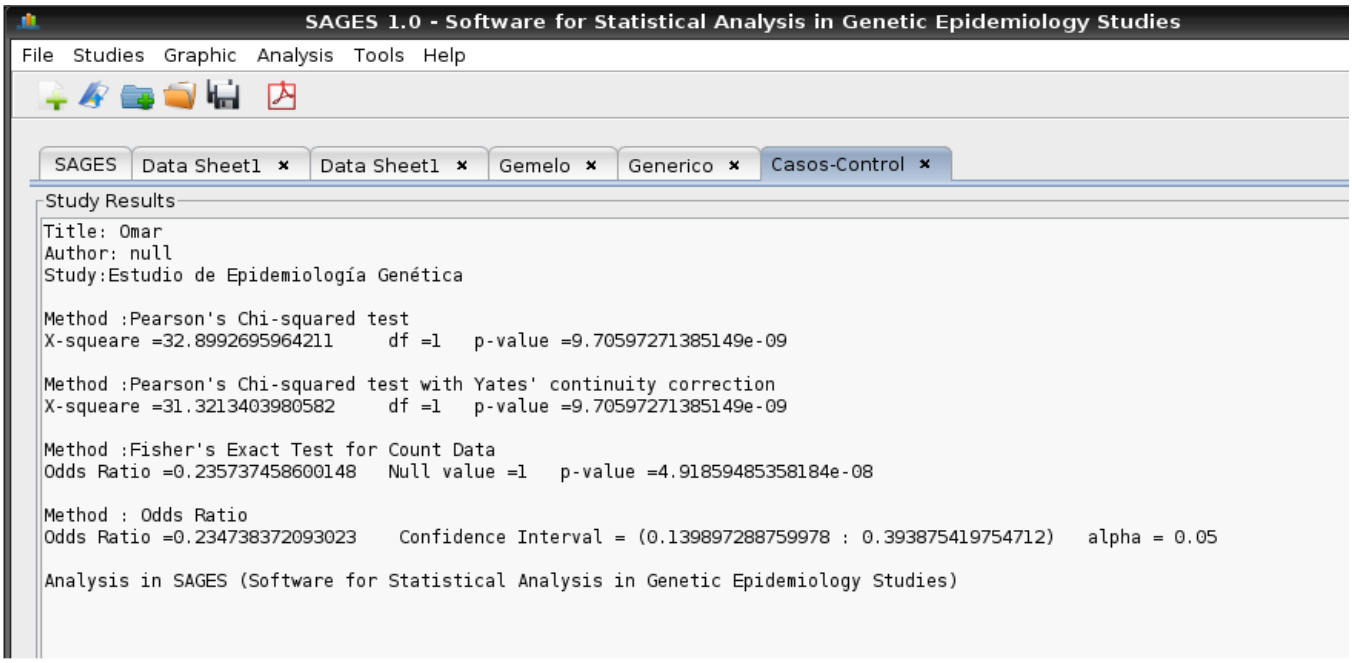


Figura 20: Interfaz de resultados: Estudio de agregación familiar para casos y controles

Esta interfaz muestra el estudio de agregación familiar para gemelos monocigóticos frente a un hermano carnal.

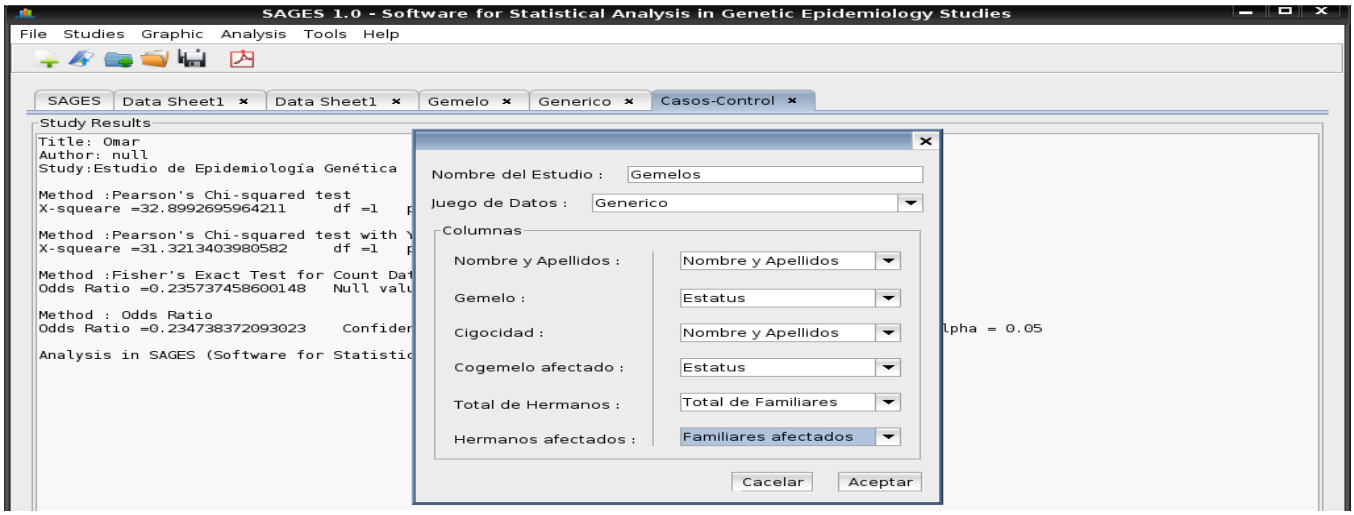


Figura 21: Interfaz estudio de agregación familiar para gemelos MC frente a un hermano carnal

Esta interfaz muestra los resultados del estudio de agregación familiar para gemelos monocigóticos frente a un hermano carnal.

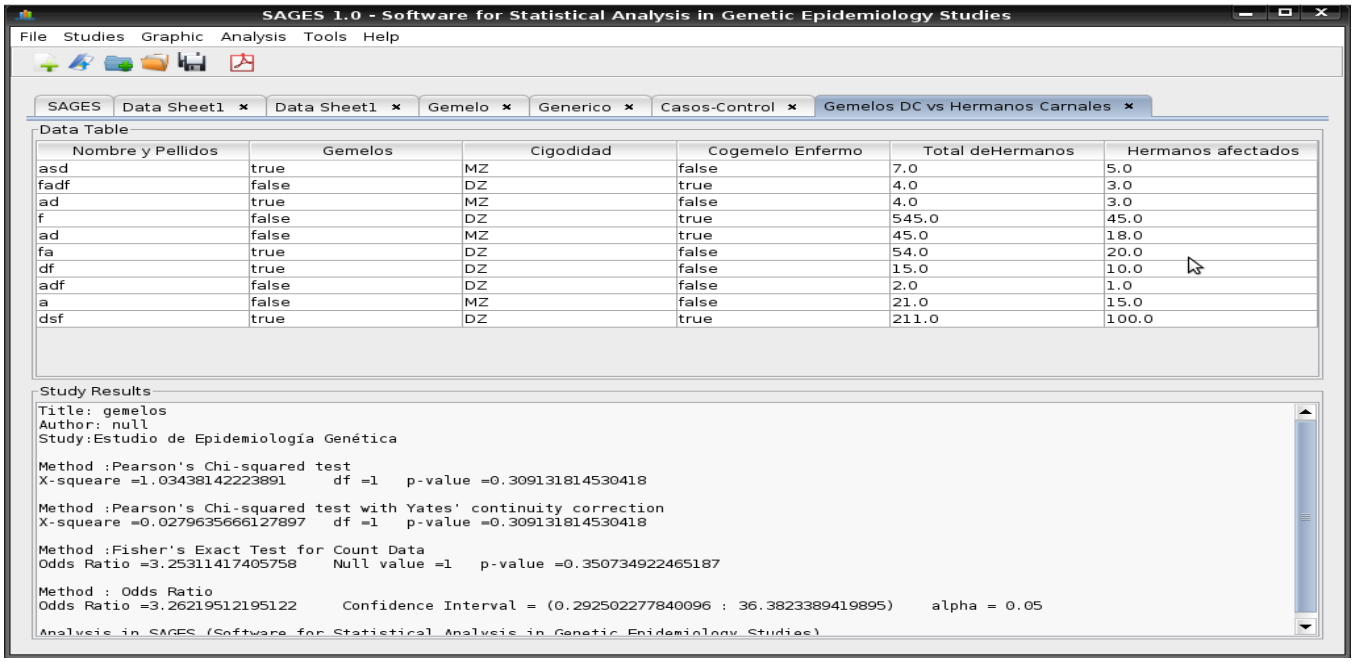


Figura 22: Interfaz de resultados: Estudio agregación familiar-Gemelos MC vs hermano carnal

4.4. Pruebas

Las pruebas son una actividad en la cual un sistema o componente es ejecutado bajo unas condiciones o requerimientos específicos, los resultados son observados y registrados, y una evaluación es hecha de algún aspecto del sistema o componente. La prueba es aplicada para diferentes tipos de objetivos, en diferentes escenarios o niveles de trabajo. A continuación se mencionan diferentes niveles de pruebas:

- **Prueba de Desarrollador:** Es la prueba diseñada e implementada por el equipo de desarrollo.
- **Prueba Independiente:** Es la prueba que es diseñada e implementada por alguien independiente del grupo de desarrolladores.
- **Prueba de Unidad:** Es la prueba enfocada a los elementos probables más pequeños del software.
- **Prueba de Integración:** Es ejecutada para asegurar que los componentes en el modelo de implementación operen correctamente cuando son combinados para ejecutar un caso de uso.
- **Prueba de Sistema:** Son las pruebas que se hacen cuando el software está funcionando como un todo.
- **Prueba de Aceptación:** Prueba de aceptación del usuario es la prueba final antes del despliegue del sistema.

Existen dos métodos fundamentales de pruebas:

Pruebas de caja negra: Pruebas que se llevan a cabo sobre la interfaz del software. El objetivo es demostrar que las funciones del software son operativas, que las entradas se aceptan de forma adecuada y se produce un resultado correcto, y que la integridad de la información externa se mantiene (no se ve el código). Se centran principalmente en los requisitos funcionales del software. Permiten encontrar:

- Funciones incorrectas o ausentes.
- Errores de interfaz.
- Errores en estructuras de datos o en accesos a las Bases de Datos externas.
- Errores de rendimiento.
- Errores de inicialización y terminación.

Pruebas de caja blanca: Se comprueban los caminos lógicos del software Se puede examinar el estado del programa en varios puntos para determinar si el estado real coincide con el esperado (es sobre el código). Requieren del conocimiento de la estructura interna del programa. Estas pruebas deben garantizar como mínimo que:

- Se ejerciten por lo menos una vez todos los caminos independientes para cada módulo.
- Se ejerciten todas las decisiones lógicas en sus vertientes verdaderas y falsa.
- Ejecuten todos los bucles en sus límites y con sus límites operacionales.
- Se ejerciten las estructuras internas de datos para asegurar su validez.

4.4.1. Diseño de casos de prueba y registro de no conformidades

Un caso de prueba se diseña según las funcionalidades descritas en los casos de usos. Este diseño se elabora previamente a realizar las pruebas funcionales a la aplicación. Se parte de la descripción de los casos de uso del sistema, como apoyo para las revisiones. Cada planilla de caso de prueba recoge la especificación de un caso de uso, dividido en secciones y escenarios, detallando las funcionalidades descritas en él y escribiendo cada variable que recoge el caso de uso. En el proceso de pruebas para la Extensión de Epidemiología Genética se hace uso del método de prueba de software de Caja Negra, por ser éste el método empleado en la liberación del producto software, aplicando el tipo de prueba del sistema.

Partiendo de la descripción de los casos de uso del sistema, como apoyo para las revisiones, se diseñó un caso de prueba asociado a cada caso de uso. Para detallar el caso de uso se utiliza una tabla, donde se desglosa esta funcionalidad en secciones y a su vez éstas en escenarios, para hacer más fructífera la ejecución de las pruebas. Esta tabla contiene los campos:

- Nombre de la sección: Se especifica el nombre de la sección [SC 1: Nombre de la sección].
- Escenarios de la sección: Se especifican los escenarios de cada sección [EC 1.1: Nombre del Escenario].
- Descripción de la funcionalidad: Se describe brevemente la funcionalidad del escenario.

El siguiente es un ejemplo donde se detalla el caso de uso Crear estudio de agregación familiar para casos población.

Nombre de la sección	Escenarios de la sección	Descripción de la funcionalidad

SC 1: Crear estudio de agregación familiar para casos y población.	EC 1.1. El genetista inserta los datos.	En este escenario el genetista inserta los datos.
	EC 1.2: El genetista inserta los datos incorrectamente.	Este escenario sigue la misma funcionalidad que el anterior verificando que todos los datos estén correctamente insertados.
	EC 1.3: El genetista no inserta los datos necesarios.	Este escenario sigue la misma funcionalidad que el primero pero verifica que todos los datos estén insertados.
	EC 1.4: El genetista no inserta variables.	Este escenario verifica que el genetista inserte variables dentro del estudio.

Tabla 20: Caso de prueba: Crear estudio de agregación familiar para casos y población

A partir de esta descripción se detallan las variables que se encuentran en la interfaz asociada al caso de uso que se le diseñó el caso de prueba.

No	Nombre del campo	Clasificación	Valor nulo	Descripción
1	Nombre del estudio	texto	no	Se introduce el nombre del estudio que va a realizar.
2	Frecuencia absoluta a partir de la proporción en la población.	números	no	Se introduce el valor de la proporción para la población.
3	Total de población.	números	no	Se introduce el total de población.
4	Frecuencia absoluta a partir de la proporción en los casos.	números	no	Se introduce el valor de la proporción para los casos.
5	Total de casos.	números	no	Se introduce el total de casos.

Tabla 21: Descripción de variables

Después de realizadas todas las pruebas, los resultados que no fueron satisfactorios pasaron a ser no conformidades y se emitieron en el registro de defectos y dificultades detectados que se encuentra en la parte final de cada diseño de caso de prueba. En la siguiente tabla se muestran algunas de las no conformidades encontradas tras la revisión del software.

Elemento	No.	No Conformidad	Aspecto correspondiente.	Etapa de detección	Clasificación.	Estado NC.
Aplicación	1	Cuando das clic en la opción adicionar variable criterio, da un error mostrando la Excepción	Crear estudio de interacción.	Pruebas de funcionalidad y confiabilidad	Significativa	PD: 9 de mayo de 2013. RA: 9 de mayo de 2013.

		ocurrida.				
Aplicación	2	La ventana que muestra como realizar el estudio no tenía el nombre de dicho estudio.	Crear estudios de agregación familiar para casos - población.	Pruebas de funcionalidad y confiabilidad	Significativa	PD: 9 de mayo de 2013. RA: 9 de mayo de 2013.
Aplicación	3	Las pestañas que se muestran en el módulo base con los resultados, no capturaban bien el nombre, asumían el mismo para todas.	Estudios de agregación familiar con todas sus variantes.	Pruebas de funcionalidad y confiabilidad	Significativa	PD: 9 de mayo de 2013. RA: 9 de mayo de 2013.
Aplicación	4	Faltó poner una tabla resumen con todos los resultados de calcular los métodos.	Crear un estudio de asociación genética.	Pruebas de funcionalidad y confiabilidad	Significativa	PD: 9 de mayo de 2013. RA: 9 de mayo de 2013.

Tabla 22: Caso de prueba: Resumen de no conformidades

A continuación se expone una gráfica que muestra como se fueron resolviendo las no conformidades encontradas por iteraciones.

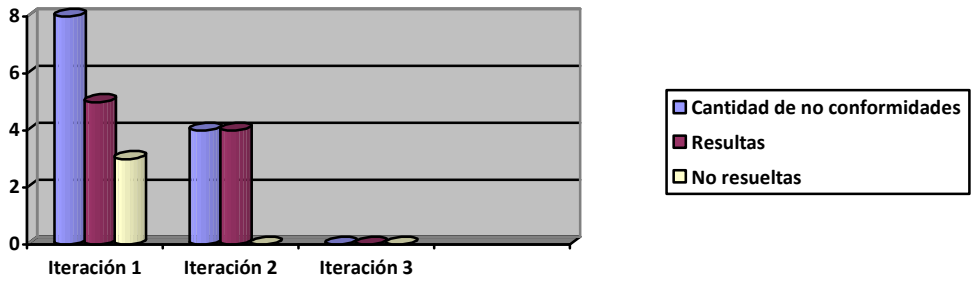


Figura 23: Gráfica que representa las iteraciones de las no conformidades

4.5. Conclusiones

En este capítulo se alcanzó la implementación del sistema, lo cual fue descrito a través de los diagramas de componentes y fragmentos de código. Además se validó el sistema a través de 6 casos de prueba, obteniéndose 8 no conformidades que fueron corregidas en tres iteraciones y revisadas por los desarrolladores.

CONCLUSIONES

1. El análisis del proceso de desarrollo de los estudios de Epidemiología Genética, permitió la correcta identificación de los requisitos de la aplicación para realizar el diseño de todas las clases mediante la utilización de patrones de diseño, lo que proporcionó una correcta implementación de la misma.
2. La aplicación de las pruebas al sistema permitió garantizar el correcto funcionamiento de la extensión, se detectaron no conformidades que fueron tratadas y finalmente solucionadas con éxito.
3. Los resultados de los estudios brindados por la extensión informática constituyen una valiosa fuente de información para la investigación y la toma de decisiones, en la prevención de estilos de vidas inadecuados involucrados en la expresión de las enfermedades que más afectan la morbimortalidad en la población cubana.
4. Con el desarrollo de la extensión para la aplicación SEEGEN-R se obtiene una herramienta que posibilita realizar nuevos estudios en el país, en el campo de la Epidemiología Genética.

RECOMENDACIONES

1. Agregar los Estudios de Adopción para ampliar las funcionalidades de la extensión.

2. Agregar los Estudios de Heredabilidad para que la herramienta obtenida sea aun más potente y centralizada.

REFERENCIAS BIBLIOGRÁFICAS

[1]. Muin J. Khoury, Terry H. Beaty, Bernice H. Cohen. Introducción. En: Responsabilidad de la obra completa. Fundamental of Genetic Epidemiology. Edición. New York: Oxford University Press, 1993.

[2]. FONTELA GONZÁLEZ, Dioleisys y GUERRA MACHADO, Leinys. EPIGEN: "Aplicación informática para el análisis estadístico en estudios de Epidemiología Genética". Trabajo de Diploma para optar por el título de Ingeniero en Ciencias de la Informática. Universidad de las Ciencias Informáticas. Ciudad Habana Ciudad Habana. 2009.

[3]. Centro Nacional de Genética Médica (CNGM). [En línea] [25-11-2012]. Disponible en: [http://www.sld.cu/sitios/genetica/verpost.php?blog=http://articulos.sld.cu/genetica&post_id=195&c=2987&tipo=2&idblog=141&p=1&n=de].

[4]. Rada G., Dr. Gabriel. EPI-CENTRO. Definiciones: Epidemiología. En: http://escuela.med.puc.cl/recursos/recepidem/introductorios4.htm. Revisado 2007

[5]. WYSZYNSKI, Diego F. La epidemiología genética: disciplina científica en expansión. Revista Panamericana de Salud Pública, Jan. 1998, vol. 3 no. 1 (Print version ISSN 1020-4989).

[6]. SEVILLA, Sergio D. Metodología de los estudios de asociación genética. Revista Insuficiencia Cardíaca, 2007, vol. 2, n.3:111-114.

[7]. KOLLINS, Scott H. PhD. Genética, neurobiología y neurofarmacología del trastorno por déficit de atención e hiperactividad (TDAH). RET, Revista de Toxicomanías., 2009, Nº. 55, 19-28.

[8]. REED, Kate. [En línea] Evidence for a genetic contribution. [14-1-2013]. Disponible en: [http://www.nchpeg.org/bssr/index.php?option=com_content&view=article&id=69&Itemid=145].

[9]. ORTEGA AZORÍN, Carolina. Interacción genético-ambiental en la modulación de adipocitoquinas y marcadores de inflamación en su asociación con obesidad y otros factores de riesgo cardiovascular en población mediterránea. Tesis doctoral, Departament de Medicina Preventiva i Salut Pública, Ciències de l'Alimentació, Toxicología i Medicina Legal, Universidad de Valencia, Valencia, 2011.

[10]. IBM**,** SPSS software, Predictive analytics software and solutions [En línea]. IBM SPSS Statistics, 2012. [10/1/2013]. Disponible en: [http://www-01.ibm.com/software/analytics/spss/products/statistics/].

[11]. CICA- Centro Informático Científico de Andalucía. [En línea]. PSPP, 2011. [14/12/2012]. Disponible en: [http://www.cica.es/Software/pspp.html].

[12]. BULMARO. [En línea].PSPP 0.6.2, 2010. [14/12/2012].Disponible en: [http://pspp.softbull.com/].

[13]. InfoStat: SOFTWARE ESTADÍSTICO. [En línea]. Infostat, 2012. [15/12/2012]. Disponible en: [http://www.infostat.com.ar/].

[14]. EXCOFFIER, Laurent, SCHNEIRDER, Stefan y ROESSLI, David. Análisis de la diversidad genética utilizando datos de marcadores moleculares: Módulo de aprendizaje. [En línea]. Programas informáticos para el análisis de la diversidad genética, 2004. [16/12/2012]. Disponible en: [http://www2.bioversityinternational.org/Publications/Molecular_Markers_Volume_2_es/PDF/IV.%20Programas%20inform%C3%A1ticos.pdf].

[15]. CASTIGLIA, Dra. Nora Inés, Epidat versión 3.1. [En línea]. Epidat, 2010-2011, [15/12/2012]. Disponible en: [http://www.consumaciencia.com.ar/epidat.html].

[16]. SANTIAGO PÉREZ, M. I., NAVEIRA BARBEITO, G. y Equipo de Epidat. EPIDAT 4.0: UNA HERRAMIENTA DE APOYO PARA LA ENSEÑANZA DE LA ESTADÍSTICA. En: IX Congreso Galego de Estatística e Investigación de Operaciones, 2009.

[17]. JACOBSON, Ivar, BOOCH, Grady y RUMBAUGH, James. El Proceso Unificado de Desarrollo de Software. Madrid: Pearson Educación, S.A, 2000. 84-7829-036-2.

[18]. LA ROSA GONZÁLEZ, Yudiel y PUPO SANTANA, Lianet. alasEPIGEN v2.0: "Aplicación informática para el análisis estadístico en estudios de Epidemiología Genética". Trabajo de Diploma para optar por el título de Ingeniero en Ciencias de la Informática. Universidad de las Ciencias Informáticas. Ciudad Habana. 2010.

[19]. Modelado de Sistemas con UML. [En línea]. [7/12/2012]. Disponible en: [http://mmc.geofisica.unam.mx/LuCAS/Tutoriales/doc-modelado-sistemas-UML/multiple-html/c12.html]

[20]. EcuRed. [En línea] Visual Paradigm. [Citado el 10/12/2012]. Disponible en: [http://www.ecured.cu/index.php/Visual_Paradigm]

[21]. MIRABAL SOSA, Mayelín, ROBINA GARCÍA, Maytee y URANGA PIÑA, Rolando. R: una herramienta poco difundida y muy útil para la investigación clínica. Revista Cubana de Investigaciones Biomédicas, 2010, v.29 n.2. Ciudad de la Habana, abril-junio.

[22]. Instituto de Estadística y Matemáticas de la WU Viena. What is R? [En línea] The R Project for statistical computing. [13-1-2013]. Disponible en: [http://www.r-project.org/about.html].

[23]. ZUKOWSKI, John. Programación Java 2. La Habana: Félix Varela, 2007.

[24]. NetBeans IDE 7.2.1 Información publicada. [En línea] [Última actualización: 06 de noviembre 2012] [Citado el 9/12/2012] Disponible en: [http://netbeans.org/community/releases/72/]

[25]. HAKAN SATMAN, Mehmet. New version: Rcaller 2.0. [En línea] Now, Rcaller has a new version, 2.0, 2011. [Fecha de consulta]. Disponible en: [http://www.mhsatman.com/rcaller/].

[26]. HAKAN SATMAN, Mehmet. Package 'Runiversal'. [En línea]. Runiversal - Package for converting R objects to Java variables and XML, 2012. [13-1-2013]. Disponible en: [http://cran.r-project.org/web/packages/Runiversal/Runiversal.pdf].

[27]. PASCUAL MORENO TAMAYO, Yoanki. Análisis y Diseño del Módulo Generador de Reportes del proyecto ONE. Trabajo de Diploma para optar por el título de Ingeniero en Ciencias de la Informática. Universidad de las Ciencias Informáticas. Ciudad Habana. 2008.

[28]. HERNÁNDEZ SARMIENTO, Diormis de la Caridad. Generador de datos. Trabajo de Diploma para optar por el título de Ingeniero en Ciencias de la Informática. Universidad de las Ciencias Informáticas. Ciudad Habana. 2008.

[29]. ALMAGUER ZALDIVAR, Jennys y CAMUÉ HERNÁNDEZ, Alexis. Plugin de recuperación del estado de entidades para la herramienta de administración de bases de datos HABD. Trabajo de Diploma para optar por el título de Ingeniero en Ciencias de la Informática. Universidad de las Ciencias Informáticas. La Habana. 2012.

[30]. ALARCÓN, RAÚL. Diseño orientado a objetos con UML. Madrid (España), Grupo EIDOS, 2000. 117 p.

[31]. PRESSMAN, R.S. Ingeniería de Software. Un enfoque práctico. 2002.

[32]. RUMBAUGH, James; JACOBSON, Ivar y BOOCH, Grady. Unified Modeling Language Reference Manual, The Pearson Higher Education, 2004.

[33]. GARCÍA PEÑALVO, Dr. Francisco José; CONDE GONZÁLEZ, Miguel Ángel y BRAVO MARTÍN, Sergio. Universidad de Salamanca – Departamento de Informática y Automática [En línea]. Ingeniería de Software: Tema 6: Diseño orientado a objetos, 2008. [15-4-2013]. Disponible en: [http://ocw.usal.es/ensenanzas-tecnicas/ingenieria-del-software/contenidos/Tema6-DOO-1pp.pdf].

BIBLIOGRAFÍA

- **Softwarelogia.** ¿Qué es un plugin? [En línea]. 2008. [11/12/2012]. Disponible en: [http://softwarelogia.com/2008/07/25/%C2%BFque-es-un-plugin/].

- **Saberia.** ¿Qué es un plugin? [En línea]. 2009 - 2012. [11/12/2012]. Disponible en: [http://www.saberia.com/2010/01/que-es-un-plugin/]

- Nueva versión: 2.0 Rcaller. [En línea]. Cambio de registro de la versión 2.0. 2011. [11/12/2012]. Disponoble en: [http://translate.google.com.cu/translate?hl=es&sl=en&u=http://www.mhsatman.com/rcaller.php&prev=/search%3Fq%3Drcaller%2Bexamples%26hl%3Des%26tbo%3Dd%26biw%3D1024%26bih%3D601&sa=X&ei=TfHHUKCFFpOG9gSEk4Ao&ved=0CDwQ7gEwAg]

- **Ivar Jacobson, Grady Booch, and Jim Rumbaugh,** Unified Software Development Process, Addison-Wesley, 1999.

- **Philippe Kruchten,** Rational Unified Process—An Introduction, Addison-Wesley, 1999.

- **Grady Booch,** Object Solutions, Addison-Wesley, 1995.

- **Ivar Jacobson, Magnus Christerson, Patrik Jonsson, and Gunnar Övergaard,** Object-Oriented Software Engineering—A Use Case Driven Approach, Wokingham, England, Addison-Wesley, 1992, 582p.

- **Ivar Jacobson, M. Griss, and P. Jonsson,** Software Reuse—Architecture, Process and Organization for Business Success, Harlow, England, AWL, 1997.

- **MIRABAL SOSA, Mayelín, ROBINA GARCÍA, Maytee y URANGA PIÑA, Rolando.** R: una herramienta poco difundida y muy útil para la investigación clínica. Revista Cubana de Investigaciones Biomédicas, 2010, v.29 n.2. Ciudad de la Habana, abril-junio.

- **Muin J. Khoury, Terry H. Beaty, Bernice H. Cohen**. Introducción. En: Responsabilidad de la obra completa. Fundamental of Genetic Epidemiology. Edición. New York: Oxford University Press, 1993.

- **FONTELA GONZÁLEZ, Dioleisys y GUERRA MACHADO, Leinys.** EPIGEN: “Aplicación informática para el análisis estadístico en estudios de Epidemiología Genética”. Ciudad Habana: s.n., 2009.

- **Centro Nacional de Genética Médica (CNGM).** [En línea] [25-11-2012]. Disponible en: [http://www.sld.cu/sitios/genetica/verpost.php?blog=http://articulos.sld.cu/genetica&post_id=195&c=2987&tipo=2&idblog=141&p=1&n=de].

- **Rada G., Dr. Gabriel**. EPI-CENTRO. Definiciones: Epidemiología. En: http://escuela.med.puc.cl/recursos/recepidem/introductorios4.htm. Revisado 2007.

- **WYSZYNSKI, Diego F.** La epidemiología genética: disciplina científica en expansión. Revista Panamericana de Salud Pública, Jan. 1998, vol. 3 no. 1 (Print version ISSN 1020-4989).

- **SEVILLA, Sergio D.** Metodología de los estudios de asociación genética. Revista Insuficiencia Cardíaca, 2007, vol. 2, n.3:111-114.

- **R. WRAY, Naomi Ph.D y M. VISSCHER, Peter.** Estimación de la heredabilidad Trait, 2008.

- **IBÁÑEZ CUADRADO, Ángela.** Genética de las adicciones. ADICCIONES, 2008. Vol.20 Núm. 2, Págs.103-110

- **LEJARRANGA, Dr. Horacio.** Heredabilidad y medioambiente en el desarrollo del niño. Arch Argent Pediatr, 2010. Vol.106 Núm. 6, Págs.532-537/ Artículo especial.

- **KOLLINS, Scott H. PhD.** Genética, neurobiología y neurofarmacología del trastorno por déficit de atención e hiperactividad (TDAH). RET, Revista de Toxicomanías., 2009, Nº. 55, 19-28.

- **REED, Kate.** [En línea] Evidence for a genetic contribution. [14-1-2013]. Disponible en: [http://www.nchpeg.org/bssr/index.php?option=com_content&view=article&id=69&Itemid=145].

- **ORTEGA AZORÍN, Carolina**. Interacción genético-ambiental en la modulación de adipocitoquinas y marcadores de inflamación en su asociación con obesidad y otros factores de riesgo cardiovascular en población mediterránea. Tesis doctoral, Departament de Medicina Preventiva i Salut Pública, Ciències de l'Alimentació, Toxicología i Medicina Legal, Universidad de Valencia, Valencia, 2011.

- **IBM,** SPSS software, Predictive analytics software and solutions [En línea]. IBM SPSS Statistics, 2012. [10/1/2013]. Disponible en: [http://www-01.ibm.com/software/analytics/spss/products/statistics/].

- **CICA**- Centro Informático Científico de Andalucía. [En línea]. PSPP, 2011. [14/12/2012]. Disponible en: [http://www.cica.es/Software/pspp.html].

- **BULMARO.** [En línea].PSPP 0.6.2, 2010. [14/12/2012].Disponible en: [http://pspp.softbull.com/].

- **InfoStat:** SOFTWARE ESTADÍSTICO. [En línea]. Infostat, 2012. [15/12/2012]. Disponible en: [http://www.infostat.com.ar/].

- **CASTIGLIA, Dra. Nora Inés,** Epidat versión 3.1. [En línea]. Epidat, 2010-2011, [15/12/2012]. Disponible en: [http://www.consumaciencia.com.ar/epidat.html].

- **SANTIAGO PÉREZ, M. I., NAVEIRA BARBEITO, G. y Equipo de Epidat.** EPIDAT 4.0: UNA HERRAMIENTA DE APOYO PARA LA ENSEÑANZA DE LA ESTADÍSTICA. En: IX Congreso Galego de Estatística e Investigación de Operacións, 2009.

- **LA ROSA GONZÁLEZ, Yudiel y PUPO SANTANA, Lianet.** alasEPIGEN v2.0: "Aplicación informática para el análisis estadístico en estudios de Epidemiología Genética". Trabajo de Diploma para optar por el título de Ingeniero en Ciencias de la Informática. Universidad de las Ciencias Informáticas. Ciudad Habana. 2010.

- JACOBSON, Ivar, BOOCH, Grady y RUMBAUGH, James. El Proceso Unificado de Desarrollo de Software. Madrid: Pearson Educación, S.A, 2000. 84-7829-036-2.

Modelado de Sistemas con UML. [En línea]. [7/12/2012]. Disponible en: [http://mmc.geofisica.unam.mx/LuCAS/Tutoriales/doc-modelado-sistemas-UML/multiple-html/c12.html]

- EcuRed. [En línea] Visual Paradigm. [Citado el 10/12/2012]. Disponible en: [http://www.ecured.cu/index.php/Visual_Paradigm]

- MIRABAL SOSA, Mayelín, ROBINA GARCÍA, Maytee y URANGA PIÑA, Rolando. R: una herramienta poco difundida y muy útil para la investigación clínica. Revista Cubana de Investigaciones Biomédicas, 2010, v.29 n.2. Ciudad de la Habana, abril-junio.

- Instituto de Estadística y Matemáticas de la WU Viena. What is R? [En línea] The R Project for statistical computing. [13-1-2013]. Disponible en: [http://www.r-project.org/about.html].

- ZUKOWSKI, John. Programación Java 2. La Habana: Félix Varela, 2007.

NetBeans IDE 7.2.1 Información publicada. [En línea] [Última actualización: 06 de noviembre 2012] [Citado el 9/12/2012] Disponible en: [http://netbeans.org/community/releases/72/]

- QUESADA ARENCIABIA, Alexis y SANTANA PÉREZ, Francisco J. Sistemas de Control de Versiones. [En línea] [Citado el 5/12/2012]. Disponible en: [http://sopa.dis.ulpgc.es/progsis/material-didactico-practico/enunciados-practicas/Sistemas_Control_Versiones.pdf].

- COLLINS SUSSMAN, Ben, FITZPATRICK, Brian W. y PILATO, C. Michael. Version Control with Subversion For Subversion 1.6. California, 2009. s. n.

- HAKAN SATMAN, Mehmet. New version: Rcaller 2.0. [En línea] Now, Rcaller has a new version, 2.0, 2011. [Fecha de consulta]. Disponible en: [http://www.mhsatman.com/rcaller/].

- HAKAN SATMAN, Mehmet. Package 'Runiversal'. [En línea]. Runiversal - Package for converting R objects to Java variables and XML, 2012. [13-1-2013]. Disponible en: [http://cran.r-project.org/web/packages/Runiversal/Runiversal.pdf].

- Angelfire. Patrones GoF. [En línea]. Patrones GoF, 2011 [8-4-2013]. Disponible en: [http://geektheplanet.net/5462/patrones-gof.xhtml].

GLOSARIO

Plugin: (Del inglés plug-in “enchufable o inserción”). Plugin es una aplicación informática que interactuando con otra aplicación le aporta una función específica a ésta última y esa aplicación adicional nombrada es ejecutada por la principal. La idea es que el nuevo componente se *enchufa* simplemente al sistema existente. Los plugins no son parches ni actualizaciones, sino propiedades añadidas a los programas originales, aparecidas por primera vez a mediados de los años 70 y conocidas también como complementos, componentes, extensiones y addons (del inglés add-on, “agregado”).

Genotipado: Por genotipado, o genotipificación o caracterización genética, se entiende el proceso de determinación del genotipo o contenido genómico, en forma de ADN, específico de un organismo biológico, mediante un procedimiento de laboratorio. En otras palabras, es la técnica de laboratorio que se utiliza para determinar la información genética de un organismo, o genotipo, y poder diferenciarlo del resto.

Morbimortalidad: Es un concepto complejo que proviene de la ciencia médica y combina dos subconceptos: morbilidad y mortalidad. La morbilidad es la presencia de un determinado tipo de enfermedad en una población. La mortalidad es la estadística sobre las muertes un una población también determinada. Así que juntando los dos subconceptos, puede entenderse que la morbimortalidad significa aquellas enfermedades causantes de de la muerte en determinadas poblaciones, espacio y tiempo.

Sesgos: El diccionario de la Real Academia Española (RAE) menciona que sesgo es la oblicuidad o torcimiento de una cosa hacia un lado. El concepto también se utiliza en sentido simbólico para mencionar una tendencia o inclinación.

Genotipo: La palabra genotipo deriva de los vocablos latinos “genus” = origen y “typus” = tipo. Se denomina genotipo al conjunto de genes que un individuo, animal o vegetal, en forma de ADN, que recibe por herencia de sus dos progenitores, formado por lo tanto, de las dos dotaciones de cromosomas, que contienen la información genética. La Genética como ciencia biológica, estudia los genotipos, y su manifestación exterior: los fenotipos.

Fenotipo: Conjunto de caracteres de un organismo que se manifiestan como resultado de la interacción entre el genotipo de dicho organismo y el medio ambiente que le rodea.

Diabetes mellitus: La diabetes mellitus (DM) es un conjunto de trastornos metabólicos que afecta a diferentes órganos y tejidos, dura toda la vida y se caracteriza por un aumento de los niveles de

glucosa (azúcar) en la sangre. Durante la digestión el organismo metaboliza los azúcares, almidones e hidratos de carbono, transformándolos en azúcares simples, esta va al torrente sanguíneo y con la ayuda de la insulina que es una hormona secretada por el páncreas, la glucosa se transforma en energía que es aprovechada por las células del cuerpo. Cuando no existe insulina o cuando esta no es producida en forma efectiva, la glucosa se acumula en sangre aumentando los niveles de azúcar, esto es la diabetes mellitus y constituye actualmente una de las principales causas de preocupación en salud pública.

SEEGEN-R: Sistema Estadístico de Epidemiología Genética- basado en R.

CNGM: Centro Nacional de Genética Médica.

IDE: Entono de Desarrollo Integrado.

GUI: Interfaz Gráfica de Usuario.

UML: Lenguaje Unificado de Modelado.

MVJ: Máquina Virtual de Java.

CVS: Sistema de Control de Versiones.

XML: Lenguaje de Marcado Extensible.

Estudios de Adopción: Los estudios de adopción miden el riesgo genético para un trastorno comparando las tasas de un trastorno entre los padres biológicos y los adoptadores de niños con una condición determinada. Si un trastorno es altamente genético por naturaleza, uno podría predecir que las tasas del trastorno serán más altas en los padres biológicos de un niño con el trastorno. Por diversas razones prácticas, los estudios de adopción son difíciles de realizar, pero los que se han llevado a cabo apoyan las bases genéticas del trastorno por déficit de atención e hiperactividad.

Estudios de Heredabilidad: Los estudios de heredabilidad, es en resumen, la cantidad de variación en un rasgo, se debe a la variación de los factores genéticos. A menudo, este término se utiliza en referencia a la semejanza entre los padres y sus hijos. En este contexto, la heredabilidad alta implica un gran parecido entre padres e hijos con respecto a un rasgo específico, mientras que la heredabilidad baja implica un bajo nivel de semejanza.