UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DETECCIÓN TEMPRANA DE RIESGO CARDIOVASCULAR USANDO TEXT MINING EN LOS CAMPOS DE TEXTO NO ESTRUCTURADO DEL REGISTRO CLÍNICO ELECTRÓNICO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CRISTIAN IGNACIO MOLINA ESPINOZA

PROFESOR GUÍA:
SR. JUAN D. VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
SR. FRANCISCO MOLINA JARA
SR. IGNACIO CALISTO LEIVA

SANTIAGO DE CHILE
2014

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR : CRISTIAN MOLINA ESPINOZA
FECHA: 02/12/2014
PROF. GUIA: SR. JUAN VELÁSQUEZ

Resumen Ejecutivo

Objetivo: Generar un modelo predictivo, basado en Machine Learning (ML) y Natural Language Processing (NLP), que a partir de signos y síntomas detectados en los campos de texto no estructurados del Registro Clínico Electrónico, pueda predecir niveles altos de riesgo cardiovascular de una persona.

Contexto: Detectar con anticipación el riesgo cardiovascular podría mejorar de gran manera el bien estar de las personas y disminuir los costos asociados su tratamiento. Actualmente, se usa el criterio de Framingham para detectar riesgo cardiovascular.

Problema: El médico puede utilizar muy poco tiempo en un paciente en la atención primaria de salud pública, por lo que no puede "sintetizar"toda la información de la historia clínica para evaluar el riesgo cardiovascular. Además, hay una tendencia a la baja en la cobertura de de los programas preventivos, donde se aplica el formulario Framingham. Luego, existe una alta probabilidad que a un gran número de personas no se les comunique a tiempo su nivel alto de riesgo cardiovascular.

Hipótesis: Existe información valiosa en los campos de texto no estructurado del registro clínico electrónico, para detectar de forma temprana Riesgo Cardiovascular.

Diseño: Se propone sintetizar de forma automática los registros de todas las atenciones de un paciente, y detectar los signos y síntomas registrados en los campos de texto no estructurado del registro clínico electrónico que permita evaluar y predecir el riesgo cardiovascular de una persona aplicando técnicas de minería de datos (text mining). Se calibró un modelo Logistic Regression, para realizar la predicción sobre Riesgo Cardiovascular.

Resultado: Se evaluó el desempeño del modelo de acuerdo a la medida AUC= 0.968 y F-Measure= 88.6 %. Además, se agrega valor al detectar riesgo cardiovascular en personas que no pertenecen al PSCV o estaban clasificados con nivel de riesgo moderado o bajo.

Conclusión: Es posible validar que existe información valiosa en los campos de texto no estructurado del registro clínico electrónico, que permite detectar de forma temprana el riesgo cardiovascular.

Fíate de Jehová de todo tu corazón, Y no te apoyes en tu propia prudencia.
Reconócelo en todos tus caminos, Y él enderezará tus veredas.
No seas sabio en tu propia opinión; Teme a Jehová, y apártate del mal;
Porque será medicina a tu cuerpo, Y refrigerio para tus huesos.
Proverbios 3:5-8

Vanidad y palabra mentirosa aparta de mí;
No me des pobreza ni riquezas; Manténme del pan necesario;
No sea que me sacie, y te niegue, y diga: ¿Quién es Jehová?
O que siendo pobre, hurte, Y blasfeme el nombre de mi Dios.
Proverbios 30:8-9

Agradecimientos

A mis padres, hermanos y hermanita que siempre e incondicionalmente han estado conmigo.

A Rita y a su familia que fueron un gran apoyo en este proceso.

A Dios por cada una de las cosas que tengo en mi vida.

CRISTIAN IGNACIO MOLINA ESPINOZA

Tabla de contenido

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Contexto

El Registro Clínico Electrónico (RCE) se está ocupando cada vez más en las prácticas de atención primaria y en la atención hospitalaria. Esta transición a registros digitales también representa una transición importante en cómo los datos del paciente están organizados y que estén accesibles para multiples usos, como minería de datos, lo que era inimaginable con los registros en papel. En particular, el RCE ofrece acceso directo a los datos, y es posible obtener una historia clínica del paciente a través de las atenciones que ha recibido. Estos datos pueden ser procesados y comparados con todo el resto de la información almacenada de los demás pacientes, así el médico podría complementar una atención clínica, al predecir los resultados de determinados tratamientos, de acuerdo a información histórica registrada, o anticiparse a diagnósticos futuros. Abriendo así, oportunidades para personalizar la toma de decisiones en la atención de cada paciente.

En paralelo, durante la última década, las técnicas de aprendizaje automático (Machine Learning) han evolucionado y ofrecen un medio para extraer rápidamente información de grandes conjuntos de datos de alta dimensionalidad. Las técnicas de Machine Learning, tiene mucha relación con los procedimientos utilizados para la minería de datos (Data Mining). Que consiste en la extracción de patrones útiles, a partir de los datos, para resolver preguntas de interés. Estas técnicas están estrechamente relacionadas con la estadística y la ingeniería, y su uso en los datos del RCE para realizar modelos predictivos plantea un desafío único. Sin embargo, su uso en Chile, con estos fines, es casi nulo.

El presente trabajo se enfoca en modelos de predicción para detectar de forma temprana el riesgo cardiovascular, a partir de los datos del registro clínico electrónico de la atención primaria de salud pública en Chile.

La importancia de trabajar en medidas preventivas y de detección temprana de riesgo se debe a que los problemas al corazón se presentan como una enfermedad progresiva grave, que por lo general

se detectan en una etapa avanzada, dejando pocas opciones para frenar la progresión.

De acuerdo a datos de la Organización Mundial de la Salud (OMS) [1], las enfermedades cardiovasculares constituyen una de las causas más importantes de discapacidad y muerte prematura en todo el mundo. El principal problema se debe a la aterosclerosis, que es una enfermedad en que la placa se deposita dentro de las arterias, que son vasos sanguíneos que llevan sangre rica en oxígeno al corazón y a otras partes del cuerpo. La placa está compuesta por grasas, colesterol, calcio y otras sustancias que se encuentran en la sangre, que con el tiempo se endurece y estrecha las arterias, lo que limita el flujo de sangre rica en oxígeno a los órganos y a otras partes del cuerpo.

La aterosclerosis progresa a lo largo de los años, de modo que cuando aparecen los síntomas, generalmente a mediana edad, suele estar en una fase avanzada. Las principales consecuencias son los episodios coronarios (infarto de miocardio) y cerebrovasculares agudos, que se producen de forma repentina y conducen a menudo a la muerte antes de que pueda dispensarse la atención médica requerida.

Por otra parte, en Chile, la prevención y el control de las enfermedades cardiovasculares son parte de los principales desafíos sanitarios del país. Esto se lleva a cabo a través de la red de servicios de salud, donde es fundamental fortalecer y ampliar la cobertura de los programas preventivos en la atención primaria (APS) y mejorar la articulación entre los distintos niveles de atención para asegurar la continuidad y una mejor calidad de la atención.

De acuerdo a lo antes mencionado, la investigación y el desarrollo de tecnologías que permitan realizar una detección temprana del riesgo, para reducir los episodios cardiovasculares y la muerte prematura en aquellas personas con alto riesgo cardiovascular, se alinea con la estrategia país de prevención.

Lo que se presenta en este trabajo es destacar lo importante de los modelos predictivos y las técnicas de Machine Learning que permiten predecir niveles altos de riesgo cardiovascular a partir del Registro Clínico Electrónico, evaluando la presencia de determinados factores de riesgo en cada persona, de acuerdo a su edad, sexo, sus de niveles de presión arterial y colesterol, entre otros. Con lo que se podría aplicar medidas preventivas a los pacientes con alto riesgo de sufrir alguna enfermedad cardiovascular.

A continuación, se resumen los principales concepto y temas tratados en el presente trabajo.

1.1.1. Registro Clínico Electrónico

Actualmente, y gracias al uso de tecnologías de información, es posible almacenar (de forma digital) los datos asociados a las atenciones de personas hechas en un establecimiento de salud. Estas tecnologías se conocen como Registro Clínico Electrónico (RCE).

El RCE está diseñado para procesar la información clínica y administrativa de un paciente, que puede contener datos demográficos, historial médico, información sobre una admisión anterior, cirugías previas o la historia obstétrica, por ejemplo. El historial médico, a su vez, incluye el malestar o enfermedad por la cual el paciente fue atendido y su historia familiar.

El tipo de datos disponibles en el RCE dependerá de quién realiza el ingreso, pero pueden incluir solicitudes y resultados de laboratorio, exámenes de radiología, prescripciones, o la generación de un informe de alta.

Gracias al Registro Clínico Electrónico, el médico tiene acceso a la información del paciente en forma inmediata, sin depender de la tradicional ficha en papel, que al no estar escrita a mano, es más entendible y ordenada, especialmente los campos de texto libre (o no estructurados) donde se indican los detalles de la atención. Adicionalmente, se mejora la calidad del proceso asistencial, evitando la duplicidad de exámenes y procedimientos que inciden en el costo final de la atención de salud.

Pero este es tan sólo el primer paso, ya que no basta sólo con tener un gran repositorio de datos clínicos. El uso de tecnologías también debe entregar información útil, a través de la generación de reportes y análisis, para mejorar la toma de decisiones. En caso contrario, el RCE se transformaría en papel digital y sólo serviría para acceder de manera un poco más eficiente a los datos, desperdiciando el gran potencial de tener los datos clínicos históricos de un gran número de personas.

En consecuencia, lo que motiva este trabajo es extender el uso de los datos del Registro Clínico Electrónico para realizar investigación y desarrollar aplicaciones que permitan apoyar la toma de decisiones tanto a nivel Operativo (Establecimientos), Táctico (Servicios de Salud) y Estratégico (Ministerio de Salud).

1.1.2. Minería de Datos en Campos de Texto

Es necesario desarrollar tecnologías y herramientas que puedan convertir una gran cantidad de datos en información útil, para poder obtener conocimiento desde esta información.

La minería de datos, en adelante Data Mining, reúne un conjunto de técnicas que permite complementar los métodos tradicionales de análisis con algoritmos más complejos, para extraer información útil a partir de una gran cantidad de datos. Uno de los objetivos de Data Mining es extraer información o reglas desconocidas a partir del procesamiento masivo de datos.

Además, es posible considerar Data Mining como una metodología interdisciplinaria, ya que relaciona tanto análisis de base de datos, estadística, machine learning [2–4], entre otras. Algunos consideran que Data Mining es el resultado de la evolución natural de las Tecnologías de Información. De acuerdo al trabajo específico que se quiera realizar y sus objetivos, Data Mining se puede dividir

en object data mining, spatial data mining, multimedia data mining, web mining y text mining. [5,6]

Este trabajo se enfocará en Text Mining, con el objetivo de recuperar información útil que los médicos registran en los campos de texto no estructurado del RCE. En términos general, cabe mencionar, que el texto es la representación más importante de información, de acuerdo a investigaciones estadísticas, estas muestran que el 80 % de la información en las organizaciones está almacenada en forma de texto [7]. Además, a través del texto se puede expresar una cantidad enorme de tipos de información, con diferentes usos y significados. Por lo que en una base de datos, los campos de texto pueden tener un enorme potencial para obtener conocimiento que aún no ha podido ser detectado. Sin embargo, si se quiere explotar estos datos con herramientas computacionales existe una barrera semántica, que tiene relación con la lingüística y el significado de las palabras, sobre todo cuando se trata de una gran cantidad de documentos.

El uso de técnicas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés) y las nuevas herramientas que permiten trabajar sobre campos de texto, permite saltar la barrera semántica para lograr extraer conocimiento a partir de los campos de texto no estructurados. [6,8,9]

1.1.3. Situación Actual y Oportunidad en la Salud Pública

Como ya se mencionaba, el gran desafío en Salud es aprovechar la gran cantidad de registros del RCE y convertir los datos en información valiosa, usando las nuevas técnicas y herramientas de Data Mining.

Este trabajo cuenta con el apoyo de la empresa SAYDEX, que posee con una base de datos de un Registro Clínico Electrónico centralizado a lo largo de todo Chile, con una cobertura superior al 70 % de total de los establecimientos de atención primaria, que de acuerdo a datos de la empresa corresponde a tener la ficha clínica electrónica de casi 9 millones de pacientes de un total de 12 millones que se atiende en la red de atención pública de salud en Chile. Esta es una ventaja competitiva a nivel nacional y un gran avance a nivel internacional, ya que generalmente en otros países cada establecimiento tiene su propio repositorio de datos que no está integrado con otros establecimientos.

Aunque Chile aún no cuenta con una Ficha Clínica Única Compartida, la base de datos en la que se realizará el presente trabajo es lo más cercano a una ficha de este tipo en Chile.

Descripción de la Empresa

La empresa se declara como co-custudio de los datos de las personas que se atienden en el sistema público de salud, siendo proveedor de soluciones y aplicaciones para el Ministerio de Salud de Chile (MINSAL) y para varias de las Redes Asistenciales del país. En este último nicho, entrega

soluciones específicas que abordan con propiedad los niveles de Atención Hospitalaria y Primaria en función al Modelo Biomédico y Biosicosocial, respectivamente. Entre otros atributos, la empresa se diferencia por [1]:

- Disponibilización de herramientas robustas que abordan efectivamente y con propiedad los modelos de atención de las personas (Biosicosocial /Biomédico).
- Conocimiento de las "reglas de negocio"presentes en el sector, del entorno cultural y condiciones de contexto.
- Involucramiento y responsabilidad en los emprendimientos alineados con el liderazgo del mandante.
- Conocimiento, metodología ajustada y experiencia en el medio puesta al servicio de los proyectos.

Las herramientas provistas por la empresa, actualmente en uso por Establecimientos de las Redes Asistenciales y por la autoridad sanitaria, son un apoyo a la gestión que permite contar con información centralizada, compartida y en línea, asegurando una operación continua y protegida de los registros que se llevan a cabo bajo estándares y normas de seguridad e interoperabilidad definidas por las autoridades.

Datos Disponibles

La empresa cuenta con una cobertura de 13 hospitales de mediana y alta complejidad, 35 de baja complejidad y más de 480 establecimientos de atención primaria (de un total de aproximadamente 600). Lo que significa que SAYDEX provee la principal herramienta de Registro Clínico Electrónico en Atención Primaria de Salud (APS) en Chile, que permite entregar información de apoyo a la gestión clínica y administrativa orientado para a los diferentes centros de salud (CESFAM, COSAM, PSR, CECOF, etc), en el marco del modelo biosicosocial de APS. La gran ventaja es que cuenta con una base de datos centralizada para todos los establecimientos del país que usan el sistema.

Entre sus principales características se destacan:

- Historia clínica, con acceso a información relevante como alergias, factores de riego, mórbidos, fármacos en uso.
- Historia de atenciones ordenadas por tipo de profesional, clasificación diagnostica, actividad, prescripción.

[1] Fuente: Sitio web de la empresa. http://www.saydex.cl/

- Ficha odontológica.
- Ficha familiar.
- Instrumentos de evaluación y genograma en 4 capas.

Por otra parte, la empresa también cuenta con un sistema de Información Hospitalario (HIS, por sus siglas en inglés) de apoyo a la gestión clínica, operacional y estratégica, que soporta la relación con el paciente y el profesional que lo atiende. Y que permite tener la Historia Clínica Compartida con la Atención Primaria (APS) a través de RAYEN.

Entre sus principales características se destacan:

- Historia clínica básica y avanzada.
- Gestión de pacientes.
- Gestión de camas en hospitalización.
- Gestión de listas de espera.
- Procesos compatibles con las patologías Auge y las Garantías explícitas de salud GES.
- Farmacia.
- Gestión de pabellones, bloques quirúrgicos.
- Consultas externas.
- Cuadro de mando, indicadores operacionales en línea.

Situación Actual en SAYDEX

Hoy en día, no se están aprovechando los datos que se almacenan gracias al sistema de Registro Clínico Electrónico. Los registros de las atenciones permanecen guardados, pero no se están procesando para extraer información.

El presente año se han dado los primero pasos en la dirección de obtener información desde los datos, con el uso de algunas herramientas de inteligencia de negocios. Sin embargo, estos esfuerzos están enfocados en análisis descriptivos, es decir, sólo muestran los datos de una forma más ordenada y visualmente más atractiva. Sin embargo, no se hacen análisis predictivos, desaprovechando toda la información que se podría extraer desde los datos históricos almacenados.

Es por esto, que la empresa apoya la iniciativa de llevar a cabo un proyecto para usar los datos disponibles para realizar estudios usando las técnicas de DataMining.

1.1.4. Descripción del Problema y Justificación

Problema

En el caso particular de las enfermedades al corazón, se observa que hay métodos para la evaluación temprana de riesgo cardiovascular, basados principalmente en las Tablas de Framingham (Ver Anexo A), que están incluidas en los EMPA (Exámenes de Medicina Preventiva para Adultos), los que son subvencionados por el Estado para que cualquier persona pueda acceder a ellos, ya que son parte de los programas de salud preventiva a través de las garantías explícitas de salud (GES). Sin embargo, muy poca gente utiliza este beneficio, ya que en la mayoría de los casos las personas acuden a un establecimiento de salud sólo cuando presentan alguna enfermedad. En el caso particular de los controles por enfermedades cardiovasculares, existe una tendencia a la baja en la cobertura del programa de salud cardiovascular (PSCV). De acuerdo a las cifras [2] entregadas por el Sistema de Información Municipal (SINIM) de la Subsecretaría de Desarrollo Regional, el número de adultos que se controlan en el programa de riesgo de enfermedades cardiovasculares en los consultorios ha caído 31 % en los últimos cinco años. Si en 2009 eran 2,5 millones de personas las que estaban bajo control, en 2013 fueron 1,7 millones, lo que no responde a ninguna lógica sanitaria, pues se trata de enfermedades crónicas que los acompañarán por el resto de sus vidas.

Por otra parte, se observa que las personas cuando se atienden por motivos que no tienen relación directa con alguna enfermedad al corazón, el médico de igual forma registra factores que se utilizan para evaluar el riesgo cardiovascular. Sin embargo, como el motivo de la consulta no corresponde a un evaluación de riesgo coronario, estos datos quedan registrados como parte de la historia clínica del paciente, pero no son utilizados posteriormente para algún tipo de análisis.

Además, no se está considerando la posibilidad de obtener factores de riesgo registrados en atenciones hechas anteriormente en el mismo establecimiento o en otros. Actualmente, la única forma de hacerlo es manualmente, es decir, que el médico lea toda la historia clínica del paciente en cada atención, pero el problema es que en los establecimientos de salud pública siempre hay una la alta demanda, lo que se traduce en que los médicos deben reducir el tiempo para atender a cada paciente, por lo que leer y revisar en detalle estos datos en cada atención consumiría gran parte del tiempo de la atención.

En resumen, se observa que gracias al Registro Clínico Electrónico es posible almacenar y acceder en forma rápida a datos relativos a la atención de los pacientes, pero no se está explotando.

La problemática planteada fue validada por médicos especialistas en cardiología en una entrevista que se realizó el 28 de Mayo del presente año, en el Hospital Clínico de la Universidad de Chile

[2] Fuente: Sitio web Colaboración Pública Salud. http://saludaps.colaboracionpublica.org

(HCUCH). En esta instancia, Luis Sepúlveda[3] mencionó la dificultad que tienen los médicos en la atención primaria con respecto a los cortos tiempos de atención por paciente, lo que hace muy complejo evaluar en detalle la historia clínica del paciente. Por otra parte, Hernán Prat[4] recalcó la importancia de poder contar con un sistema que pueda utilizar los datos que normalmente se registran en una atención, sin tener la necesidad de utilizar herramientas complementarias, donde se debe re-ingresar los datos, lo que al fin y al cabo, también utiliza el escaso tiempo de cada atención. Mencionó también, que hace unos 10 años atrás se intentó implementar un sistema experto que necesitaba que se re-ingresaran parte de los datos de la atención a dicho sistema. Aunque era muy bueno — explicaba Hernán —, poco a poco fue quedando en desuso, ya que no había tiempo para replicar la información en dicho sistema experto.

Ambos especialistas destacaron el valor que podría agregar este tipo de tecnologías para mejorar la atención a los pacientes. La situación ideal planteada por ellos es que en el mismo sistema de registro clínico electrónico se incluyan alertas con respecto a factores de riesgo registrados en atenciones anteriores, y en caso de considerar necesario, poder acceder fácilmente a dichos registros puntuales, para evaluar la situación y tomar mejores decisiones.

Justificación

El presente trabajo se enfocará en las enfermedades cardiovasculares (ECV), que de acuerdo a datos de la Organización Mundial de la Salud (OMS) constituyen una de las causas más importantes de discapacidad y muerte prematura en todo el mundo, incluyendo Chile.

El tratamiento y control de estas enfermedades están incluidas en las garantías explícitas de salud (GES o AUGE), que es el plan de salud universal para todos los afiliados de FONASA y las ISAPRES que garantiza: acceso, oportunidad de atención, protección financiera y calidad[5], por lo que desde el punto de vista económico, la cobertura de estas enfermedades representa un costo importante para el Estado.

Además, existen varias enfermedades crónicas que aumentan el riesgo cardiovascular, que también son GES, como la hipertensión arterial, enfermedad cerebrovascular isquémica, la diabetes tipo 1 y tipo 2, dislipidemia, entre otras. Cabe mencionar que estas enfermedades están en los primeros lugares del ranking de impacto financiero de patologías GES. [10]

De acuerdo a todo lo mencionado, se justifica que el presente trabajo se enfoque en las enfermedades cardiovasculares, por una mejora en este ámbito tendría un alto impacto económico.

El foco estará en la detección temprana de riesgo cardiovascular, aportando así a la medicina

[3] Cardiólogo y Jefe del Servicio de Cardiología HCUCH
[4] Cardiólogo y Director del Departamento de Cardiología HCUCH
[5] Definición del Ministerio de Salud

preventiva más que a la curativa, lo que significa dirigir los esfuerzos en mejorar la salud de las personas y no enfocarse en las enfermedades. Cabe mencionar, que gran parte de los factores claves para reducir el riesgo cardiovascular, tienen que ver solamente con el cambio de hábitos, como abandono del tabaco, dieta alimenticia, actividad física y control de peso.

El dinero gastado en prevención y promoción (como por ejemplo, programas de prevención del tabaco, programas de inmunización, detección precoz de patologías graves, etc.) puede llegar a ser más costo-efectivo que la expansión de la oferta curativa.

Para lograr lo antes mencionado, es necesario contar con sistemas que permitan la detección temprana de riesgo cardiovascular y/o que permitan apoyar la gestión del médico durante la atención para poder aplicar medidas preventivas con tiempo suficiente, en caso de existir riesgo.

El presente trabajo de título propone consolidar la historia médica y detectar factores de riesgo cardiovascular a partir del Registro Clínico Electrónico (RCE) para aprovechar la gran cantidad de datos disponibles, no sólo hacer análisis descriptivo, sino también análisis predictivo.

1.2. Objetivos

A continuación se sintetizan los objetivos del trabajo:

1.2.1. Objetivo General

Generar un modelo predictivo, basado en Machine Learning (ML) y Natural Language Processing (NLP), que a partir de signos y síntomas detectados en los campos de texto no estructurados del Registro Clínico Electrónico, pueda predecir niveles altos de riesgo cardiovascular de una persona.

1.2.2. Objetivos Específicos

1. Estudiar y analizar el estado del arte de las técnicas de Text Mining, Machine Learning y Natural Language Processing.
2. Definir un metodología para la extracción y representación estructurada de la historia clínica de cada paciente.
3. Identificar signos y síntomas relacionados a enfermedades cardiovasculares usando las técnicas estudiadas en el estado del arte.
4. Evaluar el desempeño del modelo y validar que las variables a partir del texto no estructurado agregan valor al análisis.
5. Evaluar como agrega valor el modelo propuesto, por sobre el modelo actual de evaluación de riesgo cardiovascular (Score de Framingham).
6. Validar los resultados obtenidos con expertos en la materia.

1.3. Hipótesis de Investigación

Para enunciar la hipótesis se resumen las principales observaciones que motivan este trabajo de investigación:

- Muchos signos y síntomas que se consideran como factores de riesgo para calcular la función de riesgo cardiovascular de Framingham son registrados en los campos de texto libre como hipótesis diagnóstica, diagnóstico presuntivo o diagnóstico sindromático en la historia clínica del paciente, independiente del tipo de atención, pero esta información no necesariamente es sintetizada pensando en evaluar el riesgo cardiovascular del paciente.
- Aunque la evaluación de riesgo coronario está incluida en los Exámenes de Medicina Preventiva, subvencionados por el Estado, es baja la tasa de personas que se realizan estos exámenes.

- Las personas asisten a un centro de salud cuando hay un problema, no para exámenes de medicina preventiva.

Con el propósito de lograr detectar de forma temprana el riesgo cardiovascular, usando los campos de texto del registro clínico electrónico, aplicando técnicas de Text Mining y Machine Learning, se postula la siguiente hipótesis de investigación:

"***Existe información valiosa en los campos de texto no estructurado del registro clínico electrónico que permite agregar valor a la detección temprana de Riesgo Cardiovascular***".

1.4. Metodología

El trabajo comienza con una profunda revisión bibliográfica del estado del arte de las técnicas de Text Mining y herramientas de análisis relacionadas con Machine Learning.

Luego, se utiliza como guía las etapas de la metodología CRISP-DM[6]:

- **Entendimiento del Negocio**: Se realiza una breve introducción sobre la salud pública en Chile y se entrega un contexto médico sobre las enfermedades cardiovasculares, que son el tema principal de este trabajo.
- **Entendimiento de los Datos**: Se presenta el sistema de Registro Clínico Electrónico que permite capturar los datos transaccionales de una atención en un establecimiento de salud. Además, se detalla su estructura de datos relacionada con la información que se utiliza en este trabajo.
- **Preparación de los Datos**: A partir de los datos transaccionales, se realiza un preprocesamiento de los datos para obtener variables a partir de campos estructurados y no estructurados del registro clínico electrónico.
- **Modelamiento**: Se divide en dos partes el modelamiento de los datos. Primero se realiza un análisis de contenido, que tiene como objetivo principal validar la calidad y la consistencia de los datos utilizados. Segundo, se realiza un análisis predictivo, que es el tema central de este trabajo.
- **Evaluación de Resultados**: Se trabaja con dos conjuntos de variables, uno a partir de los campos de texto no estructurado (Texto) y otro a partir de parámetros estructurados del registro clínico electrónico (No Texto). En primera instancia, se valida que el uso de las

[6]Cross Industry Standard Process for Data Mining

variables "Texto.agregan valor al análisis predictivo, utilizando las medidas de AUC, Accuracy, F-Measure, Precisión, Sensibilidad y Especificidad. Luego, se compara para cada persona que sufrió un IAM[7] el resultado obtenido con el modelo propuesto y la clasificación actual de riesgo cardiovascular (Score de Framingham), y se evalúa si el modelo permite detectar riesgo alto en las personas que no tenían una clasificación previa de riesgo cardiovascular de Framingham o que estaban calificados con riesgo moderado o bajo antes de sufrir un IAM.

Finalmente, se espera validar la hipótesis de investigación presentada.

1.5. Resultados Esperados

Se declaran los resultados que permitan cumplir con los objetivos específicos definidos en el presente trabajo:

- Construir un marco de trabajo conceptual con la revisión de las técnicas más relevantes del estado del arte relacionado con Text Mining.
- Definir una metodología para estructurar y consolidar las historias clínicas de los pacientes.
- Un documento que contenga una representación estructurada de los pacientes, donde incluya los signos y síntomas relacionados con las enfermedades cardiovasculares que aparecen en los campos de texto no estructurado de su historia clínica.
- Resumen, a nivel general, de los resultados obtenidos del análisis de contenido y del modelo predictivo propuesto.
- Resumen, en detalle, sobre el resultado del modelo predictivo en cada uno de los subgrupos de personas que ya tenían una calificación de riesgo de Framingham (especialmente en los clasificados con riesgo moderado y bajo) y las personas que al momento de registrar que sufrieron un IAM no tenían registrado ninguna clasificación de riesgo cardiovascular.

1.6. Contribución de la Memoria

A partir de los resultados obtenidos en este trabajo, se valida el uso de los datos del Registro Clínico Electrónico que tiene la empresa SAYDEX. En particular los campos de texto no estructurados que nunca habían sido utilizados para ningún tipo de análisis.

[7] Infarto Agudo al Miocardio

El modelo presentado complementa la metodología actual de predicción de Riesgo Cardiovascular Global[8], detectando Alto Riesgo en personas que no tenían registrado ningún tipo de clasificación de riesgo cardiovascular, lo que permitiría priorizar al momento de ir a buscar a personas para integrarlas al Programa de Salud Cardiovascular. Por otra parte, es posible poner énfasis en las personas con mayor probabilidad de sufrir un infarto dentro del grupo que actualmente están calificados con riesgo bajo o moderado de acuerdo a la actual clasificación de Riesgo Cardiovascular Global.

Este trabajo contribuye a los esfuerzos del Gobierno enfocados en potenciar la salud preventiva en la Atención Primaria (APS). Ya que está demostrado que el gasto realizado en APS es más costo-eficiente que el gasto en atención secundaria y terciaria, sobre cuando tiene relación a enfermedades que afectan a una gran porción de la población, como lo son las enfermedades cardiovasculares.

Por otro lado, cabe mencionar que existen investigaciones en esta misma línea [11, 12], pero que se limitan al idioma inglés. Por medio de la empresa SAYDEX, se estableció el contacto con representantes de IBM T. J. Watson Research Center en América Latina para realizar un trabajo en conjunto, quienes presentan uno de los mayores avances en análisis de texto no estructurado en la industria de la salud, pero en inglés. Este trabajo permitirá validar la calidad de los datos de texto almacenados por SAYDEX, y así lograr ser un aporte a trabajos futuros usando los datos de texto no estructurado en español usando la tecnología de IBM Watson.

El presente trabajo contribuirá a dar el punta pié inicial a investigaciones y desarrollos de productos que se enmarcan en lo que actualmente se conoce como Healthcare Analytics, en la Salud Pública en Chile.

1.7. Estructura del Contenido

El resto del documento tiene la siguiente estructura. El Capítulo 2 pressenta el Marco Teórico que entrega todas las definiciones para entender los conceptos y las técnicas utilizadas para detectar de forma temprana (*Data Mining y Machine Learning*) el riesgo cardiovascular (*Programa de Salud Cardiovascular*) analizando los campos de texto no estructurado (*Text Mining y Natural Language Processing*) de la historia clínica del paciente (*Registro Clínico Electrónico*).

Considerando la metodología de trabajo de datos CRISP-DM, la primera parte consiste en el *Entendimiento del Negocio*, por lo que se profundiza y da detalles sobre los conceptos asociados a signos y síntomas de las enfermedades cardiovasculares, que tienen mucha relación con el área de la medicina, mas no de la ingeniería. Además, se hace referencia a las estrategias de salud preventiva que utiliza la Salud Pública en Chile, en particular el Programa de Salud Cardiovascular (PSCV).

[8] Se basa en el Score de Framingham ajustado a la población chilena y considera otras variables que suman puntos al valor inicial.

Con respecto a la etapa de *Entendimiento de los datos*, se entrega información sobre su estructura y cómo estos son almacenados con el uso del registro clínico electrónico. Todo lo antes mencionado se presenta en el Capítulo 3.

En el Capítulo 4 se detalla la propuesta de investigación, el diseño la solución propuesta y los resultados esperados en cada una de las etapas planteadas.

En el Capítulo 5 se presentan las etapas restantes de la metodología CRISP-DM, que consiste en la *Preparación de los Datos* utilizados el experimento. También se presenta el *Modelamiento* hecho para analizar los datos. Finalmente se realiza la *Evaluación de los Resultados* obtenidos con su respectivo análisis. También se agrega en este capítulo una sección con discusiones y propuestas de trabajos futuros.

El Capítulo 6 presentan las conclusiones del trabajo.

Capítulo 2

Marco Conceptual

El Registro Clínico Electrónico (RCE) es sólo el primer paso en la captura y explotación de datos relacionados con la salud. Los principales desafíos están en convertir estos datos en información útil. El uso las técnicas de minería de datos y la aplicación de modelos predictivos pueden ser utilizados para generar predicciones sobre diagnósticos, riesgos, o extraer información desde datos no estructurados para evaluar los resultados de tratamientos.

En este capítulo se detalla una metodología para realizar análisis de datos (*CRIPS-DM*). Además, se presenta una recopilación de los conceptos y técnicas utilizadas para calibrar un modelo predictivo (*Data Mining y Machine Learning*) usando los campos de texto no estructurados (*Text Mining y Natural Language Processing*) del *Registro Clínico Electrónico* para detectar de forma temprana *Riesgo Cardiovascular*.

Al final del capítulo, se entrega un detalle del estado de arte con respecto a las últimas investigaciones en relación al tema.

2.1. CRISP-DM

Una de las problemáticas del análisis de datos es que si no se tiene claro el problema que se quiere resolver, se pueden obtener resultados que no tienen mucha relación con el contexto de estudio o que no agrega ningún valor a la situación actual.

Para evitar este problema, existe una metodología que considera el entendimiento del negocio como primer paso, para evitar realizar un estudios fuera de un contexto general. Esta metodología se conoce como CRoss-Industry Standard Process for Data Mining (CRISP-DM) [13], y se considera como un tipo de implementación del proceso KDD[1], que consiste en un ciclo de 6 etapas:

1. **Entendimiento del Negocio**: Es la fase inicial que se enfoca en entender los objetivos del proyecto y los requerimientos, considerando una perspectiva de negocio. Luego, se debe

[1]Knowledge Discovery in Databases

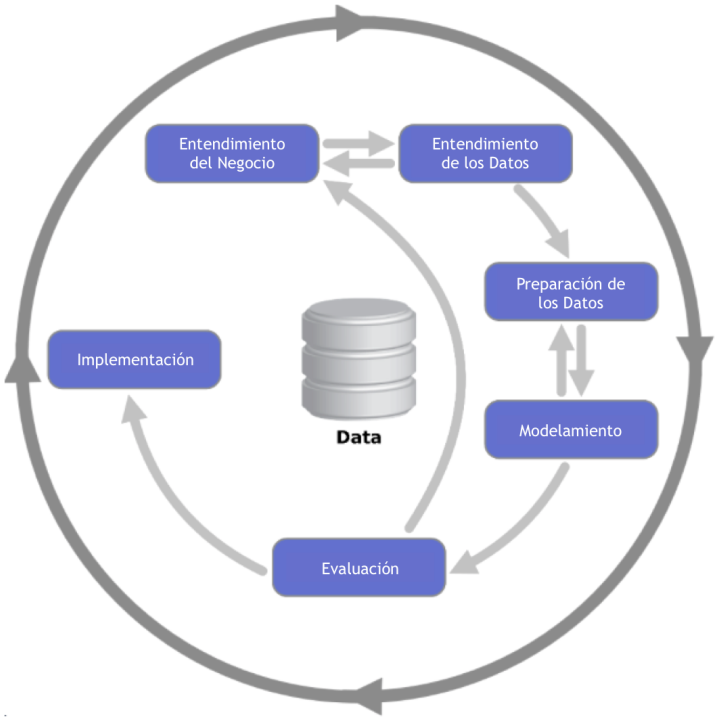


Figura 2.1: Esquema Cíclico CRISP-DM

definir los problemas a resolver y los resultados esperados basados en alcanzar los objetivos del proyecto, y no al revés, obtener resultados y luego pensar donde pueden ser útiles.

2. **Entendimiento de los Datos**: Es una fase muy importante al momento de obtener la información necesaria para el estudio, ya que permite entender el estado de los datos y si existen problemas de calidad en los registros. Este paso logra hacer el nexo entre los objetivos definidos y el cómo obtener los datos necesarios para responder a las preguntas plateadas inicialmente.

3. **Preparación de los Datos**: Consiste en todas las actividades necesarias para construir el conjunto de datos que se usarán en el estudio a partir de los registros transaccionales, como la selección, limpieza, transformación, integración y dar el formato final que será utilizado por el modelo de datos.

4. **Modelamiento**: Esta etapa es bien amplia, ya que es posible usar variadas técnicas para modelar los datos. Finalmente, y luego de iterar con la etapa anterior, se busca obtener una forma estructurada de los datos, para lograr obtener información a partir de ellos.

5. **Evaluación**: Como lo dice su nombre, se evalúan los resultados obtenidos usando todo lo antes mencionado y se comprueba que permiten alcanzar los objetivos planteados inicialmente. Teniendo en consideración esto, es posible definir los próximos pasos para una eventual

implementación de una solución, o en caso contrario, volver a la etapa inicial y redefinir los objetivos del estudio en base a lo aprendido en el proceso.

6. **Implementación**: Si los resultados obtenidos fueron validados y responden a los objetivos del estudio, se comienza con la etapa de elaborar un proyecto para llevar a cabo la implementación.

Es importante destacar que la metodología CRISP-DM es cíclica, como se muestra en la figura 2.1, y que permite pasar por cada una de las etapas más de una vez durante la ejecución del proyecto.

2.2. Data Mining y Machine Learning

Data Mining

De acuerdo a Gronescu [14], Data Mining tiene 3 ejes principales: Estadística, Inteligencia Artificial (IA, incluyendo Machine Learning) y Bases de Datos. Aunque estos tres ejes están bien especificados, es complejo dar una definición única para Data Mining. Sin embargo, la definición más usada, es probablemente la declarada en [15], donde se menciona que Data Mining "*es el proceso no trivial de extracción de información desde los datos que está presente de forma implícita, previamente desconocida y potencialmente útil para el usuario*". Esta información está presente en los datos como patrones que son muy útiles cuando se aplican para resolver problemas en un determinado contexto.

Machine Learning

Es una forma de Inteligencia Artificial, más específicamente, Machine Learning (ML) es una disciplina que consiste en el desarrollo de algoritmos que utilizan datos empíricos para llevar a cabo dos tareas:

1. Identificar o cuantificar relaciones complejas (patrones) a través de las características propias de los datos.

2. Emplear estos patrones para hacer predicciones.

En los datos es posible encontrar relaciones entre las variables observadas a través de algoritmos, los que vendrían a ser como una máquina que aprende a partir de una muestra de datos (entrenamiento) para capturar características no observadas que son relevantes a través de distribuciones de probabilidad subyacentes. Luego, es posible usar este conocimiento aprendido, para realizar decisiones más inteligentes usando datos nuevos [16].

En general, los algoritmos de Machine Learning pueden ser clasificados en muchas categorías dependiendo de los resultados que se busca obtener. Una de estas clasificaciones es:

- Aprendizaje Supervisado: Estos son algoritmos que generan una función matemática para relacionar los datos con un conjunto de salidas conocido. A este conjunto de salidas se les llama etiquetas y generalmente son el resultado de un proceso de etiquetado humano.

- Aprendizaje No Supervisado: A diferencia de lo anterior, estos algoritmos simplemente buscan modelar los datos de entradas.

2.2.1. Problema de Alta Dimensionalidad

Cuando se considera una gran cantidad de variables para analizar, existe el problema de la alta dimensionalidad de los datos. Para poder trabajar bajo estas condiciones, hay una amplia variedad de métodos que pueden clasificarse en tres categorías no excluyentes entre sí, ya que son formas complementarias que abordar el problema y, en muchos casos, es necesario recurrir a más de una estrategia.

Por ejemplo, utilizar un método de selección como paso previo a la utilización de otro tipo de método puede hacer aumentar la potencia de este último. A continuación se detallan las características generales de cada una de estas estrategias:

1. Método de reducción de la dimensión: Consiste en realizar transformaciones de los datos que permitan reducir la dimensión del espacio de valores a considerar, obteniendo un conjunto de nuevos datos, de tamaño o dimensión menor, que concentre o resuma la información relevante del conjunto de datos original. Puede ser una estrategia muy útil en conjuntos de datos con información redundante. Uno de ellos es el análisis de componentes principales (PCA por sus siglas en inglés).

2. Método de selección de variables: Consiste en seleccionar un subconjunto de las variables más relevantes, para lo que existen métodos de una variable y multivariable. En el primer caso, se establece un ranking según un criterio de asociación de cada variable con la respuesta (o etiqueta), finalmente se seleccionan las mejores. El problema de estos métodos es que no tienen en cuenta las correlaciones ni las interacciones entre variables, por lo que este método de selección no garantiza obtener el mejor subconjunto. Por otra parte, los métodos multivariables se basan en determinar el subgrupo óptimo considerando las posibles correlaciones entre ellas.

   Los métodos de selección de variables se caracterizan por dos aspectos: el algoritmo de búsqueda en el espacio de posibles subconjuntos y el criterio de evaluación de cada subconjunto. Según estas características hay tres tipos de métodos:

- Filter: No incluyen aprendizaje, es decir, no se basan en ningún modelo de clasificación o regresión, sino que utilizan criterios de evaluación independientes.
- Wrapper: Incluyen aprendizaje en la selección de variables usando como criterio de evaluación las estimaciones del error de clasificación (o predicción) basadas en un determinado modelo (o regresión), pero sin incluir conocimiento de la estructura del modelo en el criterio de evaluación. Un ejemplo de este tipo de modelos es la eliminación recursiva de variables.
- Embedded: La parte de aprendizaje y de selección son inseparables. Por ejemplo, en los modelos de clasificación basados en árboles, la selección de las variables forma parte del propio modelo de clasificación.

3. Métodos capaces de analizar un gran número de variables directamente, consiste en utilizar métodos que sean capaces de analizar directamente conjuntos de datos en los que el número de variables supera el de observaciones.

2.2.2. Modelos de Clasificación

Es una metodología para identificar combinaciones lógicas de las atributos que mejor predicen la variable respuesta según un determinado modelo de regresión. Este método explora modelos de regresión del tipo:

$$Y = \beta_0 + \sum_{j=1}^{t} \beta_j * I_{\{L_j es\_cierta\}} \tag{2.1}$$

Donde Y es la variable respuesta, y L_j es una expresión lógica, por ejemplo, "Está presente la palabra colesterol"

El objetivo del método es encontrar el mejor modelo de regresión, es decir, determinar los L_j y las β_j que permiten una mayor capacidad predictiva. Para ello se utiliza un algoritmo de exploración que permite no tener que comprobar todas las posibles expresiones lógicas, puesto que el espacio de combinaciones lógicas posibles crece de forma exponencial a medida que aumenta el número de variables.

En forma general, se define un tipo de modelo usado para abordar el problema de clasificación en dos grupos. Es decir, si se dispone de n elementos del tipo (Y_i, X_i), donde Y_i es igual a 1 o 0 según la clase a la que pertenece, y X_j es el vector de variables explicativas.

Modelo Logit

Para que el modelo entregue directamente la probabilidad de pertenecer a una u otra clase, es necesario trasformar la variable de respuesta del modelo de regresión en valores entre 0 y 1.

$$p_i = F(\beta_0 + \sum_{j=1}^{t} \beta_j^{'} * x_i) \tag{2.2}$$

Habitualmente se toma F como la función de distribución logística

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{t} \beta_j^{'} * x_i)}} \tag{2.3}$$

Se cumple que:

$$g_i = log \frac{p_i}{1 - p_i} = log \left(\frac{1}{e^{-(\beta_0 + \sum_{j=1}^{t} \beta_j^{'} * x_i)}} \right) = \beta_0 + \sum_{j=1}^{t} \beta_j^{'} * x_i \tag{2.4}$$

Al hacer esta transformación se obtiene el modelo lineal que se denomina *logit*. La variable g representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a una de las dos clases, y al ser una función lineal de las variables explicativas, facilita la estimación y la interpretación del modelo. [17]

2.2.3. Evaluación de Algoritmos Supervisados

A continuación se presentan las medidas que se utilizan en el presente trabajo para evaluar el desempeño de los algoritmos utilizados para los análisis predictivos. En este caso en particular, se trabaja con modelos de clasificación binaria, es decir, que se trabaja sólo con dos clases.

Precision, Recall, F-Measure y Accuracy

Para cualquier problema de clasificación, existe un grupo de medidas que permiten evaluar el grado de cercanía de la clase predicha con el valor real. En particular, se utilizan los conceptos de Verdadero Positivo (VP), Verdadero Negativo (VN), Falso Positivo (FP), Falso Negativo (FN) que comparan los resultados del clasificador. Estos términos se muestran en lo que se conoce como Matriz de Confusión (Ver Cuadro 2.1).

En esta matriz (también conocida como matriz de contingencia) los términos positivo y negativo corresponden al resultado esperado del clasificador, y los términos verdadero y falso corresponden a evaluar si la predicción coincide con el valor real. Considerando estas definiciones, se tiene el siguiente grupo de medidas[2] para evaluar el desempeño del clasificador:

[2]Se mantienen sus nombres en inglés, al no tener una traducción directa al español

	Real Positivo	Real Negativo
Clasificado como Positivo	Verdadero Positivo (VP)	Falso Positivo (FP)
Clasificado como Negativo	Falso Negativo (FN)	Verdadero Negativo (VN)

Cuadro 2.1: Matriz de Confusión: Resultados Clasificación vs Valores Reales para una clase.

Fuente: Elaboración Propia

$$Accuracy = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.5}$$

$$Precision = \frac{VP}{VP + FP} \tag{2.6}$$

$$Recall = \frac{VP}{VP + FN} \tag{2.7}$$

$$F(\beta) = (1 + \beta^2)\frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall} \tag{2.8}$$

$$F(1) = F - Measure = 2\frac{Precision \times Recall}{Precision + Recall} \tag{2.9}$$

En el problema de clasificación se tiene, entonces:

- Precision: Es el número de elementos que fueron correctamente clasificados como clase positiva, dividido por el número total de elementos clasificados como positivos por el algoritmo. En otras palabras, la Precision se compara con el total de elementos Predichos en esa clase.

- Recall: Es el número de elementos que fueron correctamente clasificados, dividido por el número total de elementos etiquetados que Realmente pertenecen a la clase positiva. En otras palabras el Recall se compara con el total de elementos Reales en esa clase.

  Dada estas definiciones, la Precision no dice nada sobre el número de elementos que no fueron clasificados correctamente como positivos (no considera los FN), ni el Recall dice nada sobre cuantos elementos fueron clasificados incorrectamente como positivos (no considera los FP). Por esta razón, la Precision y el Recall no deben ser usadas de forma separada. Existe una medida que combina ambas medidas, F-Measure.

- F-Measure: Corresponde a la media armónica entre Precision y Recall, que deriva de la ecuación 2.8, cuando β es uno. β corresponde a la importancia que se da al Recall por sobre la Precision.

Para complementar las descripciones anteriores, se tiene el siguiente ejemplo: dado un total de 100 pacientes, de ellos 70 no tuvieron un infarto agudo al miocardio (IAM) y 30 sí. Sólo se analizará los conceptos para una clase, los que sí sufrieron un IAM. Para la otra clase, el análisis es análogo.

- Precision: Si el modelo predice 20 pacientes como IAM, y de estos 20 sólo 15 realmente son IAM, entonces se tiene una Precision de $15/20 = 75\,\%$
- Recall: De los 15 pacientes que el modelo predice correctamente como IAM, si hay un total de 30 casos reales de IAM, entonces se tiene un Recall de $15/30 = 50\,\%$

Se pude concluir de estos indicadores que el modelo cuando predice un caso de IAM, un $75\,\%$ de las veces es verdad. Sin embargo, el bajo porcentaje de Recall quiere decir que solamente está capturando el $50\,\%$ de los casos.

Es por esto, que no se recomienda mirar estas medidas de rendimiento por separado. Para tener una mejor referencia, se calcula la media armónica de ambas medidas, lo que se conoce como $F - Measure$.

Además de problemas de clasificación, este grupo de medida también es usado en el contexto de los sistemas de Recuperación de Información. Campo de investigación que estudia y propone soluciones para facilitar el acceso a la información, seleccionando y clasificando los recursos de información más pertinentes a las necesidades de los usuarios. En este caso, un alto Recall significa que un algoritmo entrega un alto porcentaje del total de los elementos que eran necesario que fueran encontrados. Por otro lado, una alta Precision significa que el algoritmo obtuvo muy pocos resultados no deseados en la búsqueda de información.

En el contexto del problema de clasificación, la aplicación de Recall, Precision y F-Measure son cuestionadas porque no consideran la celda Verdadero Negativo de la Matriz de Confusión, ya que no aparece en ninguna de las ecuaciones. Este problema, se resuelve considerando el Accuracy (Ver ecuación 2.5). Cabe mencionar, que si se llega a considerar sólo este indicador se puede generar lo que se conoce como la paradoja del Accuracy, que consiste en que dado un modelo predictivo con un nivel más bajo de esta métrica, puede tener un mayor poder de predicción que un modelo con un Accuracy más alto. Por ejemplo, se puede dar el caso cuando la cantidad de Verdaderos Positivos es menor a la de Falsos Positivos para una determinado modelo, se puede obtener un Accuracy mayor cuando se reemplaza este modelo por otro (modelo dummy) que prediga todos los elementos como negativos. [18].

Cabe mencionar, cuando la clase más pequeña es la que importa, se sugiere usar la medida

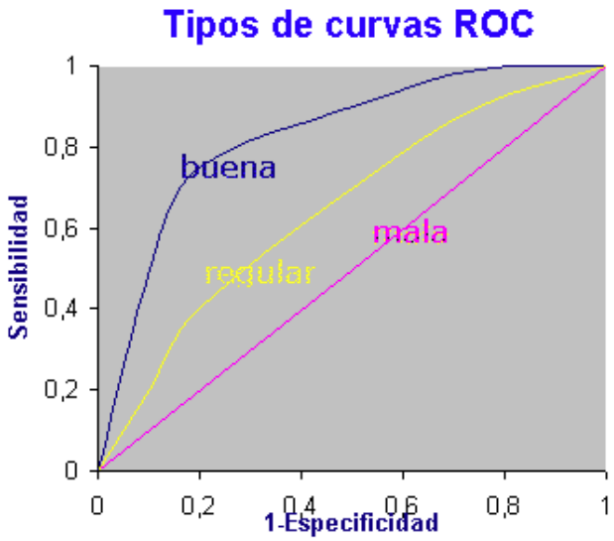


Figura 2.2: Tipos de curvas ROC

F-Measure, en lugar del Accuracy.

Otro punto importante a considerar es cuando los errores tienen diferentes costos en su aplicación real, como por ejemplo en problemas de detección de fraude, default crediticio o diagnósticos médicos; donde no es lo mismo un Falso Positivo que un Falso Negativo. Aunque desde el punto de vista de la predicción ambos son errores, en la aplicación real los Falsos Negativos son mucho más costosos. Para el caso de detección de fraude, cuando la clase positiva es fraude, un Falso Negativo es aceptar un caso de fraude. En el caso de default financiero, cuando la clase positiva representa a un cliente con alto riesgo de incumplimiento, un Falso Negativo es aceptar un mal cliente. Por último, en el caso de predecir una condición médica, si la clase positiva es "Ud. tiene riesgo cardiovascular", un Falso Negativo es dejar de diagnosticar a un persona con riesgo cardiovascular, lo que podría implicar en una descompensación del paciente por no haber sido tratado y su eventual muerte. Para modelar esta diferencia, y poder darle más importancia a la clase más costosa, se define el concepto de umbral o *threshold*. Si se utiliza una regresión logística, el resultado del modelo es un valor entre 0 y 1, y de acuerdo al umbral definido, cada valor se aproxima a 0 ó 1. Si se quiere evitar los Falsos Negativos, el umbral debe ser menor que 0,5.

Una medida que permite evaluar el desempeño del modelo para diferentes valor de *threshold* es el Área bajo la Curva ROC (AUC).

Área bajo la Curva (AUC)

Una medida de predicción global que sólo mida el error en la predicción no proporciona un buen método de evaluación. Se recomienda evaluar la capacidad predictiva en casos y controles por separado, por lo que es importante considerar dos nuevos conceptos, la sensibilidad y especificidad.

En el contexto de los modelos de predicción de condiciones médicas, se entiende por sensibilidad la proporción de personas con riesgo bien clasificadas (VP ó Recall de la Clase Positiva), y la especificidad como la proporción de sanos bien clasificados (VN ó Recall de la Clase Negativa).

Las curvas ROC (Receiver Operating Characteristic) son representaciones gráficas de la *sensibilidad* versus $1 - especificidad$, es decir, la proporción de verdaderos positivos (VP) versus la proporción de falsos positivos (FP). Cabe mencionar que cada punto en la curva ROC se determina para un valor dado del *threshold*.

Este método de evaluación de modelos de clasificación tiene dos utilidades muy importantes:

1. Permite determinar el punto de equilibrio óptimo entre sensibilidad y especificidad,

2. Da una medida de la capacidad predictiva mediante el cálculo del área bajo la curva.

El área bajo la curva ROC permite evaluar la capacidad del método para clasificar correctamente y puede entenderse como una probabilidad, donde el valor 1 significa que el método es perfecto, un valor de 0.5 indica que el método es inútil y valores intermedios miden la capacidad del método para discriminar entre casos y controles, que será mayor cuanto más se aproxime a 1.

2.2.4. Itemset Mining

El análisis de conjunto de elementos (en inglés, Itemset), consiste en encontrar patrones estadísticamente relevantes en un conjunto de datos que están representados de forma secuencial, de acuerdo a la frecuencia con que aparece este tipo de datos en escenarios de aplicaciones reales. Esta técnica constituye uno de los métodos más populares de descubrimiento de patrones.

Los patrones obtenidos, se usan en tareas de detección de dependencias funcionales, predicción de tendencias, interpretación de fenómenos y como soporte de decisiones en estrategias de producción.

Clasificación

El los problemas de clasificación toma gran importancia cuando se consideran los criterios de selección adecuados. La idea es que las transacciones o secuencias de datos con clasificaciones similares estén estrechamente relacionadas.

Esta técnica se aplica en muchos escenarios al momento de tener que reconocer patrones, donde se necesitan realizar predicciones de comportamientos basándose en registros de datos históricos.

Agrupamiento de Patrones

El objetivo es separar en grupos a las secuencias de datos, de manera que las pertenecientes a un mismo grupo sean muy similares entre sí, y al mismo tiempo sean diferentes a las de otros grupos.

Un ejemplo típico de aplicación de esta técnica, es en las transacciones comerciales donde sirve para identificar diferentes grupos de clientes con registros de compra similares, pero en el contexto del presente trabajo, para identificar grupos de pacientes con síntomas y diagnósticos similares.

Regla de Asociación

Se basa principalmente en el descubrimiento de patrones secuenciales frecuentes. Un ejemplo de uso de esta técnica es la representación de operaciones de distribución y marketing, que a partir de las secuencias y asociaciones obtenidas, ayudarían a identificar los productos de mayor promoción de acuerdo con los patrones de compra. Por ejemplo, si una pareja adquiere una nueva vivienda, en el 70 % de los casos se comprarán una cama en menos de un mes.

En el contexto del sector de la salud, se usaría para representar patrones de atención médica, como las trayectorias de los pacientes en los centros de salud, los estados evolutivos de los pacientes, los comportamientos de los síntomas, entre otros. Y así descubrir patrones en las historias de los registros médicos y mejorar el nivel de diagnóstico.

Las medidas más usadas son:

- **Support**: Se define como la proporción del total de las transacciones que contienen un determinado "itemset"(productos, signos o síntomas).

$$sup(A)$$

- **Confidence**: Es la estimación de $Pr(C|A)$, es decir, la probabilidad de observar C dado A. En el caso de reglas con el mismo valor de Confidence, es preferible la que tenga mayor Support.

$$conf(A \to C) = \frac{sup(A \bigcup C)}{sup(A)}$$

- **Lift**: Permite medir la independencia entre A y C. En este caso, lift es un valor entre $[0, +\infty[$. Valores cercanos a 1 implica que A y C son independientes, luego esta regla no es interesante. Reglas con valores mucho mayores que 1 permiten definir cuanta información entrega A sobre C. Cabe destacar que Lift mide solo co-ocurrencia (no implicancia), por lo que es una regla simétrica.

$$lift(A \to C) = \frac{conf(A \to C)}{sup(C)}$$

- **Conviction**: Permite complementar el análisis observando Confidence y Lift. Cabe destacar que Conviction se puede interpretar como la definición en lógica de la implicancia. Sus valores también se encuentran de $[0, +\infty[$. Las reglas cuyo Conviction es mucho mayor que ,

corresponden a relaciones de interés.

$$conv(A \rightarrow C) = \frac{1 - sup(C)}{1 - conf(A \rightarrow C)}$$

2.3. Text Mining y Natural Language Processing

Text Mining fue propuesto por primera vez por Ronen Feldman et al en 1995, lo que fue descrito como "El proceso de extraer patrones de interés desde una gran cantidad de texto con el propósito de obtener conocimiento". [6]

Text Mining fue también conocido como Text Data Mining, que consistía en el proceso donde conocimiento nuevo y entendible era encontrado a partir de una gran cantidad de texto. [3,6,19,20]

Una definición más clara es el proceso de analizar texto, extrayendo información y encontrando conocimiento. [19,21,22]

Las principales técnicas en Text Mining consiste en clasificación, segmentación y consolidación del texto; análisis de correlación, extracción de información, análisis de distribución, detección de tendencias, extracción de entidades/aspectos, producción de taxonomías, análisis de sentimiento, entre otros análisis. [23–25]

Para el caso de la clasificación y segmentación, esto métodos trabajan sobre un conjunto de textos. Por otro lado, la consolidación y la extracción de información trabaja se realiza sobre un documento.

En el caso particular de métodos de clasificación de texto, entre los más utilizados se encuentra Naïve Bayes, K-Nearest Neighbor, Support Vector Machines (SVM), entre otros. [6,7,9,19,22–25]

Además de métodos estadísticos y los algoritmos antes nombrados, la aplicación del procesamiento de lenguaje natural (NLP, por su sigla en inglés) es una de las herramientas más utilizadas en Text Mining. Como se menciona en [26] existen muchas definiciones para Natural Language Processing, y probablemente ninguna de estas definiciones puede satisfacer a todos. Sin embargo, como es necesario establecer una referencia, se declara que el objetivo de NLP es llevar a cabo el procesamiento de texto como lo haría un humano, es por esto que se utiliza en el área de la interacción humano-computador.

Natural Language Processing está en el campo de las ciencias de la computación, inteligencia artificial y lingüística, y procura mejorar la interacción entre computadores y el lenguaje (natural) de las personas. NLP está basado en el uso de diferentes teorías y tecnologías que en hoy en día es un área de investigación científica muy activa.

Existen tres procesamientos de texto que son necesarios para apoyar que un computador logre entender las principales ideas expresadas por una persona en un texto. Estos procesamientos son:

Tokenization, Stopwords, Stemming / Lemmatization.

Tokenization

Tokenization es el proceso de separar una cadena de texto en pequeñas partes o componentes (token), los que pueden representar elementos del texto mucho más representativos que el texto no estructurado, como palabras, frases, como también pueden haber token que no entreguen ninguna información relevante.

En general, Tokenization es aplicada previo a un análisis más complejo, donde esta lista de tokens que se obtiene es usada como elemento de entrada para otros procesamientos como NLP o Text Mining.

Aunque Tokenization es muy útil en diferentes áreas de estudio, el procedimiento siempre es el mismo. Consiste en tomar un cadena de texto, y aplicar un conjunto de reglas para finalmente generar un conjunto de tokens. Cabe mencionar, que la reglas de segmentación depende del contexto donde se esté trabajando y las necesidades particulares del estudio. Probablemente dos de los técnicas más relevantes donde se usa tokenization, están basadas en Machine Learning y en el uso de expresiones regulares, que necesitan una forma flexible de reconocer cadenas de texto.

Uno de los problemas de este procedimiento, es cuando se necesita que los tokens representen oraciones. El problema está en definir cuales son los límites de una oración. Por lo general, el principal criterio es considerar los signos de puntuación como límite de cada oración, pero el problema es cuando se usan puntuaciones para abreviaciones, o cuando otro tipo de caracteres especiales representan el final de una oración. Este problema es abordado en [27]. Cabe mencionar, que en el presente trabajo se consideraran los token como palabras.

Stopwords

En ocasiones, hay palabras muy comunes que tienen muy poco valor, es decir, entregan muy poca información semántica por si solas, por lo que es recomendable eliminarlas para el análisis. Estas palabras son llamadas Stopwords. [28]

Hay varias formas para determinar cuando una palabra es stopwords. Es muy útil tener una lista de estas palabras, esta lista es conocida como stoplist. No existe una única stoplist, sino que esta lista debe ser construida en cada caso dependiendo de los términos más frecuentes que aparecen en los documentos que se están procesando, y también de acuerdo al contexto y contenido semántico relativo a los documentos. Además, siempre es posible encontrar una lista de stopwords comunes, que no dependen necesariamente del contexto en el que se está trabajando. En general, estas listas están compuestas por palabras que no tienen un significado propio, y que siempre aparecen en

los documentos acompañando a otras palabras o como conectores, tales como los artículos, los pronombres y preposiciones.

La utilidad de tener una lista de stopwords, permite reducir la dimensionalidad de los datos que se tienen que almacenar. El dejar de considerar estas palabras, quita muy poco del contenido semántico, pero permite mejorar la eficiencia al momento de trabajar grandes cantidades de texto.

Stemming y Lemmatization

Las palabras pueden ser presentadas de multiples formas en un documento, por ejemplo, las palabras: organizar, organiza, organizador, organización. Además, existen familias de palabras relacionadas, con significados muy parecidos. En ocasiones, pareciera útil que al momento de la búsqueda de una palabra, el resultado pudiese considerar todas estas palabras relacionadas o que pertenecen a la misma familia de palabras.

Considerando esto, el objetivo de Stemming y Lemmatization es trasformar las palabras y llevarlas a su raíz (stem) o a una base común (lemma) [28], así se reduce la cantidad de dimensiones al momento de realizar el análisis.

POS Tagging

Part-of-speech tagging, o POS tagging, es el proceso que permite identificar y marcar cada palabra en una cadena de texto con su correspondiente rol que juega dentro de una oración, basado tanto en su definición como en su relación con las palabras adyacentes en al oración.

Al comienzo, POS Tagging estaba relacionado directamente con la elaboración de un corpus lingüístico[3]. El primer proceso de POS Tagging fue realizado de forma manual durante la década del 60 (Brown Corpus) [29], que corresponde a uno de los más grandes corpus, para el idioma inglés, en contexto de análisis computacional. También se han desarrollado programas que automatizan este proceso.

Tanto los métodos supervisados como los no supervisados que han sido propuestos para llevar a cabo esta tarea, cabe mencionar que los métodos supervisados son los más ampliamente utilizados. En relación a estos métodos, existen dos tipos:

- Métodos Estocásticos: Tomando el trabajo de Brown Corpus como base, se han desarrollado muchos acercamientos estadísticos. Estas técnicas incluyen, por ejemplo, el uso de Modelos de Markov Ocultos (HMM), lo que consiste en contar casos y hacer una tabla de probabilidades para determinadas secuencias, y usando algoritmos de programación dinámica, intenta resolver este problema en menos tiempo.

[3] Un corpus lingüístico es un conjunto, habitualmente muy amplio, de ejemplos reales de uso de una lengua.

- Método basado en Reglas: Básicamente, es una técnica que fue propuesta por Eric Brill en su tesis de doctorado en 1993 [30]. Esta técnica aprende desde un conjunto de patrones y luego los aplica en lugar de optimizar una cantidad estadística.

El principal objetivo, es determinar el conjunto de TAGs, en otras palabras, el sistema que permitirá realizar este etiquetado de las palabras con su respectivo part-of-speech.

Existen dos conjuntos de TAGs que son ampliamente usados. Para el caso del inglés, se usa Penn tag set, desarrollado en Penn Treebank [31], como proyecto de la Universidad de Pennsylvania.

Por otro lado, se encuentra el conjunto de TAGs conocido como EAGLES (Expert Advisory Group on Language Engineering Standards), que es ampliamente usado, dado que incluye versiones para varios idiomas, como el español. En este trabajo, se usará el conjunto de TAGs EAGLES.

Ya se presentó una recopilación de conceptos y técnicas utilizadas para calibrar un modelo predictivo de clasificación y una síntesis de los principales conceptos de *Data Mining*. Además, se detalló como extraer información a partir de los campos de texto no estructurados con el uso de *Text Mining y Natural Language Processing*.

A continuación, se entrega un marco conceptual enfocado en la tecnología del *Registro Clínico Electrónico* y un contexto general sobre *Riesgo Cardiovascular* y como se ha utilizado la tecnología para extraer información relevante a partir de los campos de texto no estructurados para la detección temprana de riesgo.

Cabe mencionar, que un acercamiento más detallado sobre las enfermedades al corazón y la estructura de datos del RCE se presenta en el Capítulo 3.

2.4. Registro Clínico Electrónico

En Latinoamérica, Brasil y Chile están adelantados en la adopción del Registro Clínico Electrónico, en tanto España, se ha convertido en un importante referente para la región. En este último más de 18 millones de historiales médicos de ciudadanos de Andalucía, Valencia y Galicia, se gestionan con sistemas digitales. Algo similar ocurre en la ciudad de Brasilia en Brasil, donde se implementó el sistema para una red de 17 hospitales que atienden a 3 millones de habitantes. En este distrito, el sistema ha ayudado a reducir en 20 % las listas de espera, 25 % la duplicación de exámenes y en cerca de 20 % la eficiencia en el manejo del stock de fármacos. [32]

RCE en Chile

La estrategia SIDRA (Sistema de Información De la Red Asistencial) es una iniciativa impulsada por el Ministerio de Salud para la implementación de tecnología digital en todo el sector, con el objetivo de fortalecer el trabajo de la Red Asistencial, dando soporte a la gestión operacional en cada nivel enfocado en mejorar la atención integral de los usuarios del Sistema Público de Salud.

Las herramientas provistas por la empresa SAYDEX, actualmente en uso por los establecimientos de las Redes Asistenciales y por la autoridad sanitaria, son un importante apoyo a la estrategia impulsada por el MINSAL y permiten compartir información clínica en tiempo real de aproximadamente 9.000.000 de personas, que corresponde a casi el 75 % del total de las personas que se atienden en Salud Pública.

Los estudios y experiencia recientes en la incorporación de RCE ha mostrado que los tiempos administrativos de los profesionales de la salud disminuyen considerablemente, recuperándolos para la atención directa a los pacientes. Un estudio del Cesfam Juan Damianovic de Punta Arenas

calculó que los jefes de programa del establecimiento ocupaban 100 horas administrativas mensuales, que correspondía a un 54,1 % de una jornada completa de un profesional al mes. Así logró recuperar esas horas con la incorporación del RCE.

Por otro lado, se tiene la posibilidad de vincular la agenda profesional con la oferta asistencial, lo que ha logrado mejorar en todos los Centros de Salud que cuentan con un RCE la capacidad de optimizar la distribución de la oferta y evaluar la real utilización de cupos médicos.

La posibilidad de crear una agenda centralizada y compartida por todos los administrativos de un establecimiento ha demostrado una disminución del 70 % del Indicador de Pacientes que No se Presentan (NSP).

Pero lo más importante de esta iniciativa es el beneficio para el paciente y su familia: la Historia Clínica única Compartida (HCC) que "sigue al paciente", en todo lugar en que este registre una atención clínica, ya sea en un hospital de alta complejidad, mediana complejidad, Centros de Salud Familiar (CESFAM), Centro de Salud Comunitario Familiar (CESCOF), etc. Esto ha significado la disminución de riesgos como reacciones adversas a medicamentos, la mejora en la información clínica para la toma de decisiones y el fin del duplicado de recetas o exámenes.

2.5. Prevención del Riesgo Cardiovascular

Las acciones preventivas se dirigen a intervenir sobre la población en general (estrategia poblacional) y personas enfermas (estrategia de alto riesgo). Estas dos estrategias son complementarias entre sí. La mayoría de los eventos cardiovasculares en una población ocurre en personas con un nivel bajo o moderado, es decir, que pueden presentar sólo una elevación ligera en los factores de riesgo. Basado en este conocimiento es que las estrategias poblacionales o de salud pública se utilizan para trasladar la distribución poblacional de los factores de riesgo hacia niveles más favorables. Con estrategias poblacionales exitosas, pequeños cambios pueden resultar en mejoras significativas en las tasas de morbi-mortalidad cardiovascular.

Rol de la Atención Primaria de Salud

En respuesta a la transición epidemiológica acelerada observada en Chile que se caracteriza por un incremento absoluto y relativo de la población adulta y de los mayores de 65 años, y el aumento en la prevalencia de las enfermedades no transmisibles y sus factores de riesgo, se han desarrollado programas de prevención y control para estas patologías en la atención primaria de salud. Entre otros, se destaca el Programa Salud Cardiovascular (PSCV), con más de un millón y medio de personas con diabetes, hipertensión, dislipidemia o tabaquismo en control. El PSCV, a cargo de un

equipo de salud multidisciplinario (médico general, enfermera y nutricionista, entre otros), utiliza un enfoque terapéutico basado en el nivel de riesgo cardiovascular absoluto, cuyo objetivo es mejorar la eficiencia en el uso de los recursos y su efectividad. Las personas inscritas en el Programa tienen derecho a las garantías explícitas de salud asociadas a la hipertensión y la diabetes, más otras prestaciones o servicios adicionales vinculados a la organización para el diagnóstico, tratamiento y seguimiento de los pacientes en PSCV.

Anualmente y por Resolución, a los establecimientos en atención primaria de salud se les fijan metas sanitarias cuyo objetivo es promover la calidad y oportunidad de la atención. Es así como para el año 2012 se monitorean y evalúan metas de cobertura efectiva para diabetes e hipertensión arterial cuyo cumplimiento está asociado a incentivos financieros. [33]

2.6. Otros trabajos que utilizan el RCE

A continuación se presentan dos trabajos que desarrollan la problemática de la detección de riesgo cardiovascular con el uso de datos del Registro Clínico Electrónico.

2.6.1. Trabajo 1: Modelamiento Predicitivo usando datos de RCE

Este trabajo [34], realizado el año 2010, tenía por objetivo modelar la detección de enfermedades cardiacas, usando Machine Learning aplicado a los datos del registro clínico electrónico. Este trabajo comparó el desempeño de varios modelos, como Logistic Regression, SVM y Boosting, considerando diferentes métodos de selección de variables en cada caso. Cabe mencionar, que este trabajo no incluyó variables obtenidas del texto no estructurado del registro clínico electrónico.

Datos utilizados

Se consideró a personas entre 50 y 79 que entre el 1 de Enero de 2003 y el 31 de Diciembre de 2006 fueron diagnosticados con una enfermedad cardiovascular y sus registros se encuentran en el sistema de la Clínica Geisinger en Estados Unidos. Se recopilaron los registros de sus atenciones hasta 2 años antes del diagnóstico de la enfermedad cardiovascular. Se obtuvo para el estudio un total de 536 casos.

Clínica Geisinger ha utilizado EpicCare EHR[4] desde 2001 y ha sido atendida por una sola compañía de laboratorios desde 1993. Los datos de pacientes de su registro clínico electrónico considera: edad, sexo, altura, peso y otros datos demográficos, estilo de vida (por ejemplo: fumador y/o alcohol), medidas clínicas (por ejemplo: la presión arterial, la función pulmonar y la densidad mineral ósea), imágenes digitales (imágenes por resonancia magnética, tomografía computarizada y rayos

[4]Software de RCE

X), todas las órdenes (es decir, laboratorios, recetas, diagnóstico por imágenes y procedimientos) con indicación requerida (es decir, medidas de la ICD-9 códigos), notas clínicas y de laboratorio.

Selección de Casos

Uno de los retos de usar los datos de RCE, para la investigación, es la caracterización de los datos a utilizar. En este caso, se definió como diagnóstico de insuficiencia cardiaca usando los siguientes criterios:

1. Aparición de diagnóstico de insuficiencia cardiaca en la lista de problemas, al menos una vez.
2. Aparición de insuficiencia cardiaca en 2 visitas de consulta externa, lo que indica coherencia en la evaluación clínica.
3. Al menos 2 medicamentos hayan sido prescritos asociados con CIE-9[5] por diagnóstico de insuficiencia cardíaca.
4. Aparición de la insuficiencia cardiaca en 1 o más visitas ambulatorias y al menos 1 medicamento recetado está asociado con CIE-9 por diagnóstico de insuficiencia cardíaca.

La fecha de diagnóstico se define como la primera aparición en los registros por un diagnóstico de insuficiencia cardíaca.

Se limita el análisis a los casos de insuficiencia cardíaca entre 50-79 años de edad, con un diagnóstico presentado entre el 1 de enero de 2003 y 31 de diciembre de 2006. Además, se excluyó a los pacientes que tuvieron su primera atención en menos de 2 años antes de la fecha indicada, para asegurarse que los datos de los pacientes estaban disponibles para el modelo predictivo.

Finalmente, se identificaron un total de 536 casos.

Variables Utilizadas

Este estudio, utilizó las siguientes variables:

1. Demográficas
   - Edad.
   - Sexo.
   - Altura.
2. Salud

[5] Clasificación Internacional de Enfermedades

- Peso.
- ¿Fumador?
- ¿Consume Alcohol?

3. Diagnósticos

   - Diabetes.
   - Fibrilación Atrial.
   - Problema Pulmunar Obstructivo.
   - Problema arterial perférico.
   - Hipertensión.
   - Accidente Cerebrovascular.
   - Infarto Agudo al Miocardio.
   - Problemas Respiratorios.

4. Datos Clínicos

   - Pulso.
   - Presión Arterial Sistólica.
   - Presión Arterial Diastólica.

5. Datos de Laboratorio

   - Hemoglobina.
   - Glucosa.
   - Microalbuminuria.
   - Hemoglobina Glicosilada (HbA1c).
   - Otros.

6. Prescripciones para Antihipertensivos.

Algunas variables fueron utilizadas como indicador (diagnóstico de diabetes) y a la vez para detectar duración (tiempo desde el primer diagnóstico de diabetes). También se incluyó otras variables secundarias, como la presión del pulso, que corresponde a la diferencia entre la presión arterial sistólica y diastólica.

Además, bajo el supuesto que la frecuencia de las visitas al médico aumenta a medida que se aproxima la fecha del diagnóstico de insuficiencia cardiaca, se crearon variables que representan la cantidad de visitas en intervalos de tiempo de 6 meses.

Por otro lado, se consideraron variables asociadas a exámenes relevantes y su valor respectivo. Por ejemplo, un examen relevante para el caso de la diabetes, que es un factor de riesgo para las enfermedades cardiovasculares, es la Hemoglobina Glicosilada (HbA1c). Para un paciente se considera si este examen fue solicitado, y además se registra el valor, cuando está disponible.

Técnicas de Machine Learning

Se comparó el desempeño de tres técnicas: Support Vector Machine (SVM), Boosting y Logistic Regression (LR). Para cada técnica de clasificación, se implementó métodos de selección de variables.

Para Logistic Regression se utilizó la selección de variables por pasos hacia adelante basado en el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). Para Boosting, se utilizó el score de importancia de cada variable con 2 umbrales diferentes. Para SVM se seleccionó el método L1-norm para la selección de variables, usando diferentes parámetros.

Resultados

El objetivo era validar la definición de insuficiencia cardiaca congestiva (ICC) basada en RCE, en contra del estudio de Framingham, el cual es el criterio que han sido ampliamente aplicado a los casos definidos de ICC de los expedientes médicos. La Aplicación de tales criterios son mano de obra intensiva. Esto se basa sobre los términos de y construcciones de las fechas para los síntomas y señales (por ejemplo:"el tiempo de circulación de 25 segundos") y es altamente dependiente sobre la fiabilidad y la integridad de la documentación (por ejemplo: "La pérdida de peso de 4,5 kg en 5 días en respuesta al tratamiento") por un médico dado y/o entre médicos. Dadas estas limitaciones, se crearon criterios que reflejaran más la práctica médica ambulatoria en un entorno EHR. Se definieron, los casos de insuficiencia cardíaca (IC) que tenían un médico de cabecera de la Clínica Geisinger (GC) durante la ventana de observación, que se define, como mínimo, hasta 12 meses antes de la fecha indicada, para asegurar los datos RCE que estaban disponibles para la predicción del modelo.

La fecha de diagnóstico se define como la primera aparición en la EHR por un diagnóstico de IC. Se compararon estos criterios con los criterios de Framingham, en una muestra aleatoria de 50 individuos que cumplían los criterios operativos de ICC. Se complementó con una revisión de la historia de cada caso, con la documentación de la primera aparición y la fecha de aparición de los mayores o menores criterios de Framingham de ICC. De los 50 casos, 42 de los 50 casos cumplían

los Criterios de Framingham para la ICC (es decir, ya sea 2 principales signos y síntomas, o 1 mayor y 2 signos y síntomas de menor importancia). Dos de los 8 casos restantes tenían documentación de 2 criterios menores. De los 42 casos que cumplieron ambos criterios, la fecha del diagnóstico era similar (es decir, más o menos 30 días) para 15, antes para 17 casos en los que usaban criterios operativos, y antes de 10 casos en los que se utilizó el criterio de Framingham.

Para el entrenamiento de los modelos se utilizó 10-folds cross-validation. Y para evaluar los resultados, el estudio ocupó la medida de área bajo la curva AUC.

Los resultados obtenidos fue que el modelo con mejor desempeño resultó ser Logistic Regression, con un AUC = 0.77 en 10-folds cross-validation.

2.6.2. Trabajo 2: Identificación Automática de Factores de Riesgo Cardiovascular, usando análisis de texto del RCE

Este trabajo [11], realizado el año 2012, utiliza la muestra de pacientes utilizados en el trabajo antes mencionado, pero se consideraron sólo criterios obtenidos de los campos de texto no estructurados del registro clínico electrónico. El foco en este trabajo fue identificar los signos y síntomas relacionados con el criterio de riesgo cardiovascular de Framingham. Aunque este trabajo utiliza campos de texto no estructurado, está desarrollado para el idioma inglés. Además, los criterios de riesgo cardiovascular de Framingham que se ocupan para la población chilena no son los mismo que los utilizados en este estudio, que se aplicó para la población de Estados Unidos.

Objetivo

La detección temprana de la insuficiencia cardiaca (IC) podría mitigar la enorme costo individual y social de esta enfermedad. La detección clínica se basa, en parte, en el reconocimiento de los múltiples signos y síntomas que comprenden los criterios diagnósticos de Framingham que están normalmente documentados, pero no necesariamente se sintetizan, por médicos de atención primaria antes de que se realicen estudios de diagnóstico más específicos. Este trabajo ha desarrollado un procedimiento de procesamiento de lenguaje natural (NLP) para identificar los signos y síntomas de Framingham entre los pacientes de atención primaria, a partir del uso de registros clínicos electrónicos, como paso previo al análisis de patrones y apoyo a la decisión clínica para la detección temprana de riesgo cardiovascular.

Diseño

Se ha desarrollado un procedimiento híbrido que realiza dos niveles de análisis: (1) A nivel de criterios mencionados, un sistema de NLP basado en reglas se construye para anotar las menciones de

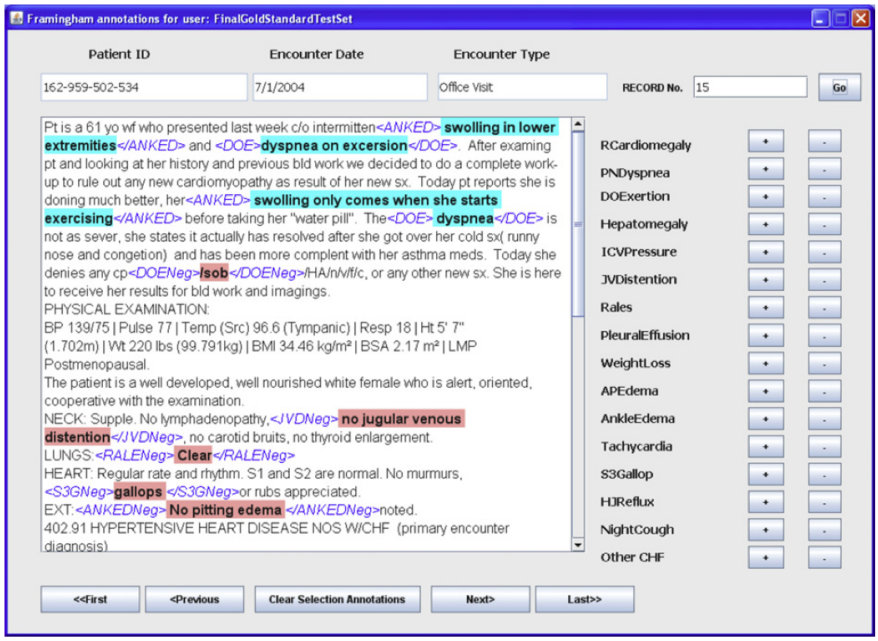


Figura 2.3: Aplicación para agregar etiquetas manualmente

criterios de Framingham. (2) A nivel de encuentro o atenciones, donde se construye un sistema para agregar etiquetas para identificar si un criterio de Framingham se afirma, niega, o es desconocido.

Resultados

Se utilizaron las medidas: Precision, Recall, and F-score para evaluar ambos procesos antes mencionados. Para evaluar la detección de criterios, y por otro lado, para evaluar el desempeño de la detección de "polaridad"para cada criterio.

Para el primer caso, se obtuvo una precision de 0.925, un recall de 0.896, y un F-score de 0.910. Para el segundo caso, el F-score fue de 0.932.

Metodología

Un cardiólogo y un lingüista analizaron un conjunto de datos de desarrollo de 65 documentos de atenciones rico en criterios de Framingham, para aprender la lingüística de criterios menciones. El lingüista también etiquetó las afirmaciones y negaciones de criterios de Framingham. Por otra parte, el experto clínico y lingüista incrementalmente miden y mejoran el rendimiento de los extractores en la documentos de desarrollo, usando una aplicación como se muestra en la figura 2.3, que permite agregar etiquetas de forma manual.

Luego, se aplicar el procesamiento automático al total de los registros a analizar, usando los siguientes pasos:

1. Procesamiento de texto básico. Abarcó tokenización, búsqueda en diccionario de conceptos,

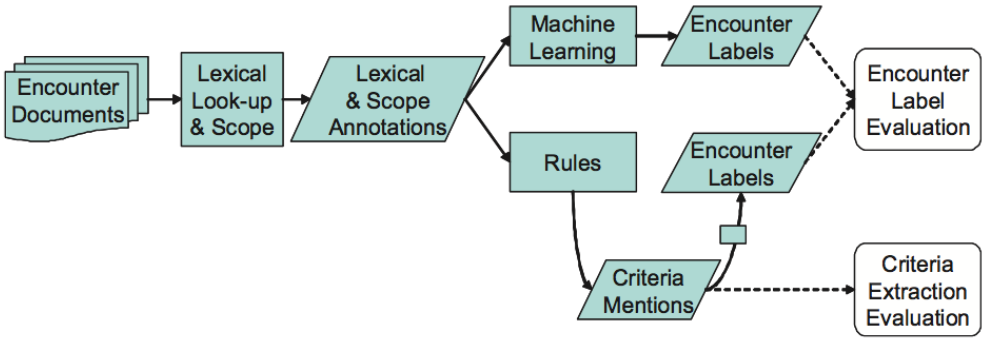


Figura 2.4: Proceso de Análisis de Texto No Estructurado

análisis morfológico y POS Tagging.

2. Diccionarios y gramáticas. Sirven para reconocer las palabras y frases usadas para expresar criterios de Framingham HF y otros indicadores posibles de insuficiencia cardiaca. Además se usa para detectar inicios del segmento, y otras estructuras sintácticas.

3. Motores de Análisis de Texto Motores. Construidos para eliminar la ambigüedad y también se utiliza para decidir cuándo fueron negados los criterios que son detectados.

En la figura 2.4 se muestra el proceso antes descrito. Este trabajo fue desarrollado por el IBM T. J. Watson Research Center, por lo que los detalles de los algoritmos utilizados no son mencionados. Sólo se entrega información sobre las aplicaciones desarrolladas por IBM para cada paso. Por ejemplo, para los análisis de texto, se usa la aplicación de IBM LanguageWare, y se utiliza también otra aplicación, IBM LanguageWare Resource Workbench (LRW), para desarrollar los diccionarios y análisis de gramática. Para la aplicación de los algoritmos que permiten realizar la clasificación se utilizar PredMED.

Capítulo 3

Caracterización de los Datos sobre Enfermedades Cardiovasculares

Considerando que hay muchos tipos de atenciones y diferentes enfermedades que se almacenan en los sistemas de Registro Clínico Electrónico, se hace necesario priorizar y enfocar el estudio en un tipo de enfermedad. Es por esto, que el presente trabajo se enfocará en las enfermedades cardiovasculares.

Para trabajar con los datos del Registro Clínico Electrónico, se utilizará la metodología CRISP-DM (Ver Figura 2.1), ya detallada en Capítulo 2. A continuación, se desarrollarán las dos primeras etapas de esta metodología que corresponden a: Entendimiento del Negocio y Entendimiento de los Datos.

En la primera sección, como parte del Entendimiento del Negocio, se realizará una introducción a la Salud Pública en Chile. Además, se presenta en detalle los puntos más relevantes asociados a las enfermedades cardiovasculares, los conceptos básicos y el detalle de sus signos y síntomas. Es muy importante este detalle, ya que el presente trabajo pretende obtener esta información del los campos de texto no estructurados del Registro Clínico Electrónico.

También, se hace una introducción al Programa de Salud Cardiovascular que ayuda a controlar a los pacientes crónicos que presentan diversos factores de riesgo cardiovascular para mantenerlos compensados. Este programa es una de las principales medidas preventivas asociada a enfermedades cardiovascular implementadas por el Ministerio de Salud.

En la segunda sección, como parte del Entendimiento de los Datos, se presenta en detalle la forma en que se registran los datos en cada una de las atenciones. Estos datos, son la principal fuente de información del presente trabajo.

Con respecto a la metodología CRISP-DM, las etapas de Preparación de Datos, Modelamiento, Evaluación e Implementación, serán abordados en los siguientes capítulos.

3.1. Entendimiento del Negocio

En esta primera sección se presentará la historia y estructura de la Salud Pública en Chile. Además, se detallará todo lo relacionado a las enfermedades cardiovasculares.

3.1.1. Salud Pública en Chile

A continuación se hace una reseña a los principales hitos de las últimas décadas de la Salud Pública en Chile, la visión estratégica para la década 2011-2020, su estructura administrativa y financiamiento.

Historia

En 1979, mediante el Decreto Ley N^{o} 2.763 se decidió la reorganización del Ministerio de Salud fusionándose el SNS (Servicio Nacional de Salud) y el SERMENA (Servicio Médico Nacional de Empleados), originando así el Sistema Nacional de Servicios de Salud que quedó compuesto por los siguientes organismos:

- Servicios de Salud, donde radicaron las funciones operativas;
- Fondo Nacional de Salud (FONASA), al que se le asignó la función financiera;
- Central de Abastecimiento (CENABAST);
- Instituto de Salud Pública (ISP).

En 1980, se inició la municipalización de los Consultorios de Atención Primaria, en la búsqueda de dar autonomía a los mismos.

En 1981 se crearon las Instituciones de Salud Previsional (ISAPRES), entidades aseguradoras privadas que, con sistemas de libre elección, permitirían optimizar la oferta de prestaciones y beneficios a sus afiliados, los que en la práctica (hasta hoy), han correspondido esencialmente al segmento de población con ingresos medios y altos.

En 1990, se planteó desarrollar una reforma de salud. Pese a lo anterior, gran parte de los esfuerzos iniciales fueron destinados a recuperar esencialmente la infraestructura hospitalaria y optimizar los mecanismos de rectoría y regulación en salud. En este sentido, se creó una superintendencia de ISAPRES destinada a fiscalizar el financiamiento y prestaciones privadas. Asimismo, se incrementaron mecanismos de control orientados a supervigilar el funcionamiento del sistema en todos sus niveles, y se establecieron instituciones abocadas a enfrentar temas emergentes como CONASIDA

(Comisión Nacional del VIH-SIDA), CONACE (Comisión para el Control de Estupefacientes) y el FONADIS (Fondo Nacional para la Discapacidad).

En el año 2002, y en vistas a generar un nueva reforma que diera respuesta a todas las necesidades pendientes, se establecieron los primeros Objetivos Sanitarios Nacionales, los cuales dieron cuenta de las prioridades que el Estado debía enfrentar para mejorar las condiciones e inequidades del sistema de salud. Ante esto, para la primera década del siglo XXI, se establecieron los siguientes objetivos:

1. Mantener los logros alcanzados: En esencia, la importante mejoría en indicadores que, como la mortalidad infantil, daban cuenta del fruto cosechado tras todos los esfuerzos aplicados hasta entonces;

2. Enfrentar el envejecimiento progresivo de la población: Con su creciente carga de patologías crónicas no transmisibles, de alto costo de atención;

3. Resolver las desigualdades: Las citadas brechas entre grupos de población de distinto nivel socioeconómico;

4. Responder adecuadamente a las expectativas de la población: Elemento continuamente señalado como un elemento de insatisfacción de la sociedad chilena para con el sistema.

Son estos planteamientos los que, a la fecha, han llevado a la materialización de iniciativas legales como el Plan AUGE (Acceso Universal con Garantías Explícitas en Salud) o GES (Garantías Explícitas en Salud), en un intento por garantizar el derecho a la salud a toda la población sin discriminación de ningún tipo y disminuir las desigualdades. [35]

Estrategia Nacional de Salud 2011-2020

Uno de los objetivos de la Estrategia Nacional de Salud 2011-2020 [35] es "Prevenir y reducir la morbilidad, la discapacidad y mortalidad prematura por afecciones crónicas no transmisibles". Entre estas enfermedades crónicas, se encuentran las Enfermedades Cardiovasculares. Como parte de este objetivo estratégico de salud, se proponen las siguientes medidas destinadas a aumentar la sobrevida de personas que presentan alto riesgo cardiovascular, debido a un IAM (Infarto Agudo al Miocardio) o ACV (Accidente Cardiovascular):

1. Mejorar la oportunidad del inicio de la atención;

2. Mejorar la calidad del tratamiento;

3. Prevención secundaria;

4. Implementar sistemas de información clínica.

En este contexto, se presenta la necesidad de mejorar los sistemas de registros clínicos, para que permitan identificar los casos incidentes y monitorear los procesos clínicos y sus resultados.

La Organización Mundial de la Salud propone que: es necesario disponer de tecnologías de información, para identificar a la población en riesgo, organizar la atención, monitorear la respuesta al tratamiento y evaluar los resultados. Además, que se requieren sistemas de comunicación, que permitan el intercambio oportuno de información con el paciente y con otros proveedores de servicios de salud que no comparten el mismo lugar físico. [36]

En Chile, existe un sistema de Registro Clínico Electrónico muy avanzando y centralizado para la gran mayoría de los establecimientos de la Atención Primaria. Sin embargo, queda mucho trabajo pendiente con respecto a las recomendaciones hechas por la OMS, ya que los sistemas de registro actuales son insuficientes para vigilar y con ello mejorar la calidad de atención y asegurar el mejor pronóstico del paciente.

Estructura

La estructura del sistema de salud en Chile cuenta con una red descentralizada de 29 Servicios de Salud Autónomos que conforman el Sistema Nacional de Servicios de Salud, coordinados por la Subsecretaría de Redes Asistenciales. De acuerdo al Departamento de Estadísticas e Información de Salud (DEIS) del Ministerio de Salud, se tiene la siguiente distribución de establecimientos en Chile:

- 105 Establecimientos hospitalarios de menor complejidad.
- 24 Establecimientos hospitalarios de mediana complejidad.
- 64 Establecimientos hospitalarios de alta complejidad.
- 322 Municipios que tienen a cargo la atención primaria de salud.

Dentro de los municipios, se encuentran distribuidos los establecimientos de Atención Primaria de Salud (APS), los que se distribuyen como se muestra en la tabla 3.1.

Con respecto a las personas, a la cobertura de FONASA corresponde a 12.427.534 personas. Por otra parte, la cobertura de las Instituciones de Salud Previsional (ISAPRE) corresponde a 2.829.554.

Cabe mencionar que la cantidad de personas que acceden a la Atención Primaria de Salud (APS), aumenta año a año, como se muestra en gráfico 3.1.

Tipo Establecimiento	Cantidad
Centros de salud familiar - CESFAM	566
Postas de salud rural - PSR	1168
Servicio atención primario de urgencia - SAPU	228
Servicio de urgencia rural	112
Centros comunitarios de salud familiar - CECOF	166

Cuadro 3.1: Red de Atención Primaria de Salud

Fuente: MINSAL - DEIS 2013

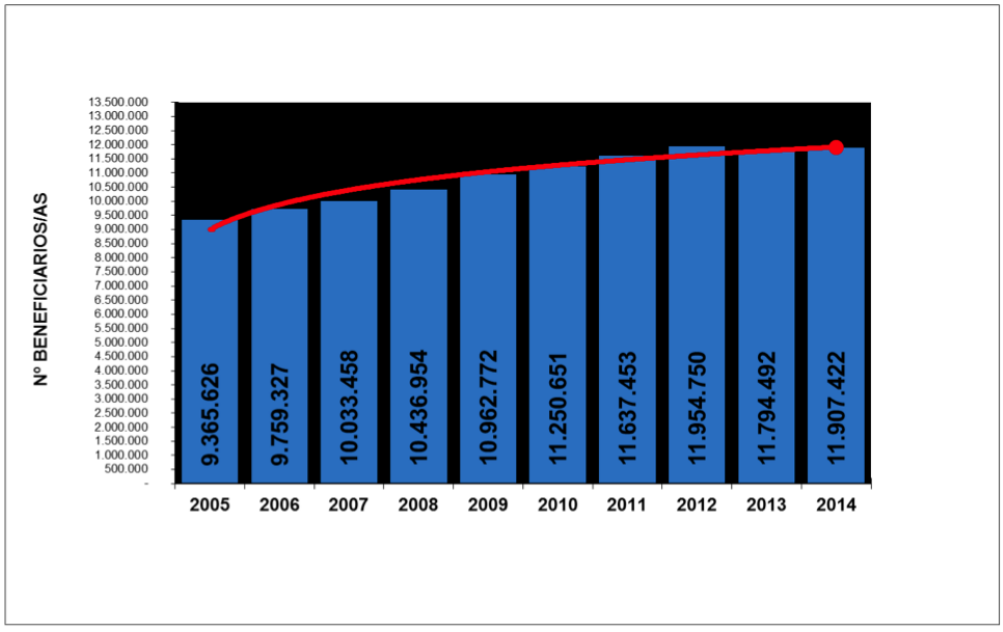


Figura 3.1: Población Objetivo para APS, 2005 a 2014

Fuente: MINSAL

Desde el punto de vista del Ministerio de Salud, se plantea como desafío fortalecer la atención primaria de salud, así evitar lo que se conoce como el hospitalocentrismo. Cuando una persona llega al hospital, es por una enfermedad de mayor gravedad y complejidad, lo que implica mayores gastos, ya sea por costos de tratamiento, medicamentos o días de hospitalización, entre otros.

Indicadores de Actividad de la Atención Primaria de Salud

La aplicación del Índice de Actividad de la Atención Primaria ha permitido evaluar la pertinencia de algunos indicadores y las dificultades en la medición de otros. Siempre en la perspectiva de perfeccionar el mecanismo de evaluación. Para el año 2014, el componente de Actividad General cuenta con 10 indicadores, el componente de Actividad con Continuidad de Atención, con dos indicadores (acceso a la atención de salud hasta las 20:00 horas y disponibilidad de fármacos trazadores) y el componente de la Actividad GES mantiene los problemas de salud a ser evaluados.

Para la selección de los ámbitos a medir, se han considerado los siguientes criterios:

- Que se enmarquen en los objetivos sanitarios vigentes.
- Que conduzcan al cumplimiento de las garantías en atención primaria en los problemas de salud incorporados al sistema GES.
- Que den cuenta de procesos de la atención primaria que enfatizan el cuidado de la salud a través del enfoque familiar y comunitario y/o su integración a la red asistencial.
- Que en su conjunto equilibren evaluación de aspectos cuantitativos y cualitativos.
- Que para la construcción de los indicadores se considere la población inscrita validada para establecimientos municipales y ONG en convenios y la beneficiaria estimada para los establecimientos dependientes de los Servicios de Salud.
- Que estén incorporadas en sistemas de registros de uso habitual (REM) y minimicen los monitoreos especiales.

De acuerdo a los criterios enumerados, en la Actividad General, se abarcan prestaciones y programas de salud que se desarrollan en el ciclo vital, que evalúan los siguientes ámbitos:

- Cobertura de Acciones Preventivas
- Oportunidad, Accesibilidad y Equidad
- Resultados en Proceso de Intervención Preventiva con Enfoque de Riesgo.

- Enfoque multidisciplinario, Enfoque familiar

La aplicación de estos indicadores tiene directa relación con el financiamiento recibido por los establecimientos de salud, como también, funciona aplicando rebajas ante los incumplimientos.

Financiamiento en Nivel Comunal

Cada municipio recibe mensualmente, del Ministerio de Salud, a través de los Servicios de Salud correspondientes, un aporte estatal, conocido como per cápita. Éste se conforma para cada comuna según criterios como población beneficiaria en la comuna, nivel socioeconómico de la población e índices de ruralidad y dificultad para acceder y prestar atención de salud. El percápita basal alcanzará este año (2013) un valor cercano a $3.500 mensuales por persona.

Además del aporte anteriormente señalado, el Estado aporta al financiamiento de la atención primaria de salud municipal, a través del propio Presupuesto Sectorial de Salud, mediante programas tales como: Chile Crece Contigo, Equidad en Salud Rural, Sistema de Urgencia Rural, Rehabilitación Integral de Base Comunitaria, Cesfam de Excelencia, entre otros.

3.1.2. Las Enfermedades Cardiovasculares

A continuación se presenta una síntesis sobre las ECV en el mundo y en Chile.

En el mundo

De acuerdo a datos de la Organización Mundial de la Salud (OMS), las enfermedades cardiovasculares constituyen una de las causas más importantes de discapacidad y muerte prematura en todo el mundo. El problema subyacente es la aterosclerosis (vasculopatía periférica), que progresa a lo largo de los años, de modo que cuando aparecen los síntomas, generalmente a mediana edad, suele estar en una fase avanzada. Las principales consecuencias, son los episodios coronarios (infarto de miocardio) y cerebrovasculares (ataque apoplético) agudos, que se producen de forma repentina y conducen a menudo a la muerte antes de que pueda dispensarse la atención médica requerida.

Es por esto, que la detección temprana del riesgo puede reducir los episodios cardiovasculares y la muerte prematura en aquellas personas con alto riesgo cardiovascular debido a uno o más factores de riesgo.

De acuerdo a recomendaciones de la OMS, la forma de prevenir los episodios cardiovasculares, es ofrecer recomendaciones basadas en la evidencia sobre cómo reducir la incidencia de primeros y sucesivos episodios clínicos de cardiopatía coronaria, enfermedad cerebrovascular y vasculopatía periférica en dos categorías de personas:

1. Prevención Primaria: Personas con factores de riesgo que aún no han presentado síntomas de enfermedad cardiovascular.

2. Prevención Secundaria: Personas con cardiopatía coronaria, enfermedad cerebrovascular o vasculopatía periférica establecidas.

Existen tablas de predicción del riesgo cardiovascular (Anexo A) que permiten estimar el riesgo cardiovascular global en la primera categoría de personas.

Las personas, incluidas en la primera categoría (prevención primaria), y que son consideradas con riesgo cardiovascular alto, de acuerdo a las tablas de predicción de riesgo, requieren intervenciones de cambio en su modo de vida, principalmente, en sus hábitos de alimentación, actividad física, abandono del tabaco y control de peso.

Las personas de la segunda categoría (prevención secundaria), siempre son consideradas con riesgo alto, y además de un cambio en su modo de vida, es necesario complementar su cuidado con un tratamiento farmacológico adecuado.

Cabe mencionar, que las medidas para evitar una enfermedad cardiovascular, es simplemente llevar una vida saludable, algo que todas las personas deberían hacer. Es decir, prevenir a tiempo y de forma objetiva a las personas con mayor riesgo, es sumamente importante para evitar este tipo de enfermedades.

En Chile

Sin embargo, el principal problema en Chile es que estas tablas no son frecuentemente utilizadas debido a que las personas no asisten a sus Exámenes de Medicina Preventiva donde se hace este tipo de evaluaciones. Además, ocurre que cuando una persona asiste a una consulta médica las atenciones suelen ser muy específicas y los pacientes son tratados sólo por el problema que tienen en ese momento, aunque muchas veces se registran los datos relevantes para poder hacer la evaluación de riesgo cardiovascular, estos datos no necesariamente son sintetizados por el médico para realizar la evaluación de riesgo.

Según las cifras del Departamento de Estadísticas e Información (DEIS) del Ministerio de Salud, anualmente mueren en Chile 97.930 personas. Del total, más de un tercio de las muertes se deben a enfermedades crónicas o también conocidas como ENT[1].

Las ENT no se transmiten de persona a persona, son de larga duración y por lo general evolucionan lentamente. Dentro de las diez primeras causas de defunción en Chile, siete son relacionadas a ENT según el DEIS. La primeras corresponden a casos de enfermedades isquémicas (reducción

[1]Enfermedades No Transmisibles

del flujo sanguíneo) del corazón y cerebrovasculares. Por lo que las enfermedades cardiovasculares son los primeras en el ranking de mobimortalidad.

El concepto de morbimortalidad, es un concepto que proviene de la ciencia médica y que combina dos subconceptos como la morbilidad y la mortalidad.

1. La morbilidad es la presencia de un determinado tipo de enfermedad en una población.

2. La mortalidad, a su vez, es la estadística sobre las muertes en una población también determinada.

Así, juntando ambos subconceptos podemos entender que la idea de morbimortalidad es determinar aquellas enfermedades causantes de la muerte en determinadas poblaciones, espacios y tiempos.

La idea de morbimortalidad tiene una utilidad principalmente estadística, ya que supone brindar información relativa a las causas de muerte en una población o grupo de personas determinadas. Luego, esta información es utilizada para analizar el porqué de la presencia de esas enfermedades particulares, su incidencia final sobre la muerte de las personas analizadas, etc. El objetivo de analizar la mobimortalidad es establecer parámetros y los medios posibles para limitar o evitar ese tipo de resultado mortales.

Finalmente y considerando: su alto impacto en tasas de mortalidad, siendo la primera causa de muertes en Chile; en los costos económicos que significan para el Estado las enfermedades crónicas, tanto en tratamiento como costos en medicamentos; y considerando los costos indirectos, que afectan la calidad de vida de las personas que sufren estas enfermedades, se considerarán las enfermedades cardiovasculares para el siguiente estudio..

Tipos de Enfermedades Cardiovasculares

A nivel general, y desde el punto de vista médico, es posible definir principalmente tres tipos de fallas al sistema circulatorio:

- Eléctrico: Que tiene relación con arritmias o todo tipo de problemas asociados al ritmo (latido) cardiaco.

- Hidráulico: Que tiene relación a problemas de presión dentro del corazón, como problemas con las válvulas internas que no permiten ejercer la presión adecuada para bombear la sangre al cuerpo. Como también problemas relacionados con las arterias obstruidas, lo que no permite la correcta circulación de la sangre.

- Mecánico: Es la condición que afecta a los músculos del corazón, que debido a algún tipo de daño, provoca un mal funcionamiento del corazón. Estos problemas se deben principalmente al daño provocado por un infarto agudo al miocardio (IAM).

El resultado de estos problemas se refleja principalmente en dos enfermedades : La Aterosclerosis y la Insuficiencia Cardiaca, las que se detallan en las siguientes secciones.

3.1.3. La Aterosclerosis

La aterosclerosis es una enfermedad en la que la placa se deposita dentro de las arterias. Las arterias son vasos sanguíneos que llevan sangre rica en oxígeno al corazón y a otras partes del cuerpo.

La placa está compuesta por grasas, colesterol, calcio y otras sustancias que se encuentran en la sangre. Con el tiempo, la placa se endurece y estrecha las arterias, con lo cual se limita el flujo de sangre rica en oxígeno a los órganos y a otras partes del cuerpo.

La aterosclerosis puede causar problemas graves, como ataque cardíaco, accidentes cerebrovasculares (derrames o ataques cerebrales) e incluso la muerte.

Las causas de la aterosclerosis no se conocen. Sin embargo, ciertas características, enfermedades o hábitos pueden elevar el riesgo de sufrir la enfermedad. Estas situaciones se llaman factores de riesgo.

Algunos factores, como la falta de actividad física, el hábito de fumar y la alimentación poco saludable, se pueden controlar. Otros no se pueden controlar, como la edad y los antecedentes familiares de enfermedades del corazón.

Algunas personas que tienen aterosclerosis no presentan signos ni síntomas. Tal vez no les diagnostiquen la aterosclerosis hasta después de haber tenido un ataque cardíaco o un accidente cerebrovascular.

El principal tratamiento para la aterosclerosis son los cambios en el estilo de vida. Es posible que también se necesiten medicinas y procedimientos médicos. Estos tratamientos, junto con la atención médica continua, pueden servirle a la persona para llevar una vida más sana.

Los tratamientos mejorados han reducido la cantidad de muertes por enfermedades relacionadas con la aterosclerosis. Estos tratamientos también han mejorado la calidad de vida de las personas que tienen estas enfermedades. Sin embargo, la aterosclerosis sigue siendo un problema frecuente de salud.

La aterosclerosis puede afectar a cualquiera de las arterias del cuerpo, incluidas las del corazón, el cerebro, los brazos, las piernas, la pelvis y los riñones. Según las arterias afectadas, pueden presentarse diferentes enfermedades.

Enfermedad coronaria

La enfermedad coronaria, conocida también como enfermedad de las arterias coronarias, es la principal causa de muerte de hombres y mujeres. Se presenta cuando la placa se deposita en las arterias coronarias. Estas arterias llevan sangre rica en oxígeno al corazón.

La placa estrecha las arterias, con lo cual el flujo sanguíneo del músculo del corazón o músculo cardíaco disminuye. Además, aumenta la probabilidad de que se formen coágulos de sangre en las arterias. Los coágulos de sangre pueden bloquear la circulación de la sangre parcial o completamente.

Si el flujo de sangre que llega al músculo cardíaco está reducido o bloqueado, se puede producir angina (dolor o molestias en el pecho) o un ataque cardíaco.

La placa también puede formarse en las arterias más pequeñas del corazón. Esta enfermedad se conoce como enfermedad coronaria microvascular. En ella la placa no causa bloqueos en las arterias como lo hace en la enfermedad coronaria.

Enfermedad de las arterias carótidas

La enfermedad de las arterias carótidas se presenta si la placa se deposita en las arterias que quedan a ambos lados del cuello (arterias carótidas). Estas arterias llevan sangre rica en oxígeno al cerebro. Si el flujo de sangre que va al cerebro está reducido o bloqueado se puede presentar un accidente cerebrovascular.

Enfermedad arterial periférica

La enfermedad arterial periférica se presenta si la placa se deposita en las principales arterias que suministran sangre rica en oxígeno a las piernas, los brazos y la pelvis.

Si el flujo de sangre a estas partes del cuerpo está reducido o bloqueado, la persona puede tener adormecida esa parte del cuerpo, sentir dolor y, a veces, tener infecciones peligrosas.

Enfermedad renal crónica

Puede presentarse enfermedad renal crónica si la placa se deposita en las arterias renales. Estas arterias llevan sangre rica en oxígeno a los riñones.

Con el tiempo, la enfermedad renal crónica causa pérdida lenta del funcionamiento de los riñones. La principal función de los riñones es eliminar los desechos y el exceso de agua del cuerpo.

Causas de la Aterosclerosis

La causa exacta de la aterosclerosis no se conoce. Sin embargo, se ha visto en estudios que la aterosclerosis es una enfermedad lenta y compleja que puede comenzar en la infancia. A medida que

la persona envejece, avanza más rápidamente.

La aterosclerosis puede comenzar cuando ciertos factores causan daños en las capas internas de las arterias. Estos factores son:

- El hábito de fumar
- Las cantidades altas de ciertas grasas y colesterol en la sangre
- La presión arterial alta
- Las cantidades altas de azúcar en la sangre debido a resistencia a la insulina o a la diabetes

La placa puede comenzar a depositarse en el lugar en que las arterias sufrieron daños. Con el tiempo, la placa se endurece y estrecha las arterias. A la larga, una zona de la placa puede romperse.

Cuando esto sucede, unos fragmentos de células llamados plaquetas se adhieren al lugar de la lesión y pueden agruparse para formar coágulos de sangre. Los coágulos estrechan las arterias aún más y limitan el flujo de sangre rica en oxígeno al cuerpo.

Según las arterias que se afecten, los coágulos de sangre pueden empeorar la angina (dolor en el pecho) o causar un ataque cardíaco o un accidente cerebrovascular.

Factores de Riesgo de Aterosclerosis

Ciertas características, enfermedades o hábitos pueden elevar el riesgo de sufrir la enfermedad. Estas situaciones se llaman factores de riesgo. Cuantos más factores de riesgo se tenga, más probabilidades hay de presentar aterosclerosis.

La mayoría de los factores de riesgo se pueden controlar, con lo cual se previene o retrasa la aparición de la aterosclerosis. Otros factores de riesgo no se pueden controlar.

Principales factores de riesgo:

1. Las concentraciones poco saludables de colesterol en la sangre. Esto abarca un colesterol LDL alto (este colesterol se conoce también como çolesterol malo") y un colesterol HDL bajo (que también se llama çolesterol bueno").

2. La presión arterial alta. La presión arterial se considera alta si permanece en 140/90 mmHg o más por un período de tiempo. Si se tiene diabetes o enfermedad renal crónica, la presión arterial alta se define como una presión de 130/80 o más ("mmHg"significa milímetros de mercurio y son las unidades en que se mide la presión arterial).

3. El hábito de fumar. El hábito de fumar puede lesionar y estrechar los vasos sanguíneos, elevar las concentraciones de colesterol y subir la presión arterial. Además, no permite que llegue suficiente oxígeno a los tejidos del cuerpo.

4. La resistencia a la insulina. Esta situación se presenta cuando el organismo no puede usar su propia insulina adecuadamente. La insulina es una hormona que ayuda a transportar el azúcar de la sangre al interior de las células, en donde se usa como fuente de energía. La resistencia a la insulina puede causar diabetes.

5. La diabetes. En esta enfermedad las concentraciones de glucosa en la sangre son demasiado altas porque el organismo no produce suficiente insulina o no usa la insulina adecuadamente.

6. El sobrepeso o la obesidad. Los términos "sobrepesoz .obesidad"se refieren a un peso corporal superior al que se considera saludable para una estatura determinada.

7. La falta de actividad física. La falta de actividad física puede empeorar otros factores de riesgo de la aterosclerosis, como las concentraciones poco saludables de colesterol en la sangre, la presión arterial alta, la diabetes y el sobrepeso o la obesidad.

8. La alimentación poco saludable. Una alimentación poco saludable puede elevar el riesgo de sufrir aterosclerosis. Los alimentos ricos en grasas saturadas, grasas trans, colesterol, sodio (sal) y azúcar pueden empeorar otros factores de riesgo de la enfermedad.

9. La edad avanzada. Al envejecer aumenta el riesgo de sufrir aterosclerosis. A medida que una persona envejece hay factores genéticos o de estilo de vida que pueden ocasionar depósitos de placa en las arterias. Para cuando la persona esté en la edad madura o sea mayor, se habrá acumulado suficiente placa como para causar signos o síntomas. En los hombres, el riesgo aumenta después de los 45 años. En las mujeres aumenta después de los 55 años.

10. Los antecedentes familiares de enfermedad coronaria de aparición temprana. Su riesgo de sufrir aterosclerosis aumenta si a su padre o a un hermano le diagnosticaron enfermedad coronaria antes de los 55 años, o si a su madre o a una hermana se la diagnosticaron antes de los 65 años.

Aunque la edad y los antecedentes familiares de enfermedad coronaria de aparición temprana son factores de riesgo, eso no quiere decir que se tenga aterosclerosis si tiene uno o ambos factores de riesgo. El control de otros factores de riesgo puede a menudo disminuir la influencia genética y prevenir la aterosclerosis, incluso en personas de edad avanzada.

Se ha visto en estudios que un número cada vez mayor de niños y adolescentes corre el riesgo de sufrir aterosclerosis. Las causas son varias, entre ellas las crecientes tasas de obesidad infantil.

Nuevos factores de riesgo

Los científicos siguen estudiando otros posibles factores de riesgo de la aterosclerosis.

Las concentraciones altas de una proteína llamada proteína C reactiva (PCR) en la sangre pueden elevar el riesgo de sufrir aterosclerosis y ataque cardíaco. Las concentraciones altas de proteína C reactiva indican que hay inflamación en el cuerpo.

La inflamación es la respuesta del organismo frente a una lesión o infección. La lesión de las paredes internas de las arterias parece desencadenar el proceso de inflamación y contribuir al crecimiento de la placa.

El índice de aterosclerosis es menor entre las personas que tienen concentraciones bajas de proteína C reactiva que entre las que tienen concentraciones altas. Se están realizando investigaciones para averiguar si al reducir la inflamación y disminuir las concentraciones de proteína C reactiva se reduce también el riesgo de sufrir aterosclerosis.

Las concentraciones altas de triglicéridos en la sangre también pueden elevar el riesgo de sufrir aterosclerosis, especialmente en las mujeres. Los triglicéridos son un tipo de grasa.

Se están llevando a cabo estudios para averiguar si en el riesgo de presentar aterosclerosis intervienen factores genéticos.

Otros factores que influyen en la aterosclerosis

Otros factores también pueden elevar el riesgo de sufrir aterosclerosis. Entre ellos se cuentan:

- La apnea del sueño. La apnea del sueño es un trastorno en el que la persona hace una o más pausas en la respiración o respira de manera superficial durante el sueño. Sin tratamiento, la apnea del sueño puede elevar el riesgo de sufrir presión arterial alta, diabetes e incluso un ataque cardíaco o un accidente cerebrovascular.

- El estrés. Se ha demostrado en investigaciones que de los factores que pueden provocar un ataque cardíaco el que más se menciona es un acontecimiento que cause alteración emocional, especialmente si se trata de uno que implique ira.

- El consumo de alcohol. Beber en exceso puede lesionar el músculo cardíaco y empeorar otros factores de riesgo de la aterosclerosis. Los hombres no deben tomar más de dos bebidas alcohólicas al día. Las mujeres no deben tomar más de una bebida alcohólica al día.

Signos y Síntomas

Por lo general, la aterosclerosis no causa signos ni síntomas hasta que estrecha gravemente una arteria o la bloquea por completo. Muchas personas no saben que tienen la enfermedad hasta que sufren una situación de urgencia médica, como un ataque cardíaco o un accidente cerebrovascular.

Algunas personas pueden tener signos y síntomas de la enfermedad. Los signos y síntomas dependen de las arterias que estén afectadas.

- Arterias coronarias

  Las arterias coronarias llevan sangre rica en oxígeno al corazón. Cuando la placa estrecha o bloquea estas arterias (en una enfermedad llamada enfermedad coronaria), un síntoma frecuente es la angina. La angina es un dolor o molestia en el pecho que se presenta cuando el músculo cardíaco no recibe suficiente sangre rica en oxígeno.

  Se puede sentir como presión o como un dolor que aprieta el pecho. El dolor también puede sentirse en los hombros, los brazos, el cuello, la mandíbula o la espalda. Puede incluso parecerse a la sensación de indigestión. El dolor tiende a empeorar con la actividad y desaparece al descansar. El estrés emocional también puede desencadenar el dolor.

  Otros síntomas de la enfermedad coronaria son la dificultad para respirar y las arritmias. Las arritmias son problemas de la rapidez o el ritmo de los latidos del corazón.

  La placa también puede formarse en las arterias más pequeñas del corazón. Esta enfermedad se conoce como enfermedad coronaria microvascular. Sus síntomas comprenden angina, dificultad para respirar, problemas para dormir, agotamiento (cansancio) y falta de energía.

- Arterias carótidas

  Las arterias carótidas llevan sangre rica en oxígeno al cerebro. Si la placa las estrecha o las bloquea (en una enfermedad llamada enfermedad de las arterias carótidas), se puede tener los síntomas de un accidente cerebrovascular o derrame cerebral. Estos síntomas pueden ser:

  1. Debilidad repentina
  2. Parálisis (incapacidad para moverse) o adormecimiento de la cara, los brazos o las piernas, especialmente en un lado del cuerpo
  3. Confusión
  4. Dificultad para hablar o para entender lo que otra persona dice
  5. Dificultad para ver por un ojo o por los dos

6. Problemas para respirar
7. Mareo, dificultad para caminar, falta de equilibrio o de coordinación y caídas inexplicables
8. Pérdida del conocimiento
9. Dolor de cabeza intenso y repentino

- Arterias periféricas

  La placa también se puede depositar en las principales arterias que llevan sangre rica en oxígeno a las piernas, los brazos y la pelvis. Esta enfermedad se llama enfermedad arterial periférica.

  Si estas arterias principales están estrechadas o bloqueadas, se puede tener adormecimiento de esas partes del cuerpo, dolor y a veces infecciones peligrosas.

- Arterias renales

  Las arterias renales llevan sangre rica en oxígeno a los riñones. Si la placa se deposita en ellas se puede presentar enfermedad renal crónica. Con el tiempo, la enfermedad renal crónica causa pérdida lenta del funcionamiento de los riñones.

  En sus inicios, a menudo la enfermedad renal no produce signos ni síntomas. A medida que empeora puede causar cansancio, cambios en la forma en que se elimina la orina (más frecuentemente o menos frecuentemente), inapetencia, náuseas (ganas de vomitar), hinchazón de las manos o los pies, picazón o adormecimiento y dificultad para concentrarse.

Diagnóstico

El médico diagnostica la aterosclerosis con base en los antecedentes médicos y familiares del paciente, el examen médico y los resultados de ciertas pruebas.

Especialistas: Si se tiene aterosclerosis, un médico de atención primaria (por ejemplo, un internista o un médico de familia) puede encargarse de su atención. Si se necesitan cuidados especiales, el médico puede recomendar a otros especialistas, por ejemplo:

- Un cardiólogo. Los cardiólogos son médicos que se especializan en el diagnóstico y tratamiento de las enfermedades y problemas del corazón. Se puede ir a un cardiólogo si se tiene enfermedad coronaria o enfermedad coronaria microvascular.

- Un especialista en medicina vascular. Se trata de un médico que se especializa en el diagnóstico y tratamiento de los problemas de los vasos sanguíneos. Se puede ver a un especialista en medicina vascular si tiene enfermedad arterial periférica.

- Un neurólogo. Este médico se especializa en el diagnóstico y tratamiento de los trastornos del sistema nervioso. Se puede ver a un neurólogo si se ha tenido un accidente cerebrovascular debido a la enfermedad de las arterias carótidas.
- Un nefrólogo. Los nefrólogos son médicos que se especializan en el diagnóstico y tratamiento de las enfermedades y problemas de los riñones. Se puede ir a un nefrólogo si se tiene enfermedad renal crónica.

Examen Médico: En este examen el médico le oirá las arterias en busca de un sonido anormal parecido a un susurro que se llama "soplo". El médico puede oír un soplo al colocar el estetoscopio sobre una arteria afectada. Un soplo puede indicar mala circulación de la sangre por depósito de placa.

El médico también puede ver si alguno de sus pulsos (por ejemplo, de la pierna o del pie) es débil o está ausente. La debilidad o ausencia del pulso puede ser un signo de que hay una arteria bloqueada.

Pruebas Diagnósticas: El médico puede recomendar una o más pruebas para diagnosticar la aterosclerosis. Estas pruebas también pueden servirle para determinar la extensión de la enfermedad y planificar el tratamiento más adecuado.

Pruebas de sangre: En las pruebas de sangre se determinan las concentraciones sanguíneas de ciertas grasas, colesterol, azúcar y proteínas. Las concentraciones anormales pueden indicar que se corre riesgo de sufrir aterosclerosis.

Electrocardiograma (ECG o EKG): El electrocardiograma es una prueba sencilla e indolora que detecta y registra la actividad eléctrica del corazón. Muestra qué tan rápido late el corazón y con qué ritmo (estable o irregular). También registra la potencia y la sincronización de los impulsos eléctricos a medida que pasan por cada parte del corazón.

Un electrocardiograma puede mostrar signos de daños cardíacos causados por la enfermedad coronaria. También puede mostrar si hubo un ataque cardíaco o si está sucediendo uno actualmente.

Radiografía de tórax: En la radiografía de tórax se obtienen imágenes de los órganos y estructuras que se encuentran dentro del pecho, entre ellos el corazón, los pulmones y los vasos sanguíneos. La radiografía de tórax puede revelar signos de insuficiencia.

Índice tobillo-humeral: Esta prueba compara la presión de la sangre en el tobillo con la presión de la sangre en el brazo para ver qué tan bien está circulando la sangre. Puede servir para diagnosticar la enfermedad arterial periférica.

Ecocardiografía: En la ecocardiografía se usan ondas sonoras para crear una imagen animada del corazón. La ecocardiografía proporciona información sobre el tamaño y la forma del corazón y

sobre cómo están funcionando las cámaras y las válvulas.

También puede identificar zonas de mala circulación en el corazón, zonas de músculo cardíaco que no se están contrayendo normalmente y lesiones anteriores del músculo cardíaco causadas por falta de circulación.

Tomografía computarizada: La tomografía computarizada crea imágenes generadas por computadora del corazón, el cerebro u otras partes del cuerpo. La prueba puede mostrar el endurecimiento y estrechamiento de las grandes arterias.

La tomografía computarizada del corazón también puede mostrar si se ha depositado calcio en las paredes de las arterias coronarias o arterias del corazón. Esto puede ser un signo temprano de la enfermedad coronaria.

Prueba de esfuerzo: Durante la prueba de esfuerzo se hará ejercicio para que el corazón trabaje mucho y lata rápidamente mientras se realizan unas pruebas cardíacas. Si no puede hacer ejercicio se le darán medicinas para que el corazón trabaje más y lata más rápidamente.

Cuando el corazón está esforzándose mucho y latiendo con rapidez necesita más sangre y oxígeno. Las arterias estrechadas por la placa no pueden suministrar suficiente sangre rica en oxígeno para satisfacer las necesidades del corazón.

La prueba de esfuerzo puede mostrar posibles signos y síntomas de la enfermedad coronaria, como:

- Alteraciones de la frecuencia cardíaca o de la presión arterial
- Sensación de falta de aliento o dolor en el pecho
- Alteraciones del ritmo cardíaco o de la actividad eléctrica del corazón

En algunas pruebas de esfuerzo se toman imágenes del corazón cuando se está haciendo ejercicio y cuando está descansando. Estas pruebas de esfuerzo con imágenes pueden mostrar qué tan bien circula la sangre en distintas partes del corazón y cómo la bombea el corazón al latir.

Angiografía: La angiografía es una prueba en la que se usan un medio de contraste y unos rayos X especiales para mostrar el interior de las arterias. Esta prueba puede mostrar si la placa está bloqueando las arterias y qué tan grave es el bloqueo.

Un tubito delgado y flexible llamado catéter se inserta en un vaso sanguíneo del brazo, la ingle (parte superior del muslo) o el cuello. A través del catéter se inyecta en las arterias un medio de contraste que se puede ver en las imágenes de rayos X. Al mirar la imagen de rayos X el médico puede ver la forma en que la sangre circula por sus arterias.

Cómo Prevenir

El tratamiento de la aterosclerosis puede consistir en cambios del estilo de vida, medicinas y procedimientos médicos o cirugía.

Los objetivos del tratamiento son:

- Aliviar los síntomas.
- Disminuir los factores de riesgo para retardar o detener el depósito de placa.
- Disminuir el riesgo de que se formen coágulos de sangre.
- Ensanchar las arterias coronarias obstruidas por la placa o dar un rodeo para evitarlas.
- Prevenir las enfermedades relacionadas con la aterosclerosis.

Cambios en el estilo de vida: Los cambios en el estilo de vida a menudo sirven para prevenir o tratar la aterosclerosis. En algunas personas estos cambios pueden ser el único tratamiento necesario.

Consumir una alimentación saludable: Una alimentación saludable forma parte importante de un estilo de vida sano. Consumir una alimentación saludable puede prevenir que se eleven la presión arterial alta y el colesterol o reducir sus valores, si están elevados. También puede ayudarle a mantenerse en un peso saludable.

Cambios terapéuticos del estilo de vida: Es posible que el médico recomiende el programa de cambios terapéuticos del estilo de vida si su colesterol es alto. Estos cambios consisten en un programa de tres partes: una alimentación saludable, actividad física y control del peso.

Según la dieta del programa, menos del 7 por ciento de sus calorías diarias deben provenir de grasas saturadas. Este tipo de grasas se encuentra principalmente en algunas carnes, productos lácteos, chocolate, productos de panadería y alimentos fritos y procesados.

Solo entre el 25 % y el 35 % de las calorías diarias deben venir de todo tipo de grasas (saturadas, trans, monoinsaturadas y poliinsaturadas).

Además, se debe consumir menos de 200 mg de colesterol al día. La cantidad de colesterol y los tipos de grasas de los alimentos preparados se encuentran en la etiqueta de información nutricional.

Los alimentos ricos en fibra soluble también forman parte de un plan de alimentación saludable. Estos alimentos impiden la absorción de colesterol en el aparato digestivo. Entre ellos están:

- Cereales integrales, como avena y salvado de avena
- Frutas, como manzanas, plátanos, naranjas, peras y ciruelas pasas

- Legumbres, como frijoles, lentejas, garbanzos, judías y habas

Una alimentación rica en frutas y verduras puede aumentar en la dieta importantes compuestos para bajar el colesterol.

Una alimentación saludable también contiene algunos tipos de pescado, como el salmón, el atún (fresco o enlatado) y las sardinas. Estos pescados son una fuente excelente de ácidos grasos omega 3, que pueden proteger al corazón de la inflamación y la formación de coágulos de sangre, y disminuir el riesgo de que se presente un ataque cardíaco. Es recomendable consumir pescado por lo menos dos veces por semana.

También se debe limitar la cantidad de sodio (sal) que consume. Esto significa elegir alimentos y condimentos con bajo contenido de sal o que no la contengan, tanto en la mesa como durante la preparación de las comidas. La etiqueta de información nutricional de los empaques de alimentos muestra la cantidad de sodio que el artículo contiene.

Limitar el consumo de bebidas alcohólicas. El exceso de alcohol eleva la presión arterial y la concentración de triglicéridos. (Los triglicéridos son un tipo de grasa que se encuentra en la sangre.) El alcohol también añade más calorías, lo cual lleva a un aumento de peso.

Los hombres no deben tomar más de dos bebidas alcohólicas al día. Las mujeres no deben tomar más de una bebida alcohólica al día. Un trago equivale a una copa de vino, un vaso de cerveza o una pequeña cantidad de licor.

Realizar actividad física: La actividad física que se practica con regularidad puede disminuir muchos factores de riesgo de la aterosclerosis, entre ellos el colesterol LDL (colesterol "malo"), la presión arterial alta y el exceso de peso.

La actividad física también puede reducir el riesgo de sufrir diabetes y puede elevar las concentraciones de colesterol HDL (el colesterol "bueno"que previene la aterosclerosis).

La salud de una persona se beneficia con dedicar tan solo 60 minutos semanales a una actividad aeróbica moderada. Para obtener mayores beneficios de salud se recomienda realizar por lo menos 150 minutos (2 horas y media) de actividad aeróbica moderada o 75 minutos (1 hora y cuarto) de actividad aeróbica intensa por semana.

Controlar el peso: Al mantenerse en un peso saludable se puede disminuir el riesgo de sufrir aterosclerosis. Una buena meta es tratar de lograr un índice de masa corporal (IMC) de menos de 25.

El IMC mide el peso en relación con la estatura y proporciona un cálculo de la grasa corporal total.

Un IMC entre 25 y 29.9 se considera sobrepeso. Un IMC de 30 o más se considera obesidad.

Para prevenir y tratar la aterosclerosis hay que fijarse la meta de tener un IMC de menos de 25. El médico o el profesional de salud que lo atiende puede ayudarle a fijarse una meta adecuada de IMC.

Dejar de fumar: Si se fuma o se usa tabaco, se debe dejar de hacerlo. El hábito de fumar puede lesionar y estrechar los vasos sanguíneos, y elevar el riesgo de sufrir aterosclerosis. Además, se debe tratar de evitar exponerse al humo que producen las personas que fuman.

Controlar el estrés: En investigaciones se ha visto que, entre los factores que pueden causar un ataque cardíaco, el que se menciona con más frecuencia es aquel acontecimiento que causa alteración emocional, especialmente si se trata de uno que implique ira. Además, algunas de las maneras en que la gente lidia con el estrés, como la bebida, el hábito de fumar o el exceso de comida, tampoco son saludables.

Aprender a controlar el estrés, relajarse y lidiar con los problemas puede mejorar la salud emocional y física. La actividad física, ciertas medicinas y la terapia de relajación también ayudan a aliviar el estrés.

3.1.4. Insuficiencia Cardiovascular

A continuación se hace referencia a las principales causas, síntomas, tratamientos relacionados con la insuficiencia cardiaca, de acuerdo a lo publicado en [37].

Causas

La insuficiencia cardíaca a menudo es una afección prolongada (crónica), aunque algunas veces se puede presentar repentinamente. Puede ser causada por muchos problemas diferentes del corazón.

La enfermedad puede afectar únicamente el lado derecho o el lado izquierdo del corazón y se denomina insuficiencia cardíaca derecha o izquierda respectivamente. Con mucha frecuencia, ambos lados del corazón resultan comprometidos.

La insuficiencia cardíaca ocurre cuando:

- El miocardio no puede bombear o expulsar la sangre del corazón adecuadamente y se denomina insuficiencia cardíaca sistólica.
- Los músculos del corazón están rígidos y no se llenan con sangre fácilmente. Esto se denomina insuficiencia cardíaca diastólica.

Estos problemas significan que el corazón ya no puede bombear suficiente sangre oxigenada al resto del cuerpo. A medida que el bombeo del corazón se vuelve menos eficaz, la sangre puede acumularse en otras áreas del cuerpo. Puede producirse una acumulación de líquido en los pulmones,

el hígado, el tracto gastrointestinal, al igual que en los brazos y las piernas. Esto se denomina insuficiencia cardíaca congestiva.

La causa más común de insuficiencia cardíaca es la arteriopatía coronaria, un estrechamiento de los pequeños vasos sanguíneos que suministran sangre y oxígeno al corazón. La hipertensión arterial que no esté bien controlada también puede llevar a que se presente insuficiencia cardíaca.

Otros problemas del corazón que pueden causar insuficiencia cardíaca son:

- Cardiopatía congénita.
- Ataque cardíaco.
- Valvulopatía cardíaca (esto puede ocurrir a partir de válvulas permeables o estrechas).
- Infección que debilita el miocardio.
- Algunos tipos de ritmos cardíacos anormales (arritmias).

También hay otras enfermedades que pueden causar o contribuir a la insuficiencia cardíaca son:

- Amiloidosis.
- Enfisema.
- Hipertiroidismo.
- Sarcoidosis.
- Anemia grave.
- Demasiado hierro en el cuerpo.
- Hipotiroidismo.

Síntomas

Los síntomas de la insuficiencia cardíaca con frecuencia empiezan de manera lenta. Al principio, pueden sólo ocurrir cuando se está muy activo. Con el tiempo, se pueden notar problemas respiratorios y otros síntomas incluso cuando se está descansando.

Los síntomas de insuficiencia cardíaca también pueden empezar de manera repentina después de un ataque cardíaco u otro problema del corazón.

Los síntomas comunes son:

- Tos.

- Fatiga, debilidad, desmayos.
- Inapetencia.
- Necesidad de orinar en la noche.
- Inflamación de los pies y los tobillos.
- Pulso irregular o rápido o una sensación de percibir los latidos cardíacos (palpitaciones).
- Dificultad respiratoria cuando se está activo o después de acostarse.
- Abdomen o hígado inflamado (agrandado).
- Hinchazón de pies y tobillos.
- Despertarse después de un par de horas debido a la dificultad respiratoria.
- Aumento de peso.

Pronóstico

A menudo, se puede controlar la insuficiencia cardíaca tomando medicamentos, cambiando el estilo de vida y tratando la afección que la causó.

La insuficiencia cardíaca puede empeorar repentinamente debido a:

- Angina.
- Comer alimentos muy salados.
- Ataque cardíaco.
- Infecciones u otras enfermedades.
- No tomar los medicamentos correctamente.

Por lo general, la insuficiencia cardíaca es una enfermedad crónica que puede empeorar con el tiempo. Algunas personas presentan insuficiencia cardíaca grave, en la cual los medicamentos, otros tratamientos y la cirugía ya no ayudan.

Factores de Riesgo

Un factor de riesgo es una característica o comportamiento de la persona que aumenta su probabilidad de contraer una enfermedad o tener una cierta afección.

Algunas de los factores que incrementan el riesgo de contraer cardiopatía y que NO se pueden cambiar son:

- Edad: El riesgo de cardiopatía aumenta con la edad.
- Sexo: Los hombres tienen un riesgo más alto de padecer cardiopatía que las mujeres que todavía están menstruando. Después de la menopausia, el riesgo en las mujeres se acerca al de los hombres.
- Genes: Si los padres padecieron cardiopatía, se tiene un riesgo más alto.

Algunos de los riesgos para cardiopatía se PUEDEN cambiar son:

- No fumar. Si efectivamente fuma, dejar de hacerlo.
- Controlar el colesterol a través de la alimentación, el ejercicio y los medicamentos, de ser necesario.
- Controlar la hipertensión arterial a través de la alimentación, el ejercicio y los medicamentos, de ser necesario.
- Controlar la diabetes a través de la alimentación, el ejercicio y los medicamentos, de ser necesario.
- Hacer ejercicio por lo menos 30 minutos al día.
- Mantener un peso saludable comiendo alimentos sanos, comiendo menos y vinculándose a un programa de pérdida de peso, si es necesario bajar de peso.
- Aprender formas saludables de hacer frente al estrés.
- Limitar la cantidad de alcohol que toma a 1 trago al día para las mujeres y 2 para los hombres.

La buena nutrición es importante para su salud cardíaca y ayuda a controlar algunos de los factores de riesgo.

- Una dieta rica en frutas, verduras y granos enteros.
- Proteínas magras, tales como pollo, pescado, frijoles y legumbres.

- Productos lácteos bajos en grasa, tales como leche al 1 % y otros artículos bajos en grasa.
- Evitar el sodio (sal) y las grasas que se encuentran en alimentos fritos, alimentos procesados y productos horneados.
- Comer menos productos animales que contengan queso, crema o huevos.
- Evitar las “grasa saturadaz cualquier cosa que contenga grasas “parcialmente hidrogenadas.º “hidrogenadas”. Estos productos generalmente están cargados de grasas poco saludables.

3.1.5. Programa de Salud Cardiovascular

El Programa de Salud Cardiovascular (PSCV) es una de las principales estrategias del Ministerio de Salud para contribuir a reducir la morbimortalidad asociada a las enfermedades cardiovasculares. Este "nace? el 2002 producto de la reorientación de los subprogramas de Hipertensión arterial (HTA) y Diabetes (DM), cuyo principal cambio fue incorporar el enfoque de riesgo cardiovascular (CV) global en el manejo de las personas bajo control, en lugar de tratar los factores de riesgo en forma separada. [38]

A contar del 2005, el manejo de la hipertensión, diabetes, infarto agudo al miocardio, se incorporan como Garantías Explícitas de Salud (GES), en tanto que el accidente cerebrovascular isquémico lo hace a partir de 2006.

Uno de los problemas que se presenta, es que la evaluación del Programa muestra que aún cuando las personas en control se clasifican según nivel de riesgo CV, esta estratificación no se expresa en un plan terapéutico y de seguimiento diferenciado.

Estrategias Complementarias

Por lo demás, es importante que el enfoque dirigido a población de alto riesgo sea complementado con estrategias de salud pública a nivel poblacional. Es decir, brindar oportunidades para hacer actividad física, reducir el contenido de sal en los alimentos procesados, aumentar los impuestos del tabaco, prohibir publicidad engañosa, entre otras cosas. Aún cuando la probabilidad de hacer un evento cardiovascular en población de bajo riesgo es menor, no existe ningún nivel de riesgo que pueda ser considerado "seguro?. Es por esto, la importancia de los esfuerzos de salud pública a nivel poblacional, ya que seguirán ocurriendo eventos en personas de bajo y moderado riesgo, que además, corresponde a la mayoría de la población.

Este tipo de enfoque poblacional de salud pública, puede reducir en forma efectiva el desarrollo de ateroesclerosis (y de paso reducir la incidencia de algunos cánceres y enfermedades respiratorias crónicas) en personas jóvenes, reduciendo la aparición de enfermedades cardiovasculares. Las estrategias poblacionales también constituyen un apoyo para aquellas personas de más alto riesgo para que hagan modificaciones en su estilo de vida.

Detección de Factores de Riesgo

La atención primaria de salud (APS) cumple un rol muy importante en la detección y registro de las personas con factores de riesgo en la prevención de las enfermedades cardiovasculares.

El Examen de Medicina Preventiva (EMP) es una de las principales herramientas disponibles para pesquisar factores de riesgo CV, como:

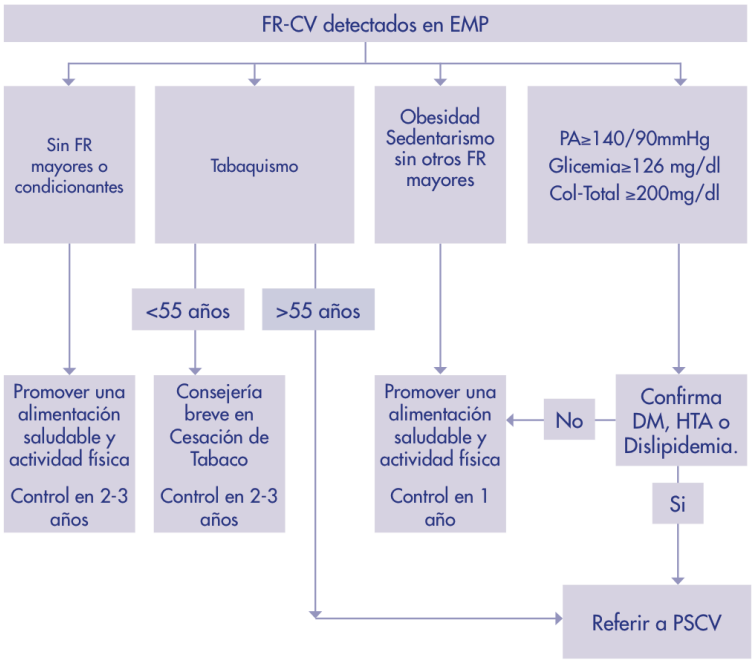


Figura 3.2: Detección de factores de RCV en el EMP

Fuente: MINSAL

- tabaquismo,
- elevación de la presión arterial,
- colesterol total y glicemia,
- obesidad.

En la figura 3.2, se señala el curso de acción a seguir ante la detección de uno o más de estos factores en el contexto del EMP o en ocasión de otro contacto entre el paciente y el sistema de salud (consulta morbilidad, hospitalización, otros).

En aquellos casos en que se confirma la condición de riesgo, se informa al paciente sobre sus derechos asociados al AUGE (en el caso de hipertensos y diabéticos), y se le invita a ingresar al PSCV. Se le solicitará los exámenes de ingreso al Programa, de tal manera que el paciente tenga todos los resultados al momento de la primera consulta médica, hito que se considera el ingreso al PSCV. El detalle de los procedimientos se encuentra en el documento [39], que entrega el MINSAL con la implementación del enfoque de riesgo en el programa de salud cardiovascular.

3.2. Entendimiento de los Datos

Se trabajará con datos del Registro Clínico Electrónico almacenados por la empresa SAYDEX. La aplicación se llama RAYEN, que contiene diversos subsistemas que pueden albergar la totalidad de los procesos de atención de un paciente. Es una herramienta amigable y ágil que permite al usuario trabajar de manera sencilla en las funcionalidades que le competen según el área en la que se desempeña dentro del establecimiento de salud, consolidando la información y disponiendo de ella en línea para una toma de decisiones oportuna.

La particularidad de esta base de datos, es que a diferencia de otras aplicaciones que realizan una instalación aislada para cada recinto, RAYEN permite tener un registro centralizado de todos los establecimientos que utilizan esta aplicación.

3.2.1. Módulos Principales

Admisión

La necesidad de registrar la información con respecto a incorporar nuevos usuarios de la Red de Atención Primaria de Salud APS, se torna crítica en la medida que aumentan las exigencias sanitarias, las que tienen por objetivo entregar un servicio de excelencia, en forma oportuna y de calidad.

Este módulo, como se muestra en la figura 3.3, permite a los establecimientos de salud administrar esta información, tanto del individuo, como de su familia. Además, permite validar el correcto registro de los datos y verificar si los inscritos no se encuentren registrados en otros establecimientos. Por último, también permite obtener informes rápidos y claros en cuanto a la población registrada en el establecimiento.

Agenda

Una de las principales herramientas que hoy en día son necesarias a la hora de llevar una buena gestión en los establecimientos de salud, es el control ágil y dinámico de las Agendas de los Profesionales. RAYEN permite al usuario de manera fácil, diseñar estas agendas y administrarlas a través de diversas funciones como cambios de vistas, bloqueos y eliminación, como se muestra en la figura 3.4.

Citas

Para poder llevar una buena gestión en los establecimientos de salud, sin duda es necesario contar con procedimientos que permitan manejar rápidamente la asignación de horas de atención a las personas que lo requieren. También es importante poder manejar la información clara y precisa

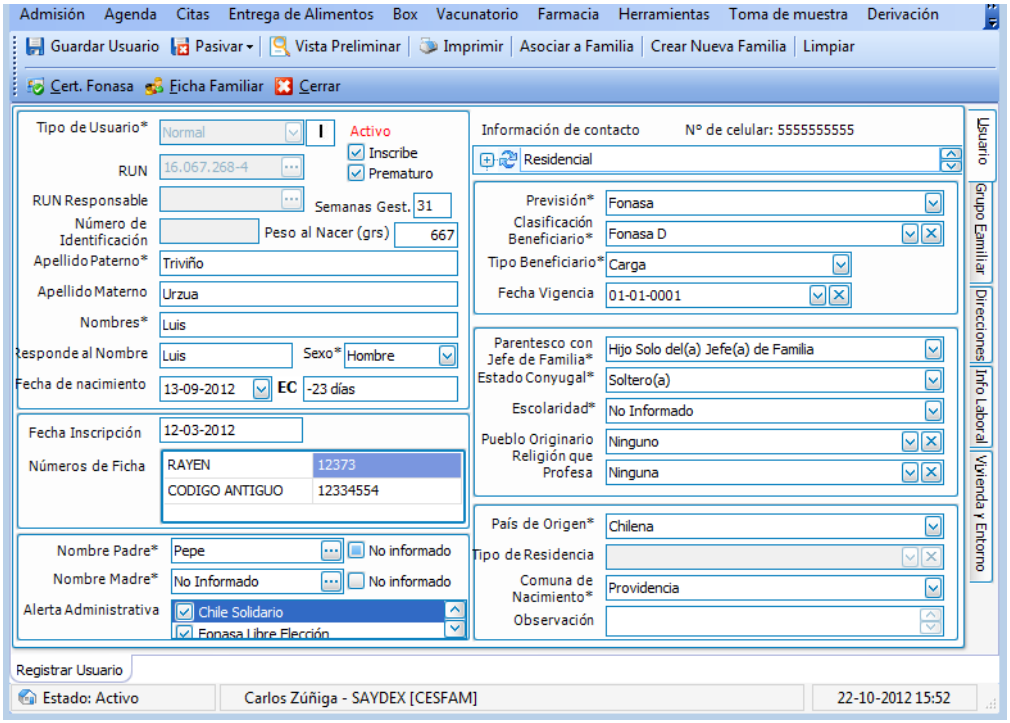


Figura 3.3: Módulo Admisión de RAYEN

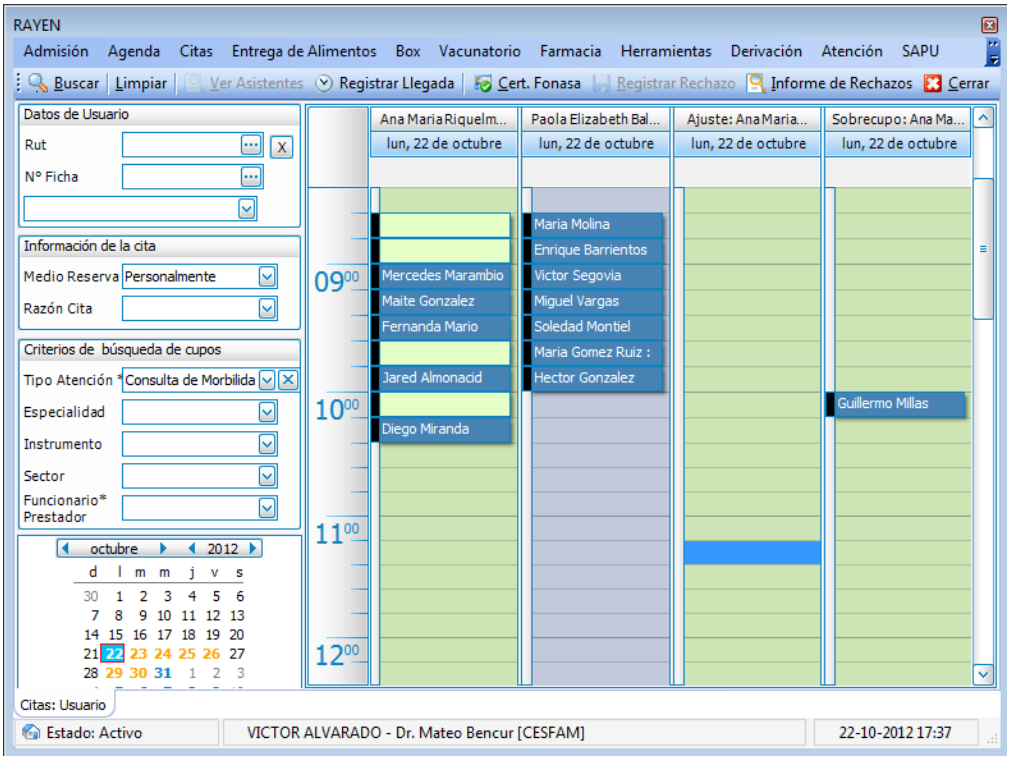


Figura 3.4: Módulo Agenda y Citas de RAYEN

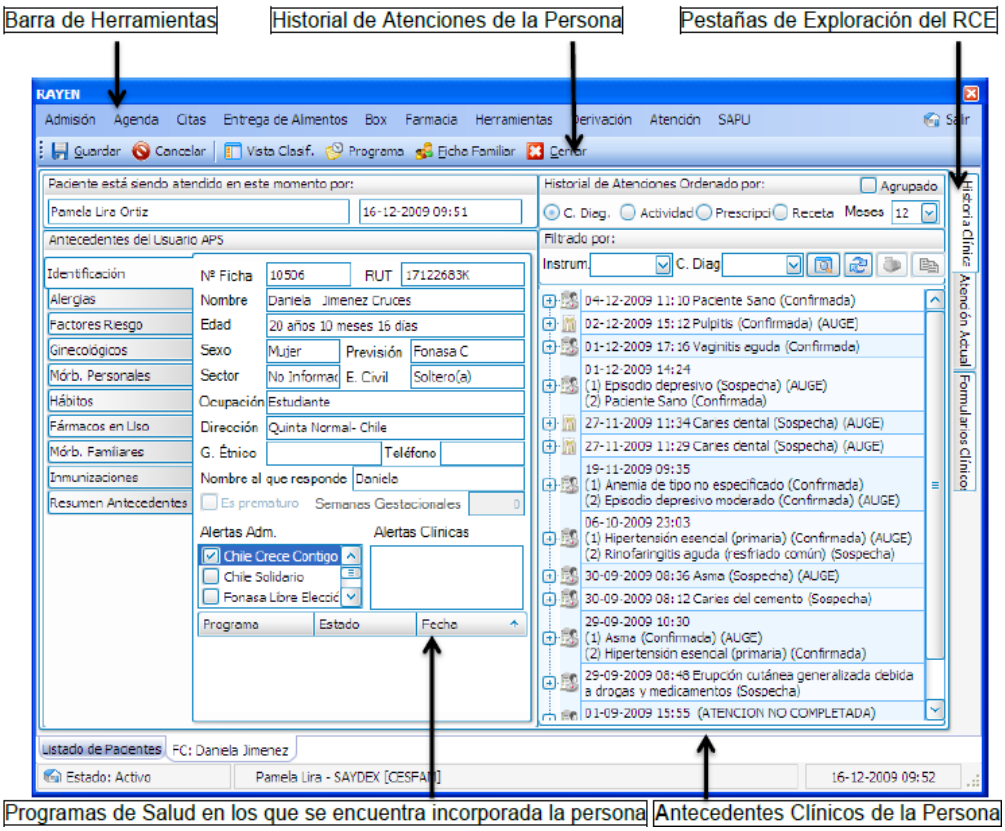


Figura 3.5: Módulo Ficha Clínica de RAYEN

para entregársela a los Funcionarios Clínicos que atienden a las personas, además de contar con la posibilidad de obtener informes estadísticos que lleven finalmente a un control ordenado de cada cita que se entrega a una persona. Para todos estos procedimientos de control, RAYEN posee el subsistema de Citas , que permite a los funcionarios Citar personas, Anular citas, Traspasar personas citadas de un Funcionario Clínico a otro, además de entregar la posibilidad de obtener diversos informes que ayuden a la gestión.

Ficha Clínica Electrónica

RAYEN consta con el Registro Clínico Electrónico, que permite llevar un completo historial de atenciones de una persona, registrar una atención actual, revisar las morbilidades personales y familiares, entre otras diversas funcionalidades que finalmente permite obtener también la información estadística de los REM (Reportes Estadísticos Mensuales). Además, permite de contar con el registro de la Población en Control (PEC).

Otros Módulos

- **VACUNATORIO**: En particular, el módulo de Vacunatorio contempla los requerimientos del registro nacional de inmunizaciones (RNI) junto con la vinculación de datos importante

de las personas que requieren la administración de alguna vacuna.

- **FARMACIA**: Uno de los procesos más complejos en la atención de pacientes es la generación de recetas por parte de los funcionarios prestadores. Pero, hay problemas cuando es imposible la interpretación por parte de los funcionarios de farmacia al no estar la receta escrita de forma clara y ordenada. Esto genera problemas a la hora del despacho y correspondientes indicaciones. Por este motivo RAYEN cuenta con un modulo de Farmacia, el que facilita el ingreso de fármacos y correspondiente posología, así como el acceso al historial de recetas, prescripciones y movimientos de fármacos e insumos. Estas características crean un sistema integral y dinámico con acceso a información en todo momento, manteniendo un control real del stock y de usuarios del sistema.

- **ENTREGA DE ALIMIENTOS**: Sin duda que a la hora de realizar la entrega de un alimento específico a una persona, se requiere agilidad y fidelidad en el registro de la información, además de poder llevar un buen control de gestión y llenar la estadística correspondiente. Por este motivo RAYEN cuenta con el Módulo de Entrega de Alimentos que por tener internamente guardada la programación tanto de PNAC como PACAM, permite la agilidad a la hora de registrar los alimentos entregados a las diferentes personas atendidas en un establecimiento. Esta herramienta, debido a su sencillez permite además ahorrar tiempo a quienes la utilizan, y a su vez les ayuda en la generación de la estadística diaria y mensual.

- **SAPU**: Actualmente, la visión de Red Asistencial de Salud, lleva a la necesidad de contar con una herramienta dinámica, que permita obtener en forma ágil los datos clínicos de una persona, independientemente del establecimiento de salud en el que se atiende comúnmente. Hoy en día, todos los establecimientos de urgencia, requieren contar con esta información, de manera precisa y completa, con el fin de evaluar al instante las condiciones clínicas de la persona y de esta forma tomar las decisiones pertinentes para la atención que se debe realizar. RAYEN cuenta con una completa herramienta, denominada SAPU, que permite satisfacer todas las necesidades planteadas anteriormente.

- **INTERCONSULTA**: Este modulo permite a los profesionales derivar a los pacientes a otros especialistas ya sea dentro del mismo establecimiento o a otro establecimiento de la red asistencial, agilizando los procesos administrativos que conllevan estos procesos.

3.2.2. Procesos Principales

Existen 3 procesos principales, que generan los datos que se utilizarán en este trabajo:

1. **Inscripción**: Corresponde cuando un usuario ingresa por primera vez al sistema, es aquí donde se registran sus datos personales. Recién ahí puede ser atendido.

2. **Atención**: Cuando una persona ya inscrita en el sistema, es posible acceder a sus datos desde cualquier establecimiento que tenga la plataforma de RCE de SAYDEX. Se puede decir, que es algo similar a la Fícha Clínica Única, que es algo por lo que se está trabajando a nivel de MINSAL. Una vez en el box de atención, el médico puede agregar todo tipo de datos clínicos, desde signos vitales, resultados de exámenes, prescripciones de medicamentos, y en particular, tiene 4 campos de texto no estructurado donde puede ingresar datos: **Anamenesis, Motivo de la Consulta, Observaciones sobre el Diagnóstico, Historia Clínica**. Son estos campos los que serán analizados en el presente trabajo.

3. **Controles**: Corresponden a atenciones posteriores, que se originan a partir de una atención preliminar. Dado un diagnóstico, el médico puede agenda próximas visitas para evaluar los resultados de los tratamientos recomendados.

A partir de los procesos antes mencionados, es posible obtener variables relevantes para el análisis:

- A partir de los datos de las inscripciones de los usuarios, es posible obtener datos demográficos de las personas, como su fecha de nacimiento (edad), sexo, dirección, entre otras.

- A partir de los datos de las atenciones, se pueden obtener datos sobre características y comportamiento o hábito de las personas, como por ejemplo, su peso, presión arterial, nivel de colesterol, si es fumador, si consume alcohol, alguna condición genética, o detalles sobre su familia y enfermedades relacionadas.

- Si se hace un análisis de varias atenciones, es posible detectar que medicamentos ha consumido el paciente, de acuerdo a prescripciones previas, también es posible determinar diagnósticos previos. Además, es posible generar variables relativas a la cantidad de veces que se ha atendido el paciente, o también hacer el análisis sobre la cantidad de veces que se ha atendido de acuerdo a un determinado diagnóstico.

- A partir del seguimiento de controles, es posible evaluar la evolución del paciente de una atención a otra. Esto permite evaluar el desempeño de los medicamentos recetados y de los tratamientos sugeridos por el médico. La información relacionada a la evolución de una enfermedad, o la efectividad de un fármaco, por lo general son registrados en los campos de texto no estructurados, ya que son campos u observaciones propias para cada persona.

En cada atención, ya sea si es primera vez que el paciente asiste al establecimiento, o se está atendiendo porque está en control por alguna enfermedad diagnosticada previamente, mucha información queda registrada en los 4 campos de texto no estructurado que se mencionaron anteriormente. Hasta ahora, la única forma de aprovechar dicha información, es que en cada atención el médico dedique minutos de la atención a leer la histórica clínica de cada paciente.

Sin embargo, en el contexto de la Salud Pública, donde las atenciones son muy breves, muchas veces el médico se enfoca en atender al paciente de acuerdo a lo que la persona dice, y no tienen suficiente tiempo para revisar toda la historia clínica. Por lo que estos datos, registrados en los campos de texto, no son aprovechados del todo.

Almacenamiento de los Datos

Los datos se almacenan en un esquema relacional, donde la principal entidad es la ATENCION. Cualquier proceso que asociado a la atención, está identificado por el ID único de la atención. El detalle de la estructura se encuentra en el Anexo B .

Los procesos asociados más importantes son:

- ANAMNESIS. Registro del relato hecho por el paciente previo al diagnostico (Texto Libre)
- DIAGNÓSTICO. Asignación de un diagnóstico hecho por el médico con una codificación CIE-10. Es posible agregar una observación en un campo de texto libre.
- ACTIVIDAD. Tipo de procedimiento aplicado en una atención.
- RECETA. Prescripciones hechas durante una atención.
- EXAMENES. Los exámenes solicitados en una atención.
- INTERCONSULTA. Solicitudes de interconsultas a establecimientos de a establecimientos de atención secundaria.

La descripción de las tablas utilizadas se encuentran en el Anexo C .

Capítulo 4

Propuesta de Investigación

En este capítulo se realiza una síntesis de los temas planteados en los capítulos anteriores y se detalla la propuesta de investigación, la que tiene por objetivo ser un aporte a las estrategias de salud preventiva en relación a las enfermedades cardiovasculares.

4.1. Detalles Propuesta de Investigación

Se presenta el marco general, dentro del cual se están desarrollando todas las soluciones de tecnología relacionadas con salud, lo que se conoce como Smarter Care [40] y su enfoque en la salud preventiva. Además, se entrega información sobre la relevancia de las enfermedades cardiovasculares y las actuales estrategias que el Estado utiliza para controlar y prevenir. Finalmente, se presenta un marco de trabajo para realizar análisis avanzado de datos a partir del RCE.

Con el contexto claramente definido, se presentan en la segunda sección de este capítulo los detalles del diseño del experimento que se realizó, se declaran los resultados esperados y cómo este trabajo puede aportar a la salud preventiva en Chile.

4.1.1. Smarter Care

El concepto de Smarter Care apunta a la atención integral e individualizada, y que permite optimizar los recursos utilizados, lo que trae como consecuencia reducir los costos asociados a los tratamientos de salud, tanto para las personas y el Estado. Todo esto, a partir de información relacionada a datos clínicos, complementada con información sobre el estilo de vida y condiciones sociales de las personas. En la figura 4.1, se presenta un esquema de este concepto. A continuación, se detalla cada uno de los puntos.

- **Datos Clínicos**, como síntomas, historia clínica, medicamentos, diagnósticos, que son excelentes variables a analizar para determinar las condiciones de salud de una persona.

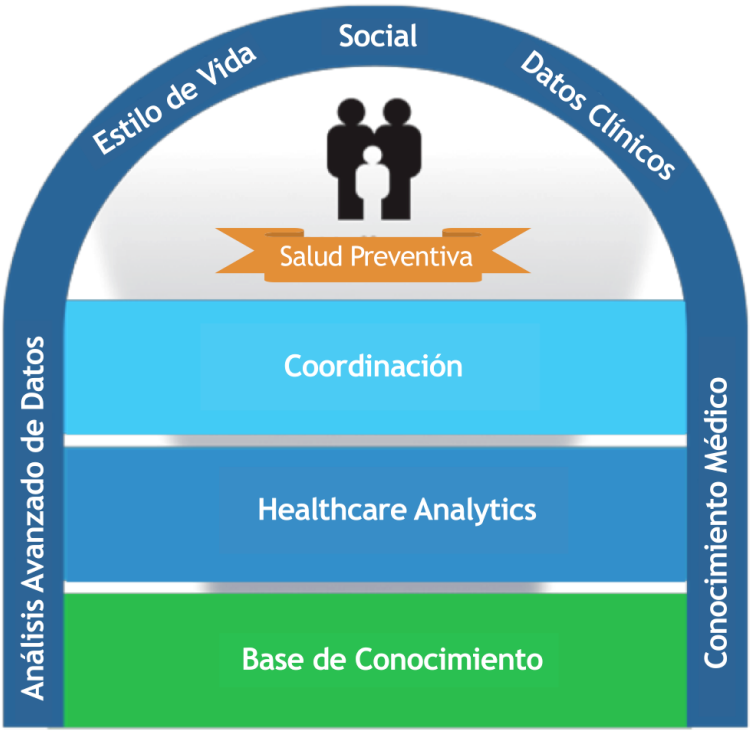


Figura 4.1: Smarter Care

Fuente: IBM

- **Condiciones Sociales**, que considera variables como el lugar de nacimiento, donde creció, donde vive actualmente, en que trabaja, entre otras cosas, pueden tener directa relación sobre sus condiciones actuales de salud y su bien estar físico y mental.
- **Estilo de Vida**, tiene un efecto directo en las condiciones de salud y el bien estar físico y mental de las personas.

Smarter Care se fundamenta en el análisis avanzado de datos disponibles de una determinada población y la base de conocimiento médica ya existente. El objetivo que se plantea Smarter Care, es lograr que a partir de una base de conocimiento y su respectivo análisis, sea posible entender lo que realmente le está ocurriendo a las personas en una determinada población. Y así, las entidades gubernamentales respectivas puedan tomar mejores decisiones con respecto a políticas de salud, sociales y estrategias preventivas, que impacten en el estilo de vida de las personas.

Se plantean tres niveles:

1. **Base de Conocimiento**: Hace referencia a tener conocimiento de las personas, de forma individual y también a nivel poblacional. Con esto, tener oportunidad de reconocer oportunidades de intervención para aplicar los tratamientos adecuados a tiempo. Las principales soluciones tecnológicas usadas en este nivel son:

- Modelos de Datos y Data Warehouse.
- Gestión de Datos Maestros (Master Data Management).
- Reportes y Paneles de Control.
- Portales de Salud.
- Monitoreo remoto de signos vitales.

2. **Healthcare Analytics**: Consiste en aplicar técnicas avanzadas de minería de datos, ya probadas y ampliamente usadas en otras industrias, pero en el contexto de la salud. Con el objetivo de ganar entendimiento, a través del análisis de los datos, que permita a los prestadores de salud tomar decisiones con una visión más completa, para lograr mejores resultados haciendo un uso eficiente y efectivo de los recursos. Las tecnologías usadas en este nivel, aportan a:

   - Análisis Poblacionales.
   - Análisis de Riesgo Poblacional e Individual.
   - Apoyo al Diagnóstico.
   - Apoyo a los protocolos médicos.
   - Reportes operacionales.

3. **Coordinación**: Consiste en intervenir en todas las áreas en las que la persona puede afectar, para bien o para mal, su salud. El objetivo es lograr convocar y coordinar a las instituciones y organizaciones involucradas y lograr su participación activa en entregar una mejor salud a las personas. En general, se plantea como objetivo, cruzar fronteras, y no sólo trabajar en organismos especializados en salud, sino de manera transversal, para ayudar a entregar un plan integrado y así para lograr un resultado óptimo y como consecuencia poder reducir los elevados costos relacionado a todas las problemáticas de salud. Se busca establecer:

   - Identificación de personas y poblaciones con mayor riesgo.
   - Estructurar planes integrales de salud.
   - Gestionar un trabajo colaborativo entre diferentes organismos.
   - Generar instrumentos para evaluar los resultados e impacto de las políticas aplicadas.

También es posible conceptualizar el término Smarter Care a través de una cadena de valor. Lo que permite graficar y describir las actividades necesarias para generar valor al usuario final y a las organismos de salud. En este caso, el usuario final corresponde a las personas, y la organización

Figura 4.2: Smarter Care: Roles

Fuente: Elaboración Propia

representa al Ministerio de Salud, incluyendo a las instituciones responsables de financiar y entregar los servicios de salud a la población. En base a esta definición se dice que se agrega valor cuando es posible disminuir los costos asociados al prestar dicho el servicio.

De acuerdo a lo antes mencionado, es importante identificar los roles que intervienen y las actividades que los relacionan, lo que permite finalmente realizar un trabajo colaborativo que agregar valor, principalmente con un impacto en disminuir los costos.

Al analizar el primer punto, es posible identificar los siguientes roles, que se muestran en la figura 4.2:

- **Salud Pública**: A través del Ministerio de Salud, que permite coordinar los esfuerzos necesarios para entregar el servicio a la población. Y por otra parte el Fondo Nacional de Salud (FONASA), que entrega el financiamiento.

- **Establecimientos de Salud**: Son las instituciones que prestan los servicios a la población, y los encargados de generar la información clínica asociada a cada paciente, la que se almacena actualmente en los sistemas de Registro Clínico Electrónico.

- **Academia**: Se hace cargo de la formación de profesionales y de realizar trabajos de investigación y desarrollo. Existe un nexo natural con las escuelas de salud pública. Sin embargo, también se han establecido relaciones con las áreas de ingeniería y tecnologías de información.

- **Personas**: Por una parte, las personas entregan su información en cada encuentro médico que se realiza en un establecimiento de salud. Sin embargo, con el aumento del uso de dispositivos móviles inteligentes, las personas, a través de aplicaciones, pueden monitorear sus signos vitales

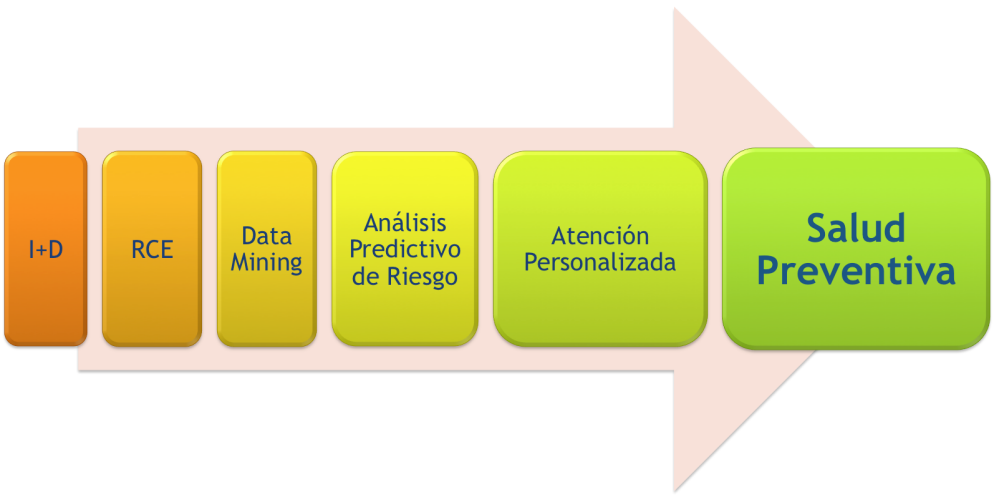


Figura 4.3: Smarter Care: Cadena de Valor

Fuente: Elaboración Propia

en tiempo real, y además pueden entregar información con respecto a su estado de salud, sin necesidad de estar físicamente frente a un médico en un establecimiento de salud. Este último punto, presenta un gran desafío desde el punto de vista tecnológico, para poder soportar la gran cantidad de información que se puede recibir, y además tener la capacidad de procesarla en tiempos razonables para entregar información útil a las personas en tiempo real.

Cabe mencionar que no se está considerando la salud privada, debido a que funcionan de manera diferente, tanto en el manejo de los datos clínicos de sus pacientes como la forma de financiamiento. Sin embargo, si se logra establecer en Chile un repositorio nacional de atenciones y la implementación de la Ficha Clínica Única Compartida, será posible integrar a todas las personas que se atienden en un establecimiento de salud, independiente si este es público o privado. Actualmente 12 millones de personas se atienen en el sistema público, lo que corresponde a más de 70 % del total de la población.

Luego de tener claro los roles presentes, es importante detallar las actividades (Figura 4.3) que son parte de esta cadena de valor, y como se relacionan entre ellas.

1. **Investigación y Desarrollo**: Que consiste en la investigación de ciencias aplicadas o ciencias básicas, para el desarrollo de soluciones que buscan incrementar la innovación y que tenga como resultado final agregar valor a las organizaciones. En particular, hay un fuerte vínculo entre la investigación y desarrollo de ciencias aplicadas en las universidades y las empresas privadas, ya que esto permite generar valor a través de los productos y servicios entregados. En este caso particular, que se trabaja con un organismo público como el Ministerio de Salud, es necesario identificar un tercer rol, además de las universidades y las empresas privadas involucradas,

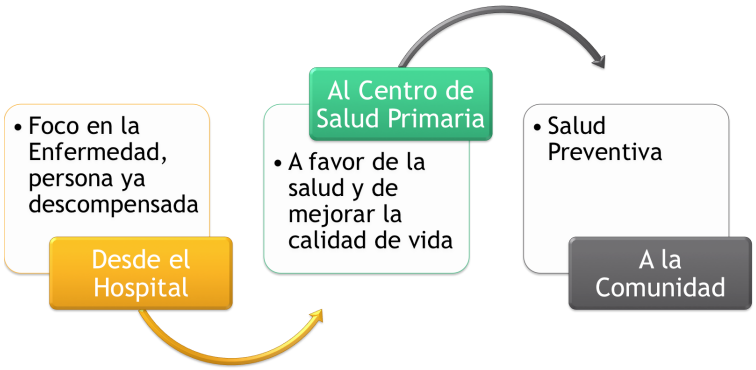


Figura 4.4: Estrategia para fortalecer la Salud Pública

Fuente: Elaboración Propia

que es el Mandante. En esta caso, los establecimientos de salud, o quienes los administren, permiten darle pertinencia a cada uno de los proyectos que se desarrollen, ya que finalmente son ellos quienes llegan al usuario final, que son las personas que se atienden en el sistema de salud.

2. **Registro Clínico Electrónico**: Como ya se mencionaba antes, los establecimientos de salud son los que generan día a día los datos clínicos asociado a los encuentros de una persona con un profesional de la salud. Es aquí donde se encuentra la materia prima para cualquier esfuerzo de estudio que se quiera realizar. En Chile, se da la particularidad que la empresa SAYDEX cubre más del 70 % de todos los establecimientos de Antención Primaria de Salud (APS), lo que se traduce en tener en sus registros a casi 10 millones de personas, de un total de 12 millones que se atienden en el sistema de salud pública. Por lo que esta empresa privada tiene un rol clave en las mejoras a la salud que puedan nacer del análisis de datos de la población.

3. **Data Mining**: Es el proceso no trivial de extracción de información desde los datos, que está presente de forma implícita, previamente desconocida y potencialmente útil para el usuario. Esta información está presente en los datos como patrones, que son muy útiles cuando son aplicados a determinados problemas o algún contexto de negocio en particular. En este caso, en el uso de los datos para el análisis predictivo de riesgo de algún problema de salud. Como ya se ha mencionado en los capítulos anteriores, este trabajo se enfoca en los problemas asociados a las Enfermedades Cardiovasculares.

4. **Análisis Predictivo de Riesgo**: Permite clasificar a personas con alto riesgo, sin tener necesariamente evidencia clínica de la enfermedad, sólo observando sus factores de riesgo presentes. Esta información permite tomar medidas con anticipación, y poder tratarla para

evitar una potencial descompensación. Esta idea es la base de la Salud Preventiva, lo que tiene un impacto directo en los costos asociados, tanto para las instituciones como las personas.

5. **Atención Personalizada**: Actualmente, las atenciones se basan en los datos estructurados que pueden son asociados a cada paciente, y como estas variables se relacionan a las estadísticas médicas. Esto permite determinar un perfil de la persona y determinar su tratamiento. Un problema, es que se está dejando fuera del análisis información valiosa que se almacena en los campos no estructurados, y actualmente los datos utilizados son sólo los que se obtienen cuando una personas está en un establecimiento de salud. Cabe mencionar, que con el uso de portales de salud, y nuevas aplicaciones en dispositivos móviles que permiten a la persona entregar información vital para determinar su estado de salud, sería posible identificar factores de riesgo, sin que la persona asista a un centro asistencial, lo que permitiría entregar una atención personalizada de acuerdo a las condiciones particulares de cada persona. Con esto es posible complementar la información de los datos clínicos, con información sobre los estilos de vida de las personas y sus condiciones sociales.

6. **Salud Preventiva**: Como meta estratégica del Ministerio de Salud, y basado en la experiencia internacional, se ha definido que es necesario fortalecer la salud pública, poniendo foco en mejorar la calidad de vida de las personas, a través de fortalecimiento de la atención primaria, en particular con los programas de salud preventiva, y así llegar a la comunidad con información relevante para las personas quienes con simples cambios de hábitos pueden disminuir en forma significativa el riesgo de sufrir alguna descompensación o llegar a padecer alguna enfermedad crónica (no congénita). Este concepto se resume en la figura 4.4.

En los últimos años se destacan diversas iniciativas poblacionales de promoción de la salud preventiva y de carácter intersectorial liderados por el gobierno de Chile. Entre ellos está el Sistema de Protección Integral a la Infancia, a través de Chile Crece Contigo (ChCC), que se basa en evidencia científica y señala la importancia de la salud en los primeros años de la vida del niño para prevenir las enfermedades no transmisibles en la vida adulta, entre ellas las cardiovasculares. Su objetivo es garantizar igualdad de condiciones para un óptimo crecimiento y desarrollo del niño durante los primeros años de vida en su etapa de mayor vulnerabilidad. La Estrategia Global contra la Obesidad (EGO-Chile), implementada durante el gobierno de la Presidenta Michelle Bachelet y más recientemente, durante el gobierno del Presidente Sebastián Piñera, el programa Elige Vivir Sano (EVS) liderado por la Primera Dama, que se trata de iniciativas de promoción de la salud con la participación de la empresa privada en coordinación con los organismos estatales. Su objetivo

	2003	2004	2005	2006	2007	2008	2009
Gasto Bolsillo	38, 9%	38,9%	39,0%	38,0%	36,6%	36,5%	34,0%*
Prestaciones Salud Mutuales	2, 3%	2,2%	2,2%	2,0%	2,1%	2,1%	2,0%
Gasto ISAPRE	19,9%	19,0%	18,8%	17,8%	18,1%	17,4%	16,6%
Gasto Público Total	38,8%	39,9%	40,0%	42,1%	43,2%	44,0%	47,4%
Gasto Total	100%	100%	100%	100%	100%	100%	100%

Figura 4.5: Gasto Total en Salud según fuente de financiamiento, 2003-2009

Fuente: Cuenta Satélite de Salud, UCSAS, DESAL, MINSAL

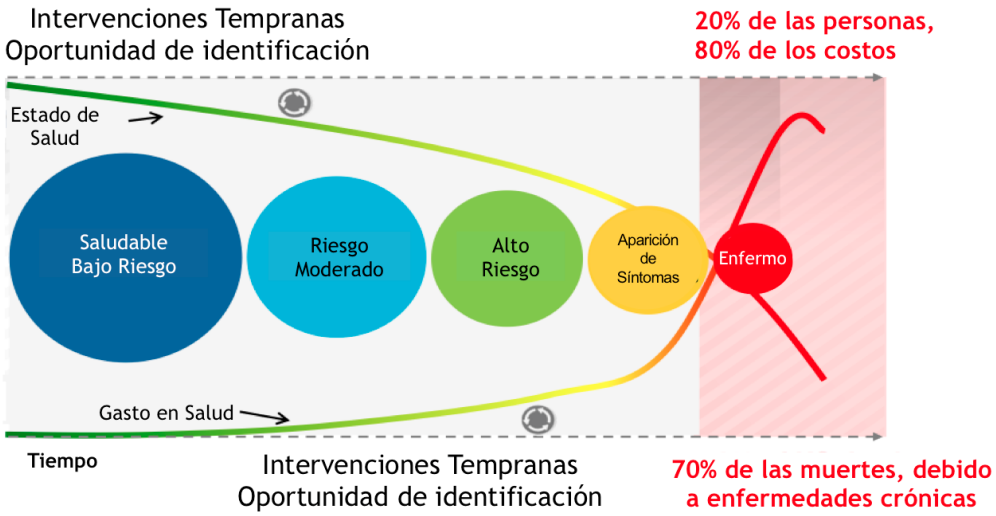


Figura 4.6: Salud Preventiva: Impacto en los Costos

es promover un estilo de vida más saludable entre los chilenos a través de la adopción de una dieta saludable y el fomento de la actividad física, para prevenir la obesidad y las enfermedades no transmisibles asociadas, como las Enfermedades Cardiovasculares.

La salud preventiva y por consiguiente las intervenciones tempranas en los pacientes con un determinado nivel de riesgo, tienen directo impacto en los costos, así también en la calidad de vida de las personas. Cabe mencionar, que del total del gasto en salud en Chile el mayor porcentaje lo tiene, en promedio, el Gasto Público y el Gasto del Bolsillo de las personas, como se muestra en la figura 4.5.

En general, las medidas que apunten a la salud preventiva y a mantener un buen estado de salud de las personas aporta a disminuir los gastos del Estado y los Gastos del Bosillo, ya que es posible evitar los costos relacionados al financiamiento de intervenciones más complejas que se presentan cuando a una persona ya se le declaró una enfermedad y llega a un centro de atención secundaria en una condición crítica ya descompensado. Esta relación, que se muestra en la figura 4.6, muestra el impacto en los costos de acuerdo al estado de salud de las personas y su nivel de riesgo.

Por todo lo antes mencionado, se entiende que las medidas que apunten a potenciar la salud preventiva o detección temprana de riesgo tienen un alto impacto en lo económico.

4.1.2. Estrategias Preventivas para las ECV

En el mundo, las enfermedades cardiovasculares (ECV) lideran las causas de muerte y de invalidez, incluido Chile. Aunque las tasas de mortalidad por estas patologías han disminuido en la última década, en nuestro país ha aumentado la importancia relativa de la ECV sobre el total de mortalidad, siendo de 15 % en 1970 y de 27 % en 2008. Similar a lo que ocurre en otros países subdesarrollados o en vías de desarrollo [41]. En Chile son la primera causa de muerte (27 % del total), y contribuyen a grandes costos en salud [42, 43]. Se estima que estas enfermedades continuarán liderando la pérdida de años/vida hasta el año 2020; actualmente mueren 17 millones de personas al año por estas enfermedades [44]. La enfermedad que es causa de la mayoría de los problemas relacionados a enfermedades cardiovasculares es la ateroesclerosis, que se desarrolla a través de los años, y se encuentra ya avanzada cuando los síntomas se manifiestan. Infartos al miocardio o cerebrales, habitualmente ocurren de forma imprevista y antes de tener acceso a un servicio de salud, por lo tanto, muchas intervenciones médicas pueden ser tardías.

Principales Factores de Riesgo

Los factores de riesgo cardiovascular clásicos tales como **hipertensión arterial, dislipidemia, diabetes, tabaquismo, obesidad y sedentarismo**, continúan siendo los de mayor impacto en la enfermedad cardiovascular, tal como se demuestra en el estudio [41] en que solo 9 factores explican el 90 % de los infartos en hombres y el 94 % en la mujer. El poder identificar aquellos individuos que están en riesgo de presentar en el futuro un evento cardiovascular permitiría poder tratarlos tempranamente y disminuir la morbimortalidad actual. Se estima que más de 50 % de los problemas que originan las ECV podrían evitarse si se logra reducir la incidencia a través de la prevención de sus factores de riesgo [45]. Las modificaciones de los factores de riesgo cardiovascular (FRCV) han mostrado reducir la mortalidad y morbilidad CV, tanto en personas aparentemente sanas (prevención primaria), como en las que ya tienen la enfermedad (prevención secundaria). [46]

Políticas de Prevención en Chile

En Chile, La Estrategia Nacional de Salud 2011-2020 [35] señala las prioridades sanitarias del Ministerio de Salud e incluye metas e indicadores para aumentar los factores protectores de la salud y mejorar la sobrevida de los pacientes que han tenido un evento cardiovascular. Las garantías

explícitas en Salud[1], y la atención primaria (APS) a través del Programa Salud Cardiovascular (PSCV) son iniciativas que contribuyen a mejorar la detección y control de las personas en riesgo.

Esta Estrategia incluye metas específicas dirigidas a reducir la prevalencia[2] de tabaquismo, el consumo perjudicial de alcohol, el sobrepeso y obesidad en población infantil y aumentar la prevalencia de actividad física en adolescentes y jóvenes; adicionalmente, se pretende aumentar la cobertura efectiva del tratamiento de las personas con hipertensión (mantener su presión arterialarterial bajo 140/90 mmHg) y diabetes (HbA1c[3] bajo 7 %), e incluye metas específicas para aumentar la sobrevida de las personas que han tenido un infarto agudo al miocardio (IAM) o ataque cerebral (ACV).

Una de las metas más desafiantes es la de "salud óptima", que propone aumentar la proporción de chilenos con al menos cinco factores protectores de los ocho siguientes:

- No fumar.
- Tener peso normal (IMC < 25).
- Realizar al menos 150 minutos de actividad física moderada a la semana.
- Consumir al menos cinco porciones de frutas o verduras al día.
- Tener presión arterial $< 120/80\ mmHg$.
- Colesterol total $< 200\ mg/dl$.
- Glicemia $< 100\ mg/dl$.

Hoy en día la proporción de chilenos que cumple con este estándar es sólo de un 16,7 %.

Para avanzar en esta línea, se pretende incorporar intervenciones educativas que contribuyan a mejorar la cobertura de los tratamientos, sobre todo, después del alta de una persona que sufrió una descompensación. También, se espera establecer mecanismos eficientes de coordinación entre el nivel de especialidad y el nivel primario de atención (APS), para dar continuidad de la atención.

En este contexto, el Ministerio de Salud declara que es imperativo ampliar, en el caso del IAM, y desarrollar, en el caso del ACV, un sistema de registros clínicos que permita identificar los casos incidentes[4] y monitorear los procesos clínicos y sus resultados. **Los sistemas de registro**

[1] Conocidas como GES o AUGE.

[2] En epidemiología, es la proporción de individuos de una población que presentan una característica en un período de tiempo.

[3] Hemoglobina glicosilada, prueba muy utilizada en la diabetes para saber si el tratamiento ha sido bueno durante los últimos tres o cuatro meses.

[4] Incidencia, corresponde al número de casos nuevos de una enfermedad en una población determinada y en un período de tiempo determinado

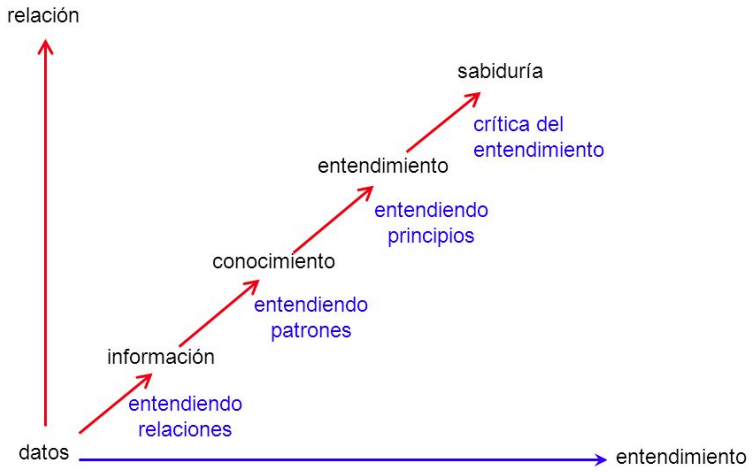


Figura 4.7: Datos, Información, Conocimiento, Entendimiento, Sabiduría

actuales son insuficientes para vigilar y con ello mejorar la calidad de atención y asegurar el mejor pronóstico del paciente. [35]

Las estrategias de prevención, para que sean efectivas, requieren que un país conozca la prevalencia de sus factores de riesgo, y su relación con incidencia de eventos cardiovasculares, como también, el riesgo atribuible a cada factor de riesgo. [47]

Recientemente se están generando datos locales y registros, tales como el registro nacional de infartos GEMI de la Sociedad de Cardiología o la Encuesta Nacional de Salud. Sin embargo hasta ahora, la mayor parte de la información que se ha utilizado proviene de los estudios de prevención cardiovascular, realizados en población europea o norteamericana

En general, un gran problema es que **no hay información disponible para poder realizar gestión**. Suele ocurrir que el gobierno tiene montones de datos, pero muy poca información y menos conocimiento a partir de dichos datos.

Es por esto que todas las medidas que apunten a potenciar la salud preventiva o detección temprana de riesgo tienen un alto impacto en lo económico.

4.1.3. Healthcare Analytics

Es ya conocida la diferenciación entre datos e información y entre información y conocimiento. En el contexto de la Salud, un punto importante en la generación de datos es el Registro Clínico Electrónico, que es donde se almacenan todos los datos relacionados a la atención de un paciente.

Cabe mencionar que actualmente el Ministerio de Salud está realizando esfuerzos en la línea de utilizar los datos del RCE para ser almacenados en un repositorio nacional de atenciones. Parte de la información relevante que se espera guardar de dicho almacén de datos (Data Warehouse) permitirá justificar y fundamentar estrategias preventivas basados en la evidencia empírica de lo

que sucede minuto a minuto en los centros de salud a lo largo de todo el país.

Será posible hacer análisis descriptivos en base a los datos históricos y analizar el impacto de medidas implementadas. También se podrá realizar análisis predictivos, principalmente relacionados a medidas preventivas y focalizando los esfuerzos en determinadas personas. En el más alto nivel, será posible realizar análisis prescriptivos, es decir, a partir del conocimiento extraído desde los datos tomar mejores decisiones con respecto a cuales son las políticas o protocolos más efectivos, de acuerdo a su potencial impacto en la población. Este conjunto de análisis avanzados en el ámbito de la salud, se conoce como Healthcare Analytics.

El punto de partida para todo esto son los datos del Registro Clínico Electrónico.

Importancia del Registro Clínico Electrónico

El RCE está diseñado para procesar la información clínica y administrativa de un paciente, que puede contener datos demográficos, historial médico, información sobre la admisión anterior, información de cirugías, entre otras cosas. El historial médico o notas clínicas, a su vez, incluye el malestar o enfermedad por la cual el paciente acudió al médico y la historia familiar. Además, dentro el RCE se pueden almacenar quién realizó el ingreso, solicitudes y resultados de laboratorio, exámenes de radiología, prescripciones, o la generación de un informe de alta.

Es decir, a partir del RCE es posible almacenar una gran cantidad de datos a partir de todo lo relacionado con los pacientes. Sin embargo, los datos son sólo una base sobre la que construir, pero por sí solos no aportan nada. Sólo cuando comienzan a tener significado y permiten dar un punto de vista comprensible, pasan a ser información útil. Y en su estado superior la información se puede convertir en conocimiento, cuando se puede ver en contexto y permite una comprensión de fenómenos y una toma inteligente de decisiones. Este flujo, desde los datos a la obtención de conocimiento, se muestra en la figura 4.7.

En los últimos años, la masificación del uso de sistemas de información clínicos, como el Registro Clínico Electrónico y algunos dispositivos de salud que permite a las personas monitorear sus signos vitales en tiempo real, han creado una explosión de información sin precedentes. Un estudio de IBM, presenta que el 93 % de los prestadores de salud mencionó la explosión de información como el factor más importante para influir en sus organizaciones dentro de los próximos cinco años [48]. Sin embargo, la abundancia de datos, es un arma de doble filo ya que tiene un enorme potencial para mejorar la toma de deciones, pero a la vez hace que sea cada vez más difícil distinguir entre que datos son esenciales y cuales no entregan información relevante. De hecho, la paradoja de datos [49], que plantea que para alcanzar el punto del conocimiento, e incluso de la información, la cantidad (y la calidad) de los datos importa, ya que con muy pocos es difícil extraer información, pero un exceso

de datos sin estructura muchas veces pasa a ser un problema, ya que el manejo de grandes volúmenes de datos es más difícil y no permite alcanzar el conocimiento esperado. Esta paradoja está siendo un obstáculo cada vez más desalentador para la creación de estrategias de análisis eficaces.

Para sobrellevar esta problemática y esta inmensa complejidad, se requiere de personas con suficiente experiencia y expertos en las materias respectivas para identificar que información es posible extraer e identificar si se cuenta con los datos necesarios, para extraer dicha información. Esto permite tomar decisiones más inteligentes, más informados y finalmente obtener mejores resultados.

Advanced Analytics

Consiste en un conjunto de técnicas, que a partir de información histórica (análisis descriptivos) permite realizar análisis predictivos, simulaciones y optimización.

Advanced Analytics se basa en principios matemáticos y que comienzan con las estadísticas descriptivas que se utilizan básicamente para resumir y contar apariciones pasadas, para utilizar esta información de una manera reactiva e intentar corregir el rumbo. Sin embargo, el uso de técnicas más avanzas, permite anticiparse a posibles resultados futuros y, o tomar acciones en el presente, para impactar el futuro. Cabe mencionar que este último punto, es el fundamento de la Salud Preventiva.

La técnica tradicional para la construcción de un modelo de predicción se basa en la comprobación de hipótesis, este es el enfoque estadístico. Por otro lado, la minería de datos (Data Mining) es una técnica para la construcción de modelos de predicción donde los datos se exploran y se utilizan para determinar qué modelo de predicción se .ªjusta"mejor.

Los organismos de salud, en el mundo, están utilizando cada vez más estos tipos de análisis para consumir, descubrir y aplicar nuevos conocimientos a partir de la información. Nuevos métodos de análisis pueden ser utilizados para impulsar mejoras prácticas clínicas y operativas para cumplir los retos que la salud presenta. Comenzando desde un punto de referencia tradicional, que consiste en el monitoreo de transacciones usando herramientas como el RCE, hacia hacia un modelo que eventualmente permitirá incorporar análisis predictivo y permitir a las organizaciones "ver el futuro". Así, a partir de un análisis predictivo, lograr una atención más personalizada, enfocada en las características propias de la persona, como sus estilos de vida, condiciones sociales, además de sus datos clínicos, y así poder predecir su comportamiento o detectar de forma temprana su posible condición de riesgo frente a algún tipo de enfermedad. (Ver Figura 4.3)

En general, el uso de Advanced Analytics en la Salud, en adelante Healthcare Analytics, tiene como principales objetivos mejorar la calidad del servicio clínico entregado a los pacientes, reducir los costos y aumentar la eficiencia, y además, aportar a mejorar las gestiones administrativas de

dentro de los establecimientos de salud. [50]

A continuación se detallan los objetivos específicos, asociados a cada uno de los objetivos generales antes mencionados:

1. Mejorar la efectividad clínica y la satisfacción de los pacientes.
    - Mejorar la calidad de las atenciones.
    - Mejorar la seguridad de los pacientes y reducir los errores médicos.
    - Mejorar la gestión de la prevención de las enfermedades.
    - Entender los perfiles médicos y su desempeño clínico.
    - Mejorar la satisfacción de las personas al ser atendidas.
2. Mejorar la efectividad operacional.
    - Reducir los costos y aumentar la eficiencia de los recursos.
    - Mejorar el cumplimiento de metas de cobertura, que resultan en pagos por buen desempeño.
    - Aumentar la velocidad de los procesos.
3. Mejorar el desempeño financiero y la gestión administrativa.
    - Mejorar la utilización de los recursos (profesionales de la salud).
    - Optimizar las cadenas de suministros (fármacos, insumos, etc.).
    - Mejorar el cumplimiento de metas administrativas.
    - Reducción de fraudes y mal uso de los recursos.

Actualmente, la visualización de datos, análisis de tendencias y predicciones a partir de los datos históricos, la estandarización de reportes son elementos de análisis que agregan mucho valor. Sin embargo, se estima que en dentro de los próximos dos años, estas necesidades van a cambiar. Mientras la visualización de datos siempre será un elemento fundamental, el énfasis estará en la realización de simulaciones o la creación de escenarios y análisis que involucren varios procesos de negocios que estén relacionados unos con otros. Por ejemplo, en el caso de las enfermedades cardiovasculares, además de realizar análisis predictivos sobre el riesgo cardiovascular, sería interesante evaluar los procedimientos médicos establecidos dentro del Programa de Salud Cardiovascular, que se encarga de controlar y mantener compensadas a las personas que sufren este tipo de enfermedad crónica.

Es posible evaluar la efectividad de los procedimientos considerando, ya que se puede obtener una trazabilidad de la historia clínica de la persona, desde que ingresó al programa, cuando y porqué se descompensó, que medicamentos utilizó, que exámenes fueron solicitados, etc.

Por otra parte, y considerando la opción de realizar simulaciones, es posible aplicar esta técnica para las campañas de vacunación que se realizan todos los años, como la campaña contra la Influenza. Considerando algunos parámetros, es posible simular un escenario para estimar la cantidad y la forma más eficiente de abastecer de las dosis a cada uno de los establecimientos del país y en tiempos oportunos.

En general, es importante tener grandes planes, seguidos de acciones discretas, para obtener los beneficios de Healthcare Analytics. Cabe mencionar, que también se requiere algunos enfoques de gestión muy específicos, por lo que presentan cinco recomendaciones para trabajar con Healthcare Analytics:

1. **Enfocarse en las oportunidades que agreguen más valor**: Buscar un desafío y darle máxima prioridad. Enfocarse en lo necesario que es realizar este esfuerzo, y no en la complejidad o sofisticación del análisis. Para tener claro los resultados esperados, es necesario tener la respuesta a preguntas como "Qué es importante", y "Por qué es importante", estas respuestas permitirán alinear los esfuerzos. Es importante definir indicadores que permitan medir el avance hasta conseguir el resultado esperado.

2. **En cada una de estas oportunidades, comenzar con preguntas, no con datos**: Comenzar con preguntas permite conducir los esfuerzos a una mejor comprensión de la información que se necesita y los datos que se pueden utilizar para generar dicha información. El foco debe estar en los resultados esperados, y después preocuparse de hacer las gestiones necesarias para obtener los datos respectivos.

3. **Integrar conocimientos para impulsar acciones y entregar valor**: Healthcare Analytics por si sólo no tiene valor. Se debe actuar de acuerdo a los conocimientos ya existentes en un determinado contexto. Hay diferentes formas de integrar conocimiento. Una forma es trabajar en una solución puntual, pero que mejore un proceso clave que involucre a muchas personas, o realizar análisis más complejos como análisis predictivos o simulaciones que permitan apoyar la toma de decisiones más complejas y estratégicas.

4. **Mantener las capacidades existentes mientras se añade otras nuevas**: Nunca se van a tener todas las capacidades necesarias para Healthcare Analytics, ya que los requerimientos siempre irán cambiando, y elevando su nivel de complejidad. Esto crea el requisito fundamental

para estas iniciativas, las que deben ser escalables, tanto en sus arquitecturas de tecnología, y en sus requerimientos de recursos humanos.

5. **Utilice una agenda de información para planificar para futuros análisis**: La mayoría de las organizaciones se desean realizar muchos proyectos de Healthcare Analytics de manera simultánea. Sin una visión integradora de cómo las piezas tienen que trabajar juntas, el resultado final podría ser un entorno tan complejo que no puede satisfacer las necesidades de la empresa o uno tan costoso de mantener que la organización debe parar y reducir la cantidad de proyectos.

Finalmente, la aplicación de **técnicas avanzadas de análisis** en la salud, Healthcare Analytics, tiene que ser complementada con una **excelente gestión de proyectos** que permita definir objetivos claros y que integre a las personas claves y necesarias que puedan realizar estas definiciones, ya que los resultados obtenidos, no sólo será un resumen de datos históricos, sino que debe ser información que entregue valor y que pueda ser utilizada para tomar decisiones en procesos estratégicos. Esta combinación de variables define si un proyecto de Healthcare Analytics es exitoso o no.

A continuación, se presenta en detalle el diseño del experimento realizado, el que permite calibrar un modelo de predicción de riesgo cardiovascular, usando los campos de texto no estructurados del registro clínico electrónico de los establecimientos de la atención primaria de salud.

4.2. Diseño del Experimiento

Para llevar a cabo el presente experimento, se estructura una plataforma de trabajo, en adelante Plataforma Healthcare Analytics, comienza con la extracción de datos, transformación y selección de las variables más relevantes de acuerdo a un contexto dado, para luego realizar los análisis predictivos.

4.2.1. Plataforma Healthcare Analytics

Esta plataforma se divide en tres módulo y cada módulo cuenta con procesos internos para el trabajo de los datos.

1. **Extracción de Información**: Módulo en el que se identifican las fuentes de datos, ya sean estructurados o no estructurados.

2. **Selección de Variables**: A partir de una representación estructurada de la unidad en estudio, en este caso, los pacientes, se aplica una base de conocimiento o contexto, para determinar

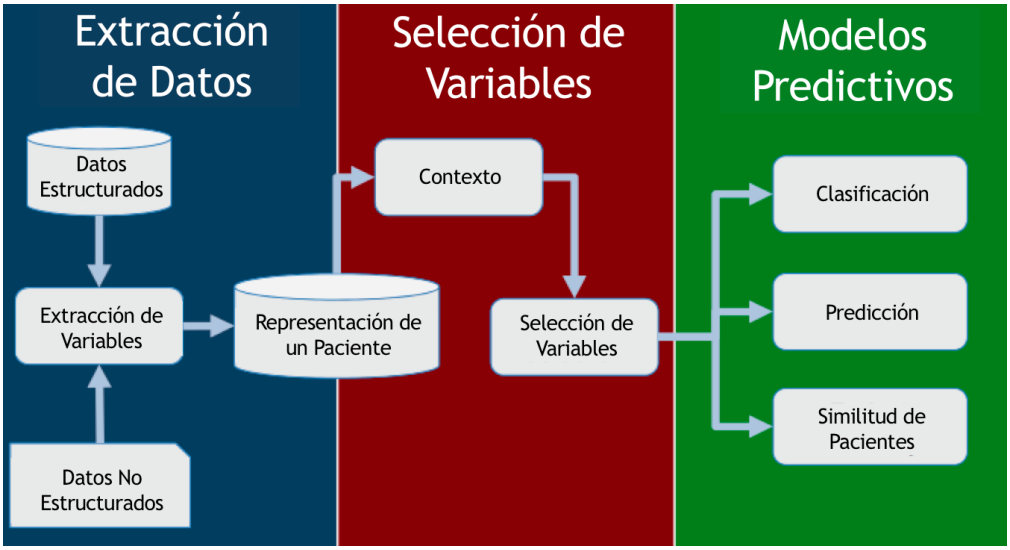


Figura 4.8: Metodología para el trabajo de los datos

cuales son las variables más relevantes para el uso de los modelos predictivos.

3. **Modelos Predictivos**: Una vez procesados los datos, y estructurados, se aplican modelos predictivos, los que permiten realizar diversos tipos de análisis.

Extracción de Datos

Corresponde al primer módulo, como se muestra en la figura 4.8, donde se identifican las fuentes de datos. El presente trabajo tiene su foco en el análisis de los datos que provienen de los campos de texto no estructurados de las notas clínicas del Registro Clínico Electrónico. Sin embargo, se utilizan fuentes de datos tanto estructurados y no estructurados.

- **Datos Estructurados**: Que provienen de campos estructurados del Registro Clínico Electrónico, como datos demográficos, asistencia a las atenciones, diagnósticos previamente asignados, entre otras cosas.
- **Datos No Estructurados**: Que provienen de los campos de texto no estructurados de las notas clínicas del Registro Clínico Electrónico. El valor agregado de estos campos, es que contienen detalles de la atención de cada paciente, los que no pueden ser almacenados en otros campos estructurados del RCE.

En resumen, se debe realizar un proceso de extracción, transformación y carga de datos (ETL, por sus siglas en inglés), y que a partir del los conjuntos de datos estructurados y no estructurados, se pueda extraer las variables que serán utilizadas para el análisis.

Finalmente, se obtiene una representación estructurada del paciente, que incluye ambas fuentes de datos que se mencionaron en un comienzo.

Selección de Variables

En el segundo módulo que se muestra en la figura 4.8, se realizan dos procesos muy importantes, y que son la antesala al análisis con los modelos predictivos. En este módulo se propone realizar un balance entre la enorme cantidad de datos disponibles, y los datos realmente necesarios. Es aquí, cuando personas que conocen muy bien el negocio deben intervenir y jugar un rol fundamental para poder discernir cuales son los datos que se deben utilizar para ser procesados posteriormente.

A partir del resultado del módulo anterior, que es la representación estructurada de un paciente, se lleva a cabo dos procesos:

1. **Contexto**: Es la base de conocimiento que debe ser aportada por parte de las personas que conocen del negocio. En este trabajo, personas con perfil clínico jugaron un fundamental en este proceso, para definir los datos más relevante a utilizar, y además cuales son las preguntas más interesantes que buscan respuesta.

2. **Selección de Variables**: Complementado la estructuración de la información de los pacientes con el contexto entregado por las personas clínicas, es posible definir criterios para seleccionar las variables que entregan mayor valor para posteriores análisis.

Se darán más detalle de cada uno de los procesos mencionados en el capítulo 5.

Modelos Predictivos

Luego de seleccionar el conjunto de datos a analizar, realizar el preprocesamiento de los datos y transformar estos datos en las variables que se utilizarán, se aplican los métodos de minería de datos propuestos para realizar tres tipos de análisis:

- **Clasificación**: Presentado como Análisis de Contenido, ya que se exploran los datos para identificar patrones. Estos patrones permite extraer información a partir de los datos, y ser un complemento al análisis predictivo propiamente tal.

  En el contexto del presente trabajo, que tiene relación con información clínica, cabe mencionar cierta diferencia con trabajos similares en otros ámbitos. Por ejemplo, cuando se trabaja en la industria financiera y se desea analizar el riesgo de crédito, este tipo de análisis, permite identificar variables que no son tan evidentes a priori, o construir "scoring", a partir de las variables más significativas que se extraen del análisis, lo que permite realizar clasificaciones de

las personas de acuerdo a su riesgo, en este caso, de pagar o no pagar su crédito. Sin embargo, en el contexto del presente trabajo, este tipo de análisis no pretende descubrir información que no se encuentre bien documentada en los libros de medicina, sino que busca validar la calidad de los datos registrados y que sean consistentes con toda la teoría médica ya bien conocida y estudiada por siglos. De todas formas, el valor de este proceso es demostrar como el análisis de contenido de los datos clínicos de una muestra de personas, permite una máquina, a través de algoritmos matemáticos, "entender"los conceptos claves de medicina en un determinado contexto.

- **Predicción**: Es el centro del análisis del presente trabajo. Como complemento al análisis de contenido antes presentado, en este proceso se utilizan todas las técnicas de Data Mining, en particular las que permiten realizar reducciones de dimensionalidad, para cambiar de espacios vectoriales, donde los modelos predictivos se puedan desempeñar mejor. En esta parte, es posible graficar como una caja negra, ya que estas técnicas transforman los datos y las variables, para lograr ajustarse de mejor forma al resultado que se busca predecir. Cabe mencionar que los métodos y algoritmos utilizados en este trabajo, permiten realizar el camino inverso y reconstruir los datos a partir de los resultados obtenidos por los modelos predictivos, a diferencia de otro tipo de métodos, como las conocidas Redes Neuronales, que suelen tener un buen desempeño, pero que realmente son una "caja negra", ya que no es posible interpretar los resultados de los coeficientes de cada variable luego de calibrar el modelo.

- **Similitud de Pacientes**: Es básicamente la agrupación de diferentes pacientes, de acuerdo a características similares en sus variables explicativas. Es posible calcular una similitud entre ellos, en el caso particular del presente trabajo, de acuerdo a los términos que aparecen en sus notas clínicas que se encuentran en los campos de texto no estructurados del registro clínico electrónico.

En el capítulo 5 se presentan los detalles de análisis realizado, técnicas utilizadas y los resultados obtenidos en cada uno de los procesos antes mencionados.

4.2.2. Resultados Esperados

De acuerdo a lo antes mencionado, y a modo de síntesis de toda la información entregada, se declaran a continuación los resultados esperados del presente proyecto, y como estos resultados agregan valor a la gestión de la salud pública en Chile.

Oportunidad que agrega valor

Para determinar que oportunidad agrega valor, es necesario responder estas dos preguntas:

- **Qué es importante**: Como en toda organización, lo más importante para cada gobierno, es lograr una distribución eficiente de los recursos, en este caso particular, para la salud. Uno de los principales problemas con respecto al financiamiento en Salud, son los recursos entregados a través de las Garantías Explícitas de Salud (GES / AUGE), que se gastan en los tratamientos de las Enfermedades Crónicas. En particular, en las Enfermedades Cardiovasculares, ya que están relacionadas a otras enfermedades con una alta prevalencia, como la Hipertensión Arterial, la Diabetes Mellitus, la Dislipidemia y el Tabaquismo.

- **Por qué es importante**: Si se logra reducir el número de casos de personas que están descompasadas o de forma equivalente, aumentar la cobertura y las proporciones de personas que se mantienen en control y compensadas a través de su incorporación a los programas de salud preventiva, esto se traduce directamente en una reducción de costos para el gobierno y en mejorar la calidad de vida de las personas. Como ya se mencionó en el anteriormente, enfocar los esfuerzos en estrategias preventivas es más costo-eficiente que esperar que la persona llegue al hospital con la enfermedad ya declarada.

La prevención mediante la gestión de la detección temprana de riesgo, permite evitar que una persona llegue a una condición de crónico (prevención primaria), o en caso de ya padecer alguna de estas enfermedades, es importante que la persona se mantenga compensada (prevención secundaria), y así reducir los costos, tanto para el Estado que debe financiar los tratamientos de una pacientes descompensados, como para la misma persona, la que puede acarrear costos en empeorar su calidad de vida o incluso fallecer.

De acuerdo a todo lo antes mencionado, el presente trabajo pretende ser un aporte a la salud preventiva, ya que se propone realizar un análisis predictivo, para detectar de forma temprana el riesgo cardiovascular. Bajo la hipótesis de que existe información valiosa en campos de texto no estructurados del registro clínico electrónico, se utilizarán estos datos para calibrar un modelo predictivo, incorporando también variables a partir de fuentes de datos estructuradas.

A modo de síntesis, se propone **Generar un modelo predictivo, basado en Machine Learning (ML) y Natural Language Processing (NLP), que a partir de signos y síntomas detectados en los campos de texto no estructurados del Registro Clínico Electrónico, pueda predecir niveles altos de riesgo cardiovascular de una persona.**

Criterio	Valor Esperado	Observación
AUC	> 0.80	Mide sensibilidad vs (1-Especificidad)
Accuracy	> 0.85	Mide la clasificación de las clase positiva y negativa
F-Meause	> 0.80	Mide la precisión y la sensibilidad (Precision y Recall)
Sensibilidad	> 0.90	También conocida como Recall Clase Positiva
Especificidad	> 0.80	También conocida como Recall Clase Negativa

Cuadro 4.1: Resultados Esperados Modelo de Predicción Temprana de Riesgo Cardiovascular

Fuente: Elaboración Propia

Preguntas a Responder

Con respecto a la utilidad del modelo predictivo, y cómo puede agregar valor a la información que existe actualmente. Además se menciona cómo medir el cumplimiento:

- El modelo predice un número aceptable de casos, de acuerdo a los criterios en la tabla 4.1.
- Es consecuente con la medida de Riesgo Cardiovascular de Framingham, es decir, predice con riesgo, a las personas clasificadas con previamente por el Score de Framingham como Alto y Muy Alto. Considerar la tabla 4.1, excepto AUC.
- Agrega valor por sobre la clasificación de Riesgo Cardiovascular de Framingham, es decir, predice con riesgo, a las personas clasificadas previamente por el Score de Framingham como Moderado y Bajo. Considerar la tabla 4.1, excepto AUC.
- Permite priorizar personas que no tienen una clasificación de Riesgo Cardiovascular de Framingham, es decir, predice con riesgo, a las personas sin tener clasificación previa por el Score de Framingham. Considerar la tabla 4.1, excepto AUC.

Conocimiento para impulsar acciones

De acuerdo a lo antes mencionado, el tema más importantes en la gestión de la salud, es la forma de financiamiento. Uno de los criterios utilizados es la evaluación del desempeño de las comunas y de los establecimientos dentro de cada una de estas, de acuerdo a un conjunto de "INDICES DE ACTIVIDADES DE LA ATENCIÓN PRIMARIA", en adelante IAAPS. Los que establecen un conjunto de ámbitos a evaluar, con sus respectivos indicadores y que funcionan aplicando rebajas en el financiamiento ante los incumplimientos. Las prestaciones que se evalúan se definen en el Decreto 94 de fecha 20 de diciembre de 2013, que determina el aporte estatal a municipalidades, firmado por: el Ministerio de Salud, el Ministerio de Hacienda y la Subsecretaría de Desarrollo Regional, respectivamente. [51]

Uno de los indicadores claves, está relacionado con la cobertura de los programas de salud preventiva y de los programas asociados a enfermedades crónicas. El conocimiento que se obtiene a partir de la información que se extrae del análisis de estos indicadores, y las falencias que se detectan, permite identificar donde poner foco y en que actividades tomas acciones con mayor prioridad.

Parte de la información que se obtiene es:

- La baja cobertura de los programas de salud preventiva.
- La baja cobertura del programa de salud cardiovascular.
- La falta de información disponible para poder gestionar las medidas preventivas y evaluar sus resultados, ya que estos indicadores no permiten tener una mirada a nivel paciente (son sólo contadores).
- El aumento en los costos que implica tener bajas tasas de cobertura. Ya que luego se traduce en un aumento de la prevalencia de las enfermedades crónicas.

El modelo de detección temprana de riesgo cardiovascular, sería una gran aporte, ya que permitiría priorizar pacientes, para saber a quienes se debe prevenir con más anticipación. Al personalizar las atenciones, de acuerdo a la gestión de casos de cada persona, y que a través de algún medio, el centro de salud se ponga en contacto con la persona, en caso de detectar un alto riesgo, sería posible aportar a mejorar la sensación de calidad de servicio. Por otra parte, puede aportar a evitar impericias de los médicos, los que pueden dejar de analizar a personas que potencialmente pueden tener alto riesgo cardiovascular, analizando su historia clínica.

Como este trabajo agrega valor

Actualmente, el método utilizado para la detección de riesgo cardiovascular son las Tablas de estratificación[5] basadas en factores de riesgo. El objetivo es poder mejorar la predicción del riesgo y ayudar a definir a quienes tratar. [52]

Aunque esta evaluación del riesgo cardiovascular ayuda a definir a quienes tratar, muchas veces al médico no le es fácil de definir, especialmente cuando el nivel de riesgo del paciente no es tan alto. El problema que existe es que, si bien la proporción de eventos coronarios es mayor en el grupo catalogado como de “alto riesgo”, dado que la población es más amplia en los niveles de riesgo moderado y bajo, el número total de eventos es mayor en estos niveles, aunque en menor proporción.

[5] Tablas de Framingham

De acuerdo a esto, es posible medir el aporte que entrega el modelo presentado en este trabajo, de acuerdo a:

- **Salud Preventiva**, dado que el modelo va en la línea de la estrategia de salud de la década que propone el gobierno. Donde se espera potenciar la atención primaria de salud, y en particular todas las medidas que aporten a la salud preventiva.
- **Agregar Valor**, al tener una alta tasa de predicción de niveles altos de riesgo cardiovascular en las personas que están clasificadas en niveles de riesgo moderado y bajo.
- **Consecuente**, al predecir con riesgo alto, a quienes ya están clasificados con riesgo cardiovascular alto y muy alto.
- **Priorizar**, al tener un alto nivel de predicción en las personas que no estaban siendo controladas a través del programa de salud cardiovascular, es decir, que no tenían ninguna clasificación de riesgo asignada antes de sufrir una crisis.

Capítulo 5

Experimento y Resultados

De acuerdo a la metodología de trabajo planteada para el presente trabajo, en este capítulo se desarrollaran las etapas restantes del CRIPS-DM, que corresponde a la Preparación de los Datos, Modelamiento y Evaluación.

Con relación a los módulos planteados en la estructuración de la Plataforma de Healthcare Analytics, la Preparación de los datos corresponde al módulo de Extracción de Datos y parte de la Selección de Variables. El módulo de Modelos Predictivos, contiene los procesos de Modelamiento y Evaluación. Con el objetivo de orientar la lectura y entender con mayor claridad la aplicación de cada una de las metodologías mencionadas, se presenta el siguiente esquema en la figura 5.1.

Cabe mencionar la diferenciación entre el Análisis de Contenido y el Análisis Predictivo. En parte, como se mencionó en el Capítulo 4, el Análisis de Contenido, en el contexto del presente trabajo, tiene por objetivo validar la calidad de los datos, permitiendo llegar a conclusiones que coincidan con la base de conocimiento médico existente y ampliamente documentada. Por otra parte, el Análisis Predictivo, tiene por objetivo lograr procesar toda la historia clínica del paciente, y de forma automatizada detectar el riesgo cardiovascular, realizando todas las transformaciones en los datos que sea necesario para lograr una mejor precisión al momento de predecir el nivel de riesgo del paciente.

5.1. Preparación de los Datos

Para este módulo, la preparación de los datos se realizó con diferentes herramientas, desde la extracción de los datos desde su fuente, el preprocesamiento y la transformación.

5.1.1. Extracción de Datos

A partir de una base de datos SQL Server, se extrajo datos tanto estructurados como no estructurados (campos de texto de las notas clínicas), para luego ser procesados en Excel y R, respecti-

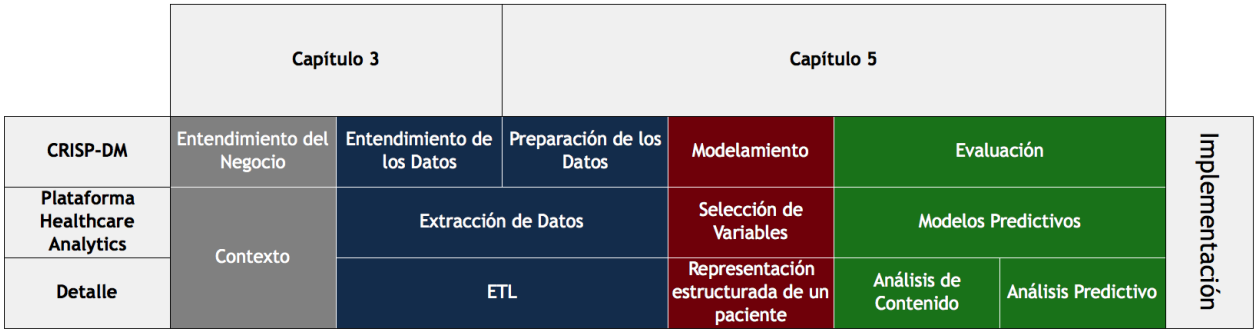


Figura 5.1: Metodologías de Trabajo

Fuente: Elaboración Propia

vamente.

Identificar los datos

Para este proceso, es sumamente importante tener claro la estructura de los datos en que se almacenan las transacciones del registro clínico electrónico. Esto se detalla en el capítulo 3.

Las tablas de la base de datos que se utilizan son:

- ATENCION: Registro de los detalles de las transacciones asociadas a una atención.
- ANAMNESIS: Registro de las notas clínicas.
- DIAGNOSTICO: Detalle de los Diagnósticos.
- NODO: Detalle de los establecimientos de salud.

En general, se pueden agrupar los campos extraídos en tres categorías: Datos Personales, Datos Demográficos, Datos de la Atención, Datos de Texto No Estructurados (Anamnesis, Motivo de Consulta y Notas Clínicas).

Los detalles de los datos extraídos son:

- Un total de 14 mil atenciones entre el año 2010 y 2013.
- Se tomó un subset de datos, que corresponde a los establecimientos de un determinado Servicio de Salud.
- Para poder realizar la comparación con el estudio [34], sólo se consideró a personas entre 50 y 79 años.
- Se obtuvo un total de 461 pacientes.

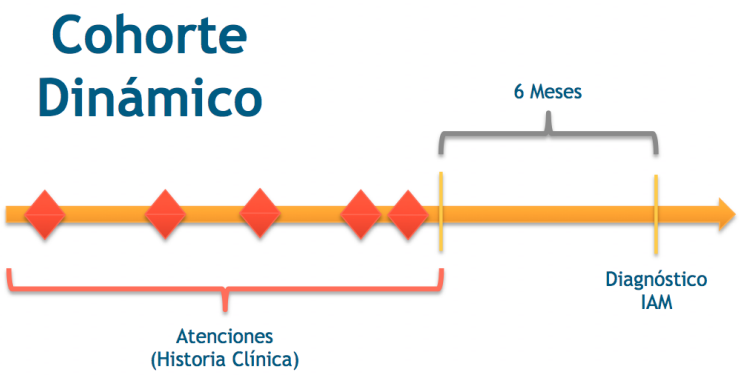


Figura 5.2: Extracción de Datos: Cohorte Dinámico

Fuente: Elaboración Propia

- Del total de pacientes, 280 tuvieron un diagnóstico de Infarto Agudo al Miocardio[1] durante el año 2013. El resto es de las personas (181), corresponde a una muestra aleatoria para balancear el set de datos total.

Cabe mencionar, que de acuerdo a la normativa actual, existe una clasificación de riesgo global, que define que si una persona sufre un infarto agudo al miocardio, independiente del nivel de riesgo que se pueda calcular utilizando las tablas de estratificación de Framingham, SIEMPRE se debe clasificar con Riesgo Cardiovascular Muy Alto.

Un punto importante es aclarar los criterios definidos para realizar la extracción de las 14 mil atenciones, que permitió definir un cohorte dinámico, como se muestra en la figura 5.2.

1. **Pacientes con Riesgo Cardiovascular**: De las personas que en una determinada fecha presentaron un diagnóstico de IAM, se tomaron todas las atenciones de dicho pacientes, realizadas 6 meses antes de la fecha de diagnóstico. Estas personas fueron clasificadas dentro de la clase positiva (Riesgo Cardiovascular Muy Alto).

2. **Pacientes Aleatorios**: Se tomó una muestra aleatoria de pacientes, que no presentaran un diagnóstico de IAM. Estas personas fueron clasificadoras dentro de la clase negativa (Sin Riesgo Cardiovascular).

   El detalle de la extracción de los datos se entrega en el Anexo E , donde se detalla la consulta a la base de datos SQL Server, que contiene los datos.

A continuación se detallan los campos que se cargaron en la sábana de datos.

[1]Se consideraron todos los códigos CIE-10 relacionados a una IAM, que corresponde a 19 diagnósticos del grupo I21

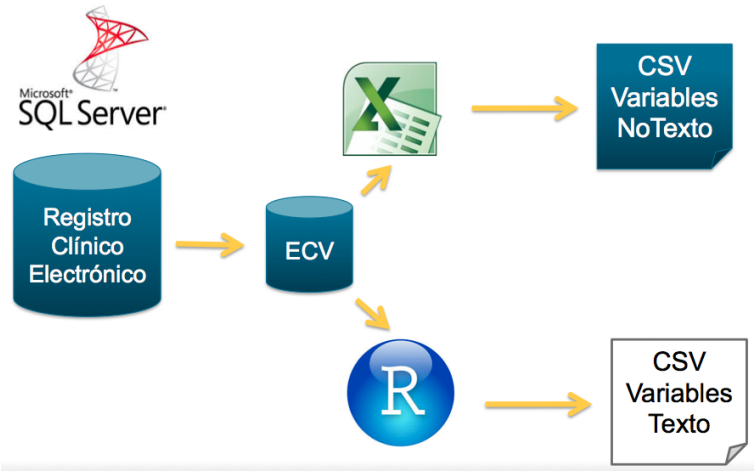


Figura 5.3: Herramientas usadas para el procesamiento de datos

Fuente: Elaboración Propia

- **Datos Personales y Demográficos**: Identificación del Paciente, Sexo, Edad.
- **Datos de la Atención**: Se consideró el campo estructurado que define el tipo de atención, si corresponde a una una atención de urgencia o no y el diagnóstico asignado por el médico. Además, para cada paciente, se tiene el campo correspondiente al Riesgo Cardiovascular, en el caso de tener disponible dicho campo. Este valor proviene de los controles de las personas que pertenecen al Programa de Salud Cardiovascular (PSCV) o que fueron evaluados con algún Examen de Medicina Preventiva (EMP).
- **Fechas**: Se puede determinar la fecha de cada atención, y en particular la fecha en que se diagnosticó a una persona con un infarto agudo al miocardio (IAM).
- **Datos de Profesionales de la Salud**: Identificación de los profesionales que realizaron la atención.
- **Datos de los Campos de Texto No Estructurados**: Se tomaron los campos donde los médicos pueden agregar sus observaciones en cada atención, en particular los campos de: Anamnesis, Motivo de Consulta y Notas Clínicas.

Preprocesamiento de los Datos

De acuerdo al esquema presentado en la figura 4.8, es necesario realizar un preprocesamiento de los datos para **extraer variables a partir de los datos**, tanto estructurados (No Texto) como no estructurados (Texto).

Para llevar a cabo este preprocesamiento, se utilizados diferentes herramientas (Figura 5.3) , como se detalla a continuación:

1. **Excel**: Se utilizó para extraer variables de comportamiento como, la cantidad de atenciones de cada paciente en determinados rangos de fecha, de acuerdo a su última atención. Para el caso de los pacientes clasificados dentro de la clase positiva (Alto Riesgo), esta última atención corresponde a su diagnóstico de IAM. Además, para cada paciente, se buscó a aparición de determinados diagnósticos en el campo estructurado correspondiente.

2. **R**: Con el uso de sus librerías "tm", y "qdap"se trabajaron específicamente los campos de texto no estructurados. Para realizar los procedimientos de limpieza de Text Mining y luego conseguir una forma estructurada de toda sus notas clínicas.

    - En primer lugar se normalizó para quitar todos los tildes de las palabras, y además quitar todo tipo de puntuación.
    - Luego se llevaron todas las palabras a minúsculas.
    - Se realizó el proceso de tokenización por palabras de cada una de las notas clínicas para cada paciente.
    - Se quitaron las StopWords, es decir, dichos términos que por lo general se repiten mucho dentro del texto pero que no agregan valor al análisis.
    - La librería "qdap", permitió sintetizar cada una de las atenciones por pacientes, y transformar sus notas clínicas en una matriz de frecuencia de términos.
    - Con el objetivo de normalizar, debido a que las personas que tenían mayor cantidad de atenciones, presentaban mayor frecuencia en cada uno de sus términos, se procesó la matriz para transformarla en una matriz binaria. En este caso, se da importancia a la aparición de un término en su historia clínica, y el peso de este posible signo o síntoma no va a depender de la cantidad de atenciones de un paciente. Esto se debe a una práctica que tienen los médicos al momento de registrar sus notas clínicas en el RCE, que es copiar y pegar las notas de una atención a otra, y solo agregar observaciones nuevas. Es por esto, que si no se aplicaba la normalización para convertir la matriz de términos a valores binarios, la muestra podría estar distorsionada por esta práctica que ocurre al momento de la atención. Con esta simple medida, se evita sobre ponderar un término que aparece en la historia clínica de un paciente.
    - Para identificar los conceptos que entregan mayor información, se aplicó POS Tagging, que permite identificar la función gramatical de las palabras dentro de un texto.

El detalle del procesamiento en R, se adjunta en el Anexo D .

El resultado final es una representación estructurada de cada paciente, que corresponde a una matriz, donde cada fila es una persona y cada una de las columnas corresponde a variables a partir de sus datos Personales, Demográficos, de Comportamiento, de la Atención y Datos de Texto No Estructurados (Anamnesis, Motivo de Consulta y Notas Clínicas).

- **Variables Personales y Demográficos**: Identificación del Paciente, Sexo, Edad.
- **Variables de Comportamiento**: Cantidad de atenciones en los intervalos de tiempo: Entre 6 a 12 meses, entre 12 a 18 meses, entre 18 y 24 meses.
- **Variables de Diagnósticos Anteriores**: Corresponden a variables binarias que identifica si dicho paciente tuvo un determinado diagnóstico en alguna de sus atenciones.
- **Variables a partir del Texto No Estructurado**: Corresponde a la fila en la matriz de términos para cada paciente, donde se consolida toda su información almacenada en texto, y se presenta cada uno de los términos que aparece en su historia clínica, consolidando todas las atenciones.
- **Variable Objetivo**: Para las personas que tuvieron un diagnóstico un diagnóstico de IAM (Riesgo Cardiovascular Muy Alto) es 1, en caso contrario 0. Esto determina al presente trabajo como un Problema de Clasificación Binaria.

5.1.2. Selección de Variables

Este módulo permite determinar cuales de todas estas variables antes definidas son realmente necesarias para el posterior análisis predictivo. Es importante mencionar que, para realizar la selección de variables es muy importante el Contexto que deben entregar las personas que tienen un conocimiento acabado de las problemáticas de negocio. En el caso particular del presente trabajo, corresponde a médicos con muchos años de experiencia en la gestión de la salud pública. Ellos son los que aportan con base de conocimiento que complementa el análisis de los datos, como se muestra en la figura 5.4. Para este trabajo se contó con el apoyo y respaldo de la Gerencia Clínica de la empresa SAYDEX.

Cabe recalcar que el foco del presente trabajo es el análisis de los campos de texto no estructurados, es por esto que se trabajó con dos subconjuntos de datos, como se muestra en la la figura 5.3. Luego, se obtuvieron variables a partir de datos estructurados (se le llamó No Texto) y datos no estructurados (Texto).

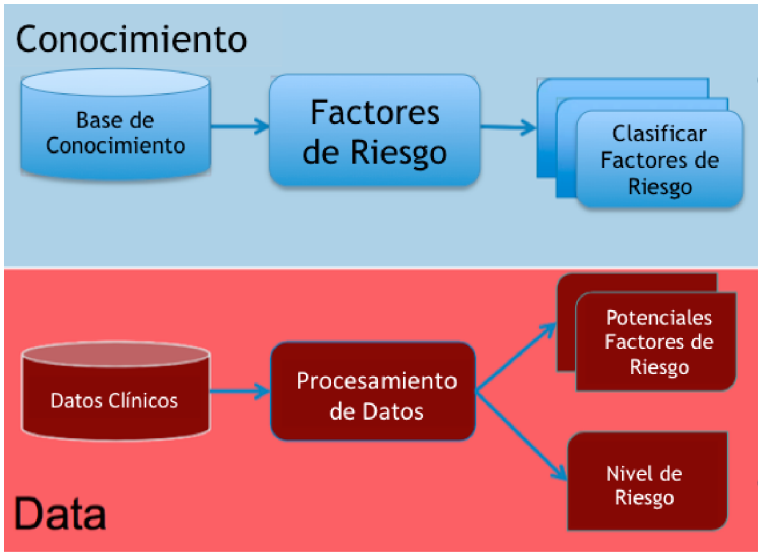


Figura 5.4: Detalle de la Extracción de los Datos

Variables No Texto

Corresponden a variables que se tomaron de los campos estructurados del registro clínico electrónico, como se define en [34]. Estas variables corresponden a:

- Variables Personales y Demográficas, que permite identificar grupos de personas de acuerdo a su sexo y edad. Es importante este conjunto de variables, ya que está demostrado desde el punto de vista médico, que son las variables más relevantes para realizar una estratificación de la población, y que tienen un valor muy explicativo al momento de analizar las enfermedades cardiovasculares. Por ejemplo, que hay estudios [53] que consideran la variable raza (por ejemplo, si una persona es afroamericano, asiático), debido a que los factores de riesgo impactan de diferente manera a personas de algún determinado origen étnico.

- Variables de Comportamiento, se agregan al análisis ya que existe evidencia que los encuentros médicos aumentan su frecuencia con respecto a la fecha de diagnóstico de alguna enfermedad cardiovascular.

- Variables de Diagnósticos Anteriores, como Diabetes Mellitus Tipo 2 (DM2), Enfermedad Pulmonar Obstructiva (EPOC), Hipertensión Arterial (HTA), Accidente Cerebro Vascular (ACV), Fibrilación Auricular, Enfermedad Vascular Periférica, que son considerados como factores de riesgo cardiovascular.

Se consideró un total de 11 variables de acuerdo a los criterios antes mencionados.

Concepto	Detalle
hta	Hipertensión Arterial
hipertension	Hipertensión Arterial
diabetes	DM2
aas	Aspirina
cardiopatia	Insuficiencia Cardiaca
insulina	DM2
tabaco	Fumador
ave	AVE
eeii	Edema, hinzachon piernas
enalapril	Farmaco para HTA
sal	IC empeora con comida salada
hidroclorotiazida	Fármaco HTA
carvedilol	Fármaco para IC
ekg	Electrocardiograma
tabaquismo	Fumador
hipotiroidismo	Causa o Constribuya a IC
ecg	Electrocardiograma
hiperplasia	Aumento tamaño organo (hígado)
ic	Insuficiencia Cardiaca
diabetico	DM2
sibilancias	Estertores, líquido en los pulmones
enalaprila	Fármaco HTA

Figura 5.5: Conceptos relacionados a Enfermedades cardiovasculares

Fuente: Elaboración Propia, validado por médicos

Variables Texto

Corresponde a la matriz de términos que resultó del preprocesamiento de los datos a partir de las notas clínicas realizadas en cada atención y que complementado con el contexto entregado por los médicos, permitió definir las variables más relevantes a partir del texto.

Cabe mencionar, que no se utilizaron todos las columnas, o conceptos, de la matriz de términos.

- Dado que se cuenta con la etiqueta POS para cada concepto, se asumió que sólo los verbos, sustantivos, adjetivos, adverbios agregan mayor valor al análisis.
- A partir de los más de 2000 conceptos que resultaron luego del filtro antes mencionado, se realizó una selección manual, agregando el punto de vista médico, para identificar las variables más importantes en relación al riesgo cardiovascular, como se muestra en la figura 5.5. El resultado fue la selección de 248 conceptos relacionados con las enfermedades cardiovasculares.

5.2. Análisis de Contenido

En esta sección se realizará el análisis de los datos que fueron estructurados como matriz, donde cada fila representa a cada persona y las 248 columnas considera a las variables a partir de los datos no estructurados de su historia médica almacenada en el Registro Clínico Electrónico. Se utilizó sólo las variables a partir del texto no estructurado, ya que ese es el foco del presente trabajo.

Este análisis es parte del tercer módulo que se muestra en la figura 4.8. En particular, los procesos relacionados a Clasificación y Similitud de los pacientes.

En general, este tipo de análisis se conoce como Aprendizaje No Supervisado. En el caso particular del presente trabajo, es necesario trabajar con técnicas que se comporten bien al manejar un gran número de variables o dimensiones, y que permitan entregar información realizando asociaciones de estas variables, a través de funciones que permitan disminuir la dimensionalidad del problema, creando lo que se conoce como variables "latentes".

5.2.1. Diseño Análisis de Contenido

Esta etapa, tienen por objetivo validar la calidad de los datos que se encuentran almacenados en el Registro Clínico Electrónico, y que estos sean consecuentes con los conocimientos de medicina ya bien estudiados, en relación a las enfermedades cardiovasculares.

Las principales técnicas utilizadas en este tipo de análisis son: Clustering, Componentes Principales y Reglas de Asociación.

Clustering, permite encontrar multiples clases o regiones en el espacio de variables en que se está trabajando, y además, representarlas de una forma más simple, combinando a todos los elementos de cada clase. De acuerdo a la técnica usada en este trabajo (K-Means), esta representación se conoco como centroide, y permite describir al conjunto de individuos pertenecientes a una misma clase o cluster.

Componentes Principales, intenta reducir la dimensionalidad del problema, con la generación de vectores que permitan representar de mejor forma la misma información. Estos vectores son conocidos como Componentes Principales.

Por otra parte, las reglas de asociación, permiten construir descripciones y relaciones más simples de los datos, cuando se tienen problemas con muchas variables o dimensiones, especialmente cuando se cuenta con variables binarias, como es el caso del presente trabajo.

Clustering

K-Means es un método que permite definir cluster y encontrar sus centros a partir de un conjunto de datos. Es necesario predeterminar el número de cluster que se quiere obtener, antes de correr el algoritmo, el que de forma iterativa mueve los centros para minimizar la varianza total de cada cluster.

Dado un número inicial de cluster, el algoritmo K-Means alterna entre dos pasos:

- Para cada centro, se identifica un subconjunto de puntos que se encuentran más cerca de él que de los otros centros. Esto forma un cluster.

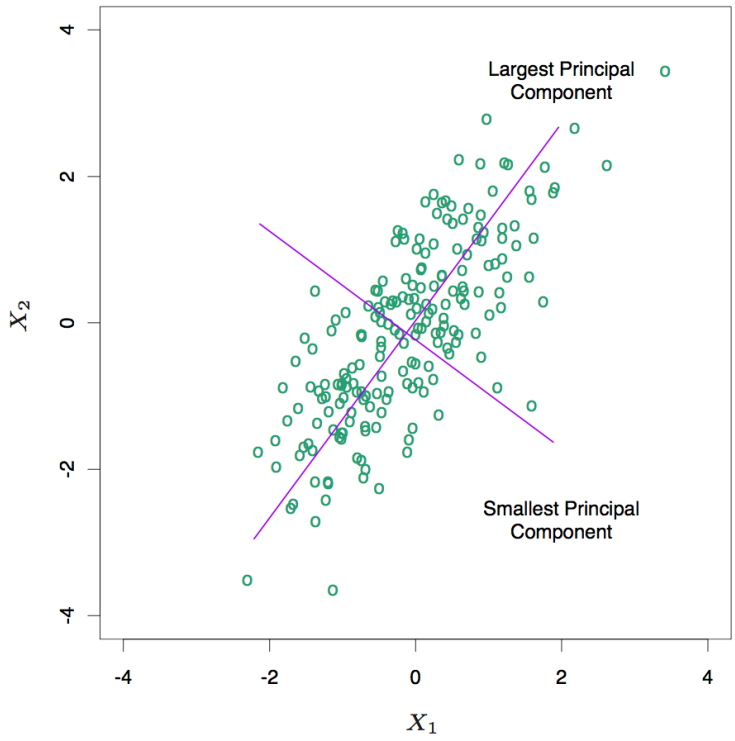


Figura 5.6: Componentes Principales en dos dimensiones

- Considerando todos los puntos pertenecientes al cluster, se recalcula un nuevo centro.

Estos dos pasos, se repiten hasta converger, es decir, hasta que la diferencia de la varianza total de cada grupo sea mínima de una iteración a otra.

La aplicación de esta técnica es para encontrar grupos de personas que presenten historias clínicas similares, ya que las variables que se utilizan acá son sólo las variables relacionas al texto de las notas clínicas hechas por los médicos. Esto, a priori, permitiría determinar perfiles de pacientes, los que deberían ser consecuentes con la teoría médica.

Análisis de Componentes Principales (PCA)

Las Componentes Principales son un conjunto de proyecciones de los datos, no correlacionados y ordenados por la varianza que permiten explicar.

La figura 5.6 permite graficar lo antes mencionado, al simplificar el análisis considerando sólo dos dimensiones. Cabe destacar que una componente (la más grande), permite explicar mejor la varianza de los datos.

Esta técnica permite interpretar las dimensiones desde otra perspectiva. Ya que los datos son los mismos, pero desde proyectados desde un punto de vista diferente, este nuevo punto de vista se define como una variable "latente", que es posible darle una interpretación si se tiene contexto

del problema. A diferencia del Clustering, que permite agrupar a las personas, esta técnica permite hacer análisis de las variables presentes, a través de las cuales se representan a las personas.

Se espera que las variables latentes (EigenVectors), permitan dar una interpretación más intuitiva, e identificar las que representen mayor variación de los datos que es donde hay más información.

Reglas de Asociación

Es una técnica que se utiliza mucho en la minería de datos que provienen de la industria comercial. El objetivo es encontrar un conjunto de variables $X = (X_1, X_2, \ldots, X_p)$ que aparecen más frecuentemente en los datos. Es muy recomendable cuando estas variables son binarias $X_j \in \{0, 1\}$, que en la industria comercial se conoce como análisis "market basket". La aplicación de esta técnica en el presente trabajo, es de forma análoga. Es decir, para una observación i, cada variable X_j toma dos valores posibles:

- $x_{ij} = 1$, para la observación (paciente) i, presenta el producto (signo o síntoma) j.
- $x_{ij} = 0$, para la observación (paciente) i, no presenta el producto (signo o síntoma) j.

A diferencia de la aplicación en la industria comercial, donde esta técnica permite definir la organización de los estantes dentro de un supermercado, o armar promociones entre productos, en el presente trabajo, permite que un computador puede entender algunos conceptos de medicina y como se relacionan entre ellos.

El valor de esta técnica en el contexto de la salud, es que permitiría armar de forma automática una base de conocimiento médico, a partir de los registros clínicos que realizan a lo largo del país.

Cada una de las reglas que se definen, se evalúan con diferentes criterios [54], entre los que se encuentran:

- **Support**: Se define como la proporción del total de las transacciones que contienen un determinado "itemset"(productos, signos o síntomas).

$$sup(A)$$

- **Confidence**: Es la estimación de $Pr(C|A)$, es decir, la probabilidad de observar C dado A. En el caso de reglas con el mismo valor de Confidence, es preferible la que tenga mayor Support.

$$conf(A \to C) = \frac{sup(A \bigcup C)}{sup(A)}$$

- **Lift**: Permite medir la independencia entre A y C. En este caso, lift es un valor entre $[0, +\infty[$. Valores cercanos a 1 implica que A y C son independientes, luego esta regla no es interesante. Reglas con valores mucho mayores que 1 permiten definir cuanta información entrega A sobre C. Cabe destacar que Lift mide solo co-ocurrencia (no implicancia), por lo que es una regla simétrica.

$$lift(A \rightarrow C) = \frac{conf(A \rightarrow C)}{sup(C)}$$

- **Conviction**: Permite complementar el análisis observando Confidence y Lift. Cabe destacar que Conviction se puede interpretar como la definición en lógica de la implicancia. Sus valores también se encuentran de $[0, +\infty[$. Las reglas cuyo Conviction es mucho mayor que , corresponden a relaciones de interés.

$$conv(A \rightarrow C) = \frac{1 - sup(C)}{1 - conf(A \rightarrow C)}$$

5.2.2. Resultados Análisis de Contenido

Este análisis se realizó usando la herramienta Rapid Miner, que permite llevar a cabo los tres análisis propuestos.

Clustering

Luego de realizar pruebas, se determinó que el número adecuado de cluster es 4. Se aplicó el algoritmo K-Means, al conjunto de personas que fueron diagnosticadas con un IAM.

A modo de representación de cada cluster, se entrega una nube de tags con los términos que representan al centro de cada cluster.

- Cluster 0 - Factores de Riesgo: Contiene a las personas que en sus notas clínicas, aparecen pocos términos. Sin embargo, aunque son pocos, los términos son justamente los principales factores de riesgo cardiovascular: Hipertensión Arterial, Diabetes, Dislipidemia, Tabaquismo. Este cluster agrupa al 54 % de total de las personas analizadas. El centro de este cluster se muestra en la figura 5.7.

- Cluster 1 - Problemas Respiratorios: Agrupa a las personas que presentan insuficiencia cardiaca. Como se presenta en el capítulo 3, una de los síntomas de la insuficiencia cardiaca es problemas asociados al sistema respiratorio, que se muestra con la presencia de disnea (dificultad al respirar) debido a la acumulación de líquido, en los pulmones. Esta acumulación de líquidos también se refleja en endemas en las extremidades inferiores (eeii). Este cluster agrupa al 11 % de total de las personas analizadas. El centro de este cluster se muestra en la figura 5.8.

Figura 5.7: Representación Gráfica Centro Cluster 0: Principales Factores de Riesgo

Fuente: Elaboración Propia

Figura 5.8: Representación Gráfica Centro Cluster 1: Problemas Respiratorios

Fuente: Elaboración Propia

- Cluster 2 - Hipertensos sin Diabetes: Representa a personas en que resalta su condición de hipertensión, pero que probablemente no presentan diabetes, debido a que el fármaco enalapril, que sirve para controlar la hipertensión y la insuficiencia cardiaca congestiva no es recomendado para personas que están en tratamiento de diabetes o de algún problema renal. Este cluster agrupa al 18 % de total de las personas analizadas. El centro de este cluster se muestra en la figura 5.9.

- Cluster 3 - Alto Riesgo: Agrupa a personas que presentan hipertensión, tabaquismo, dislipidemia y diabetes, es decir, los principales factores de riesgo. A diferencia del cluster 0, estas personas si tienen suficiente información en sus notas clínicas. Este cluster agrupa al 17 % de total de las personas analizadas. El centro de este cluster se muestra en la figura 5.10.

A partir de los resultados obtenidos, cabe destacar los Cluster 1 y Cluster 2, ya que representan perfiles de pacientes bien conocidos en personas que sufren enfermedades cardiovasculares. Sobre todo, es Cluster 1 muestra la importancia de los problemas respiratorios con relación a la insufi-

Figura 5.9: Representación Gráfica Centro Cluster 2: Hipertensos sin Diabetes

Fuente: Elaboración Propia

Figura 5.10: Representación Gráfica Centro Cluster 3: Alto Riesgo

Fuente: Elaboración Propia

ciencia cardiaca. Suele ocurrir, en muchos casos en que la persona se siente mal debido a problemas respiratorios y el médico SOLO la trata por esta causa, dejando de lado las posibilidades de una probable falla cardiovascular.

Análisis de Componentes Principales (PCA)

Este análisis, a diferencia del Clustering, lo que pretende es relacionar conceptos (no personas) a partir de los datos. El Análisis de Componentes Principales es recomendado cuando se tienen muchas variables, como es este caso, ya que permite explicar la varianza en un número reducido de nuevas variables latente o EigenVectors.

A partir de 216 [2] variables, se obtuvieron los resultados presentados en la tabla 5.1, que muestra que con sólo el 26 % de las nuevas variables o eigenvectors, es posible explicar el 80 % de la varianza. Esto muestra la importancia de este tipo de técnicas al momento de trabajar con muchas variables.

Para lograr interpretar los EigenVectors, es necesario revisar cuales son los términos más representativos en cada uno, y además, es útil revisar las diferencias entre ellos. Los resultados obtenidos destaca que cada EigenVector permite describir algún perfil o grupo de signos y síntomas que tienen relación con las enfermedades cardiovasculares.

A modo de ejemplo, se entregan representaciones gráficas de los primeros 4 eigenvectors, y una breve interpretación de cada uno:

- EigenVector 1: De la misma forma como se mostró en en análisis de Clustering, esta variable representa los principales conceptos relacionados a Problemas Respiratorios. Cabe mencionar, que realizando una comparación con los demás vectores, este representa todos los conceptos relacionados con el tabaquismo: tabaco, fuma, cigarrillos, entre otros. La representación gráfica se muestra en la figura 5.11.

- EigenVector 2: Esta dimensión representa a los conceptos relacionados con la Diabetes y la

[2] Se filtraron algunas variables de las 248 originales.

Dislipidemia. Se destacan términos como Colesterol y Triglicéridos que tienen directa relación con la manera de diagnosticar la dislipidemia. También aparecen conceptos como la glicemia y glucosa, que se relacionan con la diabetes. Al comparar este vector con el EigenVector 1, se obtiene que representa a quienes no tienen problemas respiratorios y que no incluye ningún concepto asociado al tabaquismo. Es decir, se entiende que estos dos primeros vectores son ortogonales. La representación gráfica se muestra en la figura 5.12.

- EigenVector 3: Esta variable presenta términos muy similares al EigenVector 2. En estos casos, es importante el análisis que compara ambas variables, para recalcar cuales son sus diferencias. Al realizar este análisis, se presenta que esta variable considera los casos en que se presenta la Diabetes y la Dislipidemia acompañada con signos y síntomas asociados a la insuficiencia cardiaca congestiva, la que provoca problemas en el sistema respiratorio. Es por esto que destacan conceptos como: congestiva, pulmonar, pies. En este caso, un signo común, como consecuencia de la insuficiencia cardiaca es la hinchazón en los pies, debido a la acumulación de líquidos en las extremidades inferiores (eeii). La representación gráfica se muestra en la figura 5.13.

- EigenVector 4: De la misma forma como la variable anterior se interpreta mejor al compararse con otras variables, este nuevo vector, no tiene una interpretación directa sólo observando sus conceptos, pero si se explica al momento de compararlo con otros. Se realizó el análisis y se obtuvo como resultado que este eigenvector no considera nada relacionado con la Diabetes, pero si conceptos relacionados a problemas respiratorios, como: disnea, espirometría[3], sibilancias[4], entre otros. Luego, la interpretación de esta nueva variables se obtiene de la comparación con las otras. La representación gráfica se muestra en la figura 5.14.

Cabe destacar una diferencia que se presenta entre este análisis y los resultados obtenidos del clustering. Se observa que en cada uno de los clusters, siempre destacaba el concepto de Hipertensión Arterial (HTA), lo que tiene una interpretación obvia, y es que sin importar las diferencias entre cada uno de los clusters, todas las personas tienen hipertensión cuando han sufrido un IAM.

Sin embargo, en los principales EigenVectors, en ninguno aparece algo relacionado a la Hipertensión Arterial. La explicación se debe a que el análisis de componentes principales intenta construir nuevos vectores que expliquen de mejor forma la varianza de los datos. Como se mencionó, la Hipertensión Arterial aparece en casi todas las personas que han sufrido un IAM, luego no tiene valor

[3]Examen para medir la cantidad de aire que pueden retener los pulmones de una persona
[4]Sonido que ocurre cuando el aire se desplaza a través de vías respiratorias estrechadas.

# EigenVectors	Proporción de Variables	Varianza Acumulada
4	2 %	20 %
21	10 %	50 %
57	26 %	80 %

Cuadro 5.1: Análisis de Componentes Principales: EigenVectors

Fuente: Elaboración Propia

Figura 5.11: EigenVector 1: Problemas Respiratorios

Fuente: Elaboración Propia

con respecto a agregar información al análisis. Esta es una muestra de como el análisis de Clustering se complementa con el análisis de Componentes Principales.

Por otra parte, una interpretación que se obtiene de analizar los resultados obtenidos, sólo observando los datos, es que los problemas respiratorios son relevantes al momento de analizar enfermedades cardiovasculares. La interpretación médica es que un problema al corazón como la insuficiencia cardiaca congestiva provoca que el bombeo del corazón se vuelva menos eficaz, luego la sangre puede acumularse en otras áreas del cuerpo, como en los pulmones, el hígado, el tracto gastrointestinal, al igual que en los brazos y las piernas. Principalmente, la acumulación de líquido en los pulmones es lo que provoca problemas respiratorios, lo que puede confundirse con sólo una bronquitis u otra enfermedad respiratoria.

Reglas de Asociación

Para aplicar esta técnica se usaron los 248 conceptos asociados a signos y síntomas de las enfermedades cardiovasculares. Como ya se mencionó, el objetivo de este análisis es validar la

Figura 5.12: EigenVector 2: Diabetes y Dislipidemia

Fuente: Elaboración Propia

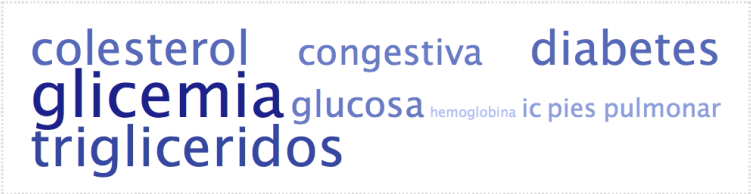


Figura 5.13: EigenVector 3: DM2 y Dislipidemia con Insuficiencia Cardiaca

Fuente: Elaboración Propia

Figura 5.14: EigenVector 4: Problemas Respiratorios, sin Diabetes

Fuente: Elaboración Propia

calidad de los datos almacenados en los campos de texto no estructurados del registro clínico electrónico, encontrando coincidencias y consecuencia entre los datos registrados y la teoría relativa a las enfermedades cardiovasculares.

La importancia de este tipo de técnica, es que si se logra validar la calidad de los datos almacenados, es posible que un computador aprenda los conceptos básicos de medicina, en un determinado contexto, sólo analizando los datos del RCE.

En la figura 5.15 se presentan las 15 reglas con mayor Confidence. La variable objetivo (Label=1), se consideró como una variable más para este análisis. Esto distorsiona un poco la medida Lift, ya que está condicionada a que sólo la mitad de la muestra total presenta un IAM, esto explica su valor cercano a 1. Sin embargo, la medida Conviction, que denota implicancia, presenta valores altos, lo que muestra el valor de la información de cada regla.

Destaca la primera regla, que relaciona el fármaco carvedilol que se usa para tratar la insuficiencia cardíaca (trastorno en el cual el corazón no puede bombear la sangre a todas las partes del cuerpo) y la hipertensión. También se usa para tratar a las personas cuyo corazón no puede bombear bien la sangre como resultado de un IAM.

En general, cabe mencionar que en estas reglas destacan los principales factores de riesgo cardiovascular, como la Hipertensión Arterial, el Tabaquismo, la Dislipidemia (Colesterol y Triglicéridos), y algunos conceptos relacionados a problemas respiratorios, como la Disnea, Torax.

Nro.	Premisa	Conclusión	Support	Confidence	Lift	Conviction
1	carvedilol	Label = 1	13.9%	98.7%	1.50	25.52
2	hta, carvedilol	Label = 1	13.0%	98.6%	1.49	23.82
3	hta, colesterol, trigliceridos	Label = 1	13.0%	98.6%	1.49	23.82
4	disnea, tabaquismo	Label = 1	12.6%	98.5%	1.49	23.14
5	hta, disnea, tabaquismo	Label = 1	11.7%	98.4%	1.49	21.43
6	tabaco, colesterol, trigliceridos	Label = 1	10.7%	98.3%	1.49	19.73
7	disnea, colesterol	Label = 1	10.5%	98.2%	1.49	19.39
8	hta, tabaco, colesterol, trigliceridos	Label = 1	10.5%	98.2%	1.49	19.39
9	aas, colesterol	Label = 1	10.2%	98.2%	1.49	18.71
10	hta, disnea, colesterol	Label = 1	10.2%	98.2%	1.49	18.71
11	hta, tabaco, torax	Label = 1	11.8%	96.9%	1.47	11.06
12	hta, dislipidemia, torax	Label = 1	11.7%	96.9%	1.47	10.89
13	hta, disnea, torax	Label = 1	11.7%	96.9%	1.47	10.89
14	tabaco, trigliceridos	Label = 1	11.5%	96.8%	1.47	10.72
15	hta, tabaco, trigliceridos	Label = 1	11.3%	96.8%	1.47	10.55

Figura 5.15: Reglas de Asociación

Fuente: Elaboración Propia

Breves Conclusiones

Se valida también que técnicas como Clustering y Análisis de Componentes Principales permiten realizar agrupaciones de personas y de conceptos, respectivamente, en un contexto predeterminado. Además, la aplicación de Reglas de Asociación, permite determinar la relación entre conceptos claves, lo que permite extraer conocimiento sobre medicina a partir de las notas clínicas que los médicos ingresan en el Registro Clínico Electrónico.

Finalmente, se destaca que a partir de más de 200 conceptos que fueron previamente seleccionados, el Análisis de Contenido permite validar su calidad, ya que son totalmente consecuentes con la teoría medica en relación a las enfermedades cardiovasculares.

El detalle del procesamiento en Rapid Miner, se entrega en en Anexo F .

5.3. Análisis Predictivo

Esta sección describe el trabajo realizado para lograr la clasificación de las personas de acuerdo a su riesgo cardiovascular, analizando parte de los datos estructurados y los campos de texto no estructurados de las notas clínicas que se encuentran en el Registro Clínico Electrónico.

Se trabaja bajo el paradigma de intentar ajustar una función, lo que permite hacer que un computador aprenda en base a datos de entrenamiento, en los que se conoce el resultado que se desea obtener. De forma simple, se tiene una función f que intenta ajustarse a un set de datos Y, minimizando el error de la expresión $Y = f(X) + \varepsilon$.

El algoritmo permite modificar la relación del resultado obtenido, dado los datos observados, e

intenta reducir el valor de la diferencia entre $y_i - \hat{f}(x_i)$, es decir, la diferencia entre el valor real y el valor predicho. Este proceso es conocido como "learning by examples". Esta tipo de análisis se conoce como Aprendizaje Supervisado.

5.3.1. Modelamiento Análisis Predictivo

Se pretende detectar signos y síntomas en los campos donde los médicos registran el motivo de la consulta, la historia clínica y la anamnesis. El valor agregado de estos campos, es que se registran observaciones sobre tratamientos y evolución de las enfermedades, lo que no se puede registrar en los campos estructurados.

A continuación se detallan las transformaciones de los datos, que luego son utilizados para calibrar un modelo de clasificación binaria.

Construcción DataSet

Esta etapa comienza a partir de los datos preprocesados, que consisten en una matriz donde cada fila representa a una persona y cada columna contiene cada una de las variables, que provienen de los datos estructurados y del texto no estructurado.

Antes de aplicar el modelo a los datos preprocesados, es necesario tener en cuenta algunos detalles, principalmente asegurarse que la base de datos esté balanceada. Es decir, dado que los datos se dividen en dos clases (con IAM y sin IAM), es importante que existan suficientes registros para cada una de estas clases.

Además, en el caso particular del presente trabajo, se está manejando un gran numero de variables (más de 250), por lo que también es importante considerar técnicas de reducción de dimensionalidad para lograr obtener mejores resultados al momento de aplicar el modelo predictivo.

Teniendo esto en consideración, se aplicaron las siguientes transformaciones a los datos:

1. Balancear la Base de Datos: El set de datos original considera 280 personas con IAM y 181 sin IAM. Se tomó una submuestra a partir de las 280 personas, lo que resultó en considerar a sólo 126 personas. Finalmente, se trabajará con un 41 % de personas con IAM, es decir, con Riesgo Cardiovascular (RCV), y con un 59 % sin RCV.

2. Análisis de Componentes Principales: Como se mencionó en la sección anterior, esta transformación permite realiza un disminución de dimensionalidad, sobre todo en problemas con una gran cantidad de variables. Se trabajó con las variables que provienen del texto no estructurado, y se redujo desde más de 200 a sólo 4 eigenvectors, los que explican el 20 % de la varianza

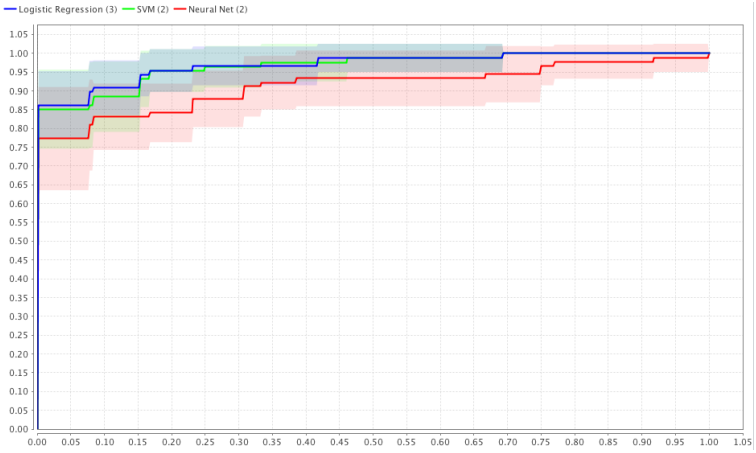


Figura 5.16: Comparación entre Modelos Predictivos

Fuente: Elaboración Propia

total. Las interpretaciones de estas nueva variables "latentes", se mencionaron en la sección anterior.

3. Inner Join: Finalmente, luego de procesar las variables del texto no estructurado, se realizó un inner join con las variables estructuradas, para construir el DataSet final que será utilizado para calibrar el modelo predictivo.

Logistic Regression

Se comparó el desempeño de tres de los modelos ampliamente usados para resolver los problemas de clasificación: Logisitic Regression, Support Vector Machine, Neural Network.

La comparación se realizó utilizando la medida Área bajo la Curva[5], como se muestra en la figura 5.16. De acuerdo a esto, el modelo que mejor comportamiento predictivo muestra con estos datos, es Logistic Regression (LR). En el capítulo 2 se entregan más detalles sobre le modelo escogido.

Cabe mencionar su función tanto explicativa y predictiva. Un ejemplo clásico, de su uso explicativo en los estudios médicos, es cuando la probabilidad que se estima puede interpretarse como una tasa de prevalencia o de incidencia. Sin embargo, su utilización en la predicción es el uso más frecuente y extendido.

Además, la LR se utiliza cuando se quiere investigar si una o varias variables explican una variable dependiente que toma un carácter cualitativo. Este hecho es muy frecuente en medicina ya que constantemente se intenta dar respuesta a preguntas formuladas en base a la presencia o ausencia de una determinada característica que no necesariamente es cuantificable sino que representa la existencia o no de un efecto de interés. Este método ha sido usado en investigaciones, como por

[5]De la Curva ROC

ejemplo, predecir el desarrollo de un "evento cardiovascular".

En general, la función logística es aquella que halla, para cada individuo según los valores de una serie de variables, la probabilidad de que presente el efecto estudiado. En este caso, la probabilidad de presentar Alto Riesgo Cardiovascular.

Análisis del Texto No Estructurado

Una vez definido el modelo a utilizar, se detalla parte de la problemática presentada en este trabajo, que consiste en validar la calidad de la información almacenada en los campos de texto no estructurado.

Para esto, se presentan tres modelos:

1. Modelo usando sólo las variables a partir del texto no estructurado: Tiene por objetivo validar la capacidad predictiva de este tipo de variables.

2. Modelo usando sólo las variables estructuradas: Replicar el análisis presentado en el estudio [34], donde se consideran variables de comportamiento del paciente, de los diagnósticos anteriores, entre otras variables estructuradas.

3. Modelo usando ambos tipos de variables: El resultado esperado es que a partir del modelo que utiliza las variables estructuradas, si se añaden las variables no estructuradas se logra mejorar el desempeño en la predicción. Con esto se concluye que considerar el texto no estructurado agrega valor al análisis de la detección temprana de Riesgo Cardiovascular.

Entrenamiento y Validación

El proceso de entrenamiento y validación a partir del DataSet que se definió anteriormente, se realizó usando el software Rapid Miner. El detalle del procesamiento se entrega en en Anexo G .

Cabe mencionar que se utilizó una combinación de métodos para realizar el procedimiento de calibración del modelo. Se consideraron 3 etapas: Entrenamiento, Testing y Validación.

A continuación, se entrega un detalle de cada una de las etapas del análisis y del procesamiento de los datos, para finalmente calibrar el modelo predictivo.

1. **Hold-out**: Se realizó una división de los datos en una proporción de 70 % para Entrenamiento y Testing, y de 30 % para Validación. Este último subconjunto de datos no será considerado para calibrar el modelo, sólo para validar los resultado obtenidos.

2. **Cross-Validation**: Este método consiste en dividir la muestra considerada para Entrenamiento y Testing en varias partes iguales. En cada iteración, se deja una de estas partes fuera,

para luego del entrenamiento, utilizar esta para testear los resultados obtenidos y realizar la calibración necesaria.

3. **Definir Threshold**: El modelo utilizado, arroja como resultado final un número entre 0 y 1, que se interpreta como una probabilidad. Sin embargo, se define un threshold o umbral, para transformar esta variable continua en una binaria. Es decir, sobre el threshold, se interpreta como 1, y en caso contrario como 0. La definición correcta de este parámetro tiene relación con los costos asociados a definir, por ejemplo, un Falso Negativo. En el contexto del presente trabajo, un Falso Negativo es definir a una persona que en realidad si presenta Riesgo Cardiovascular Alto, pero que el modelo lo predice como alguien sin riesgo. Es por esto que se considero una proporción de costos asociados de 2:1 al cometer un error en la clasificación de alguien que si tienen RCV Alto.

4. **Evaluación de Modelo**: Tanto para la etapa de Entrenamiento y Testing, como para la etapa de Validación, se utilizaron las medidas AUC, F-Measure, Sensibilidad, Especificidad y Accuracy para evaluar el desempeño del modelo. Cabe destacar que todas las medidas mencionadas, excepto AUC, se evalúan para un threshold determinado. El área bajo la curva ROC (AUC), permite evaluar la sensibilidad y la especificidad del modelo para diferentes valores del threshold. Con respecto al uso del Accuracy, es importante considerar las proporciones de cada una de las clases, ya que una base no balanceada puede distorsionar la interpretación de esta medida.

El análisis predictivo es el foco del presente trabajo, ya que este modelo permitirá cumplir el objetivo general que consiste en Generar un modelo predictivo, basado en Machine Learning y Natural Language Processing (NLP), que analice los campos de texto no estructurados del Registro Clínico Electrónico, para predecir riesgo cardiovascular, detectando signos y síntomas en su historia clínica.

5.3.2. Resultados Análisis Predictivo

Antes de comentar los resultados obtenidos, es importante tener presente cuales eran las características del modelo de acuerdo a los resultados esperados en esta parte del análisis:

- **Aceptable**: Que los resultados del modelo cumplan con las especificaciones dadas en la tabla 4.1.

- **Consecuente**: Con la clasificación de Riesgo Cardiovascular de Framingham, en los niveles Alto y Muy Alto.

	(1) Texto		(2) No Texto		(1) + (2)	
Medida	*Test*	**Validación**	*Test*	**Validación**	*Test*	**Validación**
Accuracy	*84.7%*	**83.7%**	*94.4%*	**75.0%**	*94.8%*	**90.2%**
F-Measure	*84.4%*	**83.2%**	*93.5%*	**76.3%**	*94.1%*	**88.6%**
Sensibilidad	*96.6%*	**97.4%**	*94.3%*	**97.4%**	*97.7%*	**92.1%**
Especificidad	*76.4%*	**74.1%**	*94.5%*	**59.3%**	*92.9%*	**88.9%**
AUC	*0.926*	**0.933**	*0.952*	**0.923**	*0.973*	**96.8%**

Figura 5.17: Valor Agregado por las Variables a partir del Texto No Estructurado

Fuente: Elaboración Propia

- **Agregar Valor:** El modelo permite identificar pacientes con Alto Riesgo, quienes tenían un nivel de riesgo Framingham Moderado y Bajo. Ya que en este grupo de personas es donde se produce la mayor cantidad de casos de IAM.

- **Priorizar**: Permite detectar Alto Riesgo Cardiovascular en personas que no se estaban controlando al momento de sufrir un IAM. Es decir, no tenían clasificación de Riesgo Cardiovascular de acuerdo a Framingham.

Variables del Texto No Estructurado

La primera parte de la evaluación de los resultados del modelo, consiste en validar si al considerar las variables que se obtienen a partir de los campos de texto no estructurados, estos agregan valor al momento de realizar la predicción de riesgo cardiovascular.

Los resultados se muestran en la figura 5.17. A partir de esto, se puede concluir que las variables del texto no estructurado si aportan valor al análisis predictivo de riesgo cardiovascular.

Cabe destacar que al utilizar sólo las variables No Texto, se obtiene un modelo con mucha sensibilidad, lo que se confirma con el bajo valor del F-Measure. Por otra parte, las variables a partir del texto no estructurado, por sí solo, permite tener niveles aceptables al momento de calibrar un modelo de predicción de riesgo cardiovascular.

Se observa que todas las medidas tienen mejor resultado cuando se incorpora el texto para el análisis (excepto en el caso de la Sensibilidad). En el caso que se considera sólo el texto, este presenta un mejor desempeño con respecto a las variables estructuradas (No Texto). Y lo más importante, que cuando se juntan ambos tipos de variables, se obtiene mejor resultado que cada uno por separado.

Por lo tanto, se cumple el primer resultado esperado con el modelo que utiliza los dos tipos de variables, que consiste en satisfacer los criterios de aceptabilidad del modelo, de acuerdo a la tabla 4.1.

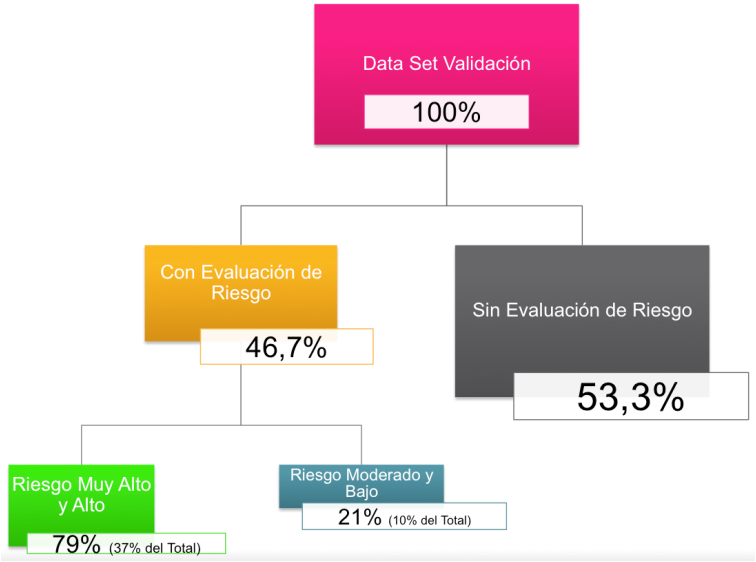


Figura 5.18: Conjunto de Validación de acuerdo a su Nivel de Riesgo Cardiovascular de Framingham

Fuente: Elaboración Propia

Comparación con Riesgo de Framingham

Parte importante, para validar los resultados del presente trabajo, es compararse con la situación actual, con la que se evalúa riesgo cardiovascular.

Luego de las subdivisiones hechas, tanto para el Entrenamiento, Testing y Validación, se utilizaron 92 personas para realizar la comparación entre los resultados obtenidos por el modelo, y su situación de acuerdo a si estaban en control cardiovascular o no, y si tenían asignado algún nivel de riesgo antes de haber sufrido un IAM.

Analizando cada uno de los casos, para saber quienes están dentro del Programa de Salud Cardiovascular, es decir, si tenían asignado algún nivel de riesgo, se obtuvo la distribución de personas como se muestra en la figura 5.18

Es importante mencionar, que esta muestra fue tomada de forma aleatoria, por lo que no necesariamente representa las proporciones reales en la población del país. Sobre todo, el grupo que tiene clasificación de riesgo moderado y bajo (con sólo un 21 % del total de personas que están en el PSCV), ya que en la población real este grupo es más grande.

Se presentan los resultados obtenidos para cada uno de los subgrupos que se muestran en la figura 5.18

- **Grupo 1 - Dentro del PSCV**: Personas que tienen asignado algún nivel de riesgo cardiovascular de Framingham. Los resultados se muestran en la figura 5.19. Se destaca la medida F-Measure es 90 %, lo que muestra un buen nivel predictivo, en el grupo de las personas que ya

Medida	Valor
Accuracy	86.0%
F-Measure	90.0%
Sensibilidad	93.1%
Especificidad	71.4%
Precision	87.1%

Figura 5.19: Evaluación Modelo en Personas dentro del PSCV

Fuente: Elaboración Propia

Medida	Valor
Accuracy	88.2%
F-Measure	92.3%
Sensibilidad	92.3%
Especificidad	75.0%
Precision	92.3%

Figura 5.20: Evaluación Modelo en Personas con Riesgo de Framingham Alto y Muy Alto

Fuente: Elaboración Propia

Medida	Valor
Accuracy	77.8%
F-Measure	75.0%
Sensibilidad	100.0%
Especificidad	66.7%
Precision	60.0%

Figura 5.21: Evaluación Modelo en Personas con Riesgo de Framingham Moderado y Bajo

Fuente: Elaboración Propia

Medida	Valor
Accuracy	93.9%
F-Measure	84.2%
Sensibilidad	88.9%
Especificidad	95.0%
Precision	80.0%

Figura 5.22: Evaluación Modelo en Personas que no pertenecen del PSCV

Fuente: Elaboración Propia

pertenecen al PSCV. Esto muestra la consistencia del modelo propuesto, con la forma actual de medir riesgo cardiovascular.

- **Grupo 1.1 - Alto Nivel de RCV**: Personas que tienen asignado niveles Altos y Muy Altos de riesgo cardiovascular de Framingham. Los resultados se muestran en la figura 5.20. Cabe mencionar que en este grupo el modelo presenta el F-Measure más alto de 92,3 %. Lo que confirma que el modelo es consecuente; con esto se obtiene uno de los resultados esperados.

- **Grupo 1.2 - Bajo Nivel de RCV**: Personas que tienen asignado niveles Moderados y Bajos de riesgo cardiovascular de Framingham. Los resultados se muestran en la figura 5.21. Aunque se muestra que los resultados obtenidos tienen una alta sensibilidad, el valor de F-Measure sigue siendo aceptable (75 %) y se complementa con un Accuracy de 77,8 %. Recordar que este grupo estaba muy por debajo de la proporción real, esto puede haber afectado el desempeño del modelo en comparación con otros grupos. De todas formas, los resultados aceptables en este grupo, permite decir que el modelo aporta en Agregar Valor a la medida actual de riesgo cardiovascular, ya que logra detectar a quienes, dentro del grupo de personas con riesgo moderado y bajo son más propensos a sufrir un IAM.

- **Grupo 2 - Sin Nivel de RCV**: Personas no que tienen asignado algún nivel de riesgo cardiovascular de Framingham. Los resultados se muestran en la figura 5.22. Este es uno de los resultados más importantes, y que permite decir que el modelo propuesto agrega valor a la situación actual. Ya que presenta muy buenos indicadores de desempeño en el grupo de personas que no estaban siendo controladas, y no tenían asignado un nivel de riesgo al momento de sufrir un IAM. Se tiene un F-Measure de 84,2 % y un Accuracy de 94 %. Además, destacar los niveles de Sensibilidad y Especificidad, lo que permite concluir que el modelo tiene un excelente comportamiento en el grupo de personas que el sistema actual de riesgo cardiovascular no logra capturar.

Con respecto al último resultado, agrega mayor valor aún dado que existen metas relacionados a indicadores de actividades de la atención primaria de salud (IAAPS), que tienen relación con la cobertura en los programas de prevención. Por lo que con los resultados del modelo, permitiría a los establecimiento priorizar a que personas debo ir a buscar en caso que no se estén controlando, pero que presentan altos niveles de riesgo, en este caso, cardiovascular.

Luego, con el análisis del Grupo 1.2 y Grupo 2, se obtienen otros de los resultados esperados, que es conseguir que el modelo agregue valor a la situación actual, y además que el nivel predictivo permita priorizar entre personas que no pertenecen a ningún programa preventivo.

Además, esto se alinea con lo antes mencionado, con respecto a personalizar la atención de salud para potenciar las estrategias de salud preventiva.

5.4. Discusiones

En esta sección se realiza un análisis de los resultados de una manera más global, considerando los objetivos del trabajo, el contexto en que se presenta (Smarter Care) y como este trabajo se relaciona con otro tipo de investigaciones y puede complementar otros tipos de análisis tradicionales en el ámbito de la salud.

5.4.1. Validación de los resultados

Los resultados fueron presentados a expertos en la materia, quienes actualmente forman parte de la Gerencia Clínica de la empresa SAYDEX:

- José Fernández: Médico Cirujano, actual Gerente Clínico de SAYDEX, quien también se ha desempeñado como Director de Centros de Salud Familiar, SAPU y CECOF, Director de Salud Municipal en la comuna de Cerro Navia, por lo que presenta una gran experiencia en lo relacionado a la gestión en salud pública, es particular atención primaria (APS).
- Inti Paredes: Médico Urólogo, actual Asesor Clínico en SAYDEX, quien fue durante el año 2008 Subdirector del Servicio de Salud Metropolitano Central y también se desempeño como Director del Hospital San Borja Arriaran. Durante el último tiempo se ha enfocado en implementaciones de soluciones tecnológicas para la salud, por lo que presenta experiencia en las necesidades de la salud pública, en especial a nivel hospitalario, y como las tecnologías pueden apoyar y mejorar las atenciones de las personas.

Los resultados fueron validados y se destacó su importancia en diferentes ámbitos de acuerdo a su impacto.

Por una parte, los resultados sirven para respaldar la utilidad y potencial que tiene el uso secundario de los datos del RCE, ya que una de las principales críticas es que los datos que son almacenados no tienen valor agregado y que la cantidad y calidad de los registros no es suficiente para realizar análisis. Por lo que los resultados obtenidos y el nivel de análisis realizado permite fortalecer y potenciar el uso del RCE y de otras las iniciativas dentro de los proyectos SIDRA del MINSAL.

Por otra parte, se considera un gran avance con respecto al uso actual que se le da a los datos del RCE, ya que a la fecha, todos los análisis no son más que contadores que sólo permiten mirar

la datos a nivel histórico. Proponer el uso de análisis predictivos es un gran avance para la salud pública en Chile.

El uso de campos del RCE que nunca se habían considerado antes, como los campos de texto no estructurados, permitirá enriquecer cualquier tipo de análisis que se pueda ocurrir mirando sólo la data estructurada. Es sabido que en los campos requeridos (campos estructurados) se registra la mínima información necesaria, sin embargo, en los campos de texto libre (no estructurados) se registra la mayor parte del detalle de la atención y las condiciones particulares de cada persona. En estos campos se registran las evoluciones de las enfermedades, la reacción a medicamentos, los resultados de tratamientos aplicados anteriormente, registros clínicos hechos en otros establecimientos y que se agregan a la historia clínica de la persona en los campos de texto libre. Por lo que este y todo trabajo que apunte a extraer conocimiento a partir de las notas clínicas tiene un valor enorme.

La empresa tiene la mayor cobertura de RCE en Chile, por lo que este tipo de soluciones tiene un enorme potencial y factibilidad de implementación efectiva en el corto plazo. Bastaría con integrar los resultados de los análisis en la interfaz del RCE para que los médicos de Chile puedan ver y utilizar los resultados de esta investigación.

5.4.2. Uso de los resultados del estudio

De acuerdo a lo mencionado en diferentes partes de este trabajo, con respecto al potencial de los resultados de un trabajo de análisis predictivo de riesgo cardiovascular, y entendiendo que los resultados del estudio no son un fin en sí mismo, sino que es importante determinar quienes y como podrán utilizar este conocimiento obtenido para tomar decisiones en favor de la salud de las personas, se resumen dos potenciales aplicaciones de los resultados obtenidos:

1. **Integración con los RCE**: Este enfoque es el más operacional, ya que será posible integrar los resultados obtenidos para cada persona dentro de la ficha clínica electrónica. Esto responde a una inquietud presentada en el primer capítulo, donde se puede mostrar en la misma pantalla que utiliza el médico para la atención del paciente la condición de riesgo y cuales fueron los factores detectados en su historia clínica. Así el médico bastaría con validar el ingreso del paciente a un programa de salud preventivo sin gastar tiempo en la revisión manual de la historia clínica.

2. **Descubrir Conocimiento**: Es sabido que hay información muy valiosa almacenada en el RCE. Sin embargo, actualmente no existe una forma directa para explorar los datos para los médicos que desean hacer investigación o simplemente explorar otros casos clínicos (manteniendo el anonimato del cada paciente) en algún contexto particular. Hoy en día, es necesario

tener conocimientos de programación y exploración de bases de datos para poder acceder a los datos 'en bruto'. Se propone complementar los resultados obtenidos, que permiten clasificar pacientes para enfocar los estudios a un grupo mucho más reducido, con otras herramientas de *Data Discovery* como la herramienta de IBM Watson Content Analytics, la que entrega una flexibilidad y usabilidad que permite que un usuario, sin conocimientos técnicos, pueda explorar los datos.

Es decir, es necesario empaquetar el modelo de clasificación de riesgo cardiovascular e integrarlo directamente con el RCE, para permitir que se alimente frecuentemente con los nuevos datos y así etiquetar a las personas de acuerdo a sus factores de riesgo.

Esta información se podrá usar para insertarla como información del paciente en el RCE, así cuando la persona se atienda en un establecimiento de salud, el médico podrá saber de forma instantánea su condición de riesgo.

Por otra parte, la clasificación hecha por el modelo, permitirá generar grupos reducidos de estudio, que complementado con una herramienta de exploración de datos permitirá llegar al detalle mismo de las notas clínicas y el texto exacto que usó el médico durante la atención del paciente. Este proceso es cíclico, ya que llegando al detalle, los médicos pueden agregar etiquetas o identificar nuevas formas de registrar en las notas clínicas, lo que permite enriquecer los datos usados por los modelos predictivos.

Se entiende que la interacción de los médicos con los datos a través de herramientas de *Data Discovery* logrará un mantenimiento continuo de la calidad de los datos que usan los modelos predictivos, que es la información que se mostrará dentro de las Fichas Clínicas Electrónicas del RCE.

En conclusión, el acceso "libre"para los médicos a los datos permitirá mantener una aplicación que beneficia directamente las personas.

5.4.3. Comparación con otros estudios

Hay dos investigaciones que abordan el problema de evaluar el riesgo cardiovascular de forma temprana, analizando los datos del registro clínico electrónico.

- El estudio [34], trabaja considerando sólo variables a partir de los campos estructurados. Utiliza Logistic Regression, y mide su desempeño usando sólo AUC igual a 0.77. Muy por debajo de los resultados que se muestran en 5.17. Cabe mencionar que los resultados no son del todo comparables, ya que los datos utilizados en ambos estudios provienen de diferentes fuentes.

- Otro estudio [12], realizado por IBM, con el uso de Watson y el registro clínico electrónico EPIC, aborda el mismo problema tratado en el presente trabajo. Ya que considera variables de los campos estructurados y las notas clínicas para la detección de riesgo cardiovascular. Dado que el trabajo tiene un foco en desarrollar un producto comercial no existen muchos detalles del "cómo", pero con el uso de la tecnología e infraestructura de IBM, se logró procesar la información de 8.500 pacientes. Se obtuvo como resultado, que incluir las notas clínicas permiten alcanzar un 85 % de Accuracy. Es difícil comparar resultados dado que el origen de datos es diferente, sin embargo, el valor del presente trabajo por sobre lo realizado en IBM, es que se enfoca en las notas clínicas en español. La mismas personas de IBM, declaran que no están en condiciones aún de utilizar su tecnología en notas clínicas en español.

- Otro estudio de IBM [11], no se enfoca en detectar riesgo cardiovascular, pero si en extraer los factores de riesgo con un análisis avanzado en el procesamiento de texto. Es exactamente en este punto, donde IBM no logra analizar las notas clínicas en español, por lo que potenciar este trabajo, en el idioma español, significaría un gran avance en la dirección de mejorar todo tipo de soluciones que se enfoquen en el análisis predictivo.

5.4.4. Aprendizaje a partir de los datos

Realizar estudios en el ámbito de la medicina no es nada nuevo. Sin embargo, con este trabajo se pretende dar el paso para complementar los análisis estadísticos tradicionales, con los modelos de minería de datos, donde SÓLO A PARTIR de los datos, es posible encontrar información que tiene relación directa con la evidencia médica.

A diferencia de la estadística tradicional, donde se establece una hipótesis en base a datos históricos, y se realizan diversos test para validar o rechazar la hipótesis planteada, la minería de datos abre las puertas a lo que se conoce como Machine Learning, donde un computador, sin haber estudiado medicina, puede comenzar a obtener patrones que tienen directa relación con la teoría médica.

Esto se mostró principalmente el la sección en que se realiza el Análisis de Contenido. Ya que a partir de análisis de clustering se pudo detectar perfiles de pacientes, y con la transformación de variables, con el uso de PCA, fue posible identificar que conceptos asociados a una enfermedad permiten identificar de mejor forma el perfil de una persona.

También es posible complementar el análisis predictivos, usando muchas más variables estructuradas, que se obtienen de los formularios de atención, las que no se incluyeron en este estudio, ya que el foco estaba en el uso de variables a partir del texto no estructurado de las notas médicas del

registro clínico electrónico.

Es decir, queda la puerta abierta para mejorar y complementar este estudio con más técnicas y ampliarlo a otro tipo de enfermedades o aplicaciones que sean de utilidad para lograr entregar una mejor calidad de servicio en la salud pública.

5.4.5. Impacto en la Gestión de la Salud Preventiva

Parte de los resultados esperados, es usar los resultados del Análisis Predictivo para ser un aporte a la gestión de la Salud Preventiva en Chile.

Esto se cumple, al lograr priorizar a pacientes que no pertenecen a algún programa de salud preventivo, pero que analizando sus notas clínicas, es posible detectar factores de riesgo. Lo mismo ocurre con las personas que aunque pertenecen a uno de estos programas, como el de Salud Cardiovascular, aparecen clasificados en niveles moderados y bajo, por lo que se oculta su condición de riesgo, la que si es detectada al momento de estudiar los datos almacenados en los campos de texto no estructurados.

Desde el punto de vista táctico, para los establecimientos de salud, quienes deben cumplir mensual y anualmente con metas sanitarias, relativas a la cobertura de los programas de salud preventiva, este tipo de soluciones permitiría identificar a que personas "Ir a Buscar". Cabe mencionar, que el correcto cumplimiento de estas metas se traduce directamente en financiamiento para el establecimiento, por lo que además de agregar valor al detectar a una persona en riesgo de una enfermedad cardiovascular, también se traduce en un beneficio económico.

En el largo plazo, lograr incorporar a personas con alto riesgo de forma temprana a las iniciativas de salud preventiva, esto resulta en una disminución de los costos que el Estado debe desembolsar en salud, ya que disminuiría la cantidad de personas enfermas crónicas que deben ser tratadas por estar descompensadas.

5.4.6. Aprovechar cada atención al máximo

Otro enfoque que se puede dar, es analizar los distintos factores de riesgo de las personas de acuerdo a su historia clínica, y que automáticamente se puedan sugerir los protocolos que correspondan. Así, por ejemplo, en el caso que una persona llegue a un establecimiento de salud a atenderse por un resfriado o una bronquitis, automáticamente se dispare una señal de alerta de acuerdo a los factores de riesgo detectados en atenciones anteriores, y que el médico pueda confirmar los protocolos y los tratamientos sugeridos, sin necesidad de hacer una evaluación muy extensa. En este caso, si la persona presentaba alto riesgo cardiovascular, se le ingresará al PSCV, además de atenderlo por su bronquitis.

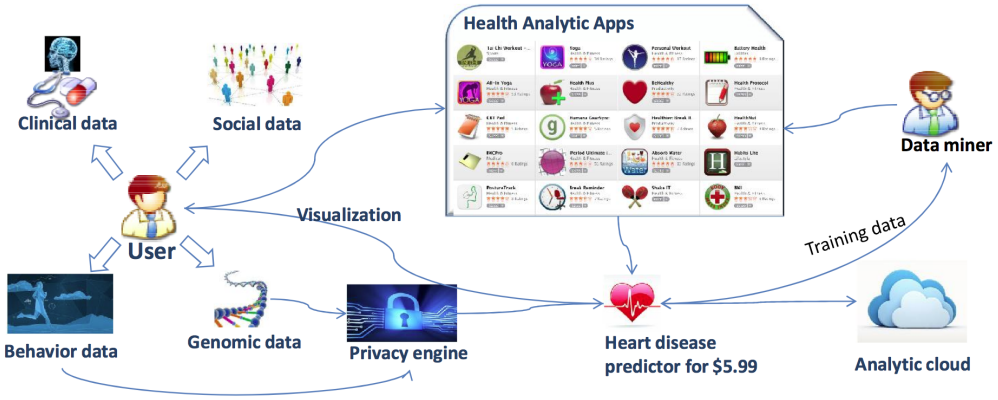


Figura 5.23: Nuevas Aplicaciones de Healthcare Analytics

Esto tiene un alto potencial en las atenciones de urgencia, ya que si analizando la historia clínica es posible detectar alto riesgo, y realizando un seguimiento del estado de su tratamiento, se detecta que no está en control hace un año, y que además no ha retirado sus fármacos en su condición de crónico, con todos estos antecedentes, que pueden ser extraídos a partir del registro clínico electrónico, es posible predecir que cuando llegue a una urgencia, esta persona esté en una crisis. Dada esta información, se podría saber que tratamiento aplicar a una persona con estas características, sin necesidad de gastar mucho tiempo en la evaluación previa, que es muchas veces en estos tiempos de espera que las personas empeoran o incluso mueren.

5.4.7. Investigaciones Relacionadas

Actualmente, en el Departamento de Ingeniería Civil Industrial, se están llevando a cabo investigaciones que permiten realizar análisis de las señales de cuerpo, para encontrar patrones y relacionarlos con el comportamiento de las personas al momento de navegar un sitio web.

Este tipo de investigaciones, aunque se enmarcan en un determinado nicho de negocio, el conocimiento que le da la base puede ser utilizado en otros ámbitos, como la salud. Actualmente, existen aplicaciones que a través de dispositivos móviles inteligentes, permiten monitorear "señales del cuerpo", o signos vitales. Por ejemplo, es posible analizar una señal que corresponde a ritmo cardiaco y detectar patrones que indique algún tipo de arritmia, que es un factor de riesgo cardiovascular.

Por lo que es totalmente posible, considerar el conocimiento utilizado en lo que se conoce como el Proyecto Akori, para que a partir de una base de datos no estructurada, que permita almacenar las señales del cuerpo, y complementado esta información con los Registro Clínicos Electrónicos, sería posible tener una monitoreo y detección de riesgo en tiempo real. Para esto, sería necesario tener una infraestructura que soporte la generación de grandes Volúmenes de datos, los que pueden ser muy Variados, y que necesitan ser procesados de forma Veloz. Esta tecnología ya existe, y se

conoce como Big Data. Este escenario se muestra en la figura 5.23

Considerando el importante rol que tiene nuestra Universidad, es de mucha importancia lograr consolidar el trabajo que se realiza en la academia, para lograr aplicarlo en el sector público, ya que esto tiene un impacto sobre muchas personas.

5.5. Trabajos Futuros

Análisis Estadístico Tradicional para la Población Chilena

Es importante recalcar que aunque este trabajo se enfocó en los datos no estructurados, es posible encontrar una gran cantidad de información valiosa a partir de los datos estructurados.

Parte de los resultados obtenidos, fue validar la teoría médica en relación a las enfermedades cardiovasculares, lo que permite confirmar la calidad de los datos no estructurados disponibles. Sin embargo, la mayoría de los estudios basados en datos estructurados, en Chile, se realizan haciendo referencia a poblaciones similares a la chilena. Con los datos disponibles, es posible ajustar estos estudios o generar nuevos con datos puros de la población nacional, incluso agrupando geográficamente a lo largo de Chile. Por lo que un valor agregado que podría entregar futuros trabajos, es generar estudios específicos para determinadas enfermedades y que se ajusten a nuestra realidad como país.

El resultado de estos estudios permitiría complementar y robustecer el set de variables "No Texto"que se considero en este trabajo.

Análisis avanzado de Texto

Es posible considerar técnicas avanzadas para el análisis de texto. Lo que permitiría obtener una interpretación semántica de las notas clínicas y lograr extraer mayor información.

Por ejemplo, poder detectar diferentes diagnósticos y lograr asociar su código CIE-10 a partir de los datos no estructurados. Lograr detectar hábitos de las personas, y como estos podría estar relacionados con sus diagnósticos. Por otra parte, poder ser un complemento a una línea de investigación que considera lograr una clasificación más acertada de un determinado diagnóstico usando las notas clínicas (SNOMED).

Una aplicación de las búsquedas semánticas en el texto no estructurado tiene relación con detectar la evolución de una enfermedad, o el impacto de un determinado tratamiento. Esto hace relación a lo que se conoce como Análisis de Sentimiento, pero enfocado en el ámbito médico, permitiría saber si la aplicación de un medicamento fue bueno o no, o si un tratamiento mejoró o empeoró la condición de una personas.

5.5.1. Similitud de Pacientes

Considerando el punto anterior, cuando se consolida la información relacionada al efecto de un determinado medicamento o protocolo en un paciente con una determinadas características físicas, es posible realizar análisis de similitud.

Es decir, cada historia clínica puede transformarse en un experimento del que se conocen los resultados. Luego, cuando una persona llega por primera vez con un conjunto de síntomas, es posible encontrar algún grupo de personas que sea similar a él, y poder recomendar los medicamentos o tratamientos que causaron mejor efecto en personas con condiciones similares.

Esto sería de gran ayuda y un gran aporte para aprovechar toda la información que se recopila día a día en los registros clínicos electrónicos.

Capítulo 6

Conclusiones

La sección de conclusiones se divide en dos partes: La confirmación de los resultados esperados y Comentarios a partir del trabajo realizado.

6.1. Resultados Esperados

En primer lugar, cabe mencionar que se da por validada la hipótesis planteada en el presente trabajo, y es posible afirmar que **Existe información valiosa en los campos de texto no estructurado del registro clínico electrónico, para detectar de forma temprana Riesgo Cardiovascular**.

Por otro lado, destaca que los resultados esperados del experimento se cumplen en sus diferentes ámbitos:

- Validar el modelo, ya que en sus resultados generales cumple con los criterios planteados.
- Por otra parte, es consecuente con respecto a la medida actual de riesgo cardiovascular, que se basa en el Score de Framingham.
- Lo más importante es que se concluye que el modelo agrega valor, ya que permite detectar con una medida F-Measure del 75 % y un Accuracy del 78 %, como se muestra en la figura 5.21.
- Otra conclusión importante, es que el modelo presentado es un aporte a la medicina preventiva ya que permite detectar a personas que no tienen clasificación de riesgo cardiovascular de Framingham, lo que permite priorizar al momento de ir a buscar a alguien para integrarlo a los programas de salud preventiva., con una medida F-Measure del 84 % y un Accuracy del 94 %, como se muestra en la figura 5.22.

Además, es posible validar dos enfoques que son presentados en la literatura relacionada a la minería de datos que es: Experimental Insight y Data Insight.

El primero de estos conceptos, se relaciona con la forma habitual de realizar estudios en medicina, y que hasta ahora permite obtener el conocimiento. Sin embargo, se mostró que con el uso de sólo los datos, estos pueden "hablar"por si solos, como se mostró en la sección de análisis de contenidos. Donde se detectaron grupos de personas que van de acuerdo a la teoría médica, como un conjunto de personas que sufrieron un infarto al corazón, y que todos sufrían alguna enfermedad al sistema respiratorio. O que a través de técnicas matemáticas de transformación que permiten explicar de mejor forma la varianza de un conjunto de datos, fue posible detectar condiciones de pacientes que logran explicar de mejor forma el riesgo cardiovascular, como se muestra en la sección de Análisis de Componentes Principales.

Esto refuerza el esquema presentado en el Capítulo 4, con respecto al concepto de Smarter Care, que se basa en estos dos pilares: Experimental Insight y Data Insight.

Por lo que se concluye que es posible comenzar a construir una plataforma de salud más inteligente, utilizando los conocimientos tanto en medicina como en análisis avanzado de datos, para lograr obtener un entendimiento más acabado de la población y chilena, y así poder llegar a entregar mejores políticas públicas en salud.

6.2. Comentarios Finales

Cabe mencionar que este tema está considerado dentro de las tendencias mundiales en salud. Considerando que las principales empresas de tecnología del mundo están dirigiendo sus investigaciones en utilizar los datos no estructurados de los registros clínicos electrónicos.

Por ejemplo, en Julio 2014 se publica la noticia en el diario La Tercera sobre el proyecto Baseline de Google, que declara que se está "investigando la posibilidad de predecir enfermedades, como el riesgo que tiene una persona de sufrir un infarto".

Por otra parte, IBM en Febrero 2014 publica un estudio el que "construye un modelo que permite identificar a apacientes con riesgo cardiovascular?El modelo incluye datos no estructurados de las notas clínicas lo que permite mejorar su accuracy (Accuracy: 85 %)".

Por lo que enfocar en esta línea las investigaciones relacionadas a minería de datos no estructurados, puede llegar a ser un gran aporte. En el sentido que la mayoría de estas empresas se enfocan en el idioma inglés.

Seguir desarrollando y mejorando este tipo de investigaciones y aprovechando la singularidad que tiene Chile de recopilar en un solo gran repositorio de información los datos de más de 70 %

de la población Chile puede ser de gran impacto tanto para las políticas públicas en nuestro país, como para el desarrollo científico en lo que se conoce como Healthcare Informatics o Healthcare Analytics.

Impacto Económico

Cabe mencionar que los resultados obtenidos pueden implicar en un impacto económico en tres ámbitos: desde el punto de vista de las personas, de los establecimiento de salud, y desde el punto de vista del Estado.

- Para las personas: El impacto económico se traduce en una mejor calidad de vida para una persona consigue tratar sus potenciales factores de riesgo con anticipación, lo que impacta en tener que gastar menos en medicamentos o tratamientos.

- Establecimientos de Salud: La cobertura en los programas de salud preventiva se traducen en incentivos económicos para los establecimientos. Con este modelo es posible priorizar a quienes ir a buscar. Por otro lado, y como efecto de mediano/largo plazo, si las personas pueden estar compensadas, no consumen tantos recursos médicos, lo que permitiría bajar la demanda de las atenciones, lo que se traduce en disminuir los costos de operación. En el caso de los establecimientos de salud secundaria y terciaria, el impacto económico es mayor, cuando se logra evitar que una persona llegue descompensada a uno de estos recintos.

- Para el Estado: Uno de los mayores gastos de Estado es en Salud, y esto se debe a que debe financiar el tratamiento de enfermedades con una alta prevalencia y que son muy costosas, ya que en el caso de las enfermedades cardiovasculares, se vuelven crónicas. Cabe mencionar, que casi el 40 % del gasto corresponde a los medicamentos que deben ser administrados a estas personas en su condición de enfermos crónicos. Por lo que las iniciativas que apoyen la Salud Preventiva, y que permitan detectar con anticipación a personas con alto riesgo de padecer este tipo de enfermedades, permitiría ahorrar una gran cantidad de dinero al Estado.

Finalmente, es importante destacar que es necesario formalizar una alianza y trabajo en conjunto con las instituciones que recopilan la vanguardia en el conocimiento, como nuestra Universidad de Chile, y así poder poder aplicarlo en mejorar una problemática real a nivel país, con el fin de ser un aporte a nuestra sociedad.

Bibliografía

[1] O. M. de la Salud, *Prevención de las enfermedades cardiovasculares*. OMS, 2008.

[2] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer Publishing, 2 ed., 2011.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2 ed., 2006.

[4] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Wiley-IEEE Press, 2 ed., 2011.

[5] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.

[6] M. Berry and J. Koqan, *Text Mining: Applications and Theory*. Wiley, 1 ed., 2010.

[7] M. Song and Y. Wu, *Handbook of Research on Text and Web Mining Technologies*. Information Science Reference, 2008.

[8] R. Feldman and J. Sanqer, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.

[9] L. Francis, *Taming Text: An Introduction to Text Mining*. Casualty Actuarial Society, 2006.

[10] M. Bustamante, P. Villarreal, and C. Cisternas, "Comparación de ranking del ministerio de salud (minsal) e impacto financiero de las 40 primeras patologías auge-ges vigentes en chile," *Revista de Administração Pública*, 2010.

[11] R. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records," *Int. J. Med. Inform.*, 2013.

[12] IBM, "Ibm predictive analytics to detect patients at risk for heart failure," *IBM News Room*, Febrero 2014.

[13] P. Chapman, "Crisp-dm 1.0," *Step-by-step data mining guide*, 2000.

[14] F. Gorunescu, "Data mining: Concepts, models and techniques," *Springer*, vol. 12, 2011.

[15] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus, "Knowledge discovery in databases: An overview," *AI magazine*, vol. 13, no. 3, p. 57, 1992.

[16] M. Wernick, Y. Yang, J. Brankov, G. Yourganov, and S. Strother, "Machine learning in medical imaging," *Signal Processing Magazine*, vol. 27, no. 4, pp. 25–38, 2010.

[17] R. Tibshirani, J. Friedman, and ï. Hastie, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Publishing, 2 ed., Agosto 2008.

[18] B. Abma, "Evaluation of requirements management tools with support for traceability- based change impact analysis," *PhD thesis, Master's thesis,University of Twente*, 2009.

[19] H. Karanikas and B. Theodoulidis, "Knowledge discovery in text and text mining software." Centre for Research in Information Management Department of Computation, 2002.

[20] B. Lent, R. Agrawal, and R. Srikant, "Discovering trends in text databases." IBM Almaden Research Center, 1997.

[21] G. Miner, J. Elder, T. Hill, R. Nisbet, D. Delen, and A. Fast, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, 2012.

[22] A. Zanasi, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. WIT Press, 2007.

[23] A. Srivastava and M. Sahami, *Text Mining: Classification, Clustering, and Applications*. Chapman and Hall/CRC, 1 ed., 2009.

[24] M. Berry, *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer Publishing, 2 ed., 2003.

[25] J. Franke, G. Nakhaeizadeh, and I. Renz, *Text Mining: Theoretical Aspects and Applications*. Physica-Verlag HD, 1 ed., 2003.

[26] E. Liddy, "Natural language processing," *In Encyclopedia of Library and Information Science*, 2001.

[27] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[28] C. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge University Press, 2008.

[29] S. DeRose, "Grammatical category disambiguation by statistical optimization," *Computational Linguistics*, vol. 14, no. 1, pp. 31–39, 1988.

[30] E. Brill, "A simple rule-based part of speech tagger," *Proceedings of the workshop on Speech and Natural Language*, pp. 112–116, 1992.

[31] M. Marcus, M. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[32] InterSystem, "Nuevas tendencias: Registro clínico electrónico," 2013.

[33] M. C. ESCOBAR, "Prevención del riesgo cardiovascular: Políticas chilenas," *Revista Médica Clínica Las Condes*, 2012.

[34] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using ehr data challenges, strategies, and a comparison of machine learning approaches," *Medical Care*, vol. 48, no. 6, 2010.

[35] G. de Chile, "Estrategia nacional de salud para el cumplimiento de los objetivos sanitarios de la década 2011-2020," *Metas 2011 - 2020*, 2011.

[36] OMS, "Preventing chronic diseases: a vital investment," *WHO global report*, 2005.

[37] M. A. Chen, *Heart Failure Overview*. U.S. National Library of Medicine, 2013.

[38] MINSAL, "Programa salud cardiovascular: Reorientación de los programas de hipertensión y diabetes.," 2002.

[39] G. de Chile, "Implementación del enfoque de riesgo en el programa de salud cardiovascular," 2009.

[40] A. J. Chamoy, "Advanced analytics to deliver impactful care management and innovative insights for research institutions," *Smarter Healthcare, Patient Care and Insights*, 2014.

[41] S. Yusuf, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study," *INTERHEART Study Investigators*, Septiembre 2004.

[42] M. de Salud, "Informes deis," *Informes DEIS*, 2004.

[43] SuperintendenciaAFP, "Informe superintendencia de afp," *Informe Superintendencia de AFP*, 2001.

[44] I. H. Foundation, "Enfermedades cardiovasculares y cerebrovasculares en las américas 2000," *Interamerican Heart Foundation*, Marzo 2001.

[45] OMS, "Informe sobre la salud en el mundo 2002: Reducir los riesgos y promover una vida sana.," *Ginebra: OMS*, 2002.

[46] E. H. Journal, "European guidelines on cardiovascular disease prevention in clinical practice," *European Heart Journal*, 2003.

[47] S. Smith, "Risk reduction therapies for patients with coronary artery disease: a call for increased implementation," *Am J Med*, 1998.

[48] IBM, "Capitalizing on complexity: Insights from the 2010 ibm global ceo study," *IBM Institute for Business Value*, 2010.

[49] A. Kaushik, *Analítica Web 2.0: El arte de analizar resultados y la ciencia de centrarse en el cliente*. Trama Equipo Editorial S. L., 2011.

[50] J. W. Cortada, D. Gordon, and B. Lenihan, "The value of analytics in healthcare: From insights to outcomes," *IBM Global Business Services*, 2012.

[51] G. de Chile, "Indice de actividad de la atención primaria (iaaps): Orientación técnica y metodológica de evaluación," 2014.

[52] S. KUNSTMANN, "Estratificación de riesgo cardiovascular en la población chilena," *Revista Médica Clínica Las Condes*, 2012.

[53] S. Yusuf, S. Reddy, S. Ounpuu, and S. Anand, "Global burden of cardiovascular diseases: Part ii: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies," *Circulation*, vol. 104, Diciembre 2001.

[54] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," *International Conference on Management of Data*, pp. 255–264, Junio 1997.

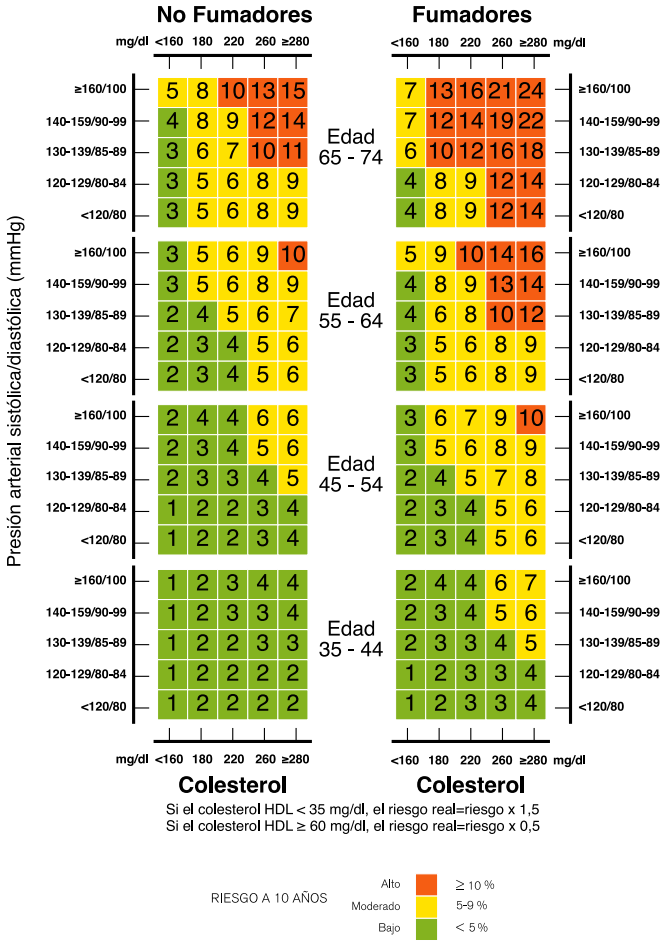
Apéndices

A . Tablas de Framingham

Se anexa el detalle de las Tablas de Framingham para la estimación de riesgo coronario a 10 años adaptadas a la población chilena.

Estas tablas permiten determinar un nivel de riesgo a partir de información clínica de una persona, según su condición de diabético, si es fumador, su edad, sexo, niveles de presión arterial y colesterol.

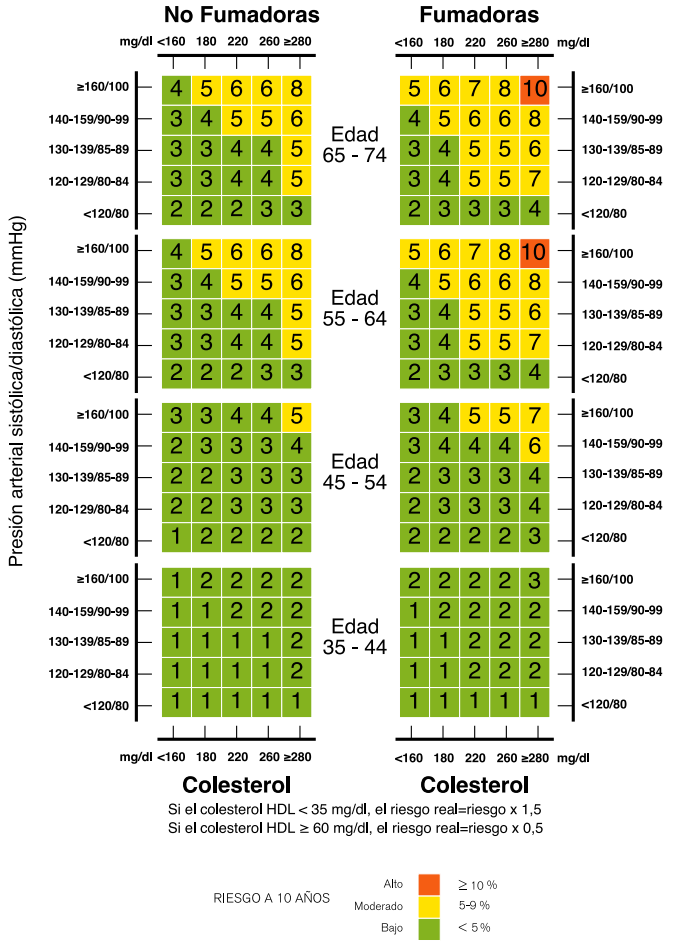
Hombres Diabéticos



Si el colesterol HDL < 35 mg/dl, el riesgo real=riesgo x 1,5
Si el colesterol HDL ≥ 60 mg/dl, el riesgo real=riesgo x 0,5

RIESGO A 10 AÑOS
Alto ≥ 10 %
Moderado 5-9 %
Bajo < 5 %

Mujeres Diabéticas



Si el colesterol HDL < 35 mg/dl, el riesgo real=riesgo x 1,5
Si el colesterol HDL ≥ 60 mg/dl, el riesgo real=riesgo x 0,5

RIESGO A 10 AÑOS
Alto ≥ 10 %
Moderado 5-9 %
Bajo < 5 %

GOBIERNO DE CHILE
MINISTERIO DE SALUD

CONICYT
COMISIÓN NACIONAL DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA

Prevención Primaria de Enfermedad Coronaria

Tablas de Framingham para la estimación de riesgo coronario a 10 años adaptadas a la población chilena.

Información disponible en:
http://www.minsal.cl/ y http://pifrecv.utalca.cl/

Proyecto FONIS SA06I20065 "Tablas de riesgo coronario para la población chilena".

Fonis
FONDO NACIONAL DE INVESTIGACIÓN Y DESARROLLO EN SALUD

UNIVERSIDAD DE TALCA

PIFRECV
Programa de Investigación Factores de Riesgo de Enfermedades Cardiovasculares

Estas tablas, basadas en el estudio de Framingham, se desarrollaron para medir el riesgo coronario (angina, infarto de miocardio silente o con síntomas, mortal o no) a 10 años y se han adaptado a las características de la población chilena siguiendo un procedimiento estándar*.

Para estimar el riesgo coronario a 10 años debe seguir la siguiente secuencia según la información que tenga el (la) usuario (a):

1 Ubicar las tablas correspondientes a la presencia o ausencia de diabetes.

2 Seleccionar la tabla que corresponda a hombre o mujer.

3 En la tabla seleccionada ubique el rango de edad en el que se encuentra el (la) usuario (a).

4 Seleccione la columna no fumador o fumador.

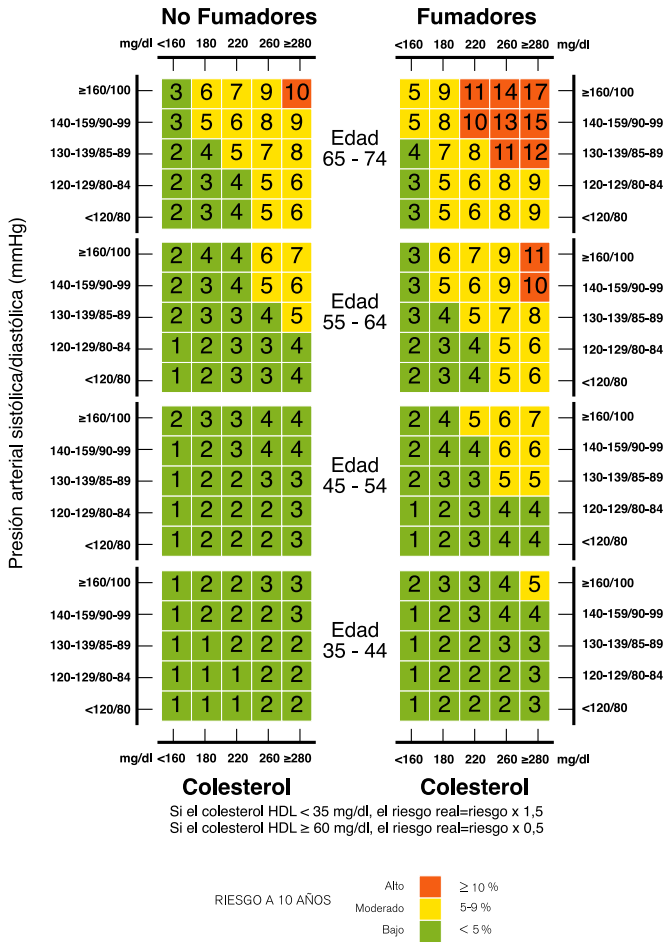
5 Busque la intersección de la presión arterial (sistólica y diastólica) con el colesterol total, ubicando la columna de colesterol con el valor más cercano al valor del usuario. Si el valor del usuario está equidistante entre dos casillas, elija el casillero de riesgo más alto.

6 El valor indicado en la casilla seleccionada muestra el riesgo coronario a 10 años y el color del fondo de la misma, pertenece al código de colores cuya leyenda se encuentra al pie de las tablas.

Estas tablas están hechas para un valor de colesterol HDL entre 35 y 59 mg/dl. Si se dispone del valor del colesterol HDL, puede corregirse el riesgo coronario hallado multiplicando por 1,5 si el valor está por debajo de 35 y por 0,5 si está por encima de 59 mg/dl.

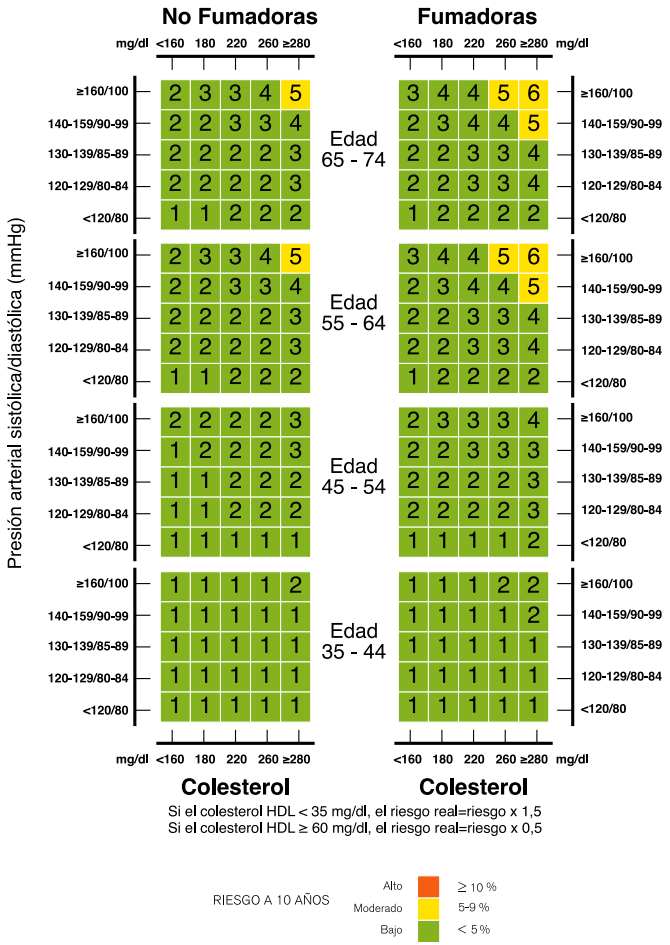
*• Wilson P, D'Agostino R, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. Circulation 1998; 97: 1837-47.
• D'Agostino R, Grundy S, Sullivan L, Wilson P. Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation. JAMA 2001; 286: 180-7.
• Marrugat J, Solanas P, D'Agostino R, Sullivan L, Ordovas J, Cordón F, Ramos R, Sala J, Masia R, Rohlfs I, Elousa R, Kannel W. Estimación del riesgo coronario en España mediante la ecuación de Framingham calibrada. Rev Esp Cardiol 2003; 56(3): 253-61.

Hombres



Si el colesterol HDL < 35 mg/dl, el riesgo real=riesgo x 1,5
Si el colesterol HDL ≥ 60 mg/dl, el riesgo real=riesgo x 0,5

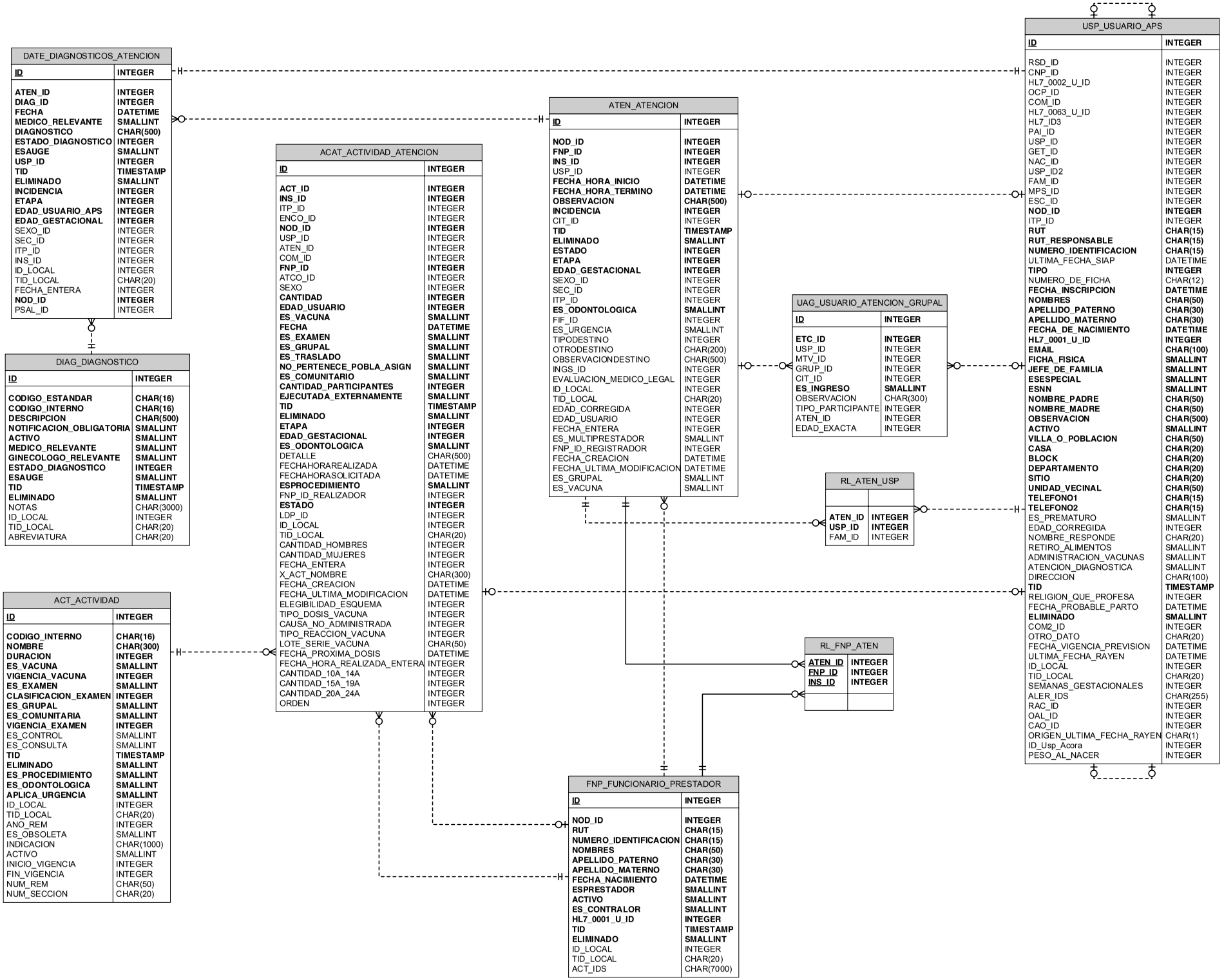
Mujeres



Si el colesterol HDL < 35 mg/dl, el riesgo real=riesgo x 1,5
Si el colesterol HDL ≥ 60 mg/dl, el riesgo real=riesgo x 0,5

B . Modelo Atenciones

Se anexa el modelo entidad relación de las tablas que almacenan los datos del registro clínico electrónico, y que fueron usadas para extraer los datos utilizados en este estudio.


DATE_DIAGNOSTICOS_ATENCION
ID INTEGER
ATEN_ID INTEGER
DIAG_ID INTEGER
FECHA DATETIME
MEDICO_RELEVANTE SMALLINT
DIAGNOSTICO CHAR(500)
ESTADO_DIAGNOSTICO INTEGER
ESAUGE SMALLINT
USP_ID INTEGER
TID TIMESTAMP
ELIMINADO SMALLINT
INCIDENCIA INTEGER
ETAPA INTEGER
EDAD_USUARIO_APS INTEGER
EDAD_GESTACIONAL INTEGER
SEXO_ID INTEGER
SEC_ID INTEGER
ITP_ID INTEGER
INS_ID INTEGER
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
FECHA_ENTERA INTEGER
NOD_ID INTEGER
PSAL_ID INTEGER
DIAG_DIAGNOSTICO
ID INTEGER
CODIGO_ESTANDAR CHAR(16)
CODIGO_INTERNO CHAR(16)
DESCRIPCION CHAR(500)
NOTIFICACION_OBLIGATORIA SMALLINT
ACTIVO SMALLINT
MEDICO_RELEVANTE SMALLINT
GINECOLOGO_RELEVANTE SMALLINT
ESTADO_DIAGNOSTICO INTEGER
ESAUGE SMALLINT
TID TIMESTAMP
ELIMINADO SMALLINT
NOTAS CHAR(3000)
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
ABREVIATURA CHAR(20)
ACT_ACTIVIDAD
ID INTEGER
CODIGO_INTERNO CHAR(16)
NOMBRE CHAR(300)
DURACION INTEGER
ES_VACUNA SMALLINT
VIGENCIA_VACUNA INTEGER
ES_EXAMEN SMALLINT
CLASIFICACION_EXAMEN INTEGER
ES_GRUPAL SMALLINT
ES_COMUNITARIA SMALLINT
VIGENCIA_EXAMEN INTEGER
ES_CONTROL SMALLINT
ES_CONSULTA SMALLINT
TID TIMESTAMP
ELIMINADO SMALLINT
ES_PROCEDIMIENTO SMALLINT
ES_ODONTOLOGICA SMALLINT
APLICA_URGENCIA SMALLINT
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
ANO_REM INTEGER
ES_OBSOLETA SMALLINT
INDICACION CHAR(1000)
ACTIVO SMALLINT
INICIO_VIGENCIA INTEGER
FIN_VIGENCIA INTEGER
NUM_REM CHAR(50)
NUM_SECCION CHAR(20)
ACAT_ACTIVIDAD_ATENCION
ID INTEGER
ACT_ID INTEGER
INS_ID INTEGER
ITP_ID INTEGER
ENCO_ID INTEGER
NOD_ID INTEGER
USP_ID INTEGER
ATEN_ID INTEGER
COM_ID INTEGER
FNP_ID INTEGER
ATCO_ID INTEGER
SEXO INTEGER
CANTIDAD INTEGER
EDAD_USUARIO INTEGER
ES_VACUNA SMALLINT
FECHA DATETIME
ES_EXAMEN SMALLINT
ES_GRUPAL SMALLINT
ES_TRASLADO SMALLINT
NO_PERTENECE_POBLA_ASIGN SMALLINT
ES_COMUNITARIO SMALLINT
CANTIDAD_PARTICIPANTES INTEGER
EJECUTADA_EXTERNAMENTE SMALLINT
TID TIMESTAMP
ELIMINADO SMALLINT
ETAPA INTEGER
EDAD_GESTACIONAL INTEGER
ES_ODONTOLOGICA SMALLINT
DETALLE CHAR(500)
FECHAHORAREALIZADA DATETIME
FECHAHORASOLICITADA DATETIME
ESPROCEDIMIENTO SMALLINT
FNP_ID_REALIZADOR INTEGER
ESTADO INTEGER
LDP_ID INTEGER
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
CANTIDAD_HOMBRES INTEGER
CANTIDAD_MUJERES INTEGER
FECHA_ENTERA INTEGER
X_ACT_NOMBRE CHAR(300)
FECHA_CREACION DATETIME
FECHA_ULTIMA_MODIFICACION DATETIME
ELEGIBILIDAD_ESQUEMA INTEGER
TIPO_DOSIS_VACUNA INTEGER
CAUSA_NO_ADMINISTRADA INTEGER
TIPO_REACCION_VACUNA INTEGER
LOTE_SERIE_VACUNA CHAR(50)
FECHA_PROXIMA_DOSIS DATETIME
FECHA_HORA_REALIZADA_ENTERA INTEGER
CANTIDAD_10A_14A INTEGER
CANTIDAD_15A_19A INTEGER
CANTIDAD_20A_24A INTEGER
ORDEN INTEGER
ATEN_ATENCION
ID INTEGER
NOD_ID INTEGER
FNP_ID INTEGER
INS_ID INTEGER
USP_ID INTEGER
FECHA_HORA_INICIO DATETIME
FECHA_HORA_TERMINO DATETIME
OBSERVACION CHAR(500)
INCIDENCIA INTEGER
CIT_ID INTEGER
TID TIMESTAMP
ELIMINADO SMALLINT
ESTADO INTEGER
ETAPA INTEGER
EDAD_GESTACIONAL INTEGER
SEXO_ID INTEGER
SEC_ID INTEGER
ITP_ID INTEGER
ES_ODONTOLOGICA SMALLINT
FIF_ID INTEGER
ES_URGENCIA SMALLINT
TIPODESTINO INTEGER
OTRODESTINO CHAR(200)
OBSERVACIONDESTINO CHAR(500)
INGS_ID INTEGER
EVALUACION_MEDICO_LEGAL INTEGER
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
EDAD_CORREGIDA INTEGER
EDAD_USUARIO INTEGER
FECHA_ENTERA INTEGER
ES_MULTIPRESTADOR SMALLINT
FNP_ID_REGISTRADOR INTEGER
FECHA_CREACION DATETIME
FECHA_ULTIMA_MODIFICACION DATETIME
ES_GRUPAL SMALLINT
ES_VACUNA SMALLINT
FNP_FUNCIONARIO_PRESTADOR
ID INTEGER
NOD_ID INTEGER
RUT CHAR(15)
NUMERO_IDENTIFICACION CHAR(15)
NOMBRES CHAR(50)
APELLIDO_PATERNO CHAR(30)
APELLIDO_MATERNO CHAR(30)
FECHA_NACIMIENTO DATETIME
ESPRESTADOR SMALLINT
ACTIVO SMALLINT
ES_CONTRALOR SMALLINT
HL7_0001_U_ID INTEGER
TID TIMESTAMP
ELIMINADO SMALLINT
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
ACT_IDS CHAR(7000)
UAG_USUARIO_ATENCION_GRUPAL
ID INTEGER
ETC_ID INTEGER
USP_ID INTEGER
MTV_ID INTEGER
GRUP_ID INTEGER
CIT_ID INTEGER
ES_INGRESO SMALLINT
OBSERVACION CHAR(300)
TIPO_PARTICIPANTE INTEGER
ATEN_ID INTEGER
EDAD_EXACTA INTEGER
RL_ATEN_USP
ATEN_ID INTEGER
USP_ID INTEGER
FAM_ID INTEGER
RL_FNP_ATEN
ATEN_ID INTEGER
FNP_ID INTEGER
INS_ID INTEGER
USP_USUARIO_APS
ID INTEGER
RSD_ID INTEGER
CNP_ID INTEGER
HL7_0002_U_ID INTEGER
OCP_ID INTEGER
COM_ID INTEGER
HL7_0063_U_ID INTEGER
HL7_ID3 INTEGER
PAI_ID INTEGER
USP_ID INTEGER
GET_ID INTEGER
NAC_ID INTEGER
USP_ID2 INTEGER
FAM_ID INTEGER
MPS_ID INTEGER
ESC_ID INTEGER
NOD_ID INTEGER
ITP_ID INTEGER
RUT CHAR(15)
RUT_RESPONSABLE CHAR(15)
NUMERO_IDENTIFICACION CHAR(15)
ULTIMA_FECHA_SIAP DATETIME
TIPO INTEGER
NUMERO_DE_FICHA CHAR(12)
FECHA_INSCRIPCION DATETIME
NOMBRES CHAR(50)
APELLIDO_PATERNO CHAR(30)
APELLIDO_MATERNO CHAR(30)
FECHA_DE_NACIMIENTO DATETIME
HL7_0001_U_ID INTEGER
EMAIL CHAR(100)
FICHA_FISICA SMALLINT
JEFE_DE_FAMILIA SMALLINT
ESESPECIAL SMALLINT
ESNN SMALLINT
NOMBRE_PADRE CHAR(50)
NOMBRE_MADRE CHAR(50)
OBSERVACION CHAR(500)
ACTIVO SMALLINT
VILLA_O_POBLACION CHAR(50)
CASA CHAR(20)
BLOCK CHAR(20)
DEPARTAMENTO CHAR(20)
SITIO CHAR(20)
UNIDAD_VECINAL CHAR(50)
TELEFONO1 CHAR(15)
TELEFONO2 CHAR(15)
ES_PREMATURO SMALLINT
EDAD_CORREGIDA INTEGER
NOMBRE_RESPONDE CHAR(20)
RETIRO_ALIMENTOS SMALLINT
ADMINISTRACION_VACUNAS SMALLINT
ATENCION_DIAGNOSTICA SMALLINT
DIRECCION CHAR(100)
TID TIMESTAMP
RELIGION_QUE_PROFESA INTEGER
FECHA_PROBABLE_PARTO DATETIME
ELIMINADO SMALLINT
COM2_ID INTEGER
OTRO_DATO CHAR(20)
FECHA_VIGENCIA_PREVISION DATETIME
ULTIMA_FECHA_RAYEN DATETIME
ID_LOCAL INTEGER
TID_LOCAL CHAR(20)
SEMANAS_GESTACIONALES INTEGER
ALER_IDS CHAR(255)
RAC_ID INTEGER
OAL_ID INTEGER
CAO_ID INTEGER
ORIGEN_ULTIMA_FECHA_RAYEN CHAR(1)
ID_Usp_Acora INTEGER
PESO_AL_NACER INTEGER

C . Diccionario de Datos

Se anexa el diccionario de datos de las tablas incluidas en el sección anterior (Anexo B).

DESCRIPCIÓN DE TABLAS POR CAMPO

ACAT_ACTIVIDAD_ATENCION		
Columna	**Tipo Dato**	**Descripción**
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada atención en la cual se registra una actividad específica.
ACT_ID	Int	Corresponde al número identificador de la actividad realizada en la atención. La identificación de cada uno de estos valores se encuentra en la tabla ACT_ACTIVIDAD.
INS_ID	Int	Corresponde al identificador del rol profesional que tenía el funcionario prestador al momento de efectuar la atención. La identificación de cada uno de estos valores se encuentra en la tabla INS_INSTRUMENTO.
ITP_ID	Int	Corresponde al número identificador de la institución previsional a la que pertenece el paciente involucrado en la atención. La identificación de cada uno de estos valores se encuentra en la tabla ITP_INSTITUCION_PREVISIONAL
ENCO_ID	Int	Corresponde al número identificador de la entidad comunitaria en la cual se realizó una atención comunitaria.
NOD_ID	Int	Corresponde al número identificador único del establecimiento o entidad de salud donde se realiza la atención. La identificación de cada uno de estos valores se encuentra en la tabla NOD_NODO.
USP_ID	Int	Corresponde al número identificador único de la persona a la cual se registró el valor. La identificación de cada uno de estos valores se encuentra en la tabla USP_USUARIO_APS.
ATEN_ID	Int	Corresponde al número identificador único de la atención en se realizó la actividad al usuario. La identificación de cada uno de estos valores se encuentra en la tabla ATEN_ATENCION.
COM_ID	Int	Corresponde al número identificador de la comuna de nacimiento del usuario al que se le realizó la actividad. Aplica cuando se administra una inmunización de Influenza H1N1.
FNP_ID	Int	Corresponde al número identificador único del funcionario prestador que realiza la atención. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
ATCO_ID	Int	Corresponde al número identificador de la atención comunitaria asociada a este registro.
SEXO_ID	Int	Corresponde al identificador único del sexo de la persona. La identificación de cada uno de estos valores se encuentra en la tabla SEX_SEXO.

FECHAHORAREALIZADA	DateTime	Corresponde a la fecha en años, meses, días y horas, en que se registra la realización o no de un procedimiento en la sala de tratamientos de Sapu o la fecha que se estipula como administración de la vacuna Influenza H1N1.
FECHAHORASOLICITADA	DateTime	Corresponde a la fecha en años, meses, días y horas, en que se solicita realizar un procedimiento en una atención en Sapu.
ESPROCEDIMIENTO	SmallInt	Indica si la actividad de la atención corresponde a un procedimiento clínico.
FNP_ID_REALIZADOR	Int	Corresponde al número de identificación único del funcionario prestador que realiza un procedimiento en una atención en Sapu. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
ESTADO	Int	Corresponde al enumerado del estado de la actividad de salud en la atención.
LDP_ID	Int	Corresponde al número identificador de la lista de precios de la valorización de la actividad realizada en Sapu. La identificación de cada uno de estos valores se encuentra en la tabla LDP_LISTA_DE_PRECIOS.
ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora.
TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
CANTIDAD_HOMBRES	Int	Corresponde al número de hombres participantes en atenciones Comunitarias
CANTIDAD_MUJERES	Int	Corresponde al número de mujeres participantes en atenciones Comunitarias
FECHA_ENTERA	Int	Corresponde a la fecha con valor entera del registro de la atención, para su posterior tratamiento de comparación.
X_ACT_NOMBRE	VarChar(300)	Corresponde a la denominación en la base de datos de la actividad realizada en la atención.
FECHA_CREACION	DateTime	Corresponde a la fecha de creación de este registro, esta es almacenada en años, meses, días y horas.
FECHA_ULTIMA_MODIFICACION	DateTime	Corresponde a la fecha de modificación de la atención, cuando es una atención individual registrada en el módulo de registro de atención.
ELEGIBILIDAD_ESQUEMA	Int	Corresponde al enumerado de los tipos de criterios de elegibilidad por los cuales se administra una dosis de Influenza H1N1.

DESCRIPCION DE TABLAS

ACAT_ACTIVIDAD_ATENCION	Contiene la información y detalle de las actividades realizadas en las distintas atenciones de salud.
ACT_ACTIVIDAD	Contiene las actividades de salud existentes, que corresponden a aquellas acciones preventivas, curativas o diagnósticas que pueden ser realizadas por funcionarios prestadores.
ANAM_ANAMNESIS	Contiene la anamnesis ingresada en una atención de salud.
ATEN_ATENCION	Contiene los registros de las atenciones realizadas en el sistema identificando entre otras cosas al usuario, funcionario prestador, fecha realización, estado, etc.
DATE_DIAGNOSTICOS_ATENCION	Contiene el detalle de los diagnósticos registrados a un usuario en una atención de salud.
DIAG_DIAGNOSTICO	Contiene la información de los distintos diagnósticos del sistema, entendiéndose por diagnóstico a la identificación de la naturaleza de una enfermedad.
FNP_FUNCIONARIO_PRESTADOR	Contiene la información y antecedentes de los funcionarios que pueden ejercer acciones en el sistema, ya sea como funcionarios prestadores o como administrativos.
USP_USUARIO_APS	Contiene la información y antecedentes de cada persona que demanda prestaciones o servicios en los establecimientos de la red asistencial de salud del ministerio.

CANTIDAD	Int	Corresponde al número de veces que se ejecuta la actividad en la atención.
EDAD_USUARIO	Int	Corresponde a la edad de la persona al momento de registrar la actividad. Esta es almacenada en años, meses, días y horas.
ES_VACUNA	SmallInt	Indica si la actividad de la atención corresponde a una inmunización.
FECHA	DateTime	Corresponde a la fecha y hora en la cual se registra la actividad dentro de la atención.
ES_EXAMEN	SmallInt	Indica si la actividad realizada en la atención corresponde a un examen diagnóstico.
ES_GRUPAL	SmallInt	Indica si la atención es grupal.
ES_TRASLADO	SmallInt	Indica si en la atención de urgencia como destino del paciente existe traslado por ambulancia de éste.
NO_PERTENECE_POBLA_ASIGN	SmallInt	Indica si la inmunización fue administrada a un usuario que pertenece a una población residente fuera de la comuna.
ES_COMUNITARIO	SmallInt	Indica si la atención de salud es comunitaria.
CANTIDAD_PARTICIPANTES	Int	Corresponde al número de personas que participan en una atención comunitaria.
EJECUTADA_EXTERNAMENTE	SmallInt	Indica si la inmunización administrada fue aplicada en un establecimiento de salud distinto al que lo registra.
TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.
ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
ETAPA	Int	Corresponde al enumerado del ciclo vital de una mujer. Aplica de 10 a 65 años.
EDAD_GESTACIONAL	Int	Corresponde a la edad de gestación de la mujer en semanas, cuando su etapa (ciclo vital) es Nº 2 o Nº 3.
ES_ODONTOLOGICA	SmallInt	Indica si la actividad realizada en la atención corresponde a una de tipo odontológica.
DETALLE	VarChar(500)	Corresponde a la especificación del artículo (medicamento o insumo) asociado a un procedimiento solicitado en Sapu o el detalle de una inmunización de Influenza H1N1.

ID_INMUNIZACION_RNI	Varchar(2000)	Corresponde al Código de la vacuna para el sistema RNI, asociada con la vacuna administrada. La identificación de cada uno de estos valores se encuentra en la tabla RL_IC_VAC.
OBSERVACION	Varchar(2000)	Corresponde a la observación en texto libre relacionada con la Actividad realizada.
ES_EXAMEN_LABORATORIO	Smalint	Indica si la actividad realizada en la atención corresponde a un examen de laboratorio.
CANTIDAD_2A	Int	Corresponde a la cantidad de personas con edades entre 2 años que participan en una actividad comunitaria.
CANTIDAD_3A	Int	Corresponde a la cantidad de personas con edades entre 3 años que participan en una actividad comunitaria.
CANTIDAD_4A	Int	Corresponde a la cantidad de personas con edades entre 4 años que participan en una actividad comunitaria.
CANTIDAD_5A	Int	Corresponde a la cantidad de personas con edades entre 5 años que participan en una actividad comunitaria.
FNP_ID_MODIFICADOR	Int	Corresponde al número identificador, del Funcionario que realizo la Modificación del registro. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
FNP_ID_ELIMINADOR	Int	Corresponde al número identificador, del Funcionario que realizo la Eliminación del registro. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
FECHA_HORA_ELIMINACION	Datetime	Corresponde a la fecha de Eliminación del registro.

ACT_ACTIVIDAD

Columna	Tipo Dato	Descripción
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada actividad de salud existente en la base de datos., es decir, cualquier acción preventiva, curativa o diagnóstica que puede ser realizada por Funcionarios Prestadores
CODIGO_INTERNO	VarChar(16)	Corresponde a una serie de caracteres que identifican a la actividad en particular.
NOMBRE	VarChar(300)	Corresponde a la denominación de la actividad en la base de datos.

TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
ANO_REM	Int	Corresponde al año de los Resúmenes estadísticos mensuales REM en que se contabiliza la actividad.
ES_OBSOLETA	SmallInt	Indica si la actividad de salud es obsoleta. Estas actividades aparecerán con el semáforo en Rojo en el cliente RAYEN
INDICACION	VarChar(1000)	Corresponde al texto que se visualiza en el semáforo al seleccionar una actividad, indicando entre otras cosas en que REM se contabiliza la actividad.
ACTIVO	SmallInt	Corresponde a un indicador de si la actividad está activa o no.
INICIO_VIGENCIA	Int	Corresponde a la fecha de inicio de vigencia de la actividad. Se indica en años, meses, días.
FIN_VIGENCIA	Int	Corresponde a la fecha de fin de vigencia de la actividad. Se indica en años, meses, días. Se utiliza para actividades utilizadas en los REM.
NUM_REM	VarChar(50)	Corresponde al número del resumen estadístico mensual en que se contabiliza la actividad en cuestión.
NUM_SECCION	VarChar(20)	Corresponde a la sección de un determinado resumen estadístico mensual en que se contabiliza la actividad.
TIPO_INMUNIZACION	Varchar(200)	Indica como observación en texto, el Tipo de inmunización que es la Actividad, cuando ésta es una vacuna (Campaña, Programática o Internacional)
DESCRIPCION	Varchar(1000)	Descripción de la actividad relacionada a las que son Inmunizaciones. Esta observación es visible por los Funcionarios al momento de registrar una Inmunización.
REQUIERE_LOTE	Smallint	Indica si la actividad requiere lote, es decir, si la actividad (Inmunización), tiene lotes asociados.

ANAM_ANAMNESIS

Columna	Tipo Dato	Descripción
ID	Int	Corresponde al identificador único de la atención en que se registró la anamnesis. La identificación de cada uno de estos valores se encuentra en la tabla

TIPO_DOSIS_VACUNA	Int	Corresponde al enumerado de las causas de la no administración de una inmunización. Aplica sólo para administración de Influenza 2010.
CAUSA_NO_ADMINISTRADA	Int	Corresponde al enumerado de las causas de la no administración de una inmunización. Aplica sólo para registro de Influenza 2010.
TIPO_REACCION_VACUNA	Int	Corresponde al enumerado de los tipos de reacción producto de la administración de una inmunización. Aplica sólo para registro de Influenza 2010.
LOTE_SERIE_VACUNA	VarChar(50)	Corresponde al número de lote o serie de las vacunas de Influenza H1N1 que están disponibles para ser aplicadas.
FECHA_PROXIMA_DOSIS	DateTime	Corresponde a la fecha en Días, meses Año en que se debe aplicar una nueva dosis de Influenza H1N1.
FECHA_HORA_REALIZADA_ENTERA	Int	Corresponde a la fecha pero en formato entero de la administración de Influenza H1N1 para facilitar las búsquedas y no sobrecargar al motor de búsqueda de la base de datos.
CANTIDAD_10A_14A	Int	Corresponde a la cantidad de personas con edades entre 10 y 14 años que participan en una actividad comunitaria.
CANTIDAD_15A_19A	Int	Corresponde a la cantidad de personas con edades entre 15y 19 años que participan en una actividad comunitaria.
CANTIDAD_20A_24A	Int	Corresponde a la cantidad de personas con edades entre 20 y 24 años que participan en una actividad comunitaria.
ORDEN	Int	Corresponde al Orden en él se despliegan las Inmunizaciones en pantalla para el Usuario.
LVN_ID	Int	Corresponde al número identificador de la asociación de la Actividad (Vacuna) con el LOTE y el NODO, de la Vacuna administrada. La identificación de cada uno de estos valores se encuentra en la tabla LVN_LOTE_VACUNA_NODO.
EAD_ID	Int	Corresponde al número identificador de la reacción adversa asociada a la vacuna administrada. La identificación de cada uno de estos valores se encuentra en la tabla EAD_EFECTO_ADVERSO.
CANTIDAD_0A_9A	Int	Corresponde a la cantidad de personas con edades entre 0 y 9 años que participan en una actividad comunitaria.
CANTIDAD_10A_19A	Int	Corresponde a la cantidad de personas con edades entre 10 y 19 años que participan en una actividad comunitaria.
CANTIDAD_20A_MASA	Int	Corresponde a la cantidad de personas con edades entre 20 y más años que participan en una actividad comunitaria.
INS_ID_REALIZADOR	Int	Corresponde al número identificador, del Instrumento del Funcionario que administró la vacuna al Usuario. La identificación de cada uno de estos valores se encuentra en la tabla INS_INSTRUMENTO.

DURACION	Int	Corresponde a la duración en minutos de la actividad.
ES_VACUNA	SmallInt	Indica si la actividad realizada corresponde a una inmunización.
VIGENCIA_VACUNA	Int	Corresponde a la cantidad de días que tiene de vigencia una vacuna.
ES_EXAMEN	SmallInt	Indica si la actividad realizada corresponde a un examen diagnóstico.
CLASIFICACION_EXAMEN	Int	Corresponde al enumerado de la clasificación del examen cuando la actividad es de tipo examen.
ES_GRUPAL	SmallInt	Indica si la actividad de salud es de cardinalidad grupal, es decir, si corresponde a un procedimiento clínico realizado por un Funcionario Prestador a dos o varios Usuarios Aps.
ES_COMUNITARIA	SmallInt	Indica si la actividad de salud es de cardinalidad comunitaria, es decir, si corresponde a un procedimiento clínico que realiza un Funcionario Prestador a Varios Usuarios APS o Especiales fuera de las dependencia del consultorio
VIGENCIA_EXAMEN	Int	Corresponde a la cantidad de días que tiene de vigencia un examen diagnóstico.
ES_CONTROL	SmallInt	Indica si la actividad de salud corresponde a un control de salud.
ES_CONSULTA	SmallInt	Indica si la actividad de salud corresponde a una consulta.
TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.
ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
ES_PROCEDIMIENTO	SmallInt	Indica si la actividad realizada corresponde a un procedimiento clínico.
ES_ODONTOLOGICA	SmallInt	Indica si la actividad de salud corresponde a un procedimiento
APLICA_URGENCIA	SmallInt	Indica si la actividad de salud se realiza en urgencia o Sapu.
ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora.

TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.
ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
ESTADO	Int	Corresponde al número identificador del enumerado que indica el estado de la atención, si esta está en estado Comenzada, completada, etc.
ETAPA	Int	Corresponde al enumerado del ciclo vital de la mujer.
EDAD_GESTACIONAL	Int	Corresponde a la edad de gestación de la mujer cuando su etapa es N° 2 o N° 3.
SEXO_ID	Int	Corresponde al identificador único del sexo de la persona. La identificación de cada uno de estos valores se encuentra en la tabla SEX_SEXO.
SEC_ID	Int	Corresponde al número identificador del sector donde fue realizada la atención, guardada en la tabla SEC_SECTOR.
ITP_ID	Int	Corresponde al número identificador de la institución previsional a la que pertenece el paciente involucrado en la atención. La identificación de cada uno de estos valores se encuentra en la tabla ITP_INSTITUCION_PREVISIONAL
ES_ODONTOLOGICA	SmallInt	Indica si la atención es de tipo odontológica.
FIF_ID	Int	Corresponde al identificador del grupo familiar al que pertenece la persona a la cual se está registrando el valor.
ES_URGENCIA	SmallInt	Indica si la atención es de urgencia.
TIPODESTINO	Int	Corresponde al número identificador del destino al cual es enviado en paciente después de esta atención, la identificación de cada uno de estos valores se encuentra en la tabla TDES_TIPO_DESTINO.
OTRODESTINO	VarChar(200)	Corresponde a un cuadro del texto del destino al cual fue enviado el paciente después de la atención, y que no estaba considerado en el enumerado TIPODESTINO.
OBSERVACIONDESTINO	VarChar(500)	Corresponde a un cuadro de texto en el cual el usuario puede ingresar una observación sobre la derivación.
INGS_ID	Int	Corresponde al enlace de esta tabla con la tabla INGS_INGRESO_SAPU en la cual se registró el ingreso del paciente a la atención.
EVALUACION_MEDICO_LEGAL	Int	Corresponde al número identificador del enumerado que indica la gravedad de las lesiones del paciente atendido.

X_ITP_NOMBRE	Varchar(50)	
FECHA_ENTERA	Int	Corresponde a la fecha con valor entera del registro de la atención, para su posterior tratamiento de comparación.
ES_MULTIPRESTADOR	SmallInt	Indica si la atención fue realizada por equipo multiprofesional o no.
FNP_ID_REGISTRADOR	Int	Corresponde al número identificador del funcionario que registra la atención en módulo de atención individual. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
FECHA_CREACION	DateTime	Corresponde a la fecha de creación de este registro, esta es almacenada en años, meses, días y horas.
FECHA_ULTIMA_CREACION	DateTime	Corresponde a la fecha de modificación de la atención, cuando es una atención individual registrada en el módulo de registro de atención.
ES_GRUPAL	Smallint	Indica si es una atención Grupal.
ES_VACUNA	Smallint	Indica a si es una atención de administración de Vacuna.
IURG_ID	Int	Corresponde al número identificador del Ingreso de Urgencia relacionado con la atención de urgencia. La identificación de cada uno de estos valores se encuentra registrada en la tabla IURG_INGRESO_URGENCIA
ES_URGENCIA_SECUNDARIA	Smallint	
ID_FUNCIONARIO_CIERRA_ATENCION_URGENCIA	Int	Corresponde al número identificador del Funcionario que realiza el cierre de la atención de urgencia. La identificación de cada uno de estos valores se encuentra registrada en la tabla FNP_FUNCIONARIO_PRESTADOR
ID_INSTRUMENTO_CIERRA_ATENCION_URGENCIA	Int	Corresponde al número identificador del Instrumento asociado al funcionario prestador que realiza el cierre de la Atención de Urgencia. La identificación de cada uno de estos valores se encuentra registrada en la tabla INS_INSTRUMENTO
LACTANCIA_MATERNA	Smallint	
OSERVACION_POST_CIERRE	Varchar(5000)	Texto Libre que se digita como observación posterior al cierre de una atención de urgencia.
FNP_ID_OBSERVACION_POST_CIERRE	Int	Corresponde al número identificador del Funcionario que realiza la observación posterior al cierre de una atención de urgencia. La identificación de cada uno de

		ATEN_ATENCION.
MOTIVO_CONSULTA	VarChar(500)	Corresponde al texto que se digita en la atención referente a la razón señalada por el usuario Aps para haber solicitado la atención.
HISTORIA_ENFERMEDAD	VarChar (32767)	Corresponde al texto que se digita en la atención con el relato histórico de lo sucedido al usuario Aps en los días semanas o meses precedentes, con relación al motivo de consulta.

ATEN_ATENCION

Columna	Tipo Dato	Descripción
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada registro de atención existente en la base de datos.
NOD_ID	Int	Corresponde al número identificador único del establecimiento o entidad de salud donde se realiza la atención. La identificación de cada uno de estos valores se encuentra en la tabla NOD_NODO.
FNP_ID	Int	Corresponde al número identificador del funcionario prestador de la atención. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
INS_ID	Int	Corresponde al número identificador del atributo profesional del funcionario prestador. La identificación de cada uno de estos valores se encuentra en la tabla INS_INSTRUMENTO.
USP_ID	Int	Corresponde al número identificador único de la persona a la cual se registro el valor. La identificación de cada uno de estos valores se encuentra en la tabla USP_USUARIO_APS.
FECHA_HORA_INICIO	DateTime	Corresponde a la fecha y hora de inicio de la atención, con la que se guardo este registro en la base de datos.
FECHA_HORA_TERMINO	DateTime	Corresponde a la fecha y hora de término de la atención, con la que se guardo este registro en la base de datos.
OBSERVACION	VarChar(500)	Corresponde a un campo de texto libre en el cual el funcionario prestador realizó una observación de la atención
INCIDENCIA	Int	Corresponde a un enumerado que indica el tipo de incidencia a la cual corresponde este registro.
CIT_ID	Int	Corresponde al número identificador de la cita en la cual se realizó la atención, guardada en la tabla CIT_CITA.

ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora
TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
EDAD_CORREGIDA	Int	Corresponde a la corrección de la edad del usuario prematuro al momento de registrar el valor en el metacampo, ésta es almacenada en años, meses, días y horas.
EDAD_USUARIO	Int	Corresponde a la edad de la persona al momento de registrar el valor en el metacampo, ésta es almacenada en años, meses, días y horas.
X_TDA_ID	Int	
X_TDA_NOMBRE	Varchar(100)	
X_INS_NOMBRE	Varchar(50)	
X_USP_NOMBRE	Varchar(150)	
X_USP_RUT	Varchar(15)	
X_USP_FECHA_NAC	DateTime	
X_FNP_NOMBRE	Varchar(150)	
X_CIT_FECHA_HORA_INICIO	DateTime	
X_CIT_FECHA_HORA_TERMINO	DateTime	
X_DURACION	Int	
X_USP_EDAD_ANOS	Int	
X_USP_EDAD_MESES	Int	
X_USP_EDAD_DIAS	Int	

ETAPA	Int	Corresponde a un enumerado de ciclo vital femenino que aplica solo para los usuarios de sexo femenino de entre 10 a 64 años.
EDAD_USUARIO_APS	Int	Corresponde a la edad de la persona al momento de serle registrado el diagnóstico. Ésta es almacenada en años, meses, días y horas.
EDAD_GESTACIONAL	Int	Corresponde a la edad gestacional del usuario cuando su ciclo vital femenino es Embarazada o Embarazada primigesta.
SEXO_ID	Int	Corresponde al identificador único del sexo de la persona al momento de serle registrado el diagnóstico. La identificación de cada uno de estos valores se encuentra en la tabla SEX_SEXO.
SEC_ID	Int	Corresponde al identificador único del sector al que pertenece la persona al momento de registrar el diagnóstico. La identificación de cada uno de estos valores se encuentra en la tabla SEC_SECTOR.
ITP_ID	Int	Corresponde al identificador único de la previsión de la persona al momento de serle registrado el diagnóstico. La identificación de cada uno de estos valores se encuentra en la tabla ITP_INSTITUCION_PREVISIONAL.
INS_ID	Int	Corresponde al identificador único del tipo de profesional o técnico que está registrando el diagnóstico. La identificación de cada uno de estos valores se encuentra en la tabla INS_INSTRUMENTO.
ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora.
TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
FECHA_ENTERA	Int	Corresponde a la fecha con valor entera del registro de la atención, para su posterior tratamiento de comparación.
NOD_ID	Int	Corresponde al número identificador del Nodo en el que se realizó el registro del diagnóstico en la atención. La identificación de cada uno de estos valores se encuentra en la tabla NOD_NODO
PSAL_ID	Int	Corresponde al número identificador del problema de salud relacionado con la clasificación diagnóstica (CIE10) que fue registrada en la atención. La identificación de cada uno de estos valores se encuentra en la tabla PSAL_PROBLEMA_SALUD.
FNP_ID	Int	Corresponde al número identificador del funcionario prestador que realizó el registro del diagnóstico asociada a la atención del paciente. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR.
ES_PRINCIPAL	Smallint	Indica si el diagnóstico registrado es el principal relacionado con la atención en el caso que se haya registrado más de una clasificación diagnóstica en la misma.

TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
ABREVIATURA	VarChar(20)	Corresponde a una abreviatura de la denominación del diagnóstico.

FNP_FUNCIONARIO_PRESTADOR

Columna	Tipo Dato	Descripción
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada registro de funcionarios existentes en la base de datos.
NOD_ID	Int	Corresponde al identificador único del centro de salud al que pertenece el funcionario. La identificación de cada uno de estos valores se encuentra en la tabla NOD_NODO.
RUT	VarChar(15)	Corresponde al Registro único nacional del funcionario.
NUMERO_IDENTIFICACION	VarChar(15)	Corresponde a un Número de Pasaporte o VISA de un funcionario que es extranjero.
NOMBRES	VarChar(50)	Corresponde a la denominación de nombre que tiene el funcionario.
APELLIDO_PATERNO	VarChar(30)	Corresponde al apellido heredado por parte de padre del funcionario.
APELLIDO_MATERNO	VarChar(30)	Corresponde al apellido heredado por parte de madre del funcionario.
FECHA_NACIMIENTO	DateTime	Corresponde a la fecha en que nació el funcionario.
ES_PRESTADOR	SmallInt	Indica si el funcionario realiza actividades de salud.
ACTIVO	SmallInt	Corresponde a un indicador de si este funcionario está activo o no.
ES_CONTRALOR	SmallInt	Corresponde a un indicador de si este funcionario es contralor o no.
HL7_0001_U_ID	Int	Corresponde al sexo del funcionario, la identificación de cada uno de estos valores se encuentra en la tabla HL7_0001_U.
TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.

		estos valores se encuentra registrada en la tabla FNP_FUNCIONARIO_PRESTADOR
FECHA_HORA_OBSERVACION_POST_CIERRE	DateTime	Corresponde a la fecha y hora del registro de la observación posterior al cierre de atención de urgencia

DATE_DIAGNOSTICOS_ATENCION

Columna	Tipo Dato	Descripción
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada atención en la cual es registrado un diagnóstico específico.
ATEN_ID	Int	Corresponde al número identificador de la atención en que se registró un diagnóstico, la identificación de cada uno de estos valores se encuentra en la tabla ATEN_ATENCION.
DIAG_ID	Int	Corresponde al número identificador del diagnóstico registrado en una atención. La identificación de cada uno de estos valores se encuentra en la tabla DIAG_DIAGNOSTICO.
FECHA	DateTime	Corresponde a la fecha de registro del diagnóstico en una atención.
MEDICO_RELEVANTE	SmallInt	Indica si el diagnóstico registrado en la atención es médico relevante.
DIAGNOSTICO	VarChar(500)	Corresponde al texto libre que se digita como complemento del diagnóstico al registrar uno nuevo.
ESTADO_DIAGNOSTICO	Int	Corresponde al enumerado que indica el estado del diagnóstico. Sospecha=0, Confirmado=1 , CierredeCasoAlta=2)
ESAUGE	SmallInt	Indica si el diagnóstico registrado en la atención fue catalogado como GES.
USP_ID	Int	Corresponde al identificador único de la persona a la cual se registró un diagnóstico. La identificación de cada uno de estos valores se encuentra en la tabla USP_USUARIO_APS.
TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.
ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
INCIDENCIA	Int	Corresponde al enumerado que indica la incidencia del diagnóstico (Nueva=1 , Repetida=2)

NO_PERTINENTE	Smallint	

DIAG_DIAGNOSTICO

Columna	Tipo Dato	Descripción
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada diagnóstico del sistema, entendiéndose por diagnóstico a la identificación de la naturaleza de una enfermedad.
CODIGO_ESTANDAR	VarChar(16)	Corresponde a la codificación internacional de enfermedades versión 10 de un diagnóstico.
CODIGO_INTERNO	VarChar(16)	Corresponde a una serie de caracteres que identifican el diagnóstico en particular.
DESCRIPCION	VarChar(500)	Corresponde a la denominación del diagnóstico en la base de datos.
NOTIFICACION_OBLIGATORIA	SmallInt	Indica si el diagnóstico es de notificación obligatoria.es decir, debe ser informado por el establecimiento de salud por tener en su mayoría relevancia epidemiológica.
ACTIVO	SmallInt	Indica si el diagnóstico se encuentra activo o no.
MEDICO_RELEVANTE	SmallInt	Indica si el diagnóstico está clasificado como médico relevante.
GINECOLOGO_RELEVANTE	SmallInt	Indica si el diagnóstico está clasificado como ginecológico relevante.
ESTADO_DIAGNOSTICO	Int	Corresponde a una serie de caracteres que identifican el diagnóstico en particular.
ESAUGE	SmallInt	Indica si el diagnóstico forma parte de las garantías explícitas de salud GES.
TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.
ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
NOTAS	VarChar(3000)	Corresponde al texto que complementa la denominación del diagnóstico.
ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora.

GET_ID	Int	Corresponde al número identificador del grupo étnico al que pertenece el usuario APS registrado, la identificación de cada uno de estos valores se encuentran en la tabla GET_GRUPO_ETNICO.
NAC_ID	Int	Corresponde al número identificador de la nacionalidad del usuario APS registrado, la identificación de cada uno de los valores se encuentra en la tabla NAC_NACIONALIDAD.
USP_ID2	Int	Corresponde al número identificador de usuario de la Madre del usuario APS registrado. Es un campo recursivo de esta misma tabla.
FAM_ID	Int	Corresponde al número identificador familiar del usuario APS registrado, la identificación de cada uno de estos valores se encuentra en la tabla FAM_FAMILIA.
MPS_ID	Int	Corresponde al número de identificación del motivo de desactivación del usuario APS registrado, la identificación de cada uno de estos valores se encuentra en la tabla MPS_MOTIVO.PASIVACION.
ESC_ID	Int	Corresponde al número de identificación de la escolaridad del usuario APS registrado, la identificación de cada uno de estos valores se encuentra en la tabla ESC_ESCOLARIDAD.
NOD_ID	Int	Corresponde al número identificador del nodo al que pertenece el usuario APS registrado, la identificación de cada uno de los valores se encuentra en la tabla NOD_NODO.
ITP_ID	Int	Corresponde al número identificador de la previsión del usuario APS registrado, la identificación de cada uno de los valores se encuentra en la tabla ITP_INSTITUCION_PREVISIONAL.
RUT	VarChar(15)	Corresponde al Rol Único Nacional del usuario APS registrado.
RUT_RESPONSABLE	VarChar(15)	Corresponde al Rol Único Nacional de la persona responsable del usuario APS registrado.
NUMERO_IDENTIFICACION	VarChar(15)	Corresponde a un Número de Pasaporte o VISA de un usuario APS registrado, que es extranjero.
ULTIMA_FECHA_SIAP	DateTime	Corresponde a la última fecha de atención del usuario APS registrado.
TIPO	Int	Corresponde a un enumerado que indica el tipo de usuario APS que se está registrando, la identificación de estos valores los encuentra en el enumerado tipoUsuarioAps.
NUMERO_DE_FICHA	VarChar(12)	Corresponde al número de ficha física del usuario APS registrado.
FECHA_INSCRIPCION	DateTime	Corresponde a la fecha de inscripción del usuario APS en este nodo.

DEPARTAMENTO	VarChar(20)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente el número o letra del departamento de residencia del usuario APS registrado.
SITIO	VarChar(20)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente el número del sitio de residencia del usuario APS registrado.
UNIDAD_VECINAL	VarChar(50)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente la unidad vecinal a la que pertenece el usuario APS registrado.
TELEFONO1	VarChar(15)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente algún número de contacto del usuario APS registrado.
TELEFONO2	VarChar(15)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente algún número de contacto del usuario APS registrado.
ES_PREMATURO	SmallInt	Corresponde a un indicador de si el usuario APS registrado es prematuro.
EDAD_CORREGIDA	Int	Corresponde a un campo numérico donde se guarda la edad exacta del usuario APS registrado.
NOMBRE_RESPONDE	VarChar(20)	Corresponde a un campo alfanumérico en el cual se guarda el nombre al cual responde el usuario APS registrado.
RETIRO_ALIMENTOS	SmallInt	Corresponde a un indicador de si el usuario APS realiza retiros de alimentos programados desde el nodo.
ADMINISTRACION_VACUNAS	SmallInt	Corresponde a un indicador de si al usuario APS registrado se le administran vacunas.
ATENCION_DIAGNOSTICA	SmallInt	Corresponde a un indicador de si el usuario APS recibe atención diagnostica.
DIRECCION	VarChar(100)	Corresponde a un campo alfanumérico donde se guarda la Vía y el número de la residencia del usuario APS registrado.
TID	Timestamp	Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia.
RELIGION_QUE_PROFESA	Int	Corresponde a un enumerado que identifica la religión que profesa el usuario APS registrado, la identificación de estos valores se encuentran en el enumerado religión.
FECHA_PROBABLE_PARTO	DateTime	Corresponde a una fecha estimativa de parto, para un usuario APS registrado en estado de gravidez.
ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
COM2_ID	Int	Corresponde al número identificador de la comuna de residencia del usuario APS registrado.

ELIMINADO	SmallInt	Representa la eliminación del registro en la base de datos, por lo tanto, no es tomado en cuenta para ninguna estadística.
ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora.
TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
ACT_IDS	VarChar(7000)	Corresponde a la desnormalización de los ID, de todas las actividades que tiene asociadas el funcionario.

USP_USUARIO_APS

Columna	Tipo Dato	Descripción
ID	Int	Corresponde a un número correlativo y único que funciona como identificador de cada registro de estado de citas existente en la base de datos.
RSD_ID	Int	Corresponde a un número de identificación de la residencia del usuario, la identificación de cada uno de estos valores se encuentra en la tabla RSD_RESIDENCIA.
CNP_ID	Int	Corresponde al número identificador del nivel previsional del usuario APS, la identificación de cada uno de estos valores se encuentra en la tabla CNP_CONVENIO_PREVISIONAL.
HL7_0002_U_ID	Int	Corresponde al número identificador del estado civil del usuario, la identificación de cada uno de estos valores se encuentra en la tabla HL7_0002_U.
OCP_ID	Int	Corresponde al número identificador de la ocupación del usuario, la identificación de cada uno de estos valores se encuentra en la tabla OCP_OCUPACION.
COM_ID	Int	Corresponde al número identificador de la comuna de nacimiento del usuario APS. La identificación de cada uno de estos valores los encuentra en la tabla COM_COMUNA.
HL7_0063_U_ID	Int	Corresponde al número identificador del parentesco con el jefe de hogar del usuario APS registrado, la identificación de cada uno de estos valores se encuentra en la tabla HL7_0063_U.
HL7_ID3	Int	Corresponde a un campo que no se utiliza.
PAI_ID	Int	Corresponde al número identificador del País del usuario temporal registrado, la identificación de cada uno de estos valores se encuentra en la tabla PAI_PAIS.
USP_ID	Int	Corresponde al número identificador de usuario del Padre del usuario APS registrado. Es un campo recursivo de esta misma tabla.

NOMBRES	VarChar(50)	Corresponde a la denominación de nombre del usuario existente en la base de datos.
APELLIDO_PATERNO	VarChar(30)	Corresponde al apellido heredado por parte de padre del usuario APS registrado.
APELLIDO_MATERNO	VarChar(30)	Corresponde al apellido heredado por parte de madre del usuario APS registrado.
FECHA_NACIMIENTO	DateTime	Corresponde a la fecha de nacimiento del usuario APS registrado.
HL7_0001_U_ID	Int	Corresponde al número identificador del sexo del usuario APS registrado, la identificación de cada uno de estos valores se encuentra en la tabla HL7_0001_U.
EMAIL	VarChar(100)	Corresponde a un campo de texto libre en el cual se puede cargar la dirección electrónica del usuario APS registrado.
FICHA_FISICA	SmallInt	Corresponde a un indicador de si existe la ficha física del usuario APS registrado.
JEFE_DE_FAMILIA	SmallInt	Corresponde a un indicador de si este usuario APS registrado es jefe de familia.
ESESPECIAL	SmallInt	Corresponde a un indicador de si este usuario APS registrado es especial.
ESNN	SmallInt	Corresponde a un indicador de si este usuario APS registrado es N.N.
NOMBRE_PADRE	VarChar(50)	Corresponde a la denominación de nombre del Padre del usuario APS registrado.
NOMBRE_MADRE	VarChar(50)	Corresponde a la denominación de nombre de la Madre del usuario APS registrado.
OBSERVACION	VarChar(50)	Corresponde a un campo de texto libre en la cual se puede cargar una observación.
ACTIVO	SmallInt	Corresponde a un indicador de si este usuario APS registrado está activo.
VILLA_O_POBLACION	VarChar(50)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente la denominación de la villa o población donde reside el usuario APS registrado.
CASA	VarChar(20)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente la dirección de residencia del usuario APS registrado.
BLOCK	VarChar(20)	Corresponde a un campo alfanumérico en el cual se guarda preferentemente el número del edificio de residencia del usuario APS registrado.

INSCRIBE	SmallInt	
ES_ADMISION_BREVE	SmallInt	Indica si el registro de la admisión del paciente fue una admisión abreviada.
ORIGEN_ADMISION_BREVE	Int	Indica el motivo de la admisión abreviada del paciente. El enumerado del origen de la admisión abreviada es: Ninguno = 0, SAPU = 1, VacunasCampaña = 2, FichaFamiliar = 3, EntregaAlimentos = 4, Vacunas = 5, UrgenciaSecundaria = 6
FNP_ID_REGISTRADOR	Int	Corresponde al identificador único del Funcionario que ha realizado la admisión abreviada al paciente. La identificación de cada uno de estos valores se encuentra en la tabla FNP_FUNCIONARIO_PRESTADOR
FECHA_NACIMIENTO_ENTERA	Int	Corresponde a la fecha de nacimiento del Usuario en formato entero (AAAAMMDD).
ACCEDE_PORTAL	SmallInt	
FACTOR_RH	Int	Corresponde al Factor RH del Grupo Sanguineo del Usuario registrado en la admisión de Urgencia. El enumerado de los distintos Factores RH son los siguientes: Ninguno = 0, Negativo = 1, Positivo = 2
GRUPO_SANGUINEO	Int	Corresponde al Grupo Sanguíneo del Usuario que fue registrado en la admisión de Urgencia. El enumerado de los distintos grupos sanguíneos son los siguientes: Ninguno = 0, O = 1, A = 2, B = 3, AB = 4
RUN_TITULAR_BENEFICIARIO	Varchar(20)	Corresponde al RUN del usuario titular registrado al momento de realizar una admisión abreviada.
CLAVE_PORTAL	Varchar(10)	

OTRO_DATO	VarChar(20)	Corresponde a un campo alfanumérico en el cual se guarda otro dato que no haya estado considerado dentro del sistema.
FECHA_VIGENCIA_PREVISION	DateTime	Corresponde a la fecha de vigencia de la previsión del usuario APS registrado.
ULTIMA_FECHA_RAYEN	DateTime	Corresponde a la fecha en la cual se realizó por última vez un registro en el sistema, al usuario APS.
ID_LOCAL	Int	Acora. Corresponde a un número correlativo y único que funciona como identificador al interior del servidor acora.
TID_LOCAL	VarChar(20)	Acora. Identificador único autogenerado en cada inserción o modificación de registro usado para controlar concurrencia en la BD del servidor acora.
SEMANAS_GESTACIONALES	Int	Corresponde a un campo numérico en el cual se guardan las semanas de gestación del usuario APS registrado en estado de gravidez.
ALER_IDS	VarChar(255)	Corresponde a los valores de alertas clínicas de los usuarios APS registrados.
RAC_ID	Int	Corresponde al número identificador de la actividad económica del usuario APS registrado, la identificación de estos valores se encuentran en la tabla RAC_RAMA_ACTIVIDAD.
OAL_ID	Int	Corresponde al número identificador del organismo administrador al cual pertenece el usuario APS registrado, la identificación de estos valores se encuentran en la tabla OAL_ORGANISMO_ADMINISTRADOR_LEY.
CAO_ID	Int	Corresponde al número identificador de la categoría ocupacional del usuario APS registrado en la base de datos, la identificación de estos valores se encuentran en la tabla CAO_CATEGORIA_OCUPACIONAL.
ORIGEN_ULTIMA_FECHA_RAYEN	VarChar(1)	Corresponde al lugar donde se registró por última vez, una atención al usuario APS.
ID_Usp_Acora	Int	
PESO_AL_NACER	Int	Corresponde al peso al nacer del Usuario inscrito en gramos.
TIPO_BENEFICIARIO	Int	Indica el Tipo de Beneficiario que es el paciente inscrito. El enumerado de los distintos Tipos de Beneficiarios son los siguientes: 1 Titular, 2 Carga.
ID_RYF	Varchar(20)	
FECHA_ULTIMA_MODIFICACION	DateTime	Corresponde a la fecha de última fecha de modificación de datos del Usuario.

D . Procesamiento de Texto en R

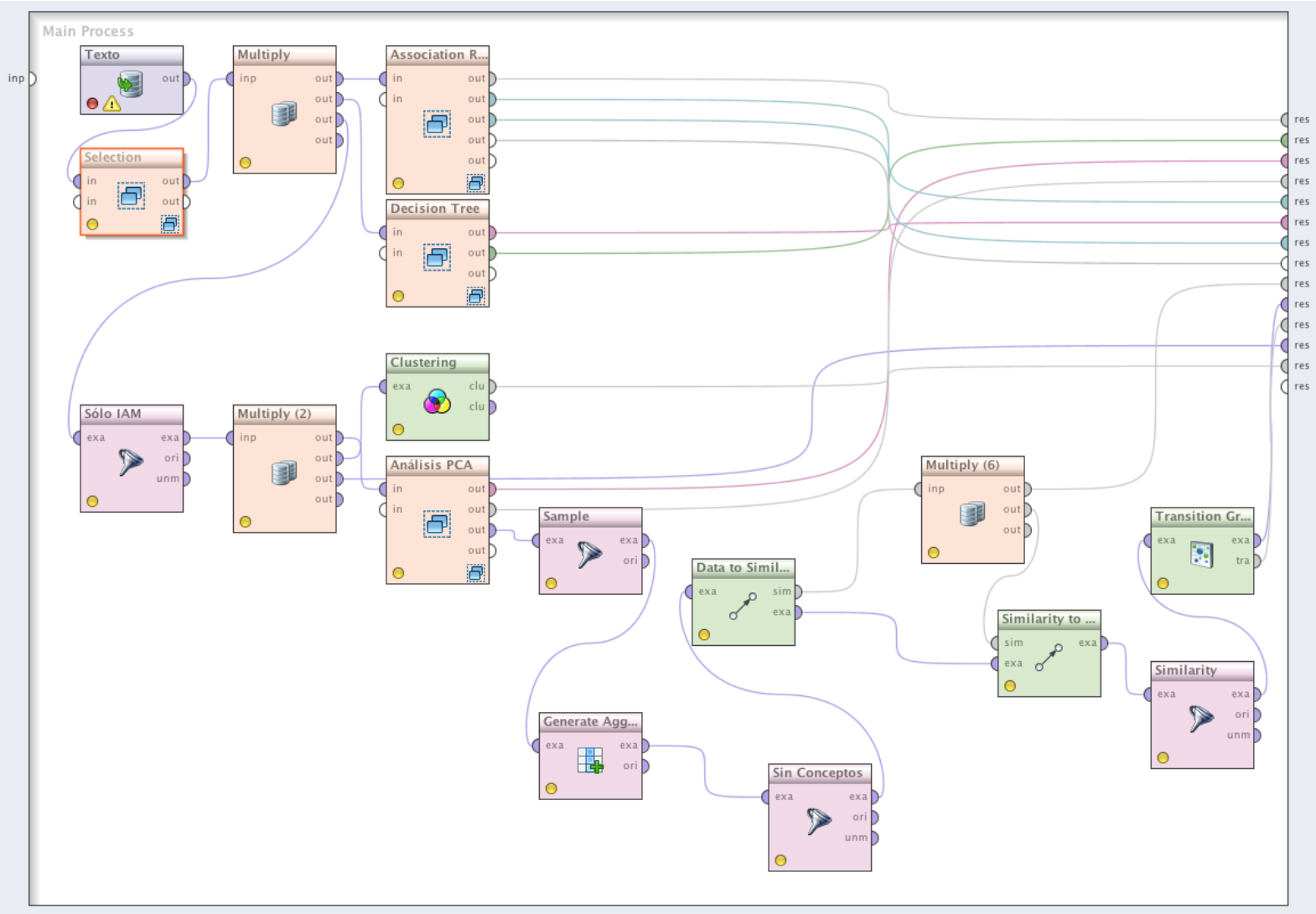
```
library("tm")
library("qdap")

file_data <-
read.csv("SS_Test109_Input_R.txt",header=T,sep="\t")
data <- as.vector(file_data)

#create corpus
text_corpus <- VCorpus(VectorSource(data))

#clean up

text_corpus <- tm_map(text_corpus, tolower)
text_corpus <- tm_map(text_corpus,
removePunctuation)
text_corpus <- tm_map(text_corpus,
function(x)removeWords(x,stopwords(kind="es")))

#Tdm
dat <- data.frame(text_corpus)
dat_wfm <- with(dat, wfm(Texto, USP_ID))
dat_output2 <-data.frame(dat_wfm)
dat_output <-data.frame(dat_tfidf)
dat_tfidf <- apply_as_tm(dat_wfm, tm:::weightTfIdf)
dat_output <-data.frame(dat_tfidf)
write.csv(dat_output, file =
"SS_Test109_Output_R.csv")
```

E. Query de Extracción

```
--
-- Texto de los diagnosticados con IAM durante el 2013

select *

from (
-- EXTRAER TEXTO COLUMNAS DATE, ATEN, DIAG, ANAM
select dat.ID,dat.ATEN_ID,tda.id as TDA_ID,tda.NOMBRE
TIPO_ACTIVIDAD,prog.DESCRIPCION
PROGRAMA,dat.DIAG_ID,dat.FECHA_ENTERA,dat.DIAGNOSTICO,dat.ESTADO_DIAGNOSTICO,
dat.USP_ID,dat.EDAD_USUARIO_APS,dat.SEXO_ID,dat.INS_ID,dat.NOD_ID,dat.FNP_ID,
aten.ES_URGENCIA_SECUNDARIA,diag.CODIGO_ESTANDAR,diag.CODIGO_INTERNO,diag.DES
CRIPCION,diag.NOTAS,ana.HISTORIA_ENFERMEDAD,ana.MOTIVO_CONSULTA
from DATE_DIAGNOSTICOS_ATENCION dat
inner join ATEN_ATENCION aten on aten.ID=dat.ATEN_ID
inner join CIT_CITA cit on cit.ID=aten.CIT_ID
inner join TDA_TIPO_DE_ATENCION tda on tda.ID=cit.TDA_ID
left join PROG_PROGRAMA_SALUD prog on prog.ID=tda.PROG_ID
inner join DIAG_DIAGNOSTICO diag on diag.ID=dat.DIAG_ID
inner join ANAM_ANAMNESIS ana on ana.ID=dat.ATEN_ID
--where dat.DIAG_ID=3827
--and dat.NOD_ID=2978
and YEAR(fecha) in (2010,2011,2012,2013)) atte

right join

(-- EXTRAE A LOS IAM DE NOd 2978,1757,938
select distinct tab.usp_id, tab.FECHA_ENTERA AS
FECHA_DIAGNOSTICO_IAM,tab.DIAG_ID,tab.DESCRIPCION
from (select
dat.ID,dat.ATEN_ID,dat.DIAG_ID,dat.FECHA_ENTERA,dat.DIAGNOSTICO,dat.ESTADO_DI
AGNOSTICO,dat.USP_ID,dat.EDAD_USUARIO_APS,dat.SEXO_ID,dat.INS_ID,dat.NOD_ID,d
at.FNP_ID,aten.ES_URGENCIA_SECUNDARIA,diag.CODIGO_ESTANDAR,diag.CODIGO_INTERN
O,diag.DESCRIPCION,diag.NOTAS,ana.HISTORIA_ENFERMEDAD,ana.MOTIVO_CONSULTA
from DATE_DIAGNOSTICOS_ATENCION dat
inner join ATEN_ATENCION aten on aten.ID=dat.ATEN_ID
inner join DIAG_DIAGNOSTICO diag on diag.ID=dat.DIAG_ID
inner join ANAM_ANAMNESIS ana on ana.ID=dat.ATEN_ID
inner join NOD_NODO nod on nod.ID=dat.NOD_ID

--CIE-10 Relacionados a IAM
where (DIAG_ID between 3827 and 3833
or DIAG_ID between 14530 and 14541)
and nod.NOD_ID = 110
--and dat.NOD_ID in (2978,1757,938)
and YEAR(fecha) in (2013)
--and (dat.FECHA='2013-06-12 15:22:49.013'
--or FECHA='2013-12-18 08:31:18.210')
) as tab

) as tab1 on tab1.USP_ID=atte.USP_ID

WHERE atte.FECHA_ENTERA < tab1.FECHA_DIAGNOSTICO_IAM
```

F . Procesamiento en Rapid Miner - Análisis de Contenido

Se presenta el esquema general del procesamiento de los datos en Rapid Miner para realizar un análisis de contenido.



G . Procesamiento en Rapid Miner - Análisis Predictivo

Se presenta el esquema general para realizar el Análisis Predictivo usando los campos de texto no estructurados del registro clínico electrónico (RCE).