

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE
TELECOMUNICACIÓN

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS MATEMÁTICAS

**MÁSTER EN TRATAMIENTO ESTADÍSTICO
COMPUTACIONAL DE LA INFORMACIÓN**



TRABAJO DE FIN DE MÁSTER

**Desarrollo y validación de una metodología para
reutilización automática de historia clínica electrónica
en el ámbito de los estudios epidemiológicos**

Diego Fernández López

Madrid, 201X

Director:Jorge Tello Guijarro

Ponente universidad:Antonio Bru Espino

Resumen

En el presente documento se desarrollan las distintas técnicas y herramientas propias de la epidemiología. Centrándonos en los distintos tipos de estudios, los modelos más usados y las precauciones que hay que tomar al interpretar los datos (sesgos existentes). Propondremos para el caso concreto de los estudios de prevalencia un método alternativo al clásico, mediante el uso de recuperación de información de historia clínica electrónica. Comparándose la precisión de ambos métodos. Finalmente comentaremos brevemente las posibilidades del medio y mostraremos un pequeño experimento donde usaremos la información recuperada de informes médicos.

Abstract In the present document the different techniques and tools of epidemiology will be developed. Focusing on the different types of studies, the most used models and the precautions to be taken in interpreting the data (existing biases).

It will be proposed for the specific case of prevalence studies an alternative method to the classic, through the use of electronic medical record information retrieval. It will be compared the accuracy of both methods.

Finally it will be briefly comment on the possibilities of the medium and show a small experiment where it will be used information retrieved from medical reports.

Índice

1. Introducción	3
2. Objetivos del trabajo y breve motivación	5
3. Contexto general general de la epidemioología	6
4. Tipos de estudios epidemiológicos:	7
4.1. Segundo Finalidad	7
4.2. Segundo unidad de análisis:	7
4.3. Direccionalidad	8
4.4. Selección de la muestra	8
4.5. Relación temporal o proximidad	8
4.6. Control de la asignación de los factores de estudio:	9
4.7. Clasificación de estudios epidemiológicos:	9
4.7.1. Estudios de Prevalencia:	9
4.7.2. Cohorte	10
4.7.3. Estudio de Casos y Controles	11
4.7.4. Ecológico o de conglomerado	13
4.7.5. Estudios experimentales:	15
5. Sesgos y valores de confusión:	16
5.1. Sesgo de selección	16
5.2. Sesgo de información	17
5.3. Sesgo de confusión	18
6. Modelos bioestadísticos más frecuentes	19
6.1. Prueba de independencia variables categóricas:	19
6.2. Modelo Lineal	20
6.3. Modelo logístico	21
6.4. Curva de Kaplan-Meier y modelos de regresión de Cox :	22
7. Modelización	27
7.1. Método clásico:	27
7.1.1. Tamaño de la muestra:	27
7.2. Método propuesto:	28
8. Estadísticos aplicables al sistema	31
8.1. Primera aproximación	31
8.2. Formulación Matricial:	34
8.3. Segunda aproximación	35
8.4. Enfermedades poco frecuentes:	37
9. Experimento	39
9.1. Comunidad 1	42
9.2. Comunidad 2	43
9.3. Comunidad 3	44
9.4. Comunidad 4	45
9.5. Comunidad 5	46
9.6. Comunidad 6	47

10. Conclusiones	48
10.1. Sesgos:	48
10.1.1. Selección:	48
10.1.2. Sesgo de información	48
10.1.3. Sesgo de confusión	48
10.2. Posibilidades del medio:	49
10.3. Correlación no implica causalidad	49
Bibliografía	50

1. Introducción

En la presente época la tecnología permite, con relativamente poco esfuerzo, el acceso y gestión de una gran cantidad de información creando lo que se ha acuñado bajo el término “Big Data”.

El big data persigue conocer la población y definirla a un nivel nunca antes descrito, encontrar relaciones entre numerosas variables y pistas para los investigadores que no podrían ver de otra manera.

¿Pero qué define el Big Data? Sin duda es un tema que suscita cierta controversia y no hay un consenso, aunque aquí señalaré 3 características:

- Muchos Datos: Gigas de información en el menor de los casos.
- Datos desestructurados: Los datos no suelen guardarse en formatos estandarizados como podría ser una tabla o un base de datos sql clásicas.
- Datos “poco fiables”: la obtención de los datos, puede tener errores que hay que limpiar. Un ejemplo en este contexto de esto reside en que muchas enfermedades se suelen escribir como acrónimos y estos acrónimos pueden llevar a dar falsos positivos. Si bien hay métodos que pueden ayudar a la eliminación de la ambigüedad es un factor a tener en cuenta.

La causa de los dos últimos casos proviene de que los datos no son creados con el objetivo de alimentar estadísticas. Por ejemplo, las fotos no se han sacado para enseñar a una red neuronal a diferenciar entre un perro y un gato, los “amigos” de Facebook no se componen para estudiar las relaciones humanas..

Un ejemplo de tipos de datos que está omnipresente y es difícil de usar por máquinas es el lenguaje natural, es decir, el lenguaje humano. Son numerosos los esfuerzos en este campo.

Savana se mueve en este último caso, al servicio de la Medicina, donde un software informático lee los documentos generados de forma rutinaria por los médicos, denominados historias médicas, en los que se reflejan la evolución enfermedades de los pacientes los cuales se escriben de forma rutinaria por los médicos usando el lenguaje natural.

Por supuesto estos datos son altamente sensibles, razón por la cual la empresa toma una serie de políticas y precauciones para anonimizar los datos de la base de datos, además de las habituales políticas de firewall, software, ...

Entre las posibilidades de este enfoque radican, el poder seguir la pista de los pacientes, saber sus antecedentes los tratamientos que se les aplican en general sus evoluciones, etc.

En el presente trabajo nos centraremos en como usar esta tecnología con el fin de estudiar la frecuencia de enfermedades, señalar cuales pueden ser los tanto las ventajas como las desventajas de reusar las historias médicas frente a los métodos “clásicos”.

Con este objetivo, veremos cuales son los métodos clásicos, entre los que destacan los estudios de cohortes, prevalencia, etc.

Señalaremos las dificultades, sesgos, más frecuentes de estos estudios, así como los modelos estadísticos usados.

Con esto trataremos el caso concreto de un estudio de prevalencia y estudiaremos como se puede realizar mediante la recuperación de datos de informes médicos usando el software **Savana**.

Mostraremos como medir el error cometido por ambos métodos y los compararemos de forma teórica.

Finalmente con el objetivo de mostrar el potencial del medio, mostraremos un estudio transversal de las enfermedades.

2. Objetivos del trabajo y breve motivación

En un primer lugar aprovechamos que los médicos deben anotar, de forma rutinaria, las historias médicas, esto es, informes que detallan los síntomas, antecedentes familiares, resultados de prueba, diagnósticos... referentes a la consulta de un paciente.

La motivación de la creación de estas historias es la de poder constatar la evolución del paciente y usar ésto en futuras decisiones médicas.

Estas historias suponen un testimonio de los sucesos médicos de la población, suponiendo una información muy útil como fuente de investigación para la medicina.

El problema de usar estas fuentes reside en que las notas carecen de una estructura fácilmente reusable como podría ser una tabla (tipo excel) o formulario a la hora de usar realizar estudios médicos, sino que se añaden en forma de texto libre, es decir expresiones del tipo “el chico le duele la cabeza”. Cualquier intento de transferir manualmente la información a una base de datos es prácticamente inviable debido a la ingente cantidad de informes que se generan.

Por otro lado, cabría la posibilidad de que los médicos llenaran informes estructurados. Si bien ha habido intentos en esta dirección, han sido rechazados por los médicos debido a la dificultad de estos sistemas (los cuales acaban conformando una cantidad ingente de desplegables) y acaban siendo una traba en el desempeño de las obligaciones de los médicos, por lo general muy ocupados.

Con estos problemas e inconvenientes en mente nació la empresa Savana, cuyo enfoque es recuperar la máxima información posible de las historias médicas usando técnicas apropiadas, la denominada programación lingüística computacional.

Esto significa que programas informáticos son capaces de recuperar y resumir la información escrita en los informes existentes (los cuales pueden ascender a varios millones), los cuales están escritos en lenguaje natural.

En este trabajo nos marcaremos distintos objetivos que consideramos útiles tanto para la empresa como para la formación del alumno que efectúa el trabajo con el objetivo de usar estos datos utilidad en el campo de la bioestadística:

1. Investigar como se efectúan los estudios sobre las poblaciones médicas y que ayudan a avanzar a la ciencia (mostrando correlaciones, avisando sobre la importancia de las enfermedades...)
2. Mostrar los modelos estadísticos más frecuentes que se usan para ayudar a interpretar los datos obtenidos en los datos médicos.
3. Mostrar las limitaciones y dificultades de los estudios, así como las principales variables que puedan invalidar o restringir la validez de los estudios (sesgos)
4. Profundizar en la capacidad de los estudios de calcular la prevalencia una enfermedad (intervalos de confianza) y modelizar como funcionaría este bajo el supuesto que tenemos ciertos niveles de error al detectar los pacientes con la enfermedad.

3. Contexto general general de la epidemioología

La epidemiología es la disciplina científica encargada de estudiar la frecuencia y distribución de fenómenos relacionados con la salud , sus determinantes en las poblaciones, y la aplicación de este estudio al control de problemas de salud.

La epidemiología no se restringe únicamente al estudio de enfermedades sino todo tipo de fenómenos relacionados con la salud , entre los que se encuentran causas de muerte como los accidentes o suicidios, hábitos de vida como el consumo de tabaco o la dieta y el uso de servicios de salud o la calidad de vida relacionada con la salud, entre otros.

Las causas de estos fenómenos abarcan todos los factores físicos, biológicos, sociales, culturales y de comportamiento que influyen sobre la salud. Los fenómenos relacionados con la salud y sus posibles causas dan lugar a algunas de las clasificaciones de las ramas de la epidemiología.

El objetivo final de la epidemiología es responder a las 3 preguntas *hipocráticas* de la prevención:

- **La enfermedad o problema de salud, ¿Se puede prevenir o controlar:**
- **Si la enfermedad se puede prevenir, ¿Cuáles son las estrategias de prevención y control más adecuadas?**
- **¿Cuál es la magnitud del beneficio de las estrategias de preventión?**

4. Tipos de estudios epidemiológicos:

Para poder responder a las preguntas hipocráticas y avanzar así en esta ciencia se efectúan distintos tipos de estudios epidemiológicos, los cuales se pueden clasificar atendiendo a distintos criterios.

En un sentido amplio, se entiende por estudio epidemiológico a cualquier actividad en la que se recurre al método epidemiológico para profundizar en el conocimiento de temas relacionados con la salud. En la práctica, la mayoría de los estudios epidemiológicos tienen como objetivo aportar información útil a la toma de decisiones en la planificación o gestión de actividades relacionadas con la salud. No obstante, tienen especial interés aquellos dirigidos a ampliar el conocimiento científico sobre un tema concreto y, cuando cumplen condiciones adecuadas para ello, pueden considerarse verdaderos estudios de investigación.

El método de investigación epidemiológica, como variante del método científico-experimental, consta de las siguientes etapas: observación y descripción de la realidad, elaboración de hipótesis, verificación de la hipótesis, y resolución e inferencia causal.

El diseño del mismo ha de constar de varias fases:

- definir y acotar el problema de estudio
- elección del diseño
- planificar las actividades para llevarlo a cabo

Atendiendo a las características de los distintos estudios se pueden clasificar atendiendo a varios criterios.

4.1. Segundo Finalidad

Los estudios epidemiológicos se clasifican según su finalidad en descriptivos y analíticos. Los estudios descriptivos son aquellos que estudian la frecuencia y distribución de los fenómenos de salud y enfermedad, mientras que los analíticos se dirigen a evaluar presuntas relaciones de causa-efecto. En otras palabras, los estudios descriptivos tratan de dar respuesta a preguntas sobre el ¿dónde?, ¿cuándo?, ¿quiénes? y ¿cómo?; mientras que los estudios analíticos tienen por objeto responder al ¿por qué? de los fenómenos de salud y enfermedad .

Entre la epidemiología descriptiva y analítica existen múltiples conexiones. Aunque la observación y descripción de la realidad en muchos casos es un fin en sí mismo, también puede ser el primer paso del método científico experimental. En los estudios analíticos es frecuente que haya un componente descriptivo inicial. Entre los estudios descriptivos suelen incluirse también aquellos que exploran presuntas relaciones de causa-efecto, pero que debido a sus limitaciones en el diseño, sus resultados no son aceptados como concluyentes. Algunos autores se refieren a estos estudios como diseños observacionales incompletos o de cribado de hipótesis.

4.2. Segundo unidad de análisis:

Según el objeto de estudio la unidad puede ser:

- el individuo

- poblaciones: (países, regiones, ciudades, distritos, familias, colegios, empresas, etc etc). Estos suelen llamarse estudios ecológicos.

Un ejemplo sería un estudio donde se contrastara la incidencia de enfermedades respiratorias de mineros frente a otras profesiones.

4.3. Direccionalidad

Según como se estudie la causalidad, Kramer y Boivin proponen clasificar los estudios entre tres tipos.

- **Hacia delante, o secuencia:** desde la causa al efecto. se evalúa la aparición del efecto. Esta es la característica principal de los estudios de cohortes.
- **Hacia atrás, o secuencia:** desde el efecto a la causa. presentan el efecto y otras no, y se evalúa la presencia de la exposición en todos.
- **Simultánea o sin direccionalidad:** La exposición y el efecto son evaluados simultáneamente en el tiempo, lo cuál es un argumento fundamental para demostrar relaciones de tipo causal. No obstante, este argumento sólo puede considerarse cumplido cuando el tiempo entre la exposición y el efecto tiene en cuenta el periodo de inducción correspondiente.

Los dos primeros tipos se engloban en los denominados estudios longitudinales cuya utilidad consiste en establecer el orden en el que se producen los acontecimientos en el tiempo, que es un argumento fundamental para demostrar relaciones de tipo causal. No obstante, este argumento sólo puede considerarse cumplido cuando el tiempo entre la exposición y el efecto tiene en cuenta el periodo de inducción correspondiente.

4.4. Selección de la muestra

Según el criterio utilizado para elegir a partir de la población diana a los sujetos que conformarán la muestra a estudiar, existen dos posibilidades:

- **Muestreo representativo:** Consiste en seleccionar una muestra representativa de la población diana.
- **Muestreo de conveniencia:** Cuando el efecto que se quieren estudiar es poco frecuente puede ser más eficiente muestrear personas que presentan el efecto, y después, buscar el grupo de comparación más adecuado. Por el contrario, cuando la exposición a estudiar es rara, puede interesar ir a buscar directamente personas expuestas y compararlas con un grupo de personas no expuestas. Esta última situación se produce en el estudio de exposiciones que ocurren en determinadas actividades laborales.

4.5. Relación temporal o proximidad

El tiempo transcurrido desde que se produjeron los hechos que se analizan hasta el momento en el que se realiza el estudio puede influir en la validez de la información. En función de esta relación temporal los diseños se clasifican en:

- **Históricos o retrospectivos:** Estudian hechos ocurridos antes del comienzo del estudio. La información puede obtenerse a partir de registros existentes, como por ejemplo en el caso de los estudios de cohortes retrospectivas, o indagando en las entrevistas sobre hechos ocurridos en el pasado.
- **Concurrentes o prospectivos:** Consideran únicamente eventos que se producen a partir del momento de inicio del estudio.
- **Mixtos:** Estudian tanto hechos históricos como concurrentes. Conviene no confundir la relación temporal y la direccionalidad. Así por ejemplo, un estudio de casos y controles puede ser prospectivo, si se realiza sobre casos incidentes, o retrospectivo, si incluye casos prevalentes.

4.6. Control de la asignación de los factores de estudio:

Los **estudios experimentales** son aquellos en los que el investigador controla la asignación de la exposición a estudio. Todos ellos son de tipo analítico, y por razones éticas, suelen limitarse exclusivamente a evaluar nuevos tratamientos y/o fármacos.

Estudios observacionales son todos aquellos en los que el investigador no controla la asignación de la exposición, limitándose a analizar factores cuya presencia o ausencia en los individuos se ha producido por un motivo independiente a la investigación.

4.7. Clasificación de estudios epidemiológicos:

A continuación explicaremos los estudios epidemiológicos estándares y donde se situarían con respecto a los criterios anteriores:

4.7.1. Estudios de Prevalencia:

En medicina se entiende por prevalencia a la fracción de la población que tiene determinada exposición o enfermedad. Por ejemplo la prevalencia de gripe en España en un instante sería la cantidad de personas que tienen gripe en ese instante partido del número de habitantes.

Así pues, se entiende por estudio de prevalencia a un tipo de investigación observacional descriptiva y analítica en el que en un único momento temporal medimos la frecuencia de la exposición y del efecto. El tipo de diseño epidemiológico que se utiliza es el transversal, por lo que los términos transversal y prevalencia se suelen utilizar como sinónimos. Estos estudios son considerados por algunos autores, en los casos de exposiciones que varían con el tiempo, como estudio de casos y controles en los que se asume que el tiempo que media entre la exposición y el inicio de la enfermedad es nulo.

Pasos de un estudio de prevalencia:

1. Seleccionamos una muestra de una población determinada. Las técnicas de muestreo más utilizadas son el muestreo aleatorio simple o el muestreo por conglomerados. En cualquier caso debemos controlar que el hecho de presentar o no el fenómeno que queremos analizar no influya sobre la probabilidad de ser seleccionado.

2. En la muestra seleccionada desconocemos quiénes están expuestos o no o quiénes que están enfermos o sanos.
3. Sin realizar ningún seguimiento se interroga a los individuos sobre su estado de exposición y enfermedad y se clasifican en las distintas categorías: enfermos-expuestos, enfermos-no expuestos, sanos-expuestos, sanos-no expuestos.
4. Se estiman la frecuencia (prevalencia) de cada categoría. Las medidas de los efectos son dos: la razón de prevalencia (aproximación al riesgo relativo de los estudios de cohortes) o la odds ratio de prevalencia (aproximación a la odds ratio de los estudios de casos y control).

4.7.2. Cohorte

Un **estudio de cohortes** es un estudio donde se identifica y sigue la evolución de uno o varios grupos con un determinado grado de exposición, con el objetivo de detectar y cuantificar la aparición del evento o enfermedad de interés a lo largo del tiempo.

La palabra “cohorte” denomina a un conjunto de personas que comparten una determinada exposición y/o “evolucionan” juntas a lo largo del tiempo.

Los estudios de cohortes se basan en el seguimiento. En el diseño de cohortes está implícita la recogida de información, a nivel individual, sobre la exposición de uno o varios grupos de personas (cohortes), y el establecimiento de un mecanismo para seguir a los miembros de estas cohortes con objeto de detectar la aparición del evento de interés y también la posible salida o abandono del estudio. El seguimiento sirve para cuantificar no sólo la frecuencia de enfermedad en las distintas cohortes (incidencia acumulada), sino el ritmo o dinámica de la enfermedad en cada una de ellas (tasa de incidencia). Además, durante el seguimiento es posible considerar e incorporar en el análisis los cambios en el nivel de exposición así como las variaciones del resto de factores de riesgo relevantes.

Por tanto, en los estudios de cohortes se valora el estado de exposición antes de que se produzca la enfermedad y permiten registrar los cambios ocurridos en el tiempo a nivel individual. La mayor parte de las enfermedades crónicas son el resultado de un proceso que se extiende a lo largo de décadas, en el que intervienen diferentes exposiciones o factores de riesgo. La naturaleza dinámica de la exposición a estos factores y su interrelación a lo largo del tiempo sólo puede ser estudiada adecuadamente en un diseño de cohortes. Esta dinámica temporal está ausente de los estudios transversales y es difícilmente investigable en los estudios de casos y controles.

Frente a los estudios caso-control, los estudios de cohortes presentan las siguientes ventajas:

- Permiten obtener información más detallada de los efectos de una determinada exposición. Posibilitan la inclusión en el estudio de uno o más eventos de interés relacionados con dicha exposición.
- Presentan menor propensión a los sesgos de recuerdo y de selección. Por un lado, la relación exposición-enfermedad es estudiada respetando la secuencia temporal. La valoración de la exposición es anterior al evento de

interés. Por otro lado, la población origen del estudio está mejor definida y es fácilmente identificable. El sesgo de selección es el punto crítico de los estudios de casos y controles, cuya validez radica en la adecuada inclusión de controles realmente representativos de la población origen de los casos.

- Facilitan el estudio de exposiciones poco frecuentes, difíciles de incluir en estudios de casos y controles. Favorecen un uso adecuado de las muestras biológicas, que son recogidas antes de que aparezca la enfermedad.
- Posibilitan la medición de la exposición de forma repetida en el tiempo. Con ello permiten valorar los cambios que se producen en la exposición y también calibrar los posibles errores de medición.
- Permiten estudiar la incidencia de la enfermedad en la población a estudio y también el riesgo (incidencia acumulada).

Entre los inconvenientes de los estudios de cohortes, respecto a los casos y controles, figuran los siguientes:

- En general requieren mucho más tiempo. El seguimiento puede durar muchos años, lo que encarece el estudio y dificulta la retención de los participantes. Los estudios de cohortes históricas no sufren dicho inconveniente, pero su calidad está determinada por las fuentes de información utilizadas.
- Cuando el evento de interés es poco frecuente (enfermedades de baja incidencia) es necesario incluir en el estudio un número elevado de participantes.
- Los estudios de cohortes basadas en grupos específicos de exposición, como pueden ser determinados sectores ocupacionales, no permiten estimar directamente la fracción de enfermedad atribuible a la exposición. Para ello, es necesario conocer cual es la frecuencia de exposición en la población general.

Vistos los pros y contras, podemos concluir que los estudios de cohortes y los caso-control son dos herramientas complementarias. Los estudios de cohortes proporcionan una población bien definida en la que identificar los casos de forma no sesgada y con información disponible de la secuencia temporal del seguimiento. Los caso-control concentran el esfuerzo en los “individuos informativos”. Por eso muchos estudios de cohortes incorporan la estrategia de utilizar diseños de caso-control dentro de la cohorte

4.7.3. Estudio de Casos y Controles

Un estudio de casos y controles es un estudio analítico observacional donde se selecciona un grupo de casos con cierta característica (ya sea enfermedad o condición), y un grupo de *controles* con gente sin dicha característica. Sobre estos grupos se mide la exposición que han tenido ante determinada sustancia y se compara. La idea es que si la prevalencia de la exposición es significativamente mayor en el grupo de los *casos* podría ser un factor de riesgo, mientras que si es menor podría ser un factor protector.

El papel de los estos estudios esta frecuentemente limitado, debido a la dificultad de interpretar el tipo de evidencia aportada.

Su principal uso ha sido el estudio de investigación en causas crónicas (canceres, enfermedades cardiovasculares...) aunque cada vez se usan más en el estudio de enfermedades transmisibles y en la evaluación de intervenciones y programas de salud.

Ventajas:

- Precio: Los estudios de casos y controles son más baratos y rápidos que los estudios de cohorte. En los estudios de casos y controles se obtienen todos los casos que aparecen en una población de base teórica en el momento de sufrir la enfermedad. En este sentido, no es necesario realizar el seguimiento de esta población, por lo que los estudios de casos y controles son más sencillos desde el punto de vista logístico y no están sujetos a problemas de pérdidas en el seguimiento.
- Selectivos: Los estudios de casos y controles son especialmente útiles en enfermedades raras o con largos períodos de latencia. En este caso, los estudios de cohorte son ineficientes, ya que es necesario seguir a un número elevado de personas durante bastante tiempo para obtener un número suficiente de casos. E incluso, los estudios de casos y controles puede llegar a ser realista. Como por ejemplo en el caso de enfermedades con incidencias inferiores a 1 caso/10.000 personas-año. En esos casos, si deseáramos investigar las causas de una de esas enfermedades mediante un estudio de cohortes, habría que seguir a 100.000 personas durante 2 años para obtener menos de 20 casos de enfermedad por término medio, lo que resulta muy ineficientes.
- Profundidad del paciente: Los estudios de casos y controles permiten estudiar una amplia variedad de posibles exposiciones. Aunque esto es también posible en estudios de cohorte, en éstos habría que determinar la exposición en todos los participantes de la cohorte, lo que puede ser ineficiente o, sencillamente muy caro. De hecho, estas consideraciones sobre la eficiencia en los estudios de cohorte llevaron al desarrollo de los diseños de casos y controles anidados y a otros diseños híbridos de muestreo dentro de una cohorte.

Inconvenientes:

- Los estudios de casos y controles no son adecuados cuando el desarrollo de la enfermedad altera los niveles de las exposiciones que se están intentando determinar. En ocasiones, puede ser difícil determinar si la exposición ha causado la enfermedad, o si la enfermedad ha modificado la exposición. Por ejemplo, durante la fase aguda de un infarto de miocardio, los niveles de colesterol total en sangre disminuyen. Por lo tanto, si se realiza un estudio de casos y controles comparando los niveles de colesterol en casos incidentes de infarto agudo de miocardio con controles sin la enfermedad, los niveles de colesterol total serán más bajos en los casos, a pesar de que numerosos estudios prospectivos han establecido que los sujetos con niveles más elevados de colesterol total tienen un mayor riesgo de padecer un infarto de miocardio.
- Los estudios de casos y controles son más propensos a sesgos de selección y de información. Por un lado, la selección de la muestra puede afectar la

asociación resultante. De hecho, una de las dificultades principales en los estudios de casos y controles es la selección de un grupo control adecuado (ver selección de controles más adelante). Por otro lado, la información sobre la exposición se recoge después de que los casos hayan desarrollado la enfermedad, por lo que frecuentemente los sujetos de estudio y los investigadores saben qué participantes son casos y quienes controles y pueden recoger o recordar la información de forma diferente.

- Los estudios de casos y controles no son eficientes para estudiar exposiciones raras. En este caso, tan sólo un número reducido de casos y de controles estarán expuestos, por lo que las estimaciones serán imprecisas.
- Los estudios de casos y controles permiten estudiar tan sólo una enfermedad (o muy pocas) a la vez. En los estudios de cohorte, una vez reclutada la cohorte y seguida en el tiempo, es posible registrar la ocurrencia de múltiples enfermedades, por lo que se puede investigar la asociación entre las exposiciones determinadas en el estudio y numerosas condiciones. Los estudios de casos y controles, por el contrario, disponen habitualmente de tan sólo una serie de casos, por lo que será necesario hacer un estudio de casos y controles para cada enfermedad que queramos investigar.
- En los estudios de casos y controles no es posible, en general, obtener estimadores de la incidencia de la enfermedad. Tan sólo proporcionan medidas relativas de efecto (en concreto, odds ratios).
- En los estudios de casos y controles no es posible actualizar la medida de la exposición (tomar medidas repetidas). En los estudios de cohortes es frecuente realizar varias visitas a los participantes durante el seguimiento, por lo que se puede volver a medir la exposición y utilizar esta información en el análisis del estudio.

4.7.4. Ecológico o de conglomerado

Los estudios ecológicos se diferencian de un estudio observacional convencional en que los niveles de exposición individual no se miden, o bien la información sobre la exposición, si se mide, no se vincula con la frecuencia de la enfermedad a nivel individual. La unidad típica de análisis estadístico es un área geográfica, como puede ser una sección censal, una provincia o un país. Para cada grupo o área, es posible estimar la distribución de exposiciones o al menos el nivel promedio de exposición, y podemos estimar la tasa de enfermedad general, pero carecemos de la información conjunta de medidas de niveles de exposición y estatus de enfermedad a nivel de individuos. Por ello, es imposible obtener la tasa de incidencia en los grupos expuestos y no expuestos a partir de datos ecológicos, debiendo recurrir a estimaciones indirectas. La estimación indirecta de efectos en los estudios ecológicos y temas claves como el control de variables de confusión, contienen complicaciones metodológicas que hacen de los mismos un área de la epidemiología con una alta especialización. La necesidad de abordar este tipo de estudios surge de la dificultad de obtener datos de alta calidad sobre exposiciones ambientales y las variables relacionadas con ellas. Por este motivo, los estudios ecológicos han sido y son utilizados frecuentemente en diversas áreas de investigación. Hace años los estudios ecológicos se encuadraban

bajo el epígrafe de simples estudios descriptivos en los que el análisis de las tasas de enfermedad se estratificaban por lugar y tiempo para explorar diferentes hipótesis prestando escasa atención a los métodos estadísticos de inferencia.

Las principales justificaciones de efectuar este tipo de estudio son:

- Bajo coste. Utilización de diferentes fuentes de datos secundarias, que se enlazan para hacer el estudio de forma agregada. Por ejemplo, los datos de mortalidad de los municipios se juntan con los datos censales y de otras encuestas.
- Limitaciones en las mediciones de los estudios individuales. En algunas situaciones y sobre todo en epidemiología ambiental, los estudios ecológicos son los únicos posibles. Ejemplo: El riesgo de cáncer en el entorno de focos contaminantes. Es muy difícil disponer de mediciones de exposición individuales. Los indicadores ‘ecológicos’ de exposición, como por ejemplo la distancia, son los únicos utilizables en muchas circunstancias.
- Limitaciones en el diseño de los estudios individuales. Si una exposición tiene una escasa variabilidad en el área de estudio, un estudio individual no tiene interés práctico. Diseñar un estudio ecológico comparando áreas puede ser una solución.
- Interés de los efectos ecológicos. Ejemplo: tratar de entender las diferencias en las tasas de enfermedad entre dos poblaciones, supone apuntar hacia un objetivo de inferencia ecológica. Esta es la situación habitual en la evaluación del impacto de procesos sociales o de políticas de intervención (ej.: legislación o programas de prevención).
- Simplicidad de análisis y presentación de resultados. En la explotación de grandes encuestas periódicas (ej.: encuestas de salud) es habitual que, aunque los datos han sido obtenidos con cuestionarios individuales, el análisis y/o resultados se muestren de forma agregada, con presentación por años, provincias o regiones como unidad de análisis.

Sus problemas metodológicos (sesgos) más importantes son:

- Sesgo intra-grupo. El sesgo puede ser originado por la presencia de sesgos intra-grupo debido a confusión, selección o información, aunque los efectos intra- grupo no se estimen. Ejemplo: si existe en cada grupo un efecto de confusión será de esperar que la estimación ecológica del efecto también esté sesgada.
- Confusión por grupo. Si la tasa de enfermedad basal en la población no expuesta varía en los grupos, especialmente si la correlación ecológica entre el nivel de exposición medio y la tasa basal es distinta de cero, se producirá un sesgo ecológico.
- Modificación del efecto por grupo (escala aditiva). Puede producirse un sesgo ecológico si la diferencia de tasas para el efecto de la exposición a nivel individual varía en los grupos.

4.7.5. Estudios experimentales:

El elemento central que define un estudio experimental sea la introducción de una intervención que altera las condiciones del mismo y es asignada a los participantes por el investigador.

Para que podamos hablar de un experimento válido desde el punto de vista científico, es preciso que además se cumplan las siguientes condiciones:

- La única razón por la que los sujetos reciben la intervención bajo estudio es el cumplimiento del protocolo del estudio.
- En el estudio existe una serie de sujetos, denominados grupo control, que no reciben la intervención cuyo efecto se desea analizar.
- La asignación de la intervención a una serie de sujetos, grupo de intervención, se lleva a cabo por un mecanismo debido al azar.

En el caso de que el último punto no se cumpla, se considerará semi-experimental, y se estará introduciendo un sesgo al variar la distribución “real” de los casos.

Estos requisitos representan condiciones sin las cuales no se pueden descartar que los efectos observados puedan deberse a factores desconocidos o no controlados y, de esta forma, poder atribuir con seguridad esos resultados a la intervención bajo estudio. Para lograr este objetivo es necesario controlar todos los factores que pueden afectar de forma relevante la condición bajo estudio; es decir, necesitamos crear unas condiciones tales que el único factor que presente variación entre los grupos de comparación sea la intervención cuyo efecto pretendemos evaluar.

Desafortunadamente, en ciencias sociales y biomédicas el número de factores que afectan a la mayoría de las condiciones de interés es tan numeroso, complejo y en muchos casos oculto o desconocido, que es imposible hacerlo uniforme. Por ejemplo, en la investigación de las causas del cáncer es imposible crear un conjunto de condiciones que conduzcan de forma invariable al desarrollo de un cáncer en un periodo fijado de tiempo, incluso aún cuando la población bajo estudio sea un grupo de ratones clonados en el laboratorio. Inevitablemente, siempre va a existir la denominada “variación biológica”, es decir, la variación que se produce en el conjunto de factores que contribuyen a que se desarrolle la condición bajo estudio. Aunque consiguiéramos mantener las mismas condiciones de temperatura, humedad, hábitat e iluminación, lo cual es bastante poco probable, es imposible hacer que todos los animales coman exactamente la misma cantidad o realicen ejercicio físico con idéntica frecuencia e intensidad, por no hablar de diferencias individuales en el metabolismo basal. En resumidas cuentas, en la investigación epidemiológica la idea de crear un duplicado exacto de un conjunto de circunstancias, en el que sólo varía un factor relevante, no es realista. No obstante, puede afirmarse que un experimento es aceptable si la variación de los factores extraños es demasiado pequeña como para afectar la condición bajo estudio de forma importante en comparación con el efecto de la intervención bajo estudio. En el ejemplo de las causas del cáncer, si la alimentación de los ratones influye en el desarrollo del cáncer, lo cual parece bastante verosímil, el control inadecuado de la misma puede suponer un problema. Si la variación en la alimentación es muy pequeña, es mucho menos probable que pueda influir sobre los resultados del experimento de forma importante.

5. Sesgos y valores de confusión:

Tanto a la hora de recoger datos como de interpretarlos usando los distintos métodos epidemiológicos es necesario tener en cuenta la presencia de sesgos.

La RAE define sesgo como “error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras.” [2]

Es decir, cualquier error sistemático en el proceso de la inferencia estadística que nos pueda alejar de la realidad.

El término sistemático excluye el error aleatorio, y al contrario que este, no se solucionaría aumentando la muestra.

Debido a esto, el repetir el estudio con mayor tamaño aumentaría la precisión pero no su validez, ya que el valor que se estaría aproximando realmente estaría igual de alejado que el valor verdadero.

A grandes rasgos hay 3 tipos de sesgos:

- Sesgo de selección.
- Sesgo de información.
- Sesgo de confusión

Aunque mostrados así puedan parecer totalmente diferenciados, en la práctica no se presentan tan diferenciados:

5.1. Sesgo de selección

El *sesgo de selección* ocurre cuando al elegir el conjunto de pacientes que conforman el estudio estos no sean representativos. Es frecuente en los estudios de casos y controles, sobre todo en la selección de controles. Los sesgos de selección pueden ocurrir en cualquier estudio epidemiológico, sin embargo, ocurren con mayor frecuencia en estudios retrospectivos y, en particular, en estudios transversales o de encuesta. En los estudios de cohorte prospectivos los sesgos de selección ocurren raramente ya que el reclutamiento y selección de la población en estudio se da antes de que ocurra el evento en estudio, así que se puede suponer que la selección de los participantes se realiza de manera independiente del evento y, en general, la participación en el estudio no puede ser influida por el evento, ya que éste aún no ha ocurrido. En contraste, la permanencia de los participantes en el estudio sí puede ser determinada por el evento, cuando esto ocurre, y es de diferente magnitud para los grupos expuesto y no expuesto, existirá la posibilidad de que los resultados se vean distorsionados por esta permanencia diferencial. Por esta razón, se recomienda maximizar las tasas de permanencia y seguimiento en los estudios de cohorte.

Los estudios transversales, y de casos y controles que se basan en casos existentes (prevalecientes), presentan importantes limitaciones relacionadas con los sesgos de selección, en particular cuando la enfermedad en estudio tiene una alta letalidad cercana al diagnóstico inicial ya que los casos existentes tienden a sobrerepresentar a los sujetos con cursos más benignos de la enfermedad. Si el factor en estudio se asocia con la letalidad, la medida de efecto derivada de un estudio de prevalencia será sesgada, dado que los casos en estudio corresponden a sobrevivientes de la enfermedad, por lo que no se representarán a los que murieron tempranamente.

A continuación veamos algunos tipos frecuentes de sesgos de selección:

1. **Voluntarios:** la *autoselección* de los integrantes de un estudio es un sesgo. Los voluntarios (autoseleccionados) suelen presentar características particulares, pues el grado de motivación de un sujeto que participa voluntariamente en una investigación puede variar sensiblemente en relación con otros sujetos.
2. **No respuesta:** Es frecuente que los pacientes no contesten. Ya sea por la negativa de los pacientes o a que no se haya podido encontrar .
3. **Abandonos del estudio:** Los pacientes del estudio pueden, por distintas circunstancias llegar a abandonar el estudio. Por lo que es importante apuntar las fechas de entrada y de salida de los mismos. Dejándonos con
4. **Trabajador sano:** Incluso antes de que los sujetos sean seleccionados para el estudio se puede producir una selección. Un ejemplo es el denominado efecto del trabajador sano. Los trabajadores suelen tener mejor salud que la población general, ya que se habrían seleccionado aquellas personas con mejor salud para obtener un puesto de trabajo y poder mantenerlo, mientras que la población general también incluye enfermos incapaces de desempeñar un trabajo. Este efecto es tanto mayor cuanto más difíciles sean las condiciones de trabajo y, por tanto, requiera más capacidades singulares
5. **Sesgo de Berkson:** El llamado sesgo o paradoja (término preferido por el autor) de Berkson se puede producir cuando los sujetos del estudio se obtienen del hospital (casos y controles hospitalarios). Las personas hospitalizadas pueden diferir de manera sistemática de la población general, de la que se pretende sean representativas, en muchos aspectos, y particularmente en cuanto a la exposición a los factores de riesgo estudiados, debido a diversos factores que influyen en la probabilidad de hospitalización.
6. **Sesgo de supervivencia:** se produce cuando se estudia una determinada patología que produce muertes precoces y en el momento del inicio del estudio esos individuos muertos ya no pueden incluirse en el grupo de los casos.
7. **Sesgo por inclusión/exclusión:** Se produce cuando se seleccionan en el grupo de control enfermedades relacionadas con la exposición, Este es un sesgo que esencialmente se produce cuando el grupo de referencia es hospitalario. Para su correcta identificación hay que proporcionar con el mayor detalle posible la patología elegida en el grupo de referencia.

5.2. Sesgo de información

El *sesgo de información* se refiere a los errores que se introducen durante la medición de la exposición, de los eventos u otras variables en la población en estudio, que se presentan de manera diferencial entre los grupos que se comparan, y que ocasionan una conclusión errónea respecto de la hipótesis que se investiga. Una posible fuente de sesgo de información puede ser cualquier factor que influya de manera diferencial sobre la calidad de las mediciones que se realizan

en los grupos expuesto y no expuesto en el contexto de los estudios de cohorte o entre los casos y controles en el contexto de los estudios de casos y controles.

Aunque no existen procedimientos libres de error de medición, no todos los errores de medición son fuente de sesgo de información. Es conveniente recordar que los errores de medición pueden ser no diferenciales cuando el grado de error del instrumento o técnica empleada es el mismo para los grupos que se comparan o diferenciales, cuando el grado de error es diferente para los grupos estudiados, el sesgo de información se refiere particularmente a este último tipo.

5.3. Sesgo de confusión

El *sesgo de confusión* es la inclusión de variables de confusión, ésto es, una variable o factor que distorsiona la medida de la asociación entre otras dos variables. El resultado de la presencia de una variable de confusión puede ser el surgimiento de un efecto donde en realidad no existe o la exageración de una asociación real (confusión positiva) o, por el contrario, la atenuación de una asociación real e incluso una inversión del sentido de una asociación real (confusión negativa).

Un caso extremo de este sesgo es la *paradoja de Simpson*, que es cuando el factor de confusión llega a invertir la conclusión de un estudio. Aún cuando no se haya incluido otro sesgo, veámoslo en un ejemplo donde se compararon dos tratamientos para cálculos de riñón:[12]

	Tratamiento A	Tratamiento B
Cálculos pequeños	93 % (81/87)	87 % (234/270)
Cálculos grandes	73 % (192/263)	69 % (55/80)
Global	78 % (273/350)	83 % (289/350)

En los datos se ve que tratamiento B tiene mayor éxito en general pero, paradójicamente, en cada subconjunto tiene mejores porcentajes de éxito el tratamiento B.

La explicación proviene de que el tratamiento A es más usado con cálculos grandes que en pequeños, y al ser el éxito en los grandes menos frecuentes que en los pequeños llega a fracasar más el tratamiento A.

La forma de combatir este tipo de efectos nos es otra que la de esforzarse en contemplar todas las variables que pudieran ser relevantes.

6. Modelos bioestadísticos más frecuentes

Con el objetivo de resumir e interpretar los datos obtenidos en un estudio, se usan distintos modelos, que simplificando la realidad, son capaces de mostrar las características de los pacientes.

Los cuales tratan de estimar una variable o riesgo de un paciente en función de otras.

A continuación explicaremos algunos de los modelos más usados, señalando sus características más destacables. Así como algunas de las precauciones que deben tomarse a la hora de ajustar estos métodos.

6.1. Prueba de independencia variables categóricas:

Supongamos que partimos de n pacientes de los cuales se miden dos variables X, Y , las cuales pueden tener distintos valores $\{(X_i, Y_j) | i \in \{1, 2, \dots, k\}, j \in \{1, 2, \dots, l\}\}$, y deseamos evaluar si las variables son independientes.

	y_1	y_2	\dots	y_l
x_1	$n_{1,1}$	$n_{1,2}$	\dots	$n_{1,l}$
x_2	$n_{2,1}$	$n_{2,2}$	\dots	$n_{2,l}$
\vdots	\vdots	\vdots	\ddots	\vdots
x_k	$n_{k,1}$	$n_{k,2}$	\dots	$n_{k,l}$

Para resolverlo nos podemos basar en un test de homogeneidad se hace un contraste de hipótesis donde la hipótesis inicial es $H_0 : X, Y$ son independientes.

Denominemos $n_{i,j}$ al número de veces que se ha dado X_i, X_j en la muestra.

Partiendo de esta hipótesis se cumple que $P(X = x, Y = y) = P(X) \cdot P(Y)$, como $P(X) \approx n_{i,\cdot} / n = \sum_j n_{i,j}$, $P(Y) \approx n_{\cdot,j} / n = \sum_i n_{i,j}$.

Por lo que se ajusta mediante el estadístico bondad de χ^2 que la distribución dada es precisamente dicha distribución:

$$\chi^2_{(k-1)(l-1)} = \sum_{i=1}^n \sum_{j=1}^k \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}} \quad (1)$$

Donde $e_{i,j}$ es la estimación del valor de $n_{i,j}$ si fuera realmente independiente:

$$e_{i,j} = \frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n} \quad (2)$$

Este se basa en teorema central del límite para aproximar que $\frac{(n_{i,j} - e_{i,j})}{\sqrt{e_{i,j}}} \approx N(0, 1)$, y que la suma de los cuadrados de normales es una χ^2 , [6], debido a la naturaleza del teorema central del límite, para que este estadístico se pueda usar n debe ser grande ($n > 30$) y las frecuencias esperadas mayores a 5.

En caso de que no se cumplan los requisitos mínimos por ser la muestra pequeña, es posible usar el test exacto de Fisher, el cual funciona sin restricciones de tamaño (si bien calcularlo exactamente en grupos grandes puede ser pesado y se recomienda el estadístico anterior).

En el caso más sencillo en que las variables X, Y tomen dos valores siguiendo la tabla:

La probabilidad exacta de observar un conjunto concreto de frecuencias a, b, c y d en una tabla 2 x 2 cuando se asume independencia y fijos los totales de filas y columnas se consideran fijos viene dada por la distribución hipergeométrica:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} \quad (3)$$

Para calcular el p-valor en la literatura estadística se calcula la lista de todas las tablas posibles y se toma la suma de los casos menos probables, si este no supera el indice de confianza (generalmente 5 %) .

A modo de ejemplo cojamos datos de una muestra que compara si hay la presencia de obesidad está relacionada con el sexo, la tabla de frecuencias es:

	Obeso	No Obeso	Total
mujer	1 (a)	4 (b)	5 (a+b)
hombre	7 (c)	2 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

[14] cuya probabilidad es

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} = \frac{5!9!8!6!}{14!1!4!7!2!} \approx 0,059 \quad (4)$$

Sobre lo que calculamos los valores de las distintas configuraciones que dejan fijo los totales parciales:

a	b	c	d	probabilidad
0	5	8	1	0.002997
1	4	7	2	0.059940
2	3	6	3	0.279720
3	2	5	4	0.419580
4	1	4	5	0.209790

Sumando los sucesos más improbables que el suceso obtenido obtenemos el p-valor:

$$p_{valor} = 0,00299 + 0,0599 > 0,05 \quad (5)$$

por lo que en este caso no disponemos de la evidencia necesaria para rechazar la hipótesis de independencia.

6.2. Modelo Lineal

Un modelo lineal supone que la variable objetivo $y, target$ de aquí en adelante, se puede aproximar Mediante las variables $\{X_i\}_{i=1}^n$ con la expresión:

$$y = \sum_{i=0}^N \beta_i \cdot X_i \quad (6)$$

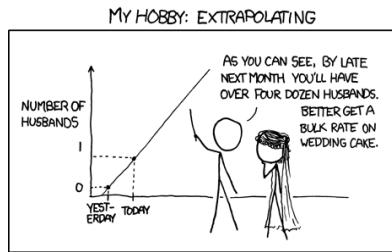
Donde β_i son los pesos que se ajustan en cada variable. Es usado cuando la variable a estimar es un variable continua. [9] como puede ser el peso (y). Por ejemplo se podría estimar el efecto de distintas dietas, teniendo en cuenta otras variables como la altura, sexo, edad, horas de ejercicio a la semana...

Los modelos lineales pueden tener limitaciones que no deberían de ser ignoradas, una de las más importantes es que solo funcionen en determinados rangos,

por ejemplo para modelizar la altura, la edad es sin duda importante en un rango de edad de 3 – 15 años, pero en individuos adultos la edad es irrelevante.

Esto nos puede llevar, en ciertas condiciones, a la conveniencia de separar en clusters a los individuos y hacer varios modelos.

Figura 1: ejemplo extremo sobre como se pueden desvirtuar los modelos de regresión ([7])



Otra dificultad de los modelos lineales es la llamada multicolinealidad, esto es, varias variables regresoras están fuertemente correlacionadas lo que puede hacer que la estimación de los parámetros β_i tenga un gran error. Hay varias estrategias para combatir este efecto, la forma más sencilla de resolver esto es excluyendo las variables más correlacionadas una a una. Debido a que las variables que generan la multicolinealidad están muy correlacionadas la eliminación de estas no afectará negativamente en el resto.

6.3. Modelo logístico

A la hora de estimar probabilidades (riesgos) el modelo lineal no suele ser el más práctico por lo cual se emplea una modificación del mismo: [11]

$$\log\left(\frac{p}{1-p}\right) = \sum_{i=0}^N \beta_i \cdot X_i \quad (7)$$

Donde X_i son variables binarias ($Y_i \in \{0, 1\}$), y representan la ocurrencia o no de un suceso. La cual, es equivalente a :

$$\frac{p}{1-p} = \prod_{i=0}^N \exp(\beta_i \cdot X_i) \quad (8)$$

Donde p es la posibilidad de que $Y = 1$.

Se llama odd-ratio a la expresión $\frac{p}{1-p}$ que se interpreta cuantas veces es más probable que ocurra la variable target a que no ocurra.

No es difícil darse cuenta de que la interpretación del valor β_i es que la variable i al ocurrir $\Rightarrow X_i = 1$ multiplica por e^{β_i} veces el odd-ratio (probabilidad de que ocurra).

Estos modelos son muy fundamentales en epidemiología, pues desempeñan un papel importante en el control de los sesgos de confusión, pero se basan en

una serie de supuestos cuyo cumplimiento no siempre es fácil comprobar. Por ejemplo, no siempre existe una relación lineal entre la variable de exposición (variable independiente, “X”) y la variable de respuesta (variable dependiente, “Y”). Cuando lo que se desea conocer es cómo una serie de factores influyen en una variable binaria o dicotómica, es decir con dos posibilidades, como por ejemplo estar sano o enfermo, responder a un tratamiento o no responder, etc. en vez de utilizar la regresión lineal, se va a utilizar la regresión logística. En este caso, al ser dicotómica la respuesta o resultado, se hablaría de regresión logística binaria.

Un ejemplo de este uso lo haremos sobre la tabla vista para ilustrar el sesgo de confusión, de la forma:

	Tratamiento A	Tratamiento B
Cálculos pequeños	93 % (81/87)	87% (234/270)
Cálculos grandes	73 % (192/263)	69 % (55/80)
Global	78 % (273/350)	83 % (289/350)

Donde las variables son C, T donde $C = 1$ si los cálculos son grandes $C = 0$ si son pequeños; y $T = 0$ si es el tratamiento A y $T = 1$ en el caso del tratamiento B.

Ajustaremos:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot C + \beta_2 \cdot T$$

Las ecuaciones son:

$$\log\left(\frac{0,93}{1-0,93}\right) = \beta_0$$

$$\log\left(\frac{0,73}{1-0,73}\right) = \beta_0 + \beta_1$$

$$\log\left(\frac{0,87}{1-0,87}\right) = \beta_0 + \beta_2$$

$$\log\left(\frac{0,69}{1-0,69}\right) = \beta_0 + \beta_1 + \beta_2$$

Usando un software obtenemos el resultado:

$$\log\left(\frac{p}{1-p}\right) = -2,18099597 + 1,152 \cdot C + 0,296 \cdot T \quad (9)$$

Donde se muestra que contribuye más el tamaño del cálculo, que el tratamiento. Si bien es más útil el tratamiento B (siendo $e^{0,296}$ veces más eficaz).

6.4. Curva de Kaplan-Meier y modelos de regresión de Cox :

Con el objetivo de estimar el tiempo de supervivencia ante una enfermedad se usan modelos de Cox los cuales se pueden ver como una extensión del modelo logístico.

En un primer lugar consideraremos T la variable del tiempo de supervivencia, de los pacientes.

Siendo $F(t) = P(T \leq t)$ se define λ con la idea:

$$\lambda(t) = \frac{\text{gente que muere en el instante } t}{\text{gente que ha sobrevivido hasta el instante } t}$$

Cuya sentido intuitivo es el riesgo que tiene un paciente de morir en un instante suponiendo que ha sobrevivido hasta el instante anterior.

$\lambda(t)$ se expresa formalmente como:

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{F(t+h) - F(t)}{h \cdot (1 - F(t))} = \frac{F'(t)}{1 - F(t)}$$

Lo que equivale:

$$\int_{t=0}^t \lambda(t) dt = \int_{t=0}^t \frac{F'(t)}{1 - F(t)} dt \Rightarrow e^{\int_{t=0}^t -\lambda(t) dt} = 1 - F(t) =: S(t) \quad (10)$$

Una hipótesis habitual es suponer que el riesgo permanezca constante, en cuyo caso la distribución será exponencial. Si bien esta distribución no siempre se ajusta bien a la realidad. Basándonos en la definición de $\lambda(t)$ podemos aproximar la distribución $S(t) = P(t < T) = 1 - F(t)$ usando la distribución empírica:

$$\hat{S}(t) = \frac{\text{numero muertos en el instante } t \text{ o antes}}{\text{número de individuos del estudio}} \quad (11)$$

Esta aproximación contiene la limitación de que no contempla la posibilidad de que los pacientes se hayan dado de baja del estudio (censura), lo cual no implica su muerte.

Como solución se usa el estimador de Kaplan-Meier:

$$\widehat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \quad (12)$$

donde $t_1 < t_2 \dots < t_n$ son los instantes de muerte registrados, d_i es el número de muertes en el instante t_i y n_i es el número de pacientes que siguen en el estudio (contemplando así la posibilidad de que los pacientes se desvinculen del estudio).

Como ejemplo usaremos esta técnica en R sobre el dataset *aml* donde se comparan el tiempo de supervivencia en pacientes con leucemia mieloide aguda tras ser operados, donde se comparan dos procedimientos sobre los ciclos de quimioterapia (variable *x*), y la variable *status* indica si en el instante *time* corresponde a un abandono (*status*=0) o a una muerte (*status*=1).

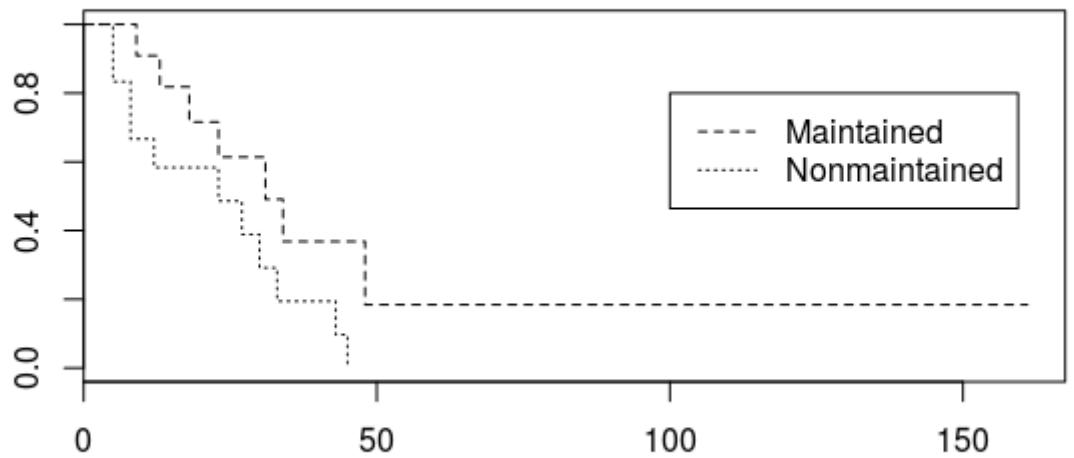
En el siguiente código se comparan los dos subgrupos:

```
> library(survival)
> survival::aml
  time status          x
1     9     1 Maintained
2    13     1 Maintained
3    13     0 Maintained
4    18     1 Maintained
5    23     1 Maintained
6    28     0 Maintained
```

```

7    31      1   Maintained
8    34      1   Maintained
9    45      0   Maintained
10   48      1   Maintained
11   161     0   Maintained
12   5       1   Nonmaintained
13   5       1   Nonmaintained
14   8       1   Nonmaintained
15   8       1   Nonmaintained
16   12      1   Nonmaintained
17   16      0   Nonmaintained
18   23      1   Nonmaintained
19   27      1   Nonmaintained
20   30      1   Nonmaintained
21   33      1   Nonmaintained
22   43      1   Nonmaintained
23   45      1   Nonmaintained
> fit = survfit(Surv(time, status) ~ x, data = aml)
> plot(fit, lty = 2:3)
> legend(100, .8, c("Maintained", "Nonmaintained"), lty = 2:3)

```



Donde las curvas representan al fracción de supervivientes esperados tras cierto tiempo.

A simple vista se ve que los pacientes con $x = \text{Maintained}$ parecen presentar mayores niveles de supervivencia. Ante lo cual nos preguntamos si las curvas son significativamente diferentes.

¿Cómo podemos comprobar que efectivamente la diferencia entre las curvas es solo fortuita?

Para comprobar formalmente que las curvas son diferentes se usa el test de **log-rank**. [15]

Cuyo cálculo consiste en tomar todos los instantes donde sucede una muerte en uno de los grupos $\{\tau_i\}_{i=1}^N$.

Para cada instante τ_i se define:

- O_j = Muertes en el instante τ_j .
- $E_j = \frac{\text{Número de pacientes en riesgo en el instante } \tau_j \text{ en el grupo 0}}{\text{Número de pacientes en riesgo en el instante } \tau_j \text{ en el grupo 1}}$
- V_j = Varianza de cada τ_j

sobre el cual el log-rank se define como:

$$Z = \frac{\sum(O_j - E_j)}{\sqrt{\sum_{j=1}^J V_j}} \quad (13)$$

Cuya distribución sigue asintóticamente una normal $N(0, 1)$ sobre el que se aplica el intervalo de confianza de dos colas. Esto es, para un $\alpha = 5\%$ rechazaríamos la hipótesis de que las distribuciones de riesgo son las mismas cuando $|Z| > 1,96$.

Sobre el *aml* el cálculo resultaría:

También es calculable con la función survdiff:

```
>survdiff(Surv(time, status) ~ x, data=aml)
Call:
survdiff(formula = Surv(time, status) ~ x, data = aml)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
x=Maintained	11	7	10.69	1.27	3.4
x=Nonmaintained	12	11	7.31	1.86	3.4

Chisq= 3.4 on 1 degrees of freedom , p= 0.0653

En este ejemplo el p-valor de 0,06 no nos daría una cierta certeza de que los son distintos. cuya expresión es:

$$\log(h(t)) = \hat{\lambda}(t) + \sum_{i=0}^N \beta_i \cdot X_i \quad (14)$$

Donde $h(t)$ es una función que estima la probabilidad de que el paciente muera en el instante t .

Y $\hat{\lambda}(t)$ es la función de riesgo base.

Este modelo se basa en la hipótesis de riesgos proporcionales, es decir, que

la proporción de riesgo mantiene proporcional con respecto al tiempo:

$$\begin{aligned}
 \log\left(\frac{\lambda(t|X_i=1)}{\lambda(t|X_i=0)}\right) &= \\
 \frac{\widehat{\lambda}(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{i-1} X_{i-1} + \beta_i \cdot 1 + \beta_{i+1} X_{i+1} + \cdots + \beta_n X_n}{\widehat{\lambda}(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{i-1} X_{i-1} + \beta_i \cdot 1 + \beta_{i+1} X_{i+1} + \cdots + \beta_n X_n} &= \beta_i \Rightarrow \\
 \log\left(\frac{\lambda(t|X_i=1)}{\lambda(t|X_i=0)}\right) &= \beta_i \Rightarrow \\
 \lambda(t|X_i=1) &= \exp(\beta_i) \lambda(t|X_i=0) \quad (15)
 \end{aligned}$$

La fórmula muestra que el cociente es independiente del tiempo. [4] Para estimar la probabilidad en las próximas t unidades de tiempo vendría dada por la expresión:

$$\begin{aligned}
 S(t) = 1 - F(t) &= e^{-\int_{t=0}^t \lambda(u) du} = \\
 e^{-\int_{t=0}^t \lambda_0(u) \cdot \sum_i x_i \beta_i} &= \\
 e^{-\int_{t=0}^t \lambda_0(u) du} \cdot e^{\exp(\sum_i x_i \beta_i (t-t_0))} &= \\
 &= S_0^{\exp(\sum_i x_i \beta_i)} \quad (16)
 \end{aligned}$$

Donde $S(t)$ se denomina la función de supervivencia.

Al utilizar la regresión de Cox es necesario verificar que se cumple dicha hipótesis. Para ello es necesario comprobar que el efecto de cada variable es constante en el tiempo. Existen varios métodos para ello. Por un lado puede utilizarse un método gráfico:

Basándonos en la ecuación 16 podemos ver que:

$$-\log(-\log(S(t|\{x_i\}_{i=1}^N))) = -\log(-\log(S_0(t))) - \sum x_i \beta_i \quad (17)$$

Esto indica que las gráficas de Supervivencia en los distintos casos deberían ser visualmente paralelas, en caso contrario se estaría vulnerando la hipótesis de que los riesgos son proporcionales.

No obstante, existen métodos estadísticos rigurosos:

- Residuos de martingala
- Residuos de deviance
- Residuos de score
- Residuos de Schoenfeld

Los cuales se pueden usar fácilmente mediante el paquete survival de R.

7. Modelización

En esta sección exploraremos la posibilidad de hacer un estudio de prevalencia usando una base de datos recuperada por los datos obtenidos mediante la recuperación de información de historias médicas, que denominaremos “método propuesto”.

En contra del método tradicional, el cual acaba funcionando como una encuesta, y por lo tanto la muestra (normalmente efectuada por médicos) y estaría fuertemente restringida, según los recursos disponibles.

7.1. Método clásico:

El proceso habitual consiste en poner a una serie de médicos a anotar los pacientes que entran en su consulta introduciendo en una tabla varias características. En base a esto se extraen las estadísticas pertinentes.

Si bien es un método que se ha usado durante mucho tiempo y da buenos resultados no está carente de ciertas dificultades:

1. Tiempo: Un estudio convencional se suele alargar durante varios meses.
2. Caro: reunir información es caro debido a los problemas logísticos propios de una encuesta, así como bonificaciones económicas que se le dan a los médicos por anotar.
3. Muestra pequeña: debido a que los recursos son limitados, no se pueden contratar a una gran cantidad de médicos. Esto será un problema a la hora de recolectar información sobre enfermedades relativamente poco frecuentes.

Dado un paciente tiene una cierta probabilidad θ de tener cierta enfermedad, una persona con dicha enfermedad tendrá una probabilidad distinta de pasar por consulta y por lo tanto de ser registrado en los informes, esto es, poner a una persona de un hospital a registrar pacientes puede suponer un sesgo.

7.1.1. Tamaño de la muestra:

Supongamos que queremos estimar con una precisión p con una confianza α ... ¿Cuántos pacientes necesitamos observar?

Debido a que cada paciente es una bernouilli X_i de media θ donde θ es la posibilidad de tener la enfermedad. Asumiendo independencia de las variables y usando el teorema central del límite el total de pacientes en una población de tamaño n sigue aproximadamente la distribución normal:

$$\sum_i X_i \approx N(\theta \cdot n, \sigma^2 = n \cdot \theta \cdot (1 - \theta)) \quad (18)$$

Esto nos lleva a que una media de una muestra de tamaño n seguirá una distribución también normal de media θ y varianza $\frac{\theta(1-\theta)}{\sqrt{n}}$

$$T := \frac{\sum X_i}{n} \quad (19)$$

Por lo cual calculamos su intervalo de confianza: $P(T \cdot \sqrt{n} \in (\sigma - z_{\alpha/2} \cdot \sqrt{\theta \cdot (1-\theta)}, \sigma + z_{\alpha/2} \cdot \sqrt{\theta \cdot (1-\theta)}) = \alpha \Rightarrow P(T \in (\sigma - z_{\alpha/2} \cdot \sqrt{\theta \cdot (1-\theta)} / \sqrt{n}, \sigma + z_{\alpha/2} \cdot \sqrt{\theta \cdot (1-\theta)} / \sqrt{n})) = \alpha$

Donde $z_{\alpha/2}$ cumple que $P(-z_{\alpha/2} < N(0,1) < z_{\alpha/2}) = \alpha$.

Si por ejemplo tomamos $\alpha = ,95$ y queremos aproximar con una precisión del $p = 2\%$ una enfermedad con una incidencia aproximada del $I = 2\%$ el error sera de:

$$\epsilon = \frac{z_{\alpha/2}\sigma(1-\sigma)}{\sqrt{n}} = p \cdot I = 0,00004 \Rightarrow 1,96 \cdot (0,02 \cdot 0,98) = 0,04 \cdot \sqrt{n} \Rightarrow n \approx 9216 \quad (20)$$

7.2. Método propuesto:

En Savana tenemos el acceso a las bases de datos de confeccionadas con las memorias médicas de varios hospitales en los cuales detectamos si un paciente tiene o no la enfermedad, desde un punto de vista probabilístico se puede resumir dicho proceso de la siguiente manera:

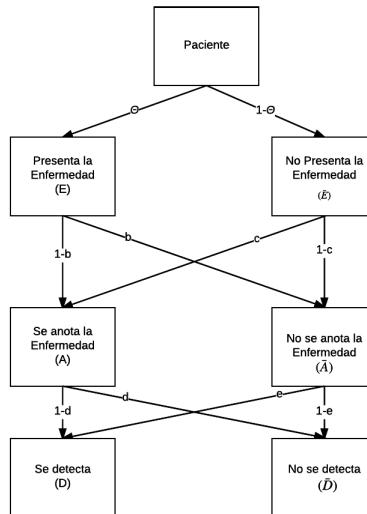


Figura 2: Esquema detección de frecuencias de conceptos

Donde las variables:

1. θ, b, c, d, e son probabilidades:

- θ : Probabilidad de que un paciente aleatorio tenga la enfermedad.
- b : Probabilidad de que dado un paciente con enfermedad esta no se anote. Esto puede ser sintomático de que no se haya diagnosticado o que no se haya generado ninguna referencia a la enfermedad en el periodo de los cuales procesamos los informes ($P(\hat{A}|E)$). Por ejemplo que revisemos los informes de 6 meses en un hospital y el paciente

no haya ido a ninguna revisión ni ningún tema relacionado con la enfermedad.

- c : Probabilidad de que dado un paciente sin enfermedad esta se anote erróneamente. Efecto de sobre diagnostico ($P(A|\bar{E})$).
- d : Probabilidad de que dado un paciente con la enfermedad anotada esta no se detecte por los algoritmos ($P(\bar{D}|A)$).
- e : Probabilidad de que dado un paciente sin la enfermedad anotada esta se detecte erróneamente por los algoritmos ($P(D|\bar{A})$)

2. X_i es el paciente un paciente que genera los informes según su enfermedad.
3. E Tiene la enfermedad.
4. A Se llega a anotar la enfermedad.
5. D Se llega a detectar la enfermedad.

Usando la definición de probabilidad condicionada:

$$P(A) = \sum_{b \in Dom(B)} P(A|B=b) \cdot P(B) \quad (21)$$

sobre la figura 2 se puede inferir cual es la distribución de final será Bernouilli.

Calculemos su media:

- $P(E) = \theta$
- $P(A) = P(A|E) \cdot P(E) + P(A|\bar{E}) \cdot P(\bar{E}) = (1-b) \cdot \theta + (c) \cdot (1-\theta)$
- $P(D) = P(D|A) \cdot P(A) + P(D|\bar{A}) \cdot P(\bar{A}) = (1-d) \cdot P(A) + (e) \cdot P(\bar{A})$

Se pueden reformular las ecuaciones notación matricial se puede observar:

$$\begin{pmatrix} P(A) \\ P(\bar{A}) \end{pmatrix} = \begin{pmatrix} 1-b & c \\ b & 1-c \end{pmatrix} \begin{pmatrix} \theta \\ 1-\theta \end{pmatrix} \quad (22)$$

$$\begin{pmatrix} P(D) \\ P(\bar{D}) \end{pmatrix} = \begin{pmatrix} 1-d & e \\ d & 1-e \end{pmatrix} \begin{pmatrix} P(A) \\ P(\bar{A}) \end{pmatrix} \quad (23)$$

Con lo que se deduce que Por lo que el número de pacientes detectados con la enfermedad dada seguirá la distribución Geométrica:

$$T = \sum_{i=1}^N D_i \sim Geom(\mu, N) \quad (24)$$

Donde $\mu = c(ab + a(-e+1)) + (-d+1)(-ab - a(-e+1) + 1)$

En el caso ideal de que la variables $b, c, d, e = 0$, es decir, siempre se registra la enfermedad y no hay ni falsos positivos ni falsos negativos. Tenemos que $\mu = \theta$ por lo que se puede aproximar a con los estadísticos de máxima verosimilitud e incluso aproximar por una normal para obtener intervalos de confianza:

$$IC = \left(\mu - \frac{z_{\alpha/2}}{\sqrt{N}}, \mu + \frac{z_{\alpha/2}}{\sqrt{N}} \right) \hat{\theta} = T/N$$

El cual es la fórmula que se suele usarse en los estudios de prevalencia cuando se quiere estimar la cantidad de pacientes se requiere usar para saber la prevalencia de una enfermedad.

Para lo cual se parte de una aproximación a priori de cual es la prevalencia dada en anteriores estudios.

Si bien, en nuestro caso En el siguiente apartado deduciremos como estimar el valor de μ en el caso de que tanto los falsos positivos como negativos sean distintos de 0.

8. Estadísticos aplicables al sistema

Por simplificación de ahora en adelante obviaremos los efectos de sobre-diagnóstico e infradiagnóstico a los cuales estarían expuestos cualquier estudio de prevalencia, quedándose las expresión en:

$$P(D) = (1 - d) \cdot P(A) + e \cdot P(\bar{A}) \quad (25)$$

lo que, usando la notación matricial es

$$\begin{pmatrix} P(D) \\ P(\bar{D}) \end{pmatrix} = \begin{pmatrix} 1 - d & e \\ d & 1 - e \end{pmatrix} \begin{pmatrix} P(A) \\ P(\bar{A}) \end{pmatrix} \quad (26)$$

$$\begin{pmatrix} P(A) \\ P(\bar{A}) \end{pmatrix} = \begin{pmatrix} \theta \\ 1 - \theta \end{pmatrix} \quad (27)$$

Las razones para esta decisión es que estas variables estarán más allá de las pretensiones de un estudio de prevalencia, esto es, aunque se pretenda lo contrario, los estudios de prevalencia aproximan en realidad el número de pacientes diagnosticados, siendo muy complicado saber los reales.

Quedándose esta incertidumbre como un sesgo más.

8.1. Primera aproximación

En la anterior sección se ha calculado la expresión de la probabilidad de que un paciente se catalogue con la enfermedad por lo que se el hecho la distribución de D_i (detectar la enfermedad en un paciente i) seguirá la distribución de Bernouilli:

$$D_i \sim Ber(P(D)) = Ber((1 - d) \cdot \theta + e \cdot (1 - \theta)) \quad (28)$$

Por lo que el número de pacientes detectados con la enfermedad dada seguirá la distribución Geométrica:

$$T = \sum_{i=1}^N D_i \sim Geom(\mu, N) \quad (29)$$

Donde $\mu = (1 - d) \cdot \theta + e \cdot (1 - \theta)$

En el caso ideal de que la variables $b, c, d, e = 0$, es decir, siempre se registra la enfermedad y no hay ni falsos positivos ni falsos negativos. Tenemos que $\mu = \theta$ por lo que se puede aproximar θ con el estadístico de máxima verosimilitud:

$$\frac{\partial \log(f(\theta, T))}{\partial \theta} = 0 \Rightarrow \quad (30)$$

$$\frac{\partial (\log \theta^T (1-\theta)^{N-T})}{\partial \theta} = 0 \Rightarrow \quad (31)$$

$$\frac{\partial (T \log(\theta) + (N-T) \log(1-\theta))}{\partial \theta} = 0 \Rightarrow \quad (32)$$

$$\frac{T}{\theta} + \frac{(T-N)}{1-\theta} = 0 \Rightarrow \quad (33)$$

$$\frac{T}{\theta} = \frac{N-T}{1-\theta} \Rightarrow \quad (34)$$

$$\hat{\theta} = T/N \quad (35)$$

En el caso más general de que las variables sean distintas de 1 en el cual tanto los falsos positivos como negativos harían sesgarse las aproximaciones.

$$0 = \frac{\partial \log(f(g(\hat{\theta}), T))}{\partial \hat{\theta}} = \frac{\partial \log(f(g(\hat{\theta}), T))}{\partial g(\hat{\theta})} \frac{\partial g(\hat{\theta})}{\partial \hat{\theta}} \Rightarrow \frac{\partial f(g(\hat{\theta}))}{\partial g(\hat{\theta})} = 0 \quad (36)$$

donde $g(\theta) = (1-d) \cdot \theta + e \cdot (1-\theta)$

Luego aplicando 30

$$0 = \frac{\partial \log(f(g(\hat{\theta}), T))}{\partial \hat{\theta}} \Rightarrow g(\hat{\theta}) = \frac{T}{N} \Rightarrow \hat{\theta} = \frac{T/N - e}{1 - d - e} \quad (37)$$

Ahora es conveniente saber cual es el comportamiento del estadístico, es decir cual es la distribución de su error su error, ante las distintas circunstancias.

Debido a la complejidad del estadístico para calcular su varianza analíticamente usaremos el método de Montecarlo, que programaremos en **Python**, para aproximar la varianza de los mismos.

De ahora en adelante asumiremos que el error en la anotación del médico es despreciable, esto es, que el que se diagnostique o no erróneamente la enfermedad y, por lo tanto, se refleje en las historias médicas, es nulo.

Esta decisión se basa en que ningún otro método de encuestas va a estar exento de dicho error, pues estos se basarán también en los diagnósticos previos de los médicos.

```
from numpy.random import binomial, normal
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt

def simulate(N,M,theta,d,e,show=True):
    mu=(1-d)*theta+e*(1-theta)
    N=20000#tamaño de la muestra (pacientes considerados en el estudio)
    M=10000#número de iteraciones de simulaciones
    T=binomial(N,mu,M)#simulamos el experimento
    #T[:]:np.array tamaño M donde T[k] es el número de enfermos en los
    #que se detecta la enfermedad
    def approx(T):
```

```

#Calcula el estadístico de theta a posteriori:
    return ((T/N-e)/(1-d-e))
s=pd.Series([approx(t) for t in T])
if show:
    s.hist()
    plt.show()
print('Media(%): ', s.mean()*100, 'error al 95% (%): ',
      s.std()*1.96*100, sep=' ')
return s

```

Esto se traduce en el modelo en que $b = c = 0$.

Tomemos por ejemplo los valores $b = c = 0, \theta = 0,03, e = 0,1, d = 0,2$ se ve claramente que el estimador es centrado e insesgado:

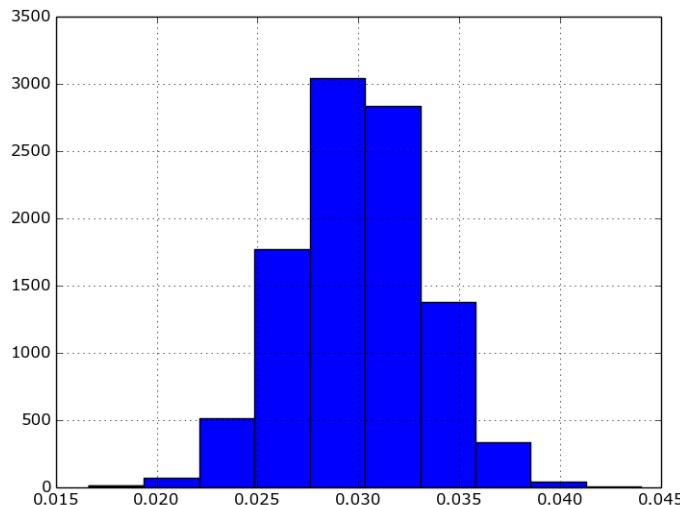
El error se cometido es aproximable a una normal:

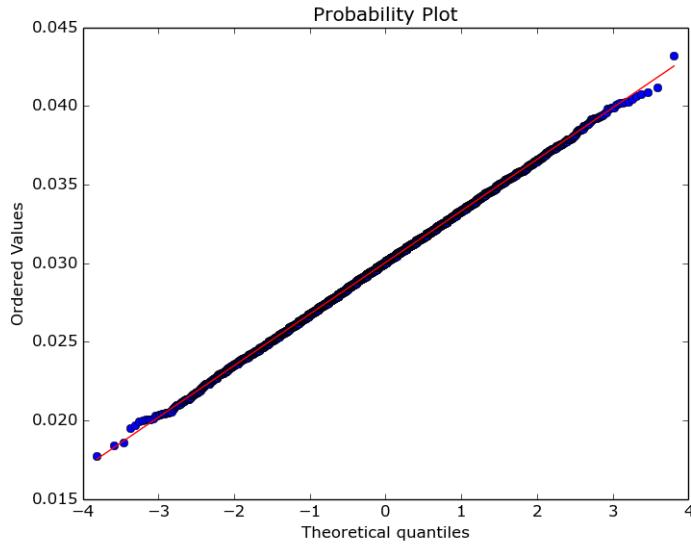
```

import stats
theta=0.03
e=.1
d=0.2
N=10000
M=10000
s=simulate(N,M,theta ,d,e ,show=False )

```

>Media(\%): 3.00473 error al 95 (\%): 0.642810683814





Como sería de esperar cuanto mayor sea el índice de falsos positivos como de falsos negativos mayor sera la varianza del estadístico, aunque este sigue siendo insesgado.

8.2. Formulación Matricial:

En el desarrollo de los estadísticos hemos encontrado que se pueden reformular los estadísticos de máxima verosimilitud usando el álgebra matricial.

Usando la expresión matricial mencionada en las anteriores subsecciones se puede intuir que se puede obtener :

$$\begin{aligned} \begin{pmatrix} \widehat{P(A)} \\ \widehat{P(\bar{A})} \end{pmatrix} &= \begin{pmatrix} 1-d & e \\ d & 1-e \end{pmatrix}^{-1} \begin{pmatrix} P(D) \\ P(\bar{D}) \end{pmatrix} \\ \Rightarrow \widehat{P(A)} &= (1-e -e) / (1-d-e) \begin{pmatrix} P(D) \\ P(\bar{D}) \end{pmatrix} = \frac{(1-e -e) \begin{pmatrix} T/N \\ 1-T/N \end{pmatrix}}{(1-d-e)} = \\ &\frac{T/N -e}{1-e-d} \quad (38) \end{aligned}$$

Lo cual corresponde con el estadístico de máxima verosimilitud.

¿Qué pasaría cuando la matriz no se pudiera invertir:

$$\begin{aligned} \det\left(\begin{pmatrix} 1-d & e \\ d & 1-e \end{pmatrix}\right) &= 0 \Rightarrow (1-d)(1-e) = e(d) \\ \Rightarrow ed - e - d + 1 &= ed \Rightarrow e + d = 1 \quad (39) \end{aligned}$$

Lo que corresponde con el caso de que la probabilidad de anotar la enfermedad no cambia en función de si existe o no la enfermedad. Pues la probabilidad de que se detecte la enfermedad es la misma en un paciente con la enfermedad y en otro que no tiene la enfermedad es la misma ($P(D|A) = P(D|\bar{A}) \Rightarrow P(D|A) =$

$P(D|\bar{A})$) por lo que no nos da ninguna información sobre si el paciente realmente tiene la enfermedad.

El IC de confianza calculado variará con respecto a las características de la matriz. Aproximando $\bar{T}/N \sim N(\mu = P(D), \sigma^2 = P(D)(1 - P(D)) = (1 - d)\theta + (1 - \theta)e + d\theta + (1 - e)(1 - \theta))$ la distribución del estadístico tendrá:

Media:

$$\begin{aligned} E[\widehat{P(A)}] &= \\ E\left[\frac{(1-e)\frac{T}{N} - e(1-\frac{T}{N})}{1-d-e}\right] &= \\ E\left[\frac{(1-e+e)\frac{T}{N} - e}{1-d-e}\right] &= \\ \frac{E[T/N] - e}{1-d-e} &= \\ \frac{(1-e-d)\theta + e - e}{1-d-e} &= \theta \quad (40) \end{aligned}$$

Con esto comprobamos lo que ya visto experimentalmente (método de Montecarlo) y es que el estadístico es insesgado.

Varianza:

$$\begin{aligned} Var(\widehat{P(A)}) &= Var\left(\frac{(1-d)\cdot\frac{T}{N} - d(1-\frac{T}{N})}{1-d-e}\right) = \\ Var\left(\frac{\frac{T}{N}-d}{1-e-d}\right) &= \frac{Var(T/N)}{(1-d-e)^2} = \\ \frac{((1-d)\theta + (1-\theta)e)((1-((1-d)\theta + (1-\theta)e)))}{(1-d-e)^2} &= \\ \theta(1-\theta) + (1-d)d\cdot\theta + (1-e)e(1-\theta) & \quad (41) \end{aligned}$$

Por lo que se puede apreciar que cuanto mayor sean el valor de d, e mayor es la varianza.

8.3. Segunda aproximación

En los anteriores puntos, hemos atacado el problema bajo el supuesto que conocíamos el error que se cometía, lo cual no es posible de conseguir, al menos sin error por lo que será necesario alguna forma de gestionar esta incertidumbre.

Para esto usaremos el concepto de *gold standar*, esto es un subconjunto de los datos clasificados por un método alternativo y que se considera fiable.

El objetivo de este conjunto será medir la precisión de nuestro modelo (falsos positivos y negativos) para aproximar el resultado.

Así pues, el proceso pasaría por medir un subconjunto de tamaño n_1 y ejecutar la detección automática de conceptos, estos valores nos permitirán aproximar los valores d, e con índices de confianza. Tras esto procesamos el total de los informes de los pacientes.

La estimación de los valores d, e se harán por el estadístico de máxima verosimilitud y tendrán un error $\epsilon_d \approx N(0, \sigma_d)$ y $\epsilon_e \approx N(0, \sigma_e)$ donde $\sigma_d = \sqrt{\frac{d(1-d)}{n_d}}$ y $\sigma_e = \sqrt{\frac{e(1-e)}{n_e}}$.

Como hemos visto ya, la estimación que hagamos será:

$$\begin{pmatrix} \widehat{P(E)} \\ \widehat{P(\bar{E})} \end{pmatrix} = \begin{pmatrix} 1 - e - \epsilon_e & d + \epsilon_d \\ e + \epsilon_e & 1 - d - \epsilon_d \end{pmatrix}^{-1} \begin{pmatrix} \widehat{T/N} \\ 1 - \widehat{T/N} \end{pmatrix} \quad (42)$$

Donde D es el número de pacientes en los que se ha detectado la enfermedad.

Para comprobar como se comportaría el error para aproximar un intervalo de confianza usaremos un método de Montecarlo. Modificando el código anterior para añadir el ruido en cada experimento:

```
from numpy.random import binomial, normal
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt

def simulate(N,M,theta,d,e,eps_d,eps_e,show=True):
    mu=(1-d)*theta+e*(1-theta)
    N=200000#tamaño de la muestra (pacientes considerados en el estudio)
    M=10000#número de iteraciones de simulaciones
    T=binomial(N,mu,M)#simulamos el experimento M veces
    #T[:np.array tamaño M donde T[k] es el enfermos en los que se detecta
    #la enfermedad en cada simulación.
    def approx(T):
        ne=+normal(0,eps_e) if eps_e>0 else 0
        nd=normal(0,eps_d) if eps_d>0 else 0

        #Calcula el estadístico de theta a posteriori:
        return ((T/N-e-ne)/(1-d-e-nd-ne))
    s=pd.Series([approx(t) for t in T])
    if show:
        s.hist()
        plt.show()
    return {'Media':s.mean()*100,
            'std':s.std()*1.96*100,
            'theta':0.03,
            'epsd':eps_d,
            'epse':eps_e}
```

Probamos varios casos manteniendo fijos el parámetro $\theta = 3\%$:

```
ans=[]
for eps_d in [0,0.1,0.01,0.001]:
    for eps_e in [0,0.1,0.01,0.001]:
        ans.append(simulate(0,0,theta,d,e,eps_d,eps_e,False))

P=pd.DataFrame(list(map(
    lambda x:[x['Media'],x['std'],x['error al 95% (%):'],x['epse'],x['theta']],
    ans
)))
P.columns=['Media(%)','std(%)','error al 95% (%):','epse','theta']
P.sort('std(%)')
```

Lo que nos da:

	Media(%)	error al 95 % (%)	epsd %	epse %	theta
12	3.000046	0.080681	0.001	0.000	0.03
0	2.999618	0.081341	0.000	0.000	0.03
8	2.999723	0.100087	0.010	0.000	0.03
15	2.998642	0.207574	0.001	0.001	0.03
3	3.000228	0.214006	0.000	0.001	0.03
11	2.999227	0.215864	0.010	0.001	0.03
4	3.035610	0.622696	0.100	0.000	0.03
7	3.031361	0.660852	0.100	0.001	0.03
10	2.971683	1.911945	0.010	0.010	0.03
14	2.992006	1.931819	0.001	0.010	0.03
2	2.972322	1.948177	0.000	0.010	0.03
6	3.012852	2.032172	0.100	0.010	0.03
9	1.922331	19.860694	0.010	0.100	0.03
13	2.052586	19.998204	0.001	0.100	0.03
1	1.965913	19.998631	0.000	0.100	0.03
5	1.852633	20.617304	0.100	0.100	0.03

Observando los datos se visualiza que el parámetro más delicado es el de e , (falso positivo) por lo que será necesario controlar esta variable para evitarlo.

8.4. Enfermedades poco frecuentes:

En el caso de enfermedades poco frecuentes la estrategia del anterior punto no sería la más conveniente. Esto es debido a que nos es que nos provoca mucho más error la generación de falsos positivos que la de falsos negativos, por lo que sería más conveniente hacer un test donde se verifica que los falsos negativos son bajos (no nos saltamos muchas enfermos) y usar como *filtro* un método alternativo por el que se corroboren todos los enfermos detectados (o una gran parte).

Por ejemplo, ante una enfermedad de una prevalencia de 0,1% es factible detectar en una gran cantidad de informes (de 1.000.000 de pacientes) y luego aplicar a anotadores que confirmen que no hay falsos positivos en los informes que indiquemos (bajo el supuesto razonable que no se generen muchos falsos positivos esta tarea debería ser razonablemente sencilla).

Asegurar la cobertura (esto es que se detecta la mayoría de los casos) no es una tarea tan sencilla, ya que es muy probable que en la creación de un gold standar escogiendo pacientes aleatorios y evaluando si se escribe que tienen la enfermedad no encuentre suficientes enfermos para estimar dicho parámetro, tambien se puede usar una base de datos generada por otros medios para ver la calidad de detección (por ejemplo algunos hospitales registran algunas enfermedades mediante la cumplimentación de questionarios electrónicos).

La solución pasará pues por usar algún método opcional que nos permita encontrar con un menor esfuerzo una muestra de pacientes en el hospital con la enfermedad (no necesariamente exhaustiva) y obtenida de forma independiente, una posibilidad (un registro previo).

Una posibilidad de medir nuestra capacidad de detección, al menos en enfermedades crónicas, es dividir en dos periodos el conjunto de los informes, por

ejemplo si tenemos 12 meses de informes tomar los 6 primeros meses en un bloque y los otros 6 en el segundo bloque (conjuntos A, B respectivamente en adelante).

De A detectar automáticamente los pacientes con la enfermedad y usar los informes de dichos pacientes del subconjunto B (los cuales tendrán la enfermedad) con el que estimar las posibilidad de que estos no se detecten.

Una vez estimado el número de falsos negativos no es difícil calcular cual sería la prevalencia, así como el intervalo de confianza de la estimación.

Finalmente se usaría para calcular el IC de estos supuestos el programa escrito en el punto anterior.

9. Experimento

No podíamos dejar pasar la ocasión de hacer un experimento que use la información recuperada por la empresa.

El objetivo es hacer un ligero estudio exploratorio sobre las correlaciones positivas entre las enfermedades, esto es, señalar enfermedades que aparecen juntas más de lo esperado en el mismo paciente. Esto puede deberse a que haya una causa común a ambas (por ejemplo las personas con hábitos deportistas tenderán a tener más lesiones de diferente índole o los fumadores tendrán más propensión a tener enfermedades de tipo respiratorio), a que una influya a la otra (una vez adquirida la diabetes la probabilidad de que se generen trombos en las extremidades aumenta), o simplemente sea una relación espúrea (al haber tantas variables, es prácticamente inevitable que alguna supere el test por casualidad).

Explicaremos los pasos efectuados.

En un primer lugar de nuestra base de datos extraemos una tabla donde vinculamos a los pacientes con las enfermedades. Con el objetivo de evitar correlaciones debido a la variable edad, nos centraremos en pacientes jóvenes, esto es, cogeremos los informes de gente que (al inicio de la muestra) tenían entre 15 y 20 años, y extraemos los 5 años posteriores. Dándonos una muestra de aproximadamente 20.000 pacientes.

Resumimos la información en una tabla de dos columnas: p : id de paciente e : id de la enfermedad diagnosticada.

```
import pandas as pd
P=pd.read_csv('data')
PCuad=P.pivot(index=P['p'],columns=P['e'],
values=np.repeat(1,rsDis.shape[0]))
```

Con lo que se genera una tabla de la manera:

Y con esta tabla donde se muestra las enfermedades de cada paciente (filas).

Sobre esto usamos es test de χ^2 visto en la sección 6.1 para cada par de enfermedades:

```
Relaciones=[]
import numpy as np
import scipy.stats as scst

Relaciones=set()
ans=Counter()
N=#numero de pacientes aquí contemplados
#tomamos las enfermedades que han afectado, al menos
# a 50 personas (nos aseguramos de no tomar enfermedades
# poco frecuentes)
corte_p_valor=.05
enf_freq=[i for i in rsCuad if rsCuad[i].sum()>=50]
for i in enf_freq:
    for j in enf_freq:
        if i<j:
            # Debido a la simetría de la relación
            # no es necesario evaluar i,j y j,i con
```

```

# if i>j nos ahorraremos esto

# Creamos la matriz 2 x 2 con las frecuencias
# de las posibilidades de tener i y j en un mismo
# paciente
n={}
I=rsCuad[i].fillna(0)
J=rsCuad[j].fillna(0)
n[(1,1)]=np.dot(I,J)

n[(1,0)]=np.dot(I,1-J)
n[(0,1)]=np.dot(1-I,J)
n[(0,0)]=N-n[(1,1)]-n[(0,1)]-n[(1,0)]
n[(1,None)]=n[(1,0)]+n[(1,1)]
n[(0,None)]=n[(0,1)]+n[(0,0)]
n[(None,1)]=n[(1,1)]+n[(0,1)]
n[(None,0)]=n[(1,0)]+n[(0,0)]
mu=n[(1,None)]*n[(None,1)]/(N**2)#media esperada
var=(mu*(1-mu))
if (scst.chi2_contingency(np.array(
    [[n[(1,1)],n[(1,0)]],
     [n[(0,1)],n[(0,0)]]])
))[1]< corte_p_valor and (n[(1,1)]/N-mu)>0:
    Relaciones.add((i,j))

```

Esto es, el programa por cada par de enfermedades calcula la tabla n :

número de pacientes con ambas enfermedades	número de pacientes con enfermedad i pero no j
número de pacientes con enfermedad j pero no i	número de pacientes con ninguna enfermedad

Y la tabla E , quedándose con aquellos que superan el test χ^2 ²¹ con una confianza de 0,05

y además la correlación es positiva, es decir, aquellas enfermedades que han ocurrido simultáneamente más de lo que cabría esperar por simple azar, (para asegurarnos de que no ocurre la relación menos de lo que es frecuente usamos la condición $(n[(1,1)]/N-mu)>0$ con lo que excluimos que se excluya la posibilidad de que se haya rechazado el test porque sea menos frecuente de lo esperado).

Con esto se forma un grafo ($U.V$) donde los nodos son las enfermedades diagnosticadas y $(u_1, u_2) \in V$ si hay evidencias de que hay correlación positiva entre ellas (el odd-ratio es positivo y se rechaza el test de χ^2) usando la librería networkx exportamos el grafo:

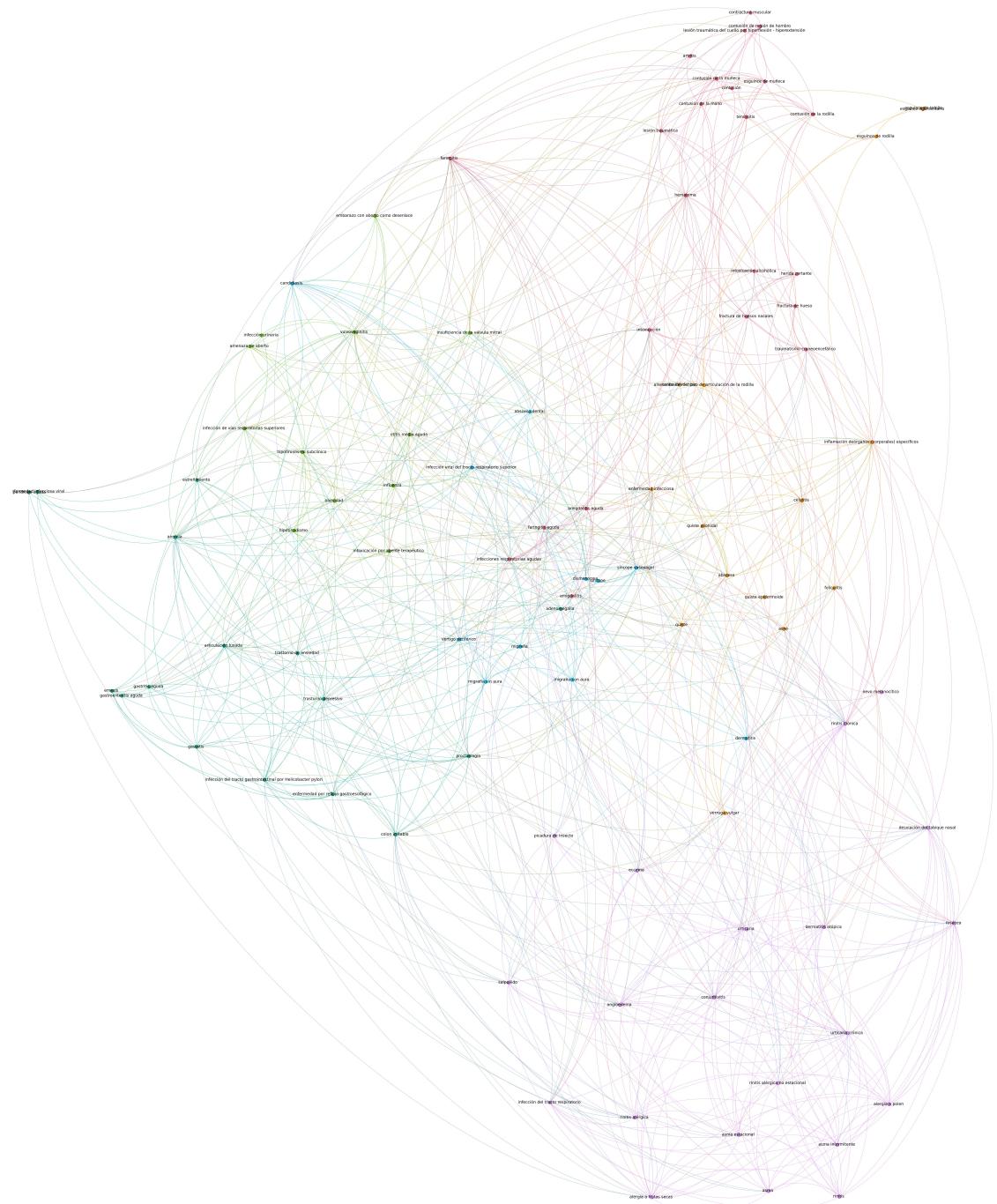
```

import networkx as nx
G=nx.Graph()
G.add_edges_from(Relaciones)
nx.write_gml(G, 'graph.gml')

```

Abriendo el grafo con el programa Gephi podemos visualizarlo:

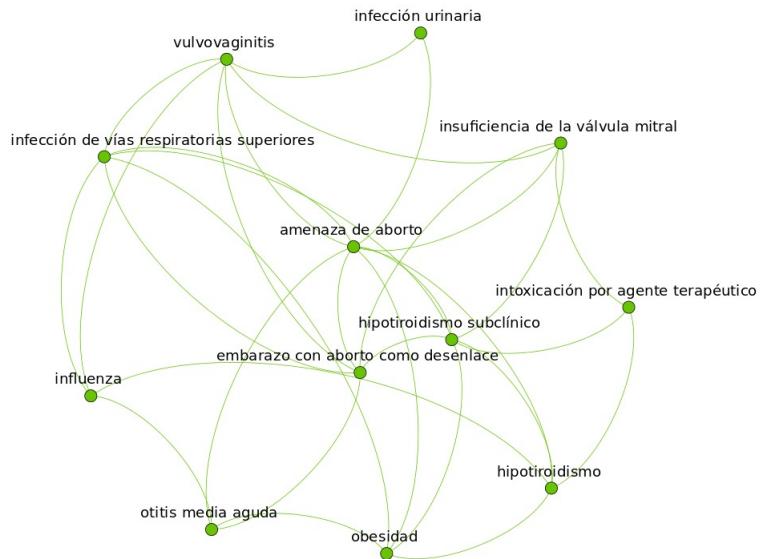
¹usamos la función chi2contingency del paquete scipy.stats de python [3]



Usando sus opciones para encontrar comunidades (esto es, grupos de enfermedades con muchas interconexión) se pueden ver algunas cosas interesantes.

Veamos comunidad a comunidad y comentaremos, con el apoyo de uno de los médicos asociados a la empresa, cada una de ellas:

9.1. Comunidad 1



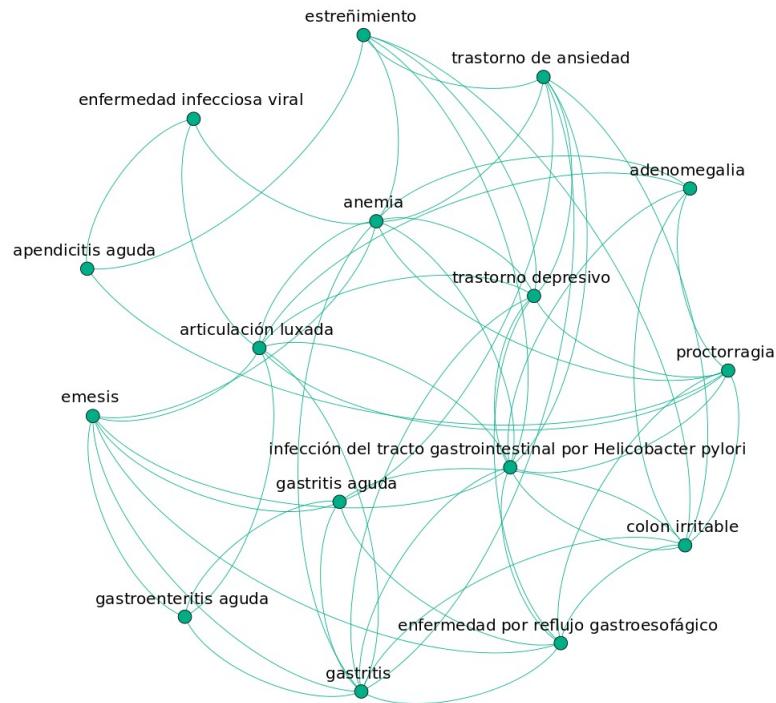
En esta comunidad se ven algunas causas que pueden ser amenazas de aborto como puede ser el hipotiroidismo, la infección urinaria, la vulvaganitis, el hipotiroidismo o la insuficiencia de la válvula mitral.

Por otro lado algunas infecciones pueden ser también amenaza.

Parece que en la comunidad se nos introduce de forma casual influenza (gripe), otitis e infección de vías superiores, las cuales si están relacionados clínicamente.

La relación entre la *intoxicación por agente terapéutico* (lo que de forma coloquial se conoce como sobredosis de medicamento) y los abortos es, en opinión de nuestro médico, intentos de aborto por parte del paciente, debido a una situación de embarazo no deseado, lo cual concuerda con el hecho de que los pacientes son jóvenes.

9.2. Comunidad 2



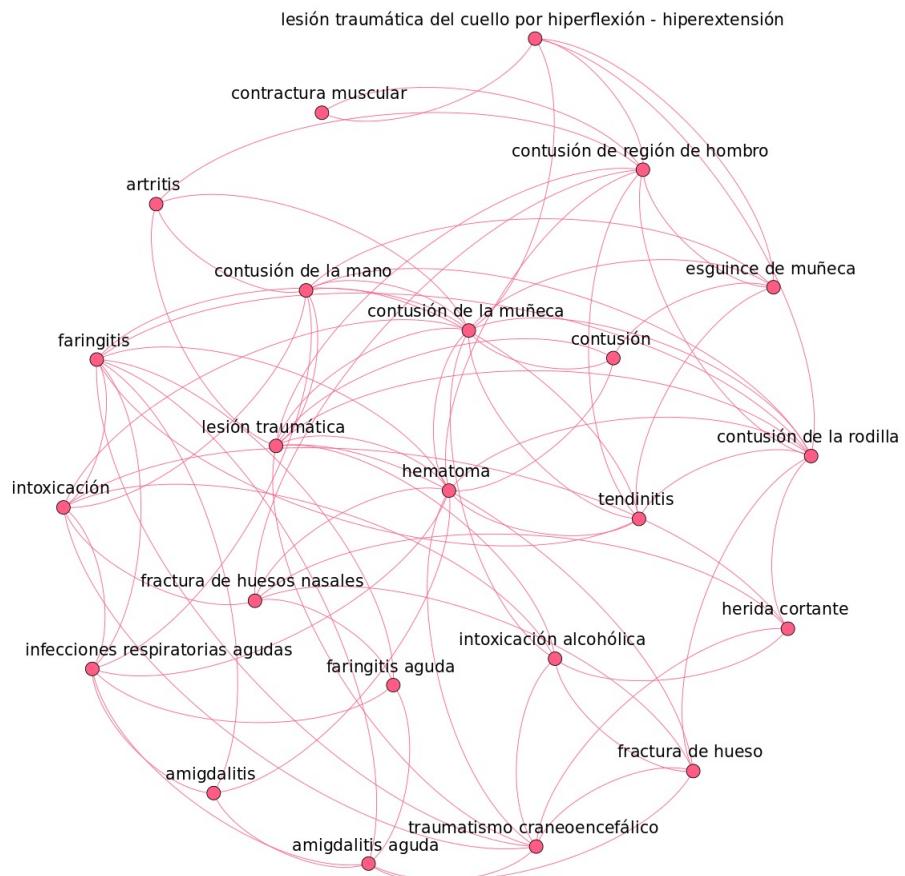
Esta comunidad está protagonizada por enfermedades de tipo gastrointestinal.

Hay un par de excepciones fácilmente explicables:

- anemia: Los problemas intestinales suelen producir por lo habitual un problema de absorción de la vitamina B-12, la cual es uno de los parámetros del “hierro en sangre”
- trastorno depresivo: las úlceras pueden producir estrés y trastornos depresivos ²

²Hasta hace relativamente pronto se consideraba que era el estrés el que causaba las úlceras, si bien los recientes estudios han demostrado que son las úlceras las causantes del estrés.

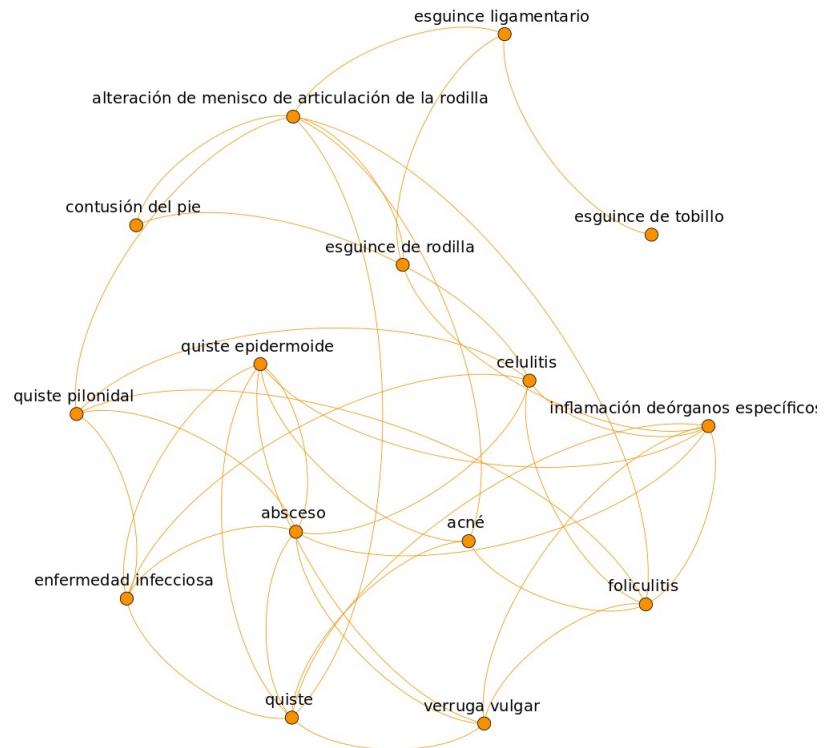
9.3. Comunidad 3



En este grafo se puede discernir por un lado la asociación de conceptos taumatológicos y de infecciones de vías altas por la parte izquierda (por tanto, dos poblaciones).

La presencia de intoxicación se debe a intoxicación etílica, ya que es causa de todo tipo de traumas.

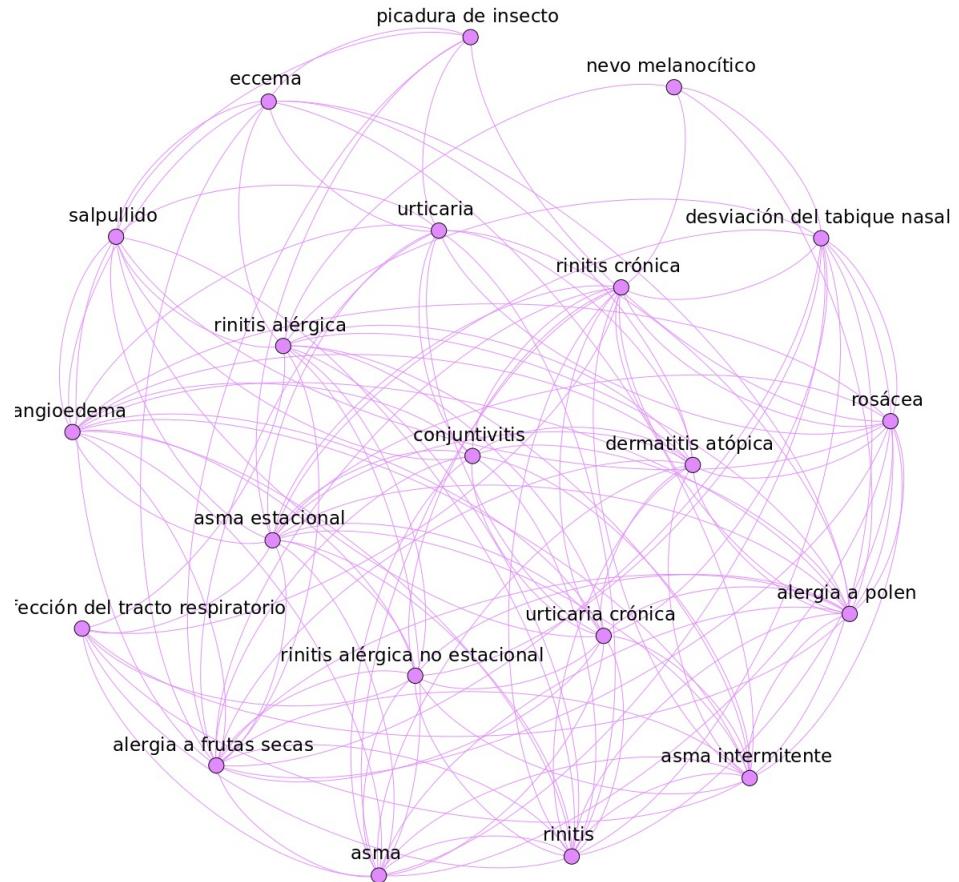
9.4. Comunidad 4



Esta comunidad se divide en dos partes:

- La parte superior agrupa las incidencias de tipo traumatólogico.
- La inferior infecciones e inflamaciones de la piel.

9.5. Comunidad 5

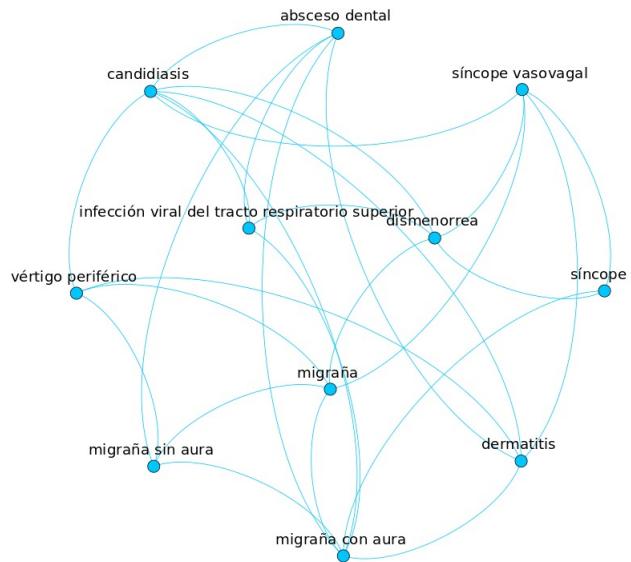


Esta comunidad agrupa muy bien elementos relacionados con las alergias.

En la parte superior derecha hay 3 elementos que podrían parecer erroneos, pero todos se explican:

- **Tabique nasal:** se asocia a la rinitis
- **Rosacea y nevo melanocito:** se asocia a otros problemas de la piel presentes en el grafo

9.6. Comunidad 6



Esta comunidad no está tan claramente definida como los anteriores.

Sino que se compone, principalmente, de enfermedades claramente más comunes en mujeres que en hombres (candidiasis,migrañas,sincopes, e infecciones urinarias). Con la excepción de el vértigo , los cuales no tienen una relación clara con el resto de enfermedades.

10. Conclusiones

En esta sección vamos a intentar definir cuales serian las ventajas y las desventajas de reutilizar las historias médicas con el objetivo de hacer estudios epidemiológicos frente al resto de métodos epidemiológicos y como puede contribuir en el marco de los distintos procesos.

10.1. Sesgos:

Todo método ecológico, a pesar de la correcta planificación y la preparación que se tome, no está exento de sesgos, y es menester ser capaces poder señalar cuales son los sesgos principales que pueden afectar el estudio.

10.1.1. Selección:

Dependiendo de donde se extraigan los informes se puede caer en el sesgo de Berkson (si es un hospital). Si bien al incluirse los informes de consulta de los pacientes este sesgo se solucionaría donde se seguirían los casos sin hospitalización.

Las únicas formas reseñables de las que este podría afectar los estudios son el sesgo de selección son los niveles de sobrediagnóstico e infradiagnóstico, que corresponderían en la terminología de la epidemiología a los sesgos de inclusión/exclusión.

10.1.2. Sesgo de información

Debido a que no tenemos acceso directo a los pacientes del estudio ni podemos, imponer de ninguna manera a los médicos las preguntas o las mediciones propias de un estudio no tendremos un criterio estandar para comparar los pacientes.

Por ejemplo, es posible que nos interese ver como son los niveles de hemoglobina en una enfermedad; pero esto solo en los pacientes que se le hagan pruebas analíticas.

Lo mismo se puede decir con parámetros tipo categórico (Fumador/No Fumador), (sobre los cuales es interesante señalar que el sistema es capaz de detectar eficazmente si estos están negados).

No obstante, es esperable, que los parámetros más revelantes aparezcan en la mayoría de los casos, pues estos serán los usados de forma sistemática para diagnosticar y valorar la enfermedad y su evolución.

Una característica interesante del Big Data es que, mientras que en un estudio cuyas preguntas han sido planificadas de antemano es imposible encontrar correlaciones con variables que no se hayan añadido al mismo.

Al añadirse en el Big Data donde se añada toda la información, sí será posible encontrar estas correlaciones *inesperadas*.

Estas relaciones pueden dar pasos a ampliar nuestros conocimientos en medicina, ayudar a hacer mejores métricas de los riesgos.

10.1.3. Sesgo de confusión

El sesgo de confusión, nace cuando en un estudio se han dejado de medir variables relevantes, lo cuál se paliaría con una correcta planificación. Por des-

gracia esto no sería posible controlar mediante el uso de las historias médicas directamente, sino que estaríamos dependientes de lo que registren los médicos, si bien las variables más relevantes para las enfermedades de los pacientes (antecedentes familiares, fumador ...) sí se suelen añadir de forma rutinaria .

Por otro lado no estaría restringido a las variables que los médicos que consideren a priori, permitiendo ir más allá de lo establecido y descubrir nuevas relaciones que hayan pasado por alto hasta ahora.

10.2. Posibilidades del medio:

El método de procesar todos las historias médicas puede dar otra información relevante sobre enfermedades más allá de las características propias de la enfermedad .

El hecho de recopilar los informes creados en el día a día permite describir cuales son los pasos de los pacientes la *realidad médica* antes de ser diagnosticadas, las pruebas que se han hecho... incluso ayudar en el desarrollo de vías médicas.

Asimismo se presentaría como un medio para combatir la denominada variante médica (ante el mismo paciente varios), lo que puede contribuir, más allá de un avance en los conocimientos, en una aplicación más eficiente de la medicina, pudiéndose explorar con facilidad los sucesos que ocurren hasta que un paciente es diagnosticado o no .

10.3. Correlación no implica causalidad

Es una crítica común entre los escépticos sobre nuestra empresa -especialmente de parte de empresas que hacen estudios epidemiológicos que usar las técnicas de **Big Data** en el campo de la medicina es inútil debido a que la correlación no implica realmente causalidad.

Si bien es razonable la objeción y es algo que podría dar la sensación de que nuestra tarea es inútil y vana, el mismo razonamiento es extensible a la epidemiología misma. Afortunadamente la experiencia ha demostrado lo contrario.

Como ejemplo cuando en Doll y Hill presentó en un estudio en un hospital mostrando que el tabaco producía cáncer de pulmón, algo que fue en su tiempo resultó inesperado,hubo críticas sobre la posibilidad de que hubiera un componente genético (sesgo de confusión sobre una variable difícilmente observable) que causará simultáneamente una mayor predisposición tanto a ser fumador como a desarrollar este tipo de cáncer.

Si bien podríamos admitir que la duda es razonable, la verdad, es que puso el foco de atención sobre una conjectura con fundamento y permitió que posteriores estudios corroborarán esto. Hoy en día nadie duda de que ser fumador es un serio factor de riesgo en esta enfermedad.

Por otro lado, aunque las correlaciones puedan no indicar una causalidad debido a que haya una variable subyacente si que ayudan a clasificar a los pacientes por riesgos lo que puede ayudar a dar juicios clínicos más acertados.

Por ejemplo aunque la conjectura de que la correlación de fumar y cáncer de pulmón solo sea debida a una causa afín, en ausencia del acceso a la variable subyacente, esta seguiría siendo una pista de cara a la asignación de un diagnóstico pues estaría indicándonos una probable predisposición al tabaquismo y por lo tanto al cáncer.

Referencias

- [1] Big data analytics in healthcare: promise and potential. http://www.isciii.es/ISCIII/es/contenidos/fd-publicaciones-isciii/fd-documentos/2009-0843_Manual_epidemiologico_ultimo_23-01-10.pdf.
- [2] Definición de sesgo (rae). <http://dle.rae.es/?id=XipMgHq>.
- [3] Documentación de la librería scipy (python). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html.
- [4] El modelo de regresión de cox. <http://deposit.ub.edu/dspace/bitstream/2445/49070/6/E1%20modelo%20de%20Cox%20de%20riesgos%20proporcionales.pdf>,.
- [5] Epidemiología y metodología aplicada a la pediatría (v): Sesgos. <https://www.aeped.es/sites/default/files/anales/50-5-21.pdf>.
- [6] Estadística no paramétrica: prueba chi-cuadrado. https://www.uoc.edu/in3/emath/docs/Chi_cuadrado.pdf.
- [7] Extrapolating gag. <https://xkcd.com/605>.
- [8] Guía práctica del curso de bioestadística aplicada a las ciencias de la salud. http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0036-3634200000050001.
- [9] Guide to biostatistics. <http://www.medpagetoday.com/lib/content/Medpage-Guide-to-Biostatistics.pdf>.
- [10] Manual docente de la escuela nacional de sanidad: Método epidemiológico. http://www.isciii.es/ISCIII/es/contenidos/fd-publicaciones-isciii/fd-documentos/2009-0843_Manual_epidemiologico_ultimo_23-01-10.pdf.
- [11] Modelos de regresión logísitca.
- [12] Paradoja de simpson. https://es.wikipedia.org/wiki/Paradoja_de_Simpson.
- [13] Sesgos en estudios epidemiológicos. http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0036-36342000000500010.
- [14] Test fisher. <http://www.fisterra.com/mbe/investiga/fisher/fisher.asp>.
- [15] Test log rank. <http://web.stanford.edu/~lutian/coursepdf/unitweek3.pdf>,.