## Resumen

Este trabajo se enmarca dentro del llamado problema de la canasta del supermercado (*market basket problem*) en el que se extrae una base de datos de transacciones para el descubrimiento de conjuntos de elementos inusualmente frecuentes. Esto implica estimar la relación entre la frecuencia de cada conjunto de elementos y una frecuencia de referencia calculada como si los conjuntos de elementos ocurrieran de forma independiente. Esta relación debe ser confiable tanto para conjuntos de elementos con frecuencias de apariciones altas y bajas. El foco del trabajo es el análisis de informes de detección de reacciones adversas a drogas obtenidos de la base de datos del sistema de reportes de la Administración de Drogas y Alimentos de los Estados Unidos (FDA).

## Introducción

Los diversos productos regulados por FDA representan aproximadamente el 25% de la economía de los Estados Unidos, se usan a diario y afectan la salud de muchos millones de personas y animales. Además de alimentos y medicamentos, incluyen, por ejemplo, suplementos nutricionales, alimentos genéticamente modificados, vacunas, corazones artificiales, láseres quirúrgicos, dispositivos utilizados para administrar medicamentos y productos biológicos, terapias genéticas, medicamentos veterinarios, alimentos para mascotas, productos de tabaco y muchos otros. Los eventos adversos asociados con estos productos son responsables de enormes costos de salud pública y financieros. Estos costos relacionados con eventos adversos afectan el desarrollo de productos de atención médica, las primas de seguros de salud y los servicios de atención médica (por ejemplo, hospitalizaciones), todo lo cual conduce a pérdidas sociales a largo plazo, como discapacidad permanente y muerte. (Agency for Healthcare Research and Quality, AHRQ., 2005)

La FDA recopila y mantiene datos que proporcionan información de seguridad para los productos regulados, una gran parte de la información que proporciona son las bases de datos de informes de problemas de seguridad presuntamente asociados con productos comercializados. El número anual de informes recibidos ha aumentado constantemente a lo largo de las décadas debido a factores como el aumento de la población, el número y tipo de productos regulados, la conciencia de la importancia de los informes y la mayor facilidad con la que se pueden enviar informes. Actualmente, la FDA recibe aproximadamente dos millones de informes de eventos adversos, errores de uso y quejas de productos cada año de consumidores, profesionales de la salud, fabricantes y otros. Estos informes se ingresan en varias bases de datos mantenidas por la FDA para análisis posteriores para identificar posibles problemas de seguridad y mejorar la comprensión de esos problemas. (Hesha J Duggirala, 2016)

## Datos

Se cuenta con información del primer trimestre de 2020 de eventos adversos, incluyendo las siguientes bases de datos:

* información demográfica y administrativa del informe;
* información de drogas y tratamiento;
* información de eventos adversos;
* información de resultados del paciente;
* información sobre la fuente de los informes

Esta base contiene información de reportes de 12.129 drogas y 12.160 eventos adversos sumando más de 10 millones de combinaciones droga - evento adverso.

## Metodología

### Preprocesamiento

Se realizó una limpieza de reportes duplicados manteniendo el último reporte para cada evento adverso y se normalizaron las unidades de medidas de las variables numéricas. Los cálculos se ven afectados directamente por la calidad de la data, en especial con respecto a los nombres de drogas y los eventos. Cualquier cambio con respecto a la forma en que está escrita puede significar que no se tome a la misma droga o evento adverso como la misma afectando las métricas. Se hizo una limpieza de los textos eliminando espacios, símbolos y llevando todas las palabras a mayúscula y clasificando las 12.129 drogas en 5.865 drogas genéricas mediante los componentes activos. Para el análisis se utilizaron dos variables demográficas como estratificadoras, el sexo y la edad de las personas. El propósito de la estratificación es evitar encontrar asociaciones espurias entre dos elementos simplemente porque ambos tienden a ser más frecuentes en el mismo estrato.

### Métodos de desproporcionalidad: PRR y EBGM

La FDA utiliza métodos de desproporcionalidad para identificar asociaciones estadísticas entre productos y eventos. (Hesha J Duggirala, 2016) Dichos métodos comparan el recuento observado de una combinación de evento de producto con un recuento "esperado". Las asociaciones de productos y eventos adversos inesperadamente altas "señalan" (Waller, 2003) que puede haber una asociación causal entre ellos. Estas asociaciones altas son denominadas como *Disproportionately Reported Combinations (DRCs)*.

Se utiliza la relación de informe proporcional (*proportional reporting ratio* o PRR) como concepto fundamental en muchos de estos métodos de desproporcionalidad. (Waller, 2003) (Finney, 1974) El PRR es el grado de informe desproporcionado de un evento adverso para un producto de interés en comparación con el informe de este mismo evento adverso para todos los demás productos en la base de datos. Por lo tanto, toda la base de datos se utiliza como un recuento de fondo "esperado".

El PRR se basa en un "supuesto de independencia", es decir, que no hay asociación entre productos y eventos mencionados en los informes. Si hay un informe desproporcionado de un evento para un producto en particular, entonces este supuesto de independencia es cuestionable, es decir, puede haber una asociación entre el evento adverso particular y el producto. Tal "asociación" es estadística; por lo tanto, no puede interpretarse como causal o relacionado con el riesgo. (Hesha J Duggirala, 2016)

Este concepto de desproporcionalidad puede mostrarse mediante una tabla de contingencia donde

* "a" son todos los informes de un evento adverso específico ("Evento Y") para el Producto (por ejemplo, un medicamento) X,
* "b" son todos los informes de todos los demás eventos adversos para el Producto X,
* "a + b" son todos los informes para el Producto X,
* "c" son todos los informes de todos los demás productos para el Evento Y,
* "d" son todos los informes de todos los demás productos para todos los demás eventos adversos, y
* "c + d" son todos los informes para todos los demás productos.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Evento Y | Demás eventos |  |
| Producto X | a | b | a + b |
| Demás productos | c | d | c + d |
|  | a + c | b + d | Total |

Siendo,

El PRR representa la relación entre el número de casos observados y los casos esperados bajo el supuesto de independencia entre productos y síntomas. Si la relación observada (N) es mayor que la relación de eventos esperada (E) entonces el Evento Y está reportado desproporcionalmente para el Producto X, utilizando el resto de la base de datos como valores “esperados”.

Este factor es la base de los métodos utilizados por la FDA pero no tiene en cuenta eventos con pequeños números observados o esperados de informes del par de interés producto-evento. Si así fuese el caso, el valor de PRR sería extremadamente alto. Por lo tanto, se emplean otros métodos estadísticos más avanzados, como el *Multi-Item Gamma Poisson Shrinker* (MGPS) que produce puntajes empíricos de la media geométrica bayesiana (EGBM). (Hesha J Duggirala, 2016) El cálculo EBGM es conceptualmente similar al del PRR, pero incorpora el encogimiento o la contracción y la estratificación bayesiana para producir puntajes de desproporcionalidad hacia el nulo, especialmente cuando hay datos limitados y un pequeño número de casos. (Szarfman, 2004)

Una diferencia importante entre las estimaciones de PRR y EBGM es que, en el caso de PRR, los eventos adversos del producto en cuestión no contribuyen al número de casos esperados, mientras que todos los eventos adversos del producto contribuyen a la expectativa cuando se usa EBGM. Las modificaciones estadísticas utilizadas en la metodología EBGM disminuyen el efecto de valores de PRR espuriamente altos, reduciendo así el número de señales de seguridad falsas positivas. (Bate, 2009) Por lo tanto, los valores EBGM proporcionan una estimación más estable de la tasa de notificación relativa de un evento para un producto en particular en relación con todos los demás eventos y productos en la base de datos que se analiza. (Szarfman, 2004)

El modelo de Bayes Empírico asume que cada cantidad observada proviene de una distribución de Poisson con media desconocida. La relación que se calcula ahora es:

para cada una de las celdas de tabla de contingencia con fila i y columna j, donde i = 1,. . . , I y j = 1,. . . , J. La función de densidad de λ viene dada por la mezcla de dos distribuciones Gamma y los parámetros pueden ser estimados utilizando la data observada:

Siendo g una distribución Gamma con parámetros y y P la probabilidad de pertenecer a la mezcla droga-producto.

La distribución de λ una vez conocidos los N datos es:

Donde E es la cantidad esperada de eventos y es la probabilidad a posterior de que λ viene de la primera mezcla droga-producto.

Para consultar más detalle respecto al desarrollo del método referirse a *Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system*. (DuMouchel W. , 1999)

### Estimación de hiperparámetros

Los hiperparámetros que se deben estimar son los correspondientes a la función de densidad λ:

La distribución marginal de cada valor observado está dada por la mezcla de dos distribuciones binomiales negativas, estas distribuciones se derivan de una mezcla de distribución de Poisson donde las medias tienen distribución Gamma. Se define entonces la función de verosimilitud, como producto de M mezclas de dos densidades binomiales negativas:

Es necesario iterar con valores iniciales dey P hasta alcanzar el valor máximo de la función de verosimilitud. Esta estimación puede realizarse utilizando toda la base de datos usando una versión compacta que se describe a continuación.

### Reducción de la cantidad de observaciones

En la práctica, los cálculos de estos hiperparámetros resultan en una gran carga computacional. Es por ello que se utilizó la metodología propuesta por DuMouchel & Pregibon de *data squashing* (DuMouchel, 2001)para compactar la base de datos. El método se basa en agrupar los datos cercanos en super puntoso *bins* y tomar como valor de las variables el valor promedio de las observaciones y sumar una nueva dimensión con el peso correspondiente a la cantidad de observaciones que están agrupadas.

Mediante esta técnica, se logró reducir la base de datos, conformada por la cantidad actual, la cantidad esperada y la nueva variable de peso, un 64%.

## Resultados

La base de datos compuesta únicamente por combinaciones droga-evento contiene 337.750 observaciones de la cuales solamente 8737 resultaron sospechosas, valor de EBGM mayor a 5, correspondiente a 0,02%. Hay 903 principios activos y 2492 eventos.

La razón por la cual la métrica de PRR debe ser reducida mediante el método de EBGM se hace evidente en la Figura 1 donde se representa la tasa de reporte PRR frente al número de casos reportados para la combinación droga-evento .

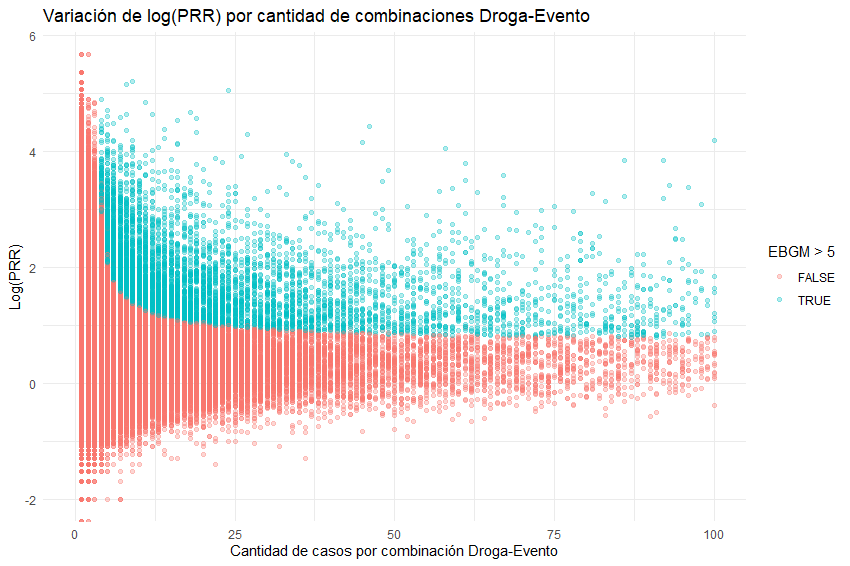


Figura 1

Los círculos celestes representan combinaciones de droga-evento con puntajes de EBGM superiores a 5, señal que son casos sospechosos.

El primer caso con EBGM mayor a 5 se corresponde cuando la cantidad de casos es igual a 4 y el log(PRR) es 2,91 (PRR = 818), cuando la cantidad de casos es 10, el primer caso sospechoso tiene log(PRR) igual a 1,68 (PRR = 20,97) y al aumentar a 100 el número de casos el primer caso sospechoso tiene log(PRR) igual a 0,88 (PRR = 7,76). Es decir, a mayor cantidad de casos, se requiere un PRR menor para considerar un caso sospechoso.

Por otro lado, el encogimiento de la métrica EBGM al tener valores de números de casos (N) pequeños el número de casos se puede visualizar en la Figura 2. En el caso de PHENYTOIN SODIUM EXTENDED - PERFORMANCE FEAR con N = 1, el valor del PRR es 460.317 (log(PRR) = 5,66) pero el valor de EGBM es 0,65. Es decir, si utilizáramos la métrica PRR sería significativo el evento, pero al usar EGBM no se toma como sospechoso. A mayores valores de N, esta diferencia no está tan marcada, por ejemplo, tomando la combinación RIVAROXABAN - VOMITING con N = 20, el PRR es 0,18 y el EGBM es 0,19 y en un caso más extremo donde N = 200 como para el caso de DALFAMPRIDINE - MULTIPLE SCLEROSIS los valores de PRR y EGBM son 28,45 y 23,89 respectivamente y para ambas métricas serían casos sospechosos.

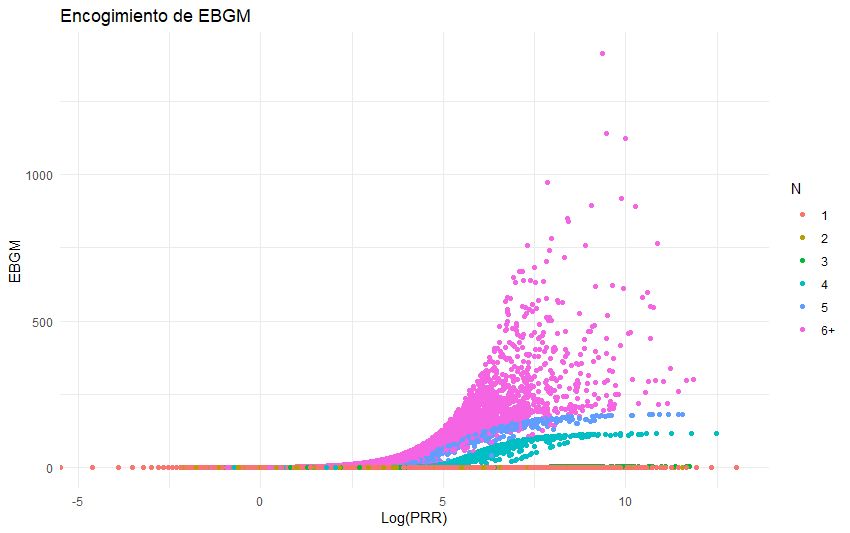


Figura 2

Dentro de las 20 combinaciones droga-evento con mayores valores de EBGM se encuentra 3 veces el componente activo Misoprostol y ningún evento se repite. *Parathyroid hormone* es la droga con mayor valor de EBGM y resulta ser un producto que sufrió un *recall*, es decir, que se tuvo que retirar del mercado. Es posible que la métrica sea elevada porque los eventos para esa droga se encuentran enmascarados por la descripción del evento, en total la droga aparece en 266 combinaciones y 190 veces es con el evento *“recalled product”.*

En la Figura 3 se ilustran las 20 combinaciones con los valores de EBGM, los intervalos de confianza y la cantidad de apariciones.

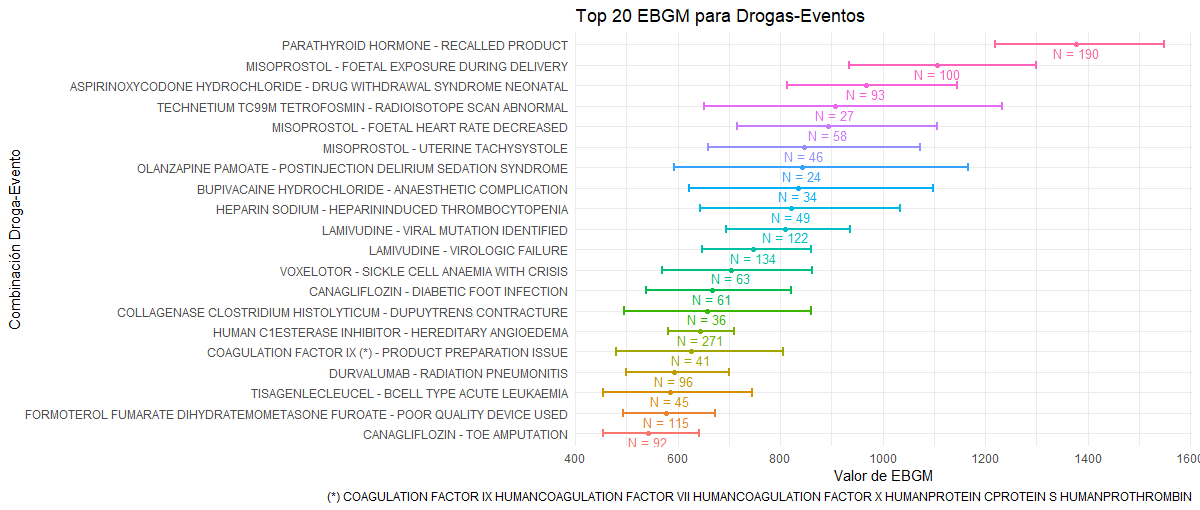


Figura 3

Al analizar los eventos adversos con mayor frecuencia, la muerte se encuentra primero con la mayor cantidad de drogas asociadas. En la Figura 4 se encuentran los demás 9 eventos adversos que completan el top 10, allí se puede encontrar el uso fuera de indicación (*off label use* y *product use in unapproved indication*), suicidio, toxicidad, interacción con otras drogas o diarrea, entre otras.

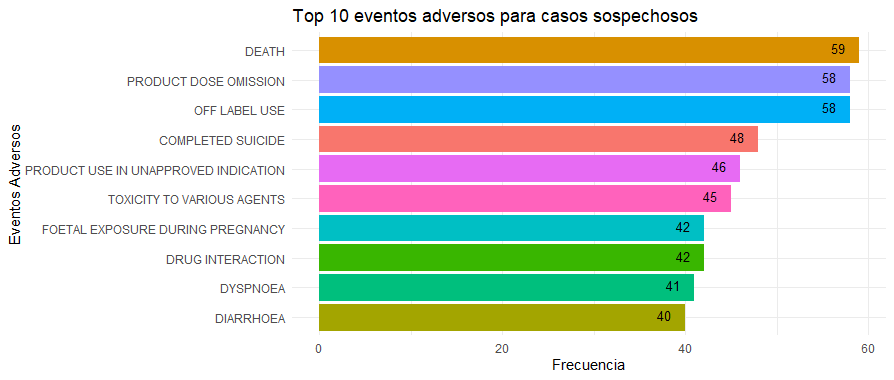


Figura 4

\*\*AGREGAR TABLA CON LOS PRIMEROS 1000\*