
Universidad de Valladolid
FACULTAD DE CIENCIAS

**TRABAJO FIN DE GRADO**

Grado en Estadística

# Evaluación de métodos para la imputación de valores no detectados en datos de expresión de citoquinas

Autor: Eduardo Herrero Velasco

Tutora: Itziar Fernández Martínez

# ÍNDICE

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# ACRÓNIMOS EMPLEADOS EN LA MEMORIA

ANOVA, Análisis de la varianza.

ECM, Error cuadrático medio.

EGF, Epidermal growth factor

FDR, F*alse Discovery Rate*

FWER, F*amily-Wise Error Rate*

IC, Intervalo de confianza.

IFNg, Interferon gamma.

IL1b, Interleukin 1 beta.

IL1RA, Interleukin 1 receptor antagonist.

IL2, IL4, IL5, IL6, IL8, IL9, IL10, IL13, IL17: Interleukin (nº).

IL12p70, Interleukin 12 (p70).

IP10, Interferon gamma induced protein 10.

KM, Kaplan-Meier.

LLOD, L*ower limit of detection.*

LOD, Limit of detection.

Log2, Transformación logaritmo en base 2.

MMP9, Matrix metallopeptidase 9

ND, No detectados.

RANTES, Regulated on activation normal T cell expressed and secreted.

RECM, Raíz del error cuadrático medio.

ROS, *Regression on Order Statistics.*

SDL, Sustitución por el límite de detección.

SDL2, Sustitución por el límite de detección/2.

S0, Sustitución por el valor 0.

TNFa, Tumour necrosis factor alpha.

VEGF, Vascular endothelial growth factor.

## RESUMEN

El análisis estadístico de datos inmunológicos puede resultar complicado dado que en ocasiones no se pueden determinar niveles cuantitativos precisos. En su lugar, es habitual tener que trabajar con valores que están por debajo de un límite de detección (valores no detectados). El objetivo de este trabajo es evaluar diferentes métodos para imputar valores no detectados en diferentes situaciones que aparecen frecuentemente en este contexto y cuya aplicación es importante para obtener resultados apropiados.

Se comparan cinco métodos: sustitución por un valor sencillo considerando 0, el límite de detección, y la mitad del límite de detección, extrapolación por el método ROS (*Regresion on Order Statistics)* y Kaplan-Meier. Para ello, se llevan a cabo, en primer lugar, un estudio de simulación que nos permite evaluar el comportamiento de los cinco métodos en diferentes escenarios. Posteriormente se analizan datos reales de expresión de citoquinas proporcionados por el Instituto de Oftalmología Aplicada (IOBA) de la Universidad de Valladolid (UVa). Con los resultados obtenidos en la simulación y en el análisis de datos reales, se puede concluir que, (i) los métodos de sustitución no son una buena elección, (ii) que el método de Kaplan-Meier funciona bien en el caso de que no exista expresión diferencial, y, (iii) que el método ROS es el que, en general, presenta los resultados más satisfactorios en diferentes escenarios, siendo una buena opción cuando el porcentaje de valores no detectados es menor que el 50%.



## ABSTRACT

The statistical analysis of immunological can be difficult because determining precise quantitative levels is occasionally not possible. In practice, however, it is usual to see observations whose values are under a detection limit (nondetects values). The purpose of this work is to evaluate different methods to impute nondetected values in different situations that appear frequently in this context and whose use is relevant to obtain proper conclusions.

Five methods are compared: single value substitution with 0, detection limit and half detection limit, extrapolation by ROS (Regression on Order Statistics) method and Kaplan-Meier. For that, first a simulation study is carried out to assess the five methods behaviors in different scenarios. Then, actual cytokines expression data provided by the Instituto de Oftalmología Aplicada (IOBA) from the Universidad de Valladolid (UVa) are analyzed. Finally, the outcome of these two analyzes allows us to conclude that (i) substitution methods are not a good option, (ii) the Kaplan-Meier method works well when there is no differential expression and (iii) the ROS method presents the most satisfactory results in different scenarios, being a good option when the non-detected value ratio is below 50%.

# 1. INTRODUCCIÓN

En esta sección se introduce, en primer lugar, el problema biológico que ha motivado este trabajo, y en segundo lugar, se hace una breve descripción de los métodos de imputación que serán comparados.

## 1.1 El problema biológico

### 1.1.1 ¿Qué es una citoquina?

Las citoquinas son proteínas de bajo peso molecular que tienen una gran importancia para controlar la actividad y el crecimiento de numerosas células del sistema inmunitario. Se encargan de la comunicación intercelular, es decir que cuando se liberan envían información de una célula a otra para que esta cumpla una determinada función. Son de gran interés, puesto que van a estar implicadas en numerosos procesos biológicos, además de tener un papel fundamental en la respuesta inflamatoria e inmune.

En condiciones normales, las citoquinas no se producen en cantidades importantes, siendo necesaria la activación de las células para que se produzcan en cantidades suficientes para ejercer sus efectos biológicos.

Básicamente, existen dos tipos de citoquinas: (i) proinflamatorias, encargadas de activar el sistema inmunológico y (ii)antiinflamatorias, que se encargan de comunicar a nuestro organismo que se ha eliminado una amenaza y hay que volver a la normalidad. Es importante que exista un equilibrio entre los dos tipos de citoquinas para que el sistema inmune funcione correctamente. Algunas citoquinas proinflamatorias son: IL-1, IL-6 o TNF, mientras que IL-4, IL-10 o TGF son ejemplos de citoquinas antiinflamatorias.

A raíz de la COVID-19, se ha puesto de moda el término “tormenta de citoquinas”, que consiste en una respuesta inflamatoria ante una infección que se descontrola, en la que las citoquinas antiinflamatorias no son capaces de contrarrestar la cantidad de citoquinas proinflamatorias, pudiendo dañar órganos de nuestro cuerpo como el pulmón, hígado e incluso el corazón. En este caso es más peligrosa la respuesta inmune de nuestro organismo contra una infección que la propia amenaza.

Las citoquinas son producidas por muchos tipos de células y sus niveles de expresión pueden medirse en diferentes tejidos, entre ellos la lágrima.

### 1.1.2 Las citoquinas en lágrima como biomarcadores

Los biomarcadores son características biológicas, fisiológicas o bioquímicas medibles, capaces de identificar procesos biológicos, patógenos o respuestas farmacológicas a una intervención terapéutica. Un buen biomarcador debe tener una alta capacidad de predicción, debe ser sensible y específico, no invasivo, fácil, económico y rápido de medir, además de aportar información relevante para la toma de decisiones.

Se ha demostrado que la lágrima es una fuente importante de material biológico, ya que contienen moléculas que incluyen proteínas, electrolitos y lípidos entre otros (Tamhane et al, 2019).

Por otra parte, los métodos utilizados para obtener lágrima son mínimamente invasivos, fáciles y económicos. Además, la proximidad de la lágrima a la superficie ocular, hace que sea muy útil como fuente de biomarcadores de enfermedades oculares cuyos síntomas se manifiestan en esta parte del ojo, como por ejemplo la enfermedad de ojo seco, el síndrome de Sjögren, alergias oculares o la enfermedad de injerto contra huésped ocular, entre otras (Enrique-de-Salamanca & Calonge, 2008). También se han utilizado en patologías extraoculares con afectación ocular, como la artritis reumatoide (Quingnon & Alfonso, 2009).

### 1.1.3 ¿Cómo se miden las citoquinas?

La expresión y actividad de las citoquinas se puede medir de forma cuantitativa a partir de muestras biológicas muy variadas: plasma, suero, lágrima, sobrenadantes de cultivo celular, etc.

Existen diferentes técnicas para cuantificar los niveles de citoquinas. Entre ellas, destacan los inmuno-ensayos, basados en el principio de que un determinado anticuerpo se unirá específicamente a las moléculas de interés. El más popular y de uso más frecuente en los laboratorios es el ensayo inmuno-enzimático ELISA, que permite cuantificar la concentración de una citoquina. Es un método muy fiable, aunque lento, necesita de muestras biológicas grandes y no es posible cuantificar distintas citoquinas simultáneamente. Una alternativa rápida y eficiente, cada vez más extendida, son los inmuno-ensayos multiplex, que, mediante técnicas basadas en citometría de flujo, permiten hacer una cuantificación masiva de citoquinas.

La tecnología que se ha utilizado para la obtención de los datos de este trabajo es de este último tipo. Concretamente se ha llevado a cabo con la tecnología Luminex® (Luminex Corporation, Austin, Texas, USA).

### 1.1.4 ¿Por qué se producen mediciones fuera de rango?

En general, cuando trabajamos con parámetros inmunológicos, en particular con niveles de citoquinas, es muy habitual que no todos los niveles cuantitativos se puedan medir con precisión. A veces la cantidad o concentración de la molécula es demasiado pequeña para que la tecnología sea capaz de discriminarla de un ruido de fondo. En estos casos, los valores que no se pueden cuantificar se llaman “no detectados” (ND).

Los valores ND estarán por debajo de un límite de detección (LOD, *limit of detection*) dependiente del instrumento de medición utilizado y diremos que son valores “por debajo del límite de detección” (LLOD*, lower limit of detection*).

### 1.1.5 Implicaciones de tener este tipo de medidas fuera de rango

Los valores ND son valores perdidos o *missing.* Concretamente, se trata de censuras por la izquierda, en las que el valor de nuestras observaciones se conoce parcialmente: están por debajo de un determinado umbral.

No tener en cuenta estas observaciones en los análisis estadísticos supondría introducir un sesgo importante en los resultados, ya que siempre eliminaríamos los valores pequeños. Es muy importante entender que, aunque una observación censurada por la izquierda no informa de un valor exacto, nos está aportando información valiosa que debe ser incorporada al tratamiento de los datos. Una forma de incorporarla será imputar un valor a estas observaciones.

Este problema suscita gran interés puesto que no es exclusivo del uso de citoquinas, sino que aparece en la cuantificación de otros parámetros inmunológicos así como en otros contextos, por ejemplo, en el tratamiento de datos medioambientales.

## 1.2 Métodos de imputación de valores no detectados

Elegir un método de imputación adecuado es muy importante y para ello hay que tener en cuenta los siguientes factores, tanto de forma individual como combinada:

- El tamaño de la muestra, el cual interesa que sea grande, ya que en muestras grandes, las estimaciones serán más precisas y se reducirá el riesgo de error.

- La distribución que siguen los datos, especialmente importante su simetría.

- El porcentaje de valores ND dentro de la muestra, siendo preferibles muestras con porcentajes pequeños, donde el riesgo de cometer errores será menor.

A continuación se describen tres aproximaciones diferentes utilizadas habitualmente para imputar los valores ND. Concretamente, el método de sustitución con tres variantes, el método ROS (*Regression on Order Statistics*) y método de Kaplan-Meier.

### 1.2.1 Método de sustitución

Es el método de imputación más simple y consiste en sustituir los valores ND por una constante. Las constantes más usadas son: 0, LOD/2 ó LOD.

A pesar de que es un método muy utilizado por su fácil implementación y porque no es necesario hacer ninguna suposición acerca de la distribución de los datos, es esperable que no de buenos resultados, puesto que al sustituir todos los valores *missing* por un mismo valor se modificará sustancialmente la información que proporcionan las observaciones no censuradas.

Es evidente que la validez de estos enfoques dependerá del porcentaje de valores desconocidos. En 2009 la USEPA *(United States Enviromental Protection Agency*) permitía el uso del método de sustitución para poder realizar estimaciones y contrastes de hipótesis siempre que la proporción de valores *missing* fuera menor que el 15%.

El error cometido en las inferencias que se llevan a cabo con los datos imputados, dependerá del valor constante que se utilice en la imputación. Así por ejemplo, cuando se sustituyen los valores *missing* por el valor 0, se introducirá un sesgo a la izquierda en la estimación de la media, mientras que cuando se sustituyen los valores *missing* por el LOD se introducirá un sesgo a la derecha, sobreestimando este parámetro. Si la sustitución se produce por el LOD/2 el sesgo sigue dependiendo del porcentaje de censuras, pero no es posible predecir su dirección (Grima - Olmedo et al, 2019).

### 1.2.2 Método ROS (Regresion on Order Statistics)

El método ROS, es un método semiparamétrico basado en el ajuste de un modelo de regresión lineal simple que utiliza los valores detectados para poder estimar la concentración de los valores ND (Helsel, 2010). El modelo ajustado se usa para generar estimaciones de las observaciones censuradas, que posteriormente se van a combinar con los valores detectados para poder estimar los parámetros que nos interesen. Se debe asumir que los datos siguen una distribución normal o lognormal (Sadegh 2008).

Una cuestión importante a tener en cuenta es que los valores imputados no tienen ningún sentido si no van acompañados del resto de valores (Mancin et al, 2017).

Para poder aplicar el método ROS se necesitan al menos 10-15 observaciones y una frecuencia de detección superior al 50%. Se recomienda usar este método cuando el conjunto de datos es grande (> 50) con menos del 50% de datos censurados. También es una buena opción con conjuntos de datos pequeños (< 50) y una proporción de datos censurados < 80% (Banta-Green, 2016).

La principal ventaja del método ROS es su robustez contra una posible especificación incorrecta de la distribución (Shoari & Dubé, 2018). Cuando el método ROS asume una distribución lognormal es una alternativa viable independientemente de la proporción de datos censurados (Quintanilla, 2017).

Los inconvenientes del método ROS son que se tratan las observaciones que se han imputado como si fueran observaciones reales, y que se subestima la varianza cuando el porcentaje de valores censurados es alto (Shoari & Dubé, 2018).

### 1.2.3 Método de Kaplan-Meier

El método de Kaplan-Meier es una técnica no paramétrica, por lo que no hace suposiciones de la distribución que siguen los datos, ni requiere una transformación de estos. Utiliza las posiciones relativas de cada observación dentro del conjunto de datos en lugar de imputar observación a observación como hace el método ROS, permitiendo de esta manera calcular estadísticos como la media o la varianza.

Este método realiza un conteo de datos por debajo de una concentración detectada y a partir de esa información estima una función de distribución. La función de distribución acumulada será una función escalonada creciente, en donde cada salto se corresponde con un valor sin censura. Para estimar estadísticos como la media o la varianza, se calcula el área bajo la función de distribución.

Para poder aplicar el método de Kaplan-Meier se necesita tener por lo menos tres observaciones detectadas y un valor detectado que sea mayor a todas las censuras. Es recomendable tener como mínimo entre 10-15 observaciones por encima del LOD, con un máximo de 60% de valores no detectados para obtener buenos resultados (Grima - Olmedo et al, 2019). Algunos autores recomiendan usar el método de Kaplan-Meier cuando la muestra contiene más de 50 observaciones, el porcentaje de datos censurados es <50% y los datos no siguen una única distribución.

Las ventajas que tiene el método de Kaplan-Meier son: su facilidad de cálculo, que los datos con los que trabaja pueden seguir varias distribuciones y su insensibilidad a valores atípicos (Shoari & Dubé, 2018).

El principal inconveniente de este método surge cuando no se detecta la observación más pequeña de los datos. Esto hace que la función de distribución acumulada permanezca de forma indefinida por debajo de dicha observación, lo que impide calcular el área por debajo de dicha función. Para solucionar este problema, se realiza la corrección del sesgo de Efron (Klein & Moeschberger, 2003). Esta corrección utiliza un valor como posible valor más bajo a detectar, permitiendo poder calcular estadísticos como la media (Helsel, 2010).

En este trabajo se va a utilizar R para imputar y comparar los distintos métodos evaluados, tanto con datos simulados como con datos reales. Concretamente, el método ROS y el método de Kaplan-Meier están implementados en el paquete NADA (Lee, 2020).

La siguiente tabla (Tabla 1), resume las principales características de los tres métodos con los que se va a trabajar.

| Método | Paquete de R | Tamaño de muestra recomendado | Porcentaje de datos censurados recomendado | Características |
|---|---|---|---|---|
| **Sustitución (0, LOD/2 ó LOD)** | Ningún paquete | n > 50 | < 15% | ·No es recomendable si se pueden aplicar técnicas más específicas.<br>·Si se usa 0 como valor de sustitución, se introduce sesgo a la izquierda en la estimación de la media.<br>· Si se usa el LOD como valor de sustitución, se sobreestima la media.<br>·Si se usa LOD/2, el sesgo depende del porcentaje de censuras. |
| **ROS** | Paquete NADA función ros() | n > 50<br>n < 50 | < 50%<br>< 80% | · Robusto contra la especificación incorrecta de la distribución y la variabilidad en la asimetría de los datos.<br>·Infraestima la varianza cuando está ante un % de valores ND alto. |
| **Kaplan-Meier** | Paquete NADA función cenfit() | n > 50 | < 50% | ·No es necesario asunciones acerca de la distribución de los datos.<br>·No tiene que realizar ninguna transformación a los datos.<br>· Robusto ante la presencia de valores atípicos.<br>· Tiene problemas cuando no se detecta la observación más pequeña de los datos. |

*Tabla 1: Tabla resumen de los diferentes métodos de imputación de valores ND que se comparan en este trabajo.*

# 2. OBJETIVOS

El principal objetivo del presente trabajo es estudiar el efecto que tiene la presencia de valores ND y la forma de imputar los mismos en el análisis de la expresión diferencial de los niveles de citoquina. Para ello se han desarrollado los siguientes objetivos complementarios:

- Comparación, a través de un estudio de simulación, de cinco métodos de imputación frecuentemente utilizados en este contexto. Se consideran distintos escenarios en los que se varía el tamaño muestral, la proporción de valores ND y el tamaño del efecto a detectar.

- Comparación de los cinco métodos evaluados utilizando un conjunto de datos reales en el que se recoge la expresión de 21 citoquinas en muestras de lágrima de hasta 308 pacientes con enfermedad de ojo seco clasificados en cuatro grupos de severidad.

# 3. RESULTADOS

## 3.1 Estudio de simulación

El propósito de este estudio de simulación es evaluar el efecto de los métodos de imputación evaluados a la hora de detectar la diferencia de expresión entre dos grupos. En adelante, se identificará a las distintas aproximaciones utilizando la siguiente notación: sustitución por 0 (S0), sustitución por LOD/2 (SDL2), sustitución por el LOD (SDL), método ROS (ROS) y método de Kaplan Meier (KM).

Para llevar a cabo la simulación, se generan conjuntos de datos con tamaños muestrales por grupo de 20, 50 y 100 individuos. Para cada tamaño muestral, se consideran cuatro escenarios atendiendo al porcentaje de no detección, concretamente se consideran el 10%, 30%, 50% y 70% de valores ND. Además, en términos de expresión diferencial se consideran cuatro situaciones: los dos grupos igualmente expresados, la expresión en uno de los grupos es 2, 4 y 8 veces mayor que en el otro. Para cada escenario se generan 1000 conjuntos de datos.

Se generan los datos a partir de una distribución normal, con una media en el grupo de referencia de 8 y desviación típica común a los dos grupos de 1, similar a lo observado en datos reales (en escala logarítmica).

### 3.1.1 Medidas de evaluación del estudio de simulación

Se denota por $B$ el número de réplicas llevadas a cabo para cada uno de los escenarios de simulación propuestos. Como medidas de evaluación se consideran las siguientes.

1. **Sesgo en la estimación de la diferencia de medias**. Sea $\delta$ la diferencia de medias entre los dos grupos y $\hat{\delta}^b$ la diferencia de medias estimada en la réplica $b$, $b = 1, \dots, B$, la estimación de Monte-Carlo para el sesgo está definida como:

$$\widehat{sesgo} = \frac{1}{B}\sum_{b=1}^{B} \hat{\delta}^b - \delta$$

2. **Probabilidad de cobertura del intervalo de confianza del 95% para la diferencia de medias**. Asumiendo varianzas iguales, el intervalo de confianza de $(1-\alpha)\cdot 100\%$ para la diferencia de medias puede construirse como:

$$\delta \in \left( \hat{\delta} \pm t_{2(n-1);1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{2}{n}} \right)$$

donde $n$ es el tamaño de cada grupo, $t_{2(n-1);1-\frac{\alpha}{2}}$ es el percentil $1-\frac{\alpha}{2}$ de una t-Student con $2(n-1)$ grados de libertad y

$$S = \sqrt{\frac{(n-1)S_1^2 + (n-1)S_2^2}{2(n-1)}} = \sqrt{\frac{S_1^2 + S_2^2}{2}}$$

es una estimación de la desviación típica común a ambos grupos obtenida a partir de las varianzas de las dos muestras, $S_1^2$ y $S_2^2$.

Se define el estimador de Monte-Carlo de la probabilidad de cobertura de este intervalo como:

$$\hat{p}_{IC} = \frac{1}{B}\sum_{b=1}^{B} I\left[ \hat{\delta}^b - t_{2(n-1);1-\frac{\alpha}{2}} \cdot S^b \cdot \sqrt{\frac{2}{n}} \leq \delta \leq \hat{\delta}^b + t_{2(n-1);1-\frac{\alpha}{2}} \cdot t_{2(n-1);1-\frac{\alpha}{2}} \cdot S^b \cdot \sqrt{\frac{2}{n}} \right]$$

donde $I[\cdot]$ representa la función indicador con valor 1 si su argumento es verdadero y 0 en caso contrario.

3. **Potencia estadística.** En cada réplica se considera el contraste t-Student para dos muestras independientes basado en el estadístico test,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{2}{n}}} \xrightarrow[H_o]{} t_{2(n-1)}$$

donde $H_o: \mu_1 - \mu_2$. El estimador de Monte-Carlo de la potencia estadística de este contraste para detectar una diferencia significativa se calcula como:

$$\hat{\beta} = \frac{1}{B}\sum_{b=1}^{B} I\left[|\hat{t}^b| > t_{2(n-1);1-\frac{\alpha}{2}}\right]$$

donde $\hat{t}^b$ representa el valor del estadístico test en la réplica $b$. Sólo se estima en aquellos escenarios en donde existe expresión diferencial.

En todos los casos, se calcula el **error del estimador de Monte-Carlo** como la desviación estándar del estimador de Monte-Carlo correspondiente. Además, como medida global se calcula el **error cuadrático medio** (ECM) definido como:

$$ECM = Var\left(\frac{1}{B}\sum_{b=1}^{B}\hat{\delta}^b\right) + \widehat{sesgo}^2$$

## 3.1.2 Los dos grupos igualmente expresados

En la tabla siguiente (Tabla 2) se muestran los resultados obtenidos en los diferentes escenarios de simulación cuando no hay expresión diferencial entre los dos grupos.

Además de los resultados con cada uno de los métodos de imputación, se muestran los resultados con los datos completos, es decir con un porcentaje del 0% de valores ND para que sirva de referencia. Además, se representa gráficamente el ECM de cada simulación.

| n | % de ND | Medida | S0 | SDL2 | SDL | ROS | K-M | Completos |
|---|---|---|---|---|---|---|---|---|
| 20 por grupo | 10% | $\widehat{sesgo}$ | 0.0213±0.8156 | 0.0145±0.5199 | 0.0078±0.2851 | -0.0191±0.2969 | -0.0082±0.2886 | 0.0103±0.3101 |
| | | $\hat{p}_{IC}$ | 0.946±0.2261 | 0.95±0.2181 | 0.955±0.2074 | 0.958±0.2007 | 0.948±0.2221 | 0.954±0.2096 |
| | | $ECM$ | 0.6656 | 0.2705 | 0.0813 | 0.0885 | 0.0834 | 0.0962 |
| | 30% | $\widehat{sesgo}$ | -0.0349±1.2843 | -0.0194±0.7344 | -0.0039±0.2404 | -0.1655±0.2744 | 0.005±0.2486 | -0.0128±0.3174 |
| | | $\hat{p}_{IC}$ | 0.93±0.2553 | 0.94±0.2376 | 0.949±0.2201 | 0.915±0.279 | 0.929±0.257 | 0.947±0.2241 |
| | | $ECM$ | 1.6507 | 0.5397 | 0.0578 | 0.1027 | 0.0618 | 0.1009 |
| | 50% | $\widehat{sesgo}$ | -0.0687±1.4059 | -0.0381±0.7731 | -0.0075±0.1856 | -0.3478±0.2436 | 0.0086±0.2062 | -0.0097±0.3139 |
| | | $\hat{p}_{IC}$ | 0.958±0.2007 | 0.954±0.2096 | 0.951±0.216 | 0.747±0.4349 | 0.911±0.2849 | 0.956±0.2052 |
| | | $ECM$ | 1.9813 | 0.5992 | 0.0345 | 0.1803 | 0.0426 | 0.0986 |
| | 70% | $\widehat{sesgo}$ | 0.0453±1.333 | 0.0251±0.7159 | 0.0049±0.1266 | -0.53±0.2476 | -0.0083±0.1901 | 0.0071±0.3104 |
| | | $\hat{p}_{IC}$ | 0.916±0.2775 | 0.925±0.2635 | 0.962±0.1913 | 0.338±0.4733 | 0.7898±0.4077 | 0.958±0.2007 |
| | | $ECM$ | 1.7791 | 0.5132 | 0.016 | 0.3422 | 0.0362 | 0.0964 |
| 50 por grupo | 10% | $\widehat{sesgo}$ | 0.0136±0.5254 | 0.0105±0.3306 | 0.0073±0.1759 | -0.03±0.1826 | -0.0084±0.1771 | 0.0081±0.194 |
| | | $\hat{p}_{IC}$ | 0.953±0.2117 | 0.955±0.2074 | 0.958±0.2007 | 0.958±0.2007 | 0.955±0.2074 | 0.953±0.2117 |
| | | $ECM$ | 0.2762 | 0.1094 | 0.031 | 0.0342 | 0.0314 | 0.0377 |
| | 30% | $\widehat{sesgo}$ | -0.0297±0.7761 | -0.0151±0.4423 | -6e-04±0.1461 | -0.1779±0.1659 | -5e-04±0.1478 | -0.004±0.1967 |
| | | $\hat{p}_{IC}$ | 0.951±0.216 | 0.956±0.2052 | 0.96±0.1961 | 0.843±0.364 | 0.95±0.2181 | 0.958±0.2007 |
| | | $ECM$ | 0.6032 | 0.1959 | 0.0213 | 0.0592 | 0.0219 | 0.0387 |
| | 50% | $\widehat{sesgo}$ | -0.0113±0.8674 | -0.0078±0.4762 | -0.0044±0.112 | -0.3612±0.1537 | 0.0061±0.1169 | -0.0055±0.1935 |
| | | $\hat{p}_{IC}$ | 0.948±0.2221 | 0.949±0.2201 | 0.955±0.2074 | 0.414±0.4928 | 0.941±0.2357 | 0.956±0.2052 |
| | | $ECM$ | 0.7525 | 0.2268 | 0.0126 | 0.1541 | 0.0137 | 0.0375 |
| | 70% | $\widehat{sesgo}$ | 0.0224±0.8523 | 0.0112±0.4587 | 0±0.0836 | -0.5588±0.1747 | -7e-04±0.099 | 0.0089±0.2015 |
| | | $\hat{p}_{IC}$ | 0.944±0.23 | 0.948±0.2221 | 0.941±0.2357 | 0.023±0.15 | 0.897±0.3041 | 0.946±0.2261 |
| | | $ECM$ | 0.7269 | 0.2105 | 0.007 | 0.3428 | 0.0098 | 0.0407 |
| 100 por grupo | 10% | $\widehat{sesgo}$ | 0.0117±0.3706 | 0.0096±0.2376 | 0.0075±0.1314 | -0.0342±0.1355 | -0.0077±0.1317 | 0.0086±0.1443 |
| | | $\hat{p}_{IC}$ | 0.942±0.2339 | 0.948±0.2221 | 0.949±0.2201 | 0.941±0.2357 | 0.947±0.2241 | 0.948±0.2221 |
| | | $ECM$ | 0.1375 | 0.0565 | 0.0173 | 0.0195 | 0.0174 | 0.0209 |
| | 30% | $\widehat{sesgo}$ | 0.0312±0.534 | 0.0204±0.304 | 0.0096±0.1006 | -0.1709±0.1148 | -0.0095±0.1009 | 0.0128±0.1317 |
| | | $\hat{p}_{IC}$ | 0.965±0.1839 | 0.965±0.1839 | 0.962±0.1913 | 0.745±0.4361 | 0.963±0.1889 | 0.962±0.1913 |
| | | $ECM$ | 0.2861 | 0.0928 | 0.0102 | 0.0424 | 0.0103 | 0.0175 |
| | 50% | $\widehat{sesgo}$ | 0.0189±0.6105 | 0.0105±0.3364 | 0.0021±0.081 | -0.3565±0.1094 | -0.002±0.0829 | 0.0056±0.1377 |
| | | $\hat{p}_{IC}$ | 0.95±0.2181 | 0.949±0.2201 | 0.951±0.216 | 0.121±0.3263 | 0.946±0.2261 | 0.951±0.216 |
| | | $ECM$ | 0.3731 | 0.1133 | 0.0066 | 0.1391 | 0.0069 | 0.019 |
| | 70% | $\widehat{sesgo}$ | 0.0117±0.6036 | 0.0065±0.3238 | 0.0012±0.0564 | -0.5722±0.1354 | -0.0018±0.0643 | 0.0026±0.1436 |
| | | $\hat{p}_{IC}$ | 0.947±0.2241 | 0.945±0.2281 | 0.965±0.1839 | 0±0 | 0.925±0.2635 | 0.943±0.232 |
| | | $ECM$ | 0.3644 | 0.1049 | 0.0032 | 0.3458 | 0.0041 | 0.0206 |

*Tabla 2: Resultados obtenidos para los datos simulados con dos grupos igualmente expresados.*

Fijándonos en el ECM representado en la Figura 1, se observa que el peor método para cualquier tamaño y porcentaje de ND es el de sustitución por el valor 0, con ECM muy grandes.

El método ROS tiende a infra-estimar la diferencia de medias cuando aumenta el porcentaje de ND, viéndose muy afectada la probabilidad de cobertura, que sufre un acelerado descenso a medida que aumenta este porcentaje. Esto se traduciría en un mayor número de falsos positivos en el contraste de comparación de medias, especialmente con un porcentaje de ND superior a 50.

Los dos mejores métodos en este escenario con dos grupos igualmente expresados son Kaplan-Meier y el método de sustitución por el LOD, con resultados muy similares.

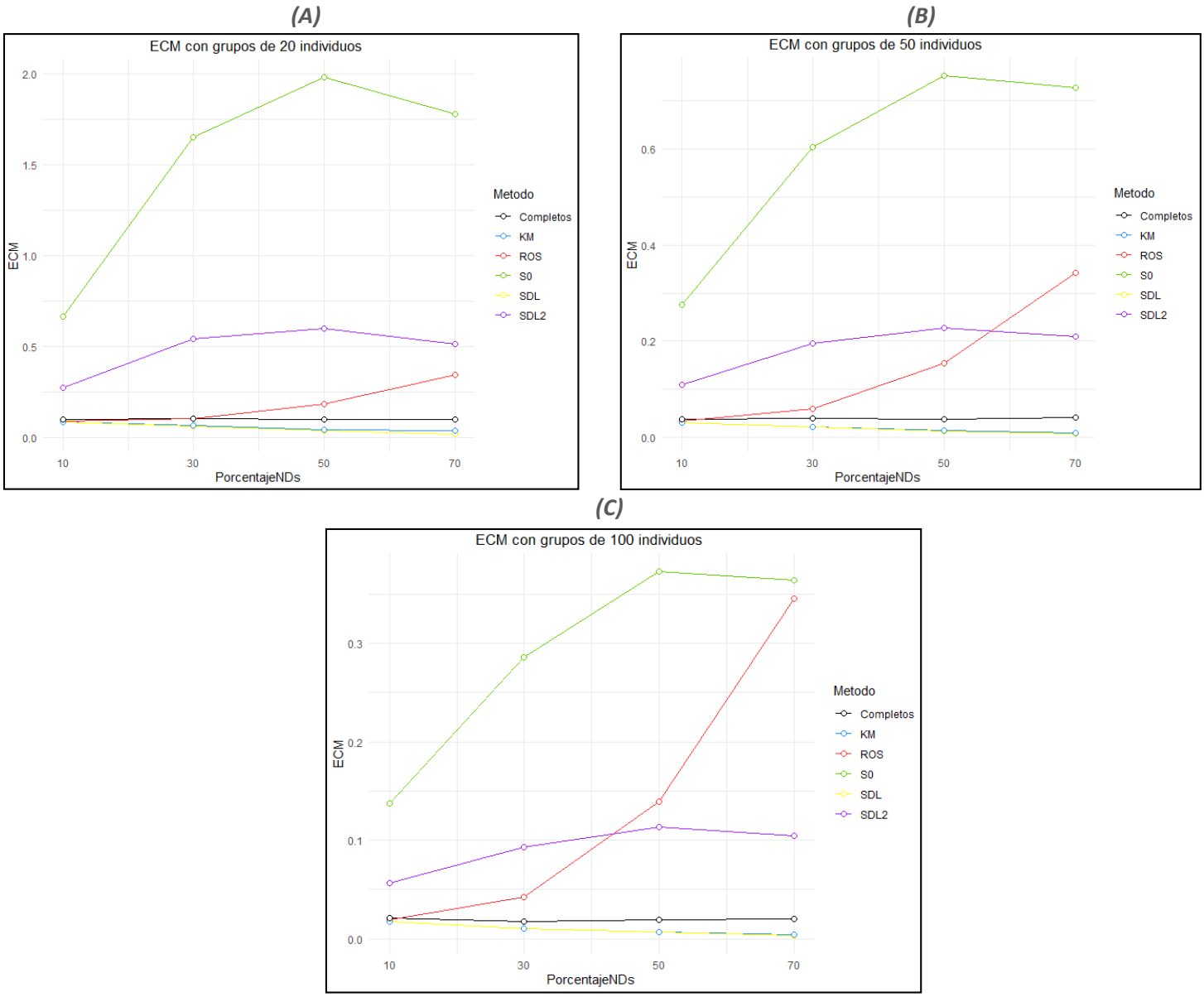



*Figura 1: ECM para los datos simulados con dos grupos igualmente expresados y tamaño muestral de (A) 20 individuos por grupo; (B) 50 individuos por grupo y (C) 100 individuos por grupo.*

### 3.1.3 Doble expresión en uno de los grupos

Se simula la situación en la que uno de los grupos tiene el doble de expresión que el otro.

| n | % de ND | Medida | S0 | SDL2 | SDL | ROS | K-M | Completos |
|---|---|---|---|---|---|---|---|---|
| 20 por grupo | 10% | $\widehat{sesgo}$ | 0.9776±0.6516 | 0.4501±0.451 | -0.0774±0.3067 | -0.0381±0.3261 | -1.9231±0.3092 | 0.0057±0.3209 |
| | | $\hat{p}_{IC}$ | 0.877±0.3286 | 0.917±0.276 | 0.923±0.2667 | 0.926±0.2619 | 0.917±0.276 | 0.942±0.2339 |
| | | $\hat{\beta}$ | 0.656±0.4753 | 0.79±0.4075 | 0.87±0.3365 | 0.855±0.3523 | 0.871±0.3354 | 0.874±0.332 |
| | | $ECM$ | 1.3803 | 0.406 | 0.1001 | 0.1078 | 3.7939 | 0.103 |
| | 30% | $\widehat{sesgo}$ | 2.2982±1.1576 | 1.0026±0.6689 | -0.293±0.2486 | -0.2512±0.3253 | -1.6967±0.2606 | -0.002±0.3119 |
| | | $\hat{p}_{IC}$ | 0.552±0.4975 | 0.701±0.458 | 0.778±0.4158 | 0.859±0.3482 | 0.734±0.4421 | 0.953±0.2117 |
| | | $\hat{\beta}$ | 0.73±0.4442 | 0.767±0.423 | 0.83±0.3758 | 0.641±0.4799 | 0.824±0.381 | 0.87±0.3365 |
| | | $ECM$ | 6.6219 | 1.4527 | 0.1476 | 0.1689 | 2.9468 | 0.0973 |
| | 50% | $\widehat{sesgo}$ | 2.8111±1.404 | 1.1569±0.7711 | -0.4973±0.192 | -0.5682±0.3119 | -1.4535±0.2248 | 0.0053±0.3183 |
| | | $\hat{p}_{IC}$ | 0.449±0.4976 | 0.674±0.469 | 0.294±0.4558 | 0.579±0.494 | 0.2345±0.4239 | 0.946±0.2261 |
| | | $\hat{\beta}$ | 0.761±0.4267 | 0.78±0.4145 | 0.746±0.4355 | 0.218±0.4131 | 0.6563±0.4752 | 0.872±0.3343 |
| | | $ECM$ | 9.8737 | 1.933 | 0.2842 | 0.4201 | 2.1631 | 0.1013 |
| | 70% | $\widehat{sesgo}$ | 2.3501±1.2175 | 0.8208±0.6519 | -0.7086±0.1266 | -1.0557±0.2308 | -1.4516±1.7183 | 0.0107±0.3087 |
| | | $\hat{p}_{IC}$ | 0.523±0.4997 | 0.822±0.3827 | 0.011±0.1044 | 0.11±0.313 | 0.0094±0.0963 | 0.958±0.2007 |
| | | $\hat{\beta}$ | 0.752±0.4321 | 0.739±0.4394 | 0.588±0.4924 | 0.006±0.0773 | 0.4468±0.4975 | 0.879±0.3263 |
| | | $ECM$ | 7.0054 | 1.0987 | 0.5182 | 1.1677 | 5.0598 | 0.0954 |
| 50 por grupo | 10% | $\widehat{sesgo}$ | 0.9495±0.4151 | 0.4301±0.2868 | -0.0894±0.1969 | -0.0653±0.2102 | -1.9121±0.198 | -0.0065±0.2044 |
| | | $\hat{p}_{IC}$ | 0.549±0.4978 | 0.759±0.4279 | 0.897±0.3041 | 0.915±0.279 | 0.895±0.3067 | 0.946±0.2261 |
| | | $\hat{\beta}$ | 0.971±0.1679 | 0.992±0.0891 | 0.997±0.0547 | 0.995±0.0706 | 0.997±0.0547 | 0.999±0.0316 |
| | | $ECM$ | 1.0739 | 0.2672 | 0.0468 | 0.0484 | 3.6952 | 0.0418 |
| | 30% | $\widehat{sesgo}$ | 2.2552±0.7477 | 0.9795±0.4302 | -0.2962±0.157 | -0.2896±0.2041 | -1.6985±0.1586 | -0.0127±0.1965 |
| | | $\hat{p}_{IC}$ | 0.195±0.3964 | 0.421±0.494 | 0.498±0.5002 | 0.671±0.4701 | 0.472±0.4995 | 0.952±0.2139 |
| | | $\hat{\beta}$ | 0.979±0.1435 | 0.991±0.0945 | 0.993±0.0834 | 0.951±0.216 | 0.994±0.0773 | 0.999±0.0316 |
| | | $ECM$ | 5.6449 | 1.1445 | 0.1124 | 0.1256 | 2.9101 | 0.0388 |
| | 50% | $\widehat{sesgo}$ | 2.7935±0.8609 | 1.1464±0.4748 | -0.5007±0.1243 | -0.6112±0.185 | -1.4827±0.1348 | 0.0037±0.1994 |
| | | $\hat{p}_{IC}$ | 0.126±0.332 | 0.358±0.4797 | 0.028±0.1651 | 0.14±0.3472 | 0.024±0.1531 | 0.951±0.216 |
| | | $\hat{\beta}$ | 0.984±0.1255 | 0.987±0.1133 | 0.983±0.1293 | 0.488±0.5001 | 0.965±0.1839 | 0.999±0.0316 |
| | | $ECM$ | 8.5446 | 1.5396 | 0.2662 | 0.4078 | 2.2165 | 0.0398 |
| | 70% | $\widehat{sesgo}$ | 2.2233±0.7793 | 0.753±0.4199 | -0.7173±0.0846 | -1.1003±0.1434 | -1.2406±0.3268 | -0.0058±0.2011 |
| | | $\hat{p}_{IC}$ | 0.215±0.411 | 0.619±0.4859 | 0±0 | 0±0 | 0.001±0.0317 | 0.951±0.216 |
| | | $\hat{\beta}$ | 0.966±0.1813 | 0.974±0.1592 | 0.946±0.2261 | 0.022±0.1468 | 0.7685±0.422 | 0.999±0.0316 |
| | | $ECM$ | 5.5502 | 0.7433 | 0.5216 | 1.2313 | 1.6459 | 0.0405 |
| 100 por grupo | 10% | $\widehat{sesgo}$ | 0.9603±0.2875 | 0.44±0.1992 | -0.0803±0.1371 | -0.0644±0.1465 | -1.9205±0.1372 | 0.0032±0.1433 |
| | | $\hat{p}_{IC}$ | 0.223±0.4165 | 0.519±0.4999 | 0.892±0.3105 | 0.904±0.2947 | 0.892±0.3105 | 0.944±0.23 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.0049 | 0.2333 | 0.0252 | 0.0256 | 3.707 | 0.0205 |
| | 30% | $\widehat{sesgo}$ | 2.3001±0.5147 | 1.0107±0.2974 | -0.2786±0.1102 | -0.2824±0.1397 | -1.7198±0.1115 | 0.008±0.1373 |
| | | $\hat{p}_{IC}$ | 0.016±0.1255 | 0.113±0.3168 | 0.274±0.4462 | 0.472±0.4995 | 0.26±0.4389 | 0.947±0.2241 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 0.999±0.0316 | 1±0 | 1±0 |
| | | $ECM$ | 5.5553 | 1.11 | 0.0898 | 0.0992 | 2.97 | 0.0189 |
| | 50% | $\widehat{sesgo}$ | 2.7407±0.6174 | 1.1191±0.341 | -0.5026±0.089 | -0.6274±0.1308 | -1.4881±0.091 | -0.0016±0.1419 |
| | | $\hat{p}_{IC}$ | 0.011±0.1044 | 0.098±0.2975 | 0.001±0.0316 | 0.009±0.0945 | 0.001±0.0316 | 0.945±0.2281 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 0.778±0.4158 | 1±0 | 1±0 |
| | | $ECM$ | 7.8926 | 1.3686 | 0.2605 | 0.4107 | 2.2228 | 0.0202 |
| | 70% | $\widehat{sesgo}$ | 2.2326±0.5511 | 0.7585±0.2955 | -0.7155±0.0573 | -1.1038±0.0979 | -1.2593±0.0691 | -0.0042±0.1423 |
| | | $\hat{p}_{IC}$ | 0.032±0.1761 | 0.361±0.4805 | 0±0 | 0±0 | 0±0 | 0.951±0.216 |
| | | $\hat{\beta}$ | 0.999±0.0316 | 1±0 | 1±0 | 0.052±0.2221 | 0.976±0.1531 | 1±0 |
| | | $ECM$ | 5.288 | 0.6627 | 0.5153 | 1.228 | 1.5907 | 0.0203 |

*Tabla 3: Resultados obtenidos para los datos simulados con doble expresión en uno de los grupos.*

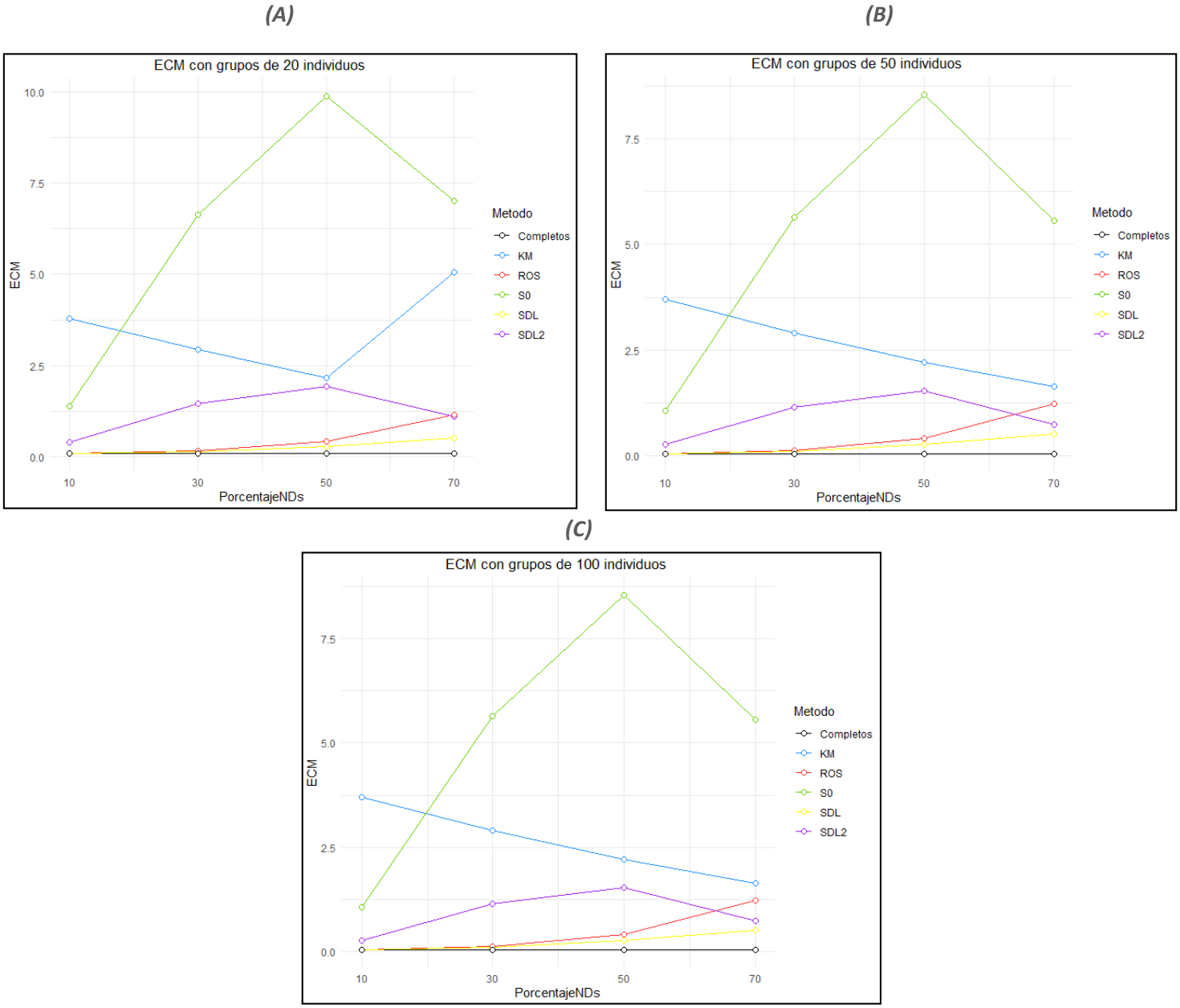



*Figura 2: ECM para los datos simulados con doble expresión en uno de los grupos y tamaño muestral de (A) 20 individuos por grupo; (B) 50 individuos por grupo y (C) 100 individuos por grupo.*

Observando la Tabla 3 se puede comprobar el aumento que se produce en el sesgo al usar cualquiera de los métodos respecto a la situación anterior, donde había dos grupos igualmente expresados. Los métodos de sustitución por 0 y Kaplan-Meier son los que mayor aumento sufren, siendo el sesgo de estos dos métodos muy superior al del resto de métodos, sobre todo cuando el porcentaje de ND es bajo.

Se puede ver en las tablas que cuando el porcentaje de ND aumenta, la probabilidad de cobertura del método ROS disminuye rápidamente.

Como cabía esperar, a medida que aumenta el tamaño se produce un aumento en la potencia estadística obtenida con cualquiera de los métodos de imputación.

Una diferencia importante que se puede apreciar en el gráfico del ECM es que los métodos que mejores resultados proporcionan cuando el porcentaje de ND es < 50% son el método de ROS y el método de sustitución por el LOD. Cuando el porcentaje de ND es alto (70%) y se trabaja con grupos grandes, el método de sustitución por el LOD/2 mejora mucho y llega a ser una buena opción junto a la sustitución por el LOD.

Con grupos más grandes, el método de Kaplan-Meier proporciona un ECM mucho menor al trabajar con un porcentaje de ND elevado (70%) que el que se obtenía con grupos de tamaño 20. Cuando se usa este método, la probabilidad de cobertura es muy pequeña, salvo con porcentajes de ND bajos y tamaños muestrales pequeños.

### 3.1.4 Cuatro veces más de expresión en uno de los grupos

Se simula la situación en la que uno de los grupos tiene cuatro veces más de expresión que el otro. De aquí en adelante en las figuras se representa $RECM = \sqrt{ECM}$, ya que existen ECM con valores muy elevados que no permiten apreciar bien las diferencias.

Los peores métodos son la sustitución por 0 y el método de Kaplan-Meier, ambos con un sesgo y un ECM grande. Cuando la tasa de no detección es del 70% Kaplan-Meier tiene un ECM muy alto, llegando a ser el peor método cuando se trabaja con grupos de tamaño 20 o 50. Destacar el gran descenso que se produce en el ECM con este método a medida que aumenta el tamaño de los grupos, incluso cuando el porcentaje de ND es alto (70%).

| n | % de ND | Medida | S0 | SDL2 | SDL | ROS | K-M | Completos |
|---|---|---|---|---|---|---|---|---|
| 20 por grupo | 10% | $\widehat{sesgo}$ | 1.2601±0.3823 | 0.5637±0.3342 | -0.1328±0.3127 | -0.0473±0.3227 | -3.866±0.3129 | -0.0186±0.3092 |
| | | $\hat{p}_{IC}$ | 0.883±0.3216 | 0.937±0.2431 | 0.902±0.2975 | 0.942±0.2339 | 0.901±0.2988 | 0.955±0.2074 |
| | | $\hat{\beta}$ | 0.996±0.0632 | 0.997±0.0547 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.7341 | 0.4294 | 0.1154 | 0.1063 | 15.0439 | 0.096 |
| | 30% | $\widehat{sesgo}$ | 3.8977±0.7468 | 1.7139±0.4715 | -0.47±0.2907 | -0.1997±0.3496 | -3.5125±0.3001 | 0.005±0.2988 |
| | | $\hat{p}_{IC}$ | 0.054±0.2261 | 0.224±0.4171 | 0.498±0.5002 | 0.872±0.3343 | 0.469±0.4993 | 0.969±0.1734 |
| | | $\hat{\beta}$ | 0.999±0.0316 | 1±0 | 1±0 | 0.999±0.0316 | 1±0 | 1±0 |
| | | $ECM$ | 15.7501 | 3.1597 | 0.3054 | 0.1621 | 12.4278 | 0.0893 |
| | 50% | $\widehat{sesgo}$ | 5.1424±1.1859 | 2.0696±0.6597 | -1.0031±0.2237 | -0.7502±0.3401 | -2.9174±0.732 | -0.0149±0.3173 |
| | | $\hat{p}_{IC}$ | 0.027±0.1622 | 0.144±0.3513 | 0.018±0.133 | 0.413±0.4926 | 0.0086±0.0924 | 0.941±0.2357 |
| | | $\hat{\beta}$ | 0.998±0.0447 | 0.998±0.0447 | 0.999±0.0316 | 0.964±0.1864 | 0.9645±0.1851 | 1±0 |
| | | $ECM$ | 27.8506 | 4.7186 | 1.0563 | 0.6784 | 9.0469 | 0.1009 |
| | 70% | $\widehat{sesgo}$ | 3.6911±0.8115 | 1.0787±0.4319 | -1.5337±0.1218 | -1.6679±0.1739 | -7.0728±5.112 | -0.0178±0.3198 |
| | | $\hat{p}_{IC}$ | 0.141±0.3482 | 0.576±0.4944 | 0±0 | 0.002±0.0447 | 0±0 | 0.949±0.2201 |
| | | $\hat{\beta}$ | 0.997±0.0547 | 0.997±0.0547 | 0.979±0.1435 | 0±0 | 0.9179±0.2748 | 1±0 |
| | | $ECM$ | 14.2828 | 1.3501 | 2.3671 | 2.8122 | 76.1567 | 0.1026 |
| 50 por grupo | 10% | $\widehat{sesgo}$ | 1.3032±0.2246 | 0.6003±0.2106 | -0.1025±0.2071 | -0.0295±0.2136 | -3.8969±0.2068 | 0.0072±0.2062 |
| | | $\hat{p}_{IC}$ | 0.111±0.3143 | 0.572±0.495 | 0.883±0.3216 | 0.922±0.2683 | 0.883±0.3216 | 0.933±0.2501 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.7488 | 0.4048 | 0.0534 | 0.0465 | 15.2282 | 0.0426 |
| | 30% | $\widehat{sesgo}$ | 3.8681±0.4939 | 1.693±0.3147 | -0.4821±0.188 | -0.2576±0.2283 | -3.5118±0.1901 | -0.0066±0.1985 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.009±0.0945 | 0.184±0.3877 | 0.692±0.4619 | 0.177±0.3819 | 0.962±0.1913 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 15.206 | 2.9652 | 0.2678 | 0.1185 | 12.3686 | 0.0395 |
| | 50% | $\widehat{sesgo}$ | 5.1962±0.6978 | 2.0987±0.3903 | -0.9987±0.1387 | -0.7928±0.1948 | -2.953±0.1585 | 0.0091±0.2027 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.003±0.0547 | 0±0 | 0.036±0.1864 | 0±0 | 0.949±0.2201 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 27.4871 | 4.557 | 1.0166 | 0.6665 | 8.7453 | 0.0412 |
| | 70% | $\widehat{sesgo}$ | 3.7094±0.5368 | 1.092±0.2881 | -1.5254±0.0825 | -1.6959±0.1101 | -4.0843±3.9223 | -0.0104±0.2035 |
| | | $\hat{p}_{IC}$ | 0.002±0.0447 | 0.231±0.4217 | 0±0 | 0±0 | 0±0 | 0.955±0.2074 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 0.04±0.1961 | 0.8539±0.3535 | 1±0 |
| | | $ECM$ | 14.0481 | 1.2755 | 2.3336 | 2.8884 | 32.0659 | 0.0415 |
| 100 por grupo | 10% | $\widehat{sesgo}$ | 1.2903±0.1604 | 0.5904±0.1502 | -0.1094±0.1475 | -0.0438±0.1526 | -3.891±0.1474 | 8e-04±0.1463 |
| | | $\hat{p}_{IC}$ | 0.003±0.0547 | 0.181±0.3852 | 0.833±0.3732 | 0.91±0.2863 | 0.833±0.3732 | 0.95±0.2181 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.6906 | 0.3712 | 0.0337 | 0.0252 | 15.1612 | 0.0214 |
| | 30% | $\widehat{sesgo}$ | 3.887±0.3421 | 1.7087±0.2181 | -0.4697±0.1324 | -0.2634±0.1576 | -3.5281±0.1333 | 0.0043±0.1355 |
| | | $\hat{p}_{IC}$ | 0±0 | 0±0 | 0.039±0.1937 | 0.509±0.5002 | 0.037±0.1889 | 0.956±0.2052 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 15.2259 | 2.9671 | 0.2381 | 0.0942 | 12.4651 | 0.0184 |
| | 50% | $\widehat{sesgo}$ | 5.1289±0.5197 | 2.0648±0.2895 | -0.9992±0.1032 | -0.8302±0.1409 | -2.9782±0.1096 | -0.0051±0.1396 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.002±0.0447 | 0±0 | 0±0 | 0±0 | 0.955±0.2074 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 26.5758 | 4.3474 | 1.0091 | 0.709 | 8.8815 | 0.0195 |
| | 70% | $\widehat{sesgo}$ | 3.7177±0.3701 | 1.0975±0.1983 | -1.5227±0.0584 | -1.7093±0.0732 | -2.5733±1.5471 | 0.0011±0.1392 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.021±0.1435 | 0±0 | 0±0 | 0±0 | 0.952±0.2139 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 0.459±0.4986 | 0.9341±0.2483 | 1±0 |
| | | $ECM$ | 13.9579 | 1.2438 | 2.322 | 2.9271 | 9.0155 | 0.0194 |

*Tabla 4: Resultados obtenidos para los datos simulados con cuatro veces más de expresión en uno de los grupos.*

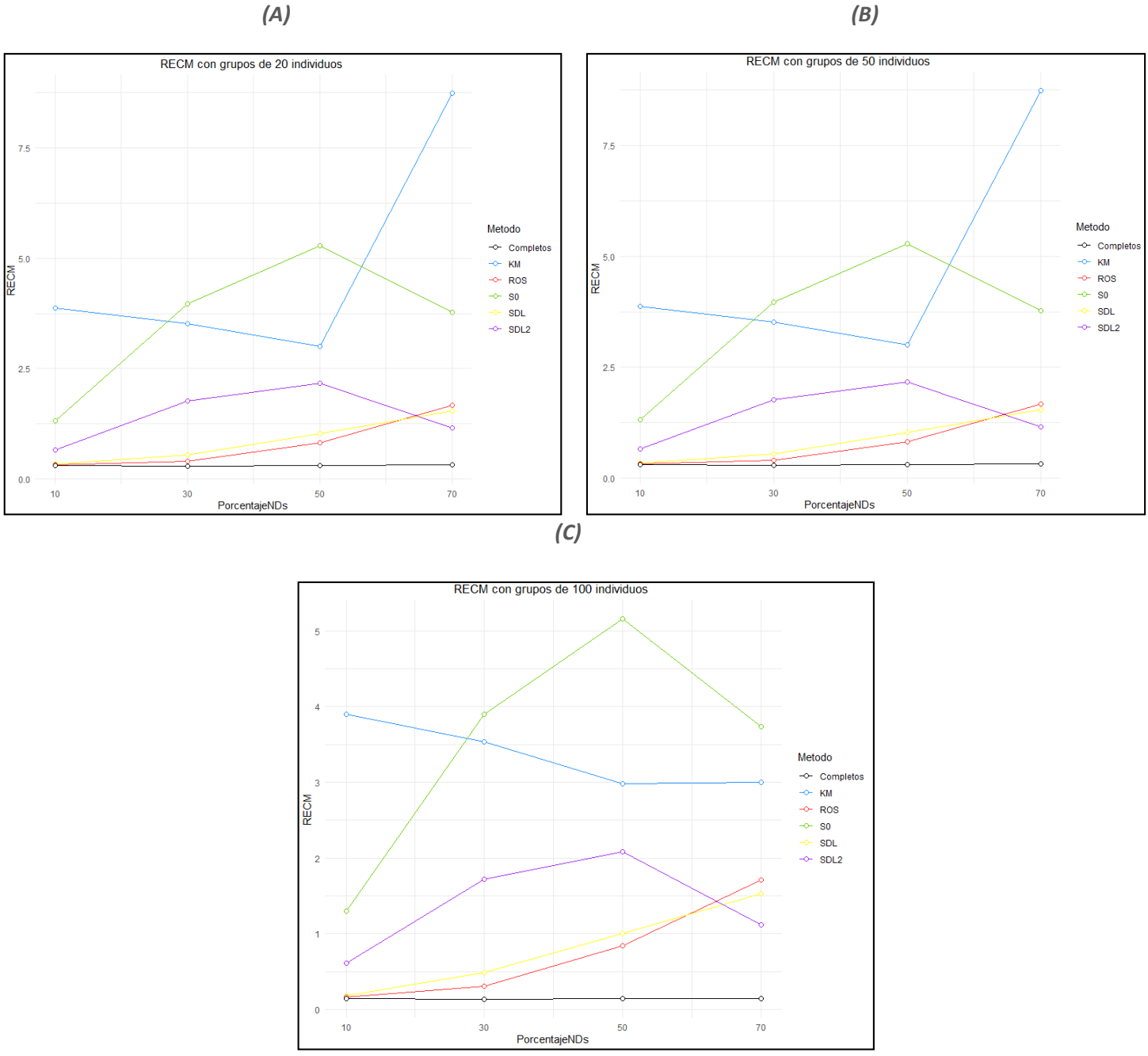



*Figura 3: ECM para los datos simulados con cuatro veces más de expresión en uno de los grupos y tamaño muestral de (A) 20 individuos por grupo; (B) 50 individuos por grupo y (C) 100 individuos por grupo.*

Todos los métodos tienen una potencia estadística muy elevada en los diferentes casos, a excepción del método ROS cuando el porcentaje de datos ND es del 70%. En este caso en el grupo de no expresión, prácticamente todos los valores serán ND y se imputan utilizando los datos del otro grupo, por lo que es esperable que la diferencia entre ellos sea menor, lo que afecta a la potencia.

La probabilidad de cobertura es baja en todos los métodos. Destaca como incluso para porcentajes de ND pequeños esta probabilidad es muy pequeña, siendo el método ROS el que mejor se comporta en estos casos.

Cuando observamos los gráficos de la Figura 3, se pueden observar varios cambios respecto a lo que se había observado hasta el momento. Ahora cuando el porcentaje de ND es < 50%, los dos mejores métodos siguen siendo los mismos, pero por primera vez el método ROS es ligeramente mejor que el de sustitución por el LOD. Por otro lado, cuando estamos ante un porcentaje alto (70%), el mejor resultado se obtiene con la sustitución por el LOD/2.

### 3.1.5 Ocho veces más de expresión en uno de los grupos.

Trabajando con una diferencia de expresión de ocho veces más, la probabilidad de cobertura es muy pequeña en todos los métodos, en algunos de ellos incluso con un porcentaje de ND muy bajo.

A medida que aumenta el tamaño de los grupos, el ECM de los métodos ROS y sustitución por el LOD disminuye cuando se trabaja con porcentajes de ND pequeños.

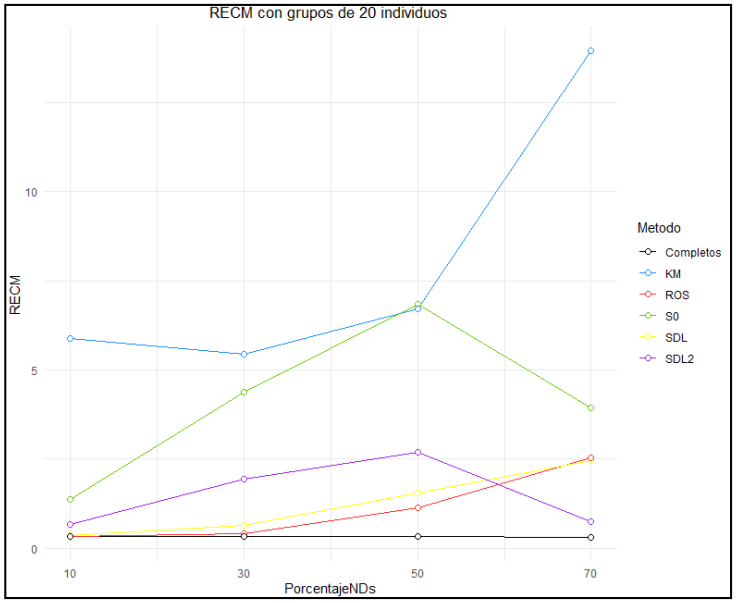
Los peores métodos siguen siendo Kaplan-Meier y la sustitución por 0, con unos sesgos y ECM muy altos. Se puede ver como los resultados obtenidos con el método de Kaplan-Meier han ido empeorando a medida que aumenta la diferencia de expresión entre grupos.

El mejor método es de nuevo ROS cuando el porcentaje de ND es < 50% y cuando es > 50% el método de sustitución por el LOD/2 es la mejor opción.

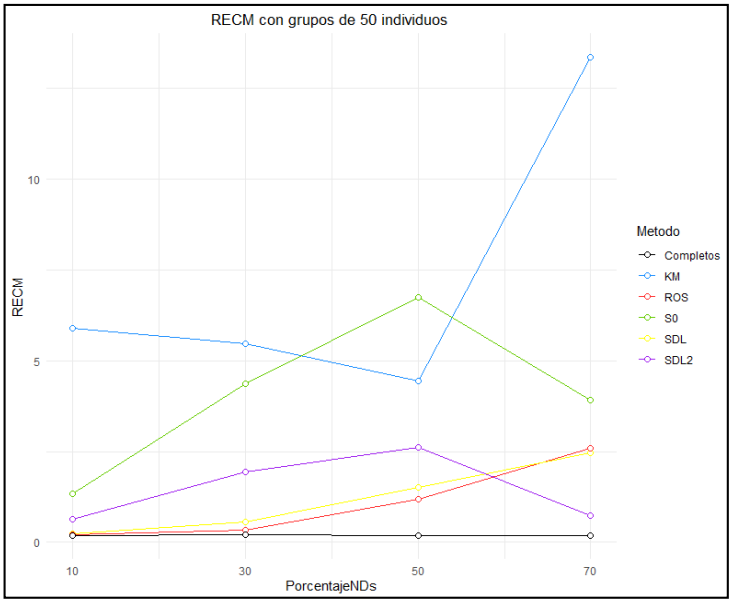
| n | % de ND | Medida | S0 | SDL2 | SDL | ROS | K-M | Completos |
|---|---|---|---|---|---|---|---|---|
| 20 por grupo | 10% | $\widehat{sesgo}$ | 1.3258±0.2917 | 0.6021±0.3057 | -0.1215±0.3219 | -1e-04±0.3311 | -5.8747±0.3218 | -0.0074±0.3169 |
| | | $\hat{p}_{IC}$ | 0.86±0.3472 | 0.922±0.2683 | 0.904±0.2947 | 0.939±0.2395 | 0.898±0.3028 | 0.954±0.2096 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.8428 | 0.456 | 0.1184 | 0.1097 | 34.6161 | 0.1005 |
| | 30% | $\widehat{sesgo}$ | 4.3582±0.3737 | 1.9054±0.3299 | -0.5473±0.3394 | -0.2048±0.3549 | -5.4306±0.3426 | -0.0048±0.3144 |
| | | $\hat{p}_{IC}$ | 0.002±0.0447 | 0.051±0.2201 | 0.43±0.4953 | 0.848±0.3592 | 0.388±0.4875 | 0.95±0.2181 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 19.1333 | 3.7395 | 0.4148 | 0.1679 | 29.6093 | 0.0989 |
| | 50% | $\widehat{sesgo}$ | 6.7799±0.8986 | 2.6363±0.521 | -1.5072±0.2933 | -1.0886±0.3326 | -5.7024±3.5649 | 0.0047±0.3255 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.012±0.1089 | 0.001±0.0316 | 0.126±0.332 | 0±0 | 0.957±0.203 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 0.9881±0.1085 | 1±0 |
| | | $ECM$ | 46.7747 | 7.2217 | 2.3578 | 1.2958 | 45.226 | 0.1059 |
| | 70% | $\widehat{sesgo}$ | 3.9163±0.311 | 0.726±0.1839 | -2.4643±0.1313 | -2.5339±0.1357 | -13.7174±2.5071 | 0.0015±0.3093 |
| | | $\hat{p}_{IC}$ | 0.052±0.2221 | 1±0 | 0±0 | 0±0 | 0±0 | 0.962±0.1913 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 0±0 | 1±0 | 1±0 |
| | | $ECM$ | 15.4341 | 0.5609 | 6.0901 | 6.4393 | 194.4536 | 0.0956 |
| 50 por grupo | 10% | $\widehat{sesgo}$ | 1.3315±0.1809 | 0.6132±0.1885 | -0.1051±0.1981 | -2e-04±0.2048 | -5.8936±0.1981 | 0.0087±0.1946 |
| | | $\hat{p}_{IC}$ | 0.05±0.2181 | 0.541±0.4986 | 0.884±0.3204 | 0.949±0.2201 | 0.883±0.3216 | 0.964±0.1864 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.8056 | 0.4115 | 0.0503 | 0.0419 | 34.7737 | 0.038 |
| | 30% | $\widehat{sesgo}$ | 4.367±0.2189 | 1.9171±0.2001 | -0.5328±0.2105 | -0.2473±0.2147 | -5.4595±0.2114 | 0.0025±0.2016 |
| | | $\hat{p}_{IC}$ | 0±0 | 0±0 | 0.137±0.344 | 0.691±0.4623 | 0.132±0.3387 | 0.949±0.2201 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 19.1189 | 3.7154 | 0.3282 | 0.1072 | 29.8507 | 0.0406 |
| | 50% | $\widehat{sesgo}$ | 6.7248±0.5271 | 2.6073±0.3114 | -1.5102±0.188 | -1.1606±0.2097 | -4.3968±0.5903 | -0.015±0.1937 |
| | | $\hat{p}_{IC}$ | 0±0 | 0±0 | 0±0 | 0.001±0.0316 | 0±0 | 0.957±0.203 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 0.9989±0.0325 | 1±0 |
| | | $ECM$ | 45.5006 | 6.895 | 2.316 | 1.391 | 19.6804 | 0.0377 |
| | 70% | $\widehat{sesgo}$ | 3.9194±0.2012 | 0.7244±0.1169 | -2.4705±0.0833 | -2.5795±0.0777 | -12.7816±3.8207 | 0.0102±0.1974 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.991±0.0945 | 0±0 | 0±0 | 0±0 | 0.955±0.2074 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 0.083±0.276 | 0.9977±0.0479 | 1±0 |
| | | $ECM$ | 15.4019 | 0.5385 | 6.1102 | 6.6596 | 177.9679 | 0.0391 |
| 100 por grupo | 10% | $\widehat{sesgo}$ | 1.3199±0.1338 | 0.6036±0.1382 | -0.1126±0.1443 | -0.0151±0.1493 | -5.8868±0.1442 | -8e-04±0.1436 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.133±0.3397 | 0.844±0.363 | 0.934±0.2484 | 0.843±0.364 | 0.94±0.2376 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 1.76 | 0.3835 | 0.0335 | 0.0225 | 34.6757 | 0.0206 |
| | 30% | $\widehat{sesgo}$ | 4.3681±0.161 | 1.9152±0.1489 | -0.5378±0.1558 | -0.2691±0.1604 | -5.4586±0.1559 | -0.004±0.1447 |
| | | $\hat{p}_{IC}$ | 0±0 | 0±0 | 0.011±0.1044 | 0.475±0.4996 | 0.011±0.1044 | 0.948±0.2221 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 19.1062 | 3.69 | 0.3135 | 0.0981 | 29.8202 | 0.0209 |
| | 50% | $\widehat{sesgo}$ | 6.7401±0.3964 | 2.6218±0.2337 | -1.4964±0.135 | -1.1911±0.1408 | -4.4482±0.1516 | -0.002±0.1457 |
| | | $\hat{p}_{IC}$ | 0±0 | 0±0 | 0±0 | 0±0 | 0±0 | 0.947±0.2241 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 | 1±0 |
| | | $ECM$ | 45.5855 | 6.9285 | 2.2575 | 1.4387 | 19.8096 | 0.0212 |
| | 70% | $\widehat{sesgo}$ | 3.9209±0.1418 | 0.726±0.0813 | -2.4688±0.0599 | -2.5954±0.0501 | -11.5327±4.7873 | -3e-04±0.1442 |
| | | $\hat{p}_{IC}$ | 0±0 | 0.126±0.332 | 0±0 | 0±0 | 0±0 | 0.943±0.232 |
| | | $\hat{\beta}$ | 1±0 | 1±0 | 1±0 | 1±0 | 0.9962±0.0616 | 1±0 |
| | | $ECM$ | 15.3933 | 0.5337 | 6.0985 | 6.7388 | 155.9218 | 0.0208 |

*Tabla 5: Resultados obtenidos para los datos simulados con ocho veces más de expresión en uno de los grupos.*

*(A)*




*(B)*




*(C)*

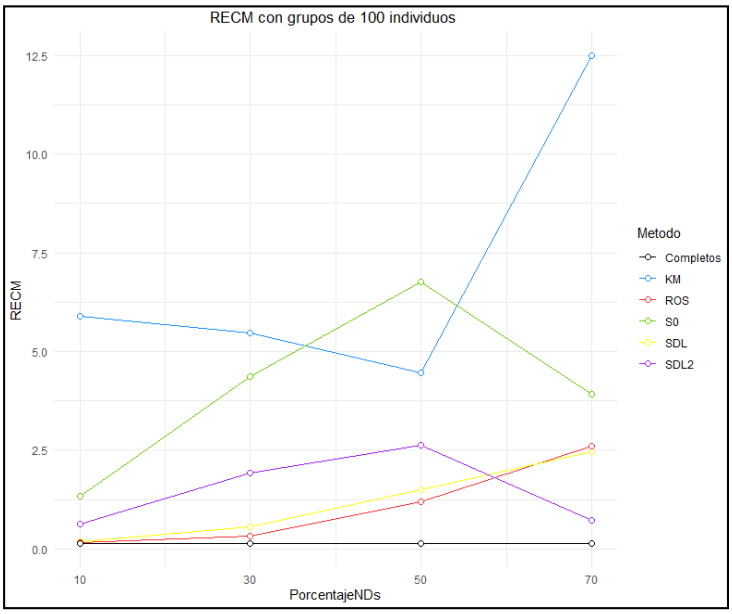



*Figura 4: ECM para los datos simulados con ocho veces más de expresión en uno de los grupos y tamaño muestral de (A) 20 individuos por grupo; (B) 50 individuos por grupo y (C) 100 individuos por grupo.*

## 3.2 Estudio con datos reales de citoquinas en lágrima

El conjunto de datos con el que vamos a trabajar contiene información de 308 individuos, a los que se les ha medido la concentración de hasta 21 citoquinas en lágrima. Se trata de una muestra de pacientes de la enfermedad de ojo seco, tratados y evaluados en el Instituto de Oftalmobiología Aplicada (IOBA) de la UVa. Se consideran cuatro grupos de individuos atendiendo a la severidad de la enfermedad (sano, leve, moderado y severo).

### 3.2.1 Citoquinas evaluadas

Hay mucho interés en el estudio de la concentración de citoquinas en lágrima como posibles biomarcadores de la enfermedad de ojo seco y, por ello, existe una gran cantidad de estudios al respecto. A pesar de que en muchos de estos estudios, las conclusiones son contradictorias, en la siguiente tabla (Tabla 6) se resumen las citoquinas analizadas en este trabajo junto con su comportamiento esperado para esta patología según un meta-análisis recientemente publicado (Roda et al, 2020).

Para tres citoquinas: Fractalkina, RANTES y MMP9, no se tiene información en (Roda et al, 2020). Sin embargo, MMP9 se considera sobre-expresada, ya que en la actualidad existe un dispositivo ampliamente utilizado en la práctica clínica para la identificación de pacientes de ojo seco llamado InflammaDry ® basado en los niveles de esta molécula. En el caso de VEGF, se han obtenido resultados contradictorios en los diferentes estudios incluidos en el meta-análisis. Las citoquinas etiquetadas como “posiblemente” muestran indicios de expresión diferencial, pero no han sido verificados con el meta-análisis.

| Acrónimo | Nombre completo | Resultados en la enfermedad de ojo seco |
|---|---|---|
| EGF | Epidermal growth factor | Posiblemente infra-expresada |
| Eotaxin | Eotaxin o CCL11 | Posiblemente sobre-expresada |
| Fractalkina | Fractalkine o CX3CL1 | Sin información |
| IFNg | Interferon gamma | Sobre-expresada |
| IL1b | Interleukin 1 beta | Sobre-expresada |
| IL1RA | Interleukin 1 receptor antagonist | Posiblemente sobre-expresada |
| IL2 | Interleukin 2 | Posiblemente sobre-expresada |
| IL4 | Interleukin 4 | Posiblemente sobre-expresada |
| IL5 | Interleukin 5 | Posiblemente sobre-expresada |
| IL6 | Interleukin 6 | Sobre-expresada |
| IL8 | Interleukin 8 | Sobre-expresada |
| IL9 | Interleukin 9 | Posiblemente sobre-expresada |
| IL10 | Interleukin 10 | Sobre-expresada |
| IL12p70 | Interleukin 12 (p70) | Posiblemente sobre-expresada |
| IL13 | Interleukin 13 | Posiblemente sobre-expresada |
| IL17 | Interleukin 17 | Posiblemente sobre-expresada |
| IP10 | Interferon gamma induced protein 10 o CXCL10 | Posiblemente infra-expresada |
| MMP9 | Matrix metallopeptidase 9 | Sobre-expresada |
| RANTES | Regulated on activation normal T cell expressed and secreted o CCL5 | Sin información |
| TNFa | Tumour necrosis factor alpha | Sobre-expresada |
| VEGF | Vascular endothelial growth factor | Contradicción entre diferentes estudios |

*Tabla 6: Expresión diferencial de las diferentes citoquinas analizadas para la enfermedad de ojo seco según los resultados obtenidos en el meta-ánalisis (Roda et al, 2020).*

### 3.2.2 Transformación de los niveles de concentración de citoquina

La distribución de los niveles de concentración de citoquina, al igual que la de otros parámetros inmunológicos, suele ser asimétrica a la derecha. Ocurre que la mayoría de las citoquinas se expresan a niveles bajos o muy bajos, cercanos a cero con unos pocos individuos con valores altos. Por eso, es muy habitual llevar a cabo una transformación logarítmica antes de realizar cualquier análisis de datos, ya que esta transformación corrige la asimetría de este tipo de distribuciones.

Entre todas las transformaciones logarítmicas posibles, una de las más populares en este ámbito es la transformación logaritmo en base 2 (log2), y la razón tiene que ver con facilitar la interpretación de los resultados.

En un experimento con niveles de expresión, para medir cambios en una molécula, es muy frecuente utilizar el fold-change (FC), definido como un cociente entre la concentración en una muestra objetivo, respecto a la concentración en una muestra tomada como referencia. Trabajar con cocientes tiene desventajas en cuanto a la comparación de moléculas sobre- e infra-expresadas. Un valor de FC en el rango $[1,\infty)$ indica sobre-expresión en la muestra objetivo respecto de la de referencia, y un valor de FC en el rango $[0,1]$ indica infra-expresión en la muestra objetivo respecto de la de referencia.

En la escala logarítmica, el FC se convierte en una diferencia, y un determinado valor y su recíproco serán simétricos. Valores positivos indicarán sobre-expresión, negativos, infra-expresión y una molécula expresada a nivel constante tendrá un log2(FC) de 0. La base 2 de la transformación implica que cambios de una unidad en la escala log2 doblan su valor en la escala original, lo que facilita interpretar los resultados.

Por este motivo previamente al análisis de los datos reales de citoquinas, se lleva a cabo una transformación log2.

### 3.2.3 Métodos estadísticos para evaluar el cambio de expresión entre grupos de enfermedad

Para evaluar el efecto de los grupos de severidad en los niveles de cada citoquina utilizamos el análisis de la varianza (ANOVA) de una vía, considerando

como variable respuesta la concentración de dicha citoquina y como factor el grupo, con cuatro niveles (Genser et al, 2007).

Las condiciones de aplicación de esta metodología incluyen:

- Independencia, que está garantizada por el diseño muestral.
- Normalidad, que se ha verificado en todos los casos usando el contraste de Shapiro-Wilk.
- Homogeneidad de varianza, para cuya evaluación se utiliza el test de Fligner. Esta condición es esencial en este caso, puesto que los grupos no están balanceados, por tanto, si no es posible asumirla se utiliza la corrección de Welch a la hora de realizar el ANOVA.

Para llevar a cabo las comparaciones post-hoc y determinar entre que pares de grupos se encuentran las diferencias estadísticamente significativas, se utiliza el contraste T-Student para dos muestras independientes con la corrección de Bonferroni para solucionar el problema de las comparaciones múltiples.

El análisis citoquina a citoquina requiere realizar un ANOVA por cada una de las moléculas evaluadas, y esto hace que vuelva a aparecer el problema de las comparaciones múltiples. A medida que aumentan el número de contrastes, la probabilidad de que se obtengan falsos positivos aumenta y esto debe ser corregido. Existen dos opciones para corregir este problema. La primera es controlar la FWER (Family-Wise Error Rate), definida como la probabilidad de tener uno o más falsos positivos en el conjunto de contrastes realizados. La segunda tiene que ver con controlar la FDR (False Discovery Rate) (Benjamini y Hochberg, 1995), definida como la proporción de falsos positivos entre todas las citoquinas inicialmente identificadas como diferencialmente expresadas. Los criterios basados en el control de FWER pueden ser muy restrictivos debido al gran número de contrastes realizados, por lo que en este contexto es más popular el control de la FDR. En este trabajo, se utilizará esta corrección para ajustar los p-valores obtenidos.

### 3.2.4 Resultados

En las Tabla 7 se muestra la tasa de valores ND para cada citoquina y grupo. Es importante aclarar que en esta base de datos el número de individuos para los que se dispone de niveles de concentración de cada citoquina es variable. La razón es que los individuos no han sido analizados a la vez en el tiempo, y en algunos experimentos no se han determinado todas las citoquinas.

| Citoquina | Severos | | Moderados | | Leves | | Sanos | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | % ND | N | % ND | N | % ND | N | % ND | N | % ND |
| **EGF** | 29 | 6.90% | 121 | 7.43% | 104 | 13.46% | 54 | 5.56% | 308 | 9.09% |
| **Eotaxin** | 8 | 37.5% | 23 | 52.17% | 57 | 80.70% | 33 | 87.88% | 121 | 74.38% |
| **Fractalkina** | 15 | 0% | 59 | 10.17% | 82 | 8.54% | 53 | 13.21% | 209 | 9.57% |
| **IFNg** | 28 | 60.71% | 93 | 67.74% | 90 | 80% | 34 | 88.24% | 245 | 74.29% |
| **IL1b** | 28 | 53.57% | 93 | 63.44% | 90 | 55.56% | 53 | 43.40% | 264 | 55.68% |
| **IL1RA** | 28 | 0% | 93 | 1.08% | 90 | 1.11% | 52 | 1.92% | 263 | 1.14% |
| **IL2** | 25 | 68% | 85 | 75.29% | 80 | 66.25% | 41 | 58.54% | 231 | 68.40% |
| **IL4** | 12 | 41.67% | 28 | 57.14% | 31 | 45.16% | 38 | 78.95% | 109 | 59.63% |
| **IL5** | 3 | 100% | 8 | 87.50% | 41 | 58.54% | 31 | 58.06% | 83 | 62.65% |
| **IL6** | 29 | 10.34% | 121 | 23.14% | 104 | 37.50% | 54 | 44.44% | 308 | 30.52% |
| **IL8** | 29 | 0% | 121 | 2.48% | 104 | 5.76% | 54 | 1.85% | 308 | 3.24% |
| **IL9** | 3 | 100% | 8 | 100% | 10 | 100% | 12 | 100% | 33 | 100% |
| **IL10** | 928 | 42.86 | 93 | 52.69% | 90 | 57.78% | 53 | 58.49% | 264 | 54.55% |
| **IL12p70** | 25 | 32% | 85 | 61.18% | 80 | 65% | 22 | 77.27% | 129 | 60.85% |
| **IL13** | 15 | 46.67% | 36 | 50% | 41 | 78.05% | 31 | 80.65% | 123 | 66.67% |
| **Il17** | 28 | 85.71% | 93 | 90.32% | 90 | 85.56% | 34 | 97.06% | 245 | 88.98% |
| **IP10** | 28 | 3.57& | 93 | 3.22% | 90 | 1.11% | 34 | 0% | 245 | 2.04% |
| **MMP9** | 14 | 0% | 57 | 8.77% | 30 | 20% | 22 | 18.18% | 123 | 12.20% |
| **RANTES** | 28 | 7.14% | 93 | 17.20% | 90 | 26.67% | 34 | 38.24% | 245 | 22.45% |
| **TNFa** | 28 | 39.29% | 93 | 62.37% | 90 | 76.67% | 53 | 69.81% | 264 | 66.29% |
| **VEGF** | 16 | 18.75% | 65 | 24.62% | 90 | 41.11% | 34 | 64.71% | 205 | 78% |

*Tabla 7: Tasa de valores ND para cada citoquina, de forma total y por grupos.*
*N: número de individuos evaluados.*

Como se ha visto en las simulaciones, no tiene mucho sentido imputar los valores ND en los casos en los que su tasa de ND es superior al 50%, por tanto, de entre las 21 citoquinas inicialmente consideradas, se seleccionan las 9 moléculas cuya proporción de ND es menor que el 50%. Concretamente: EGF, Fractalkina, IL1RA, IL6, IL8, IP10, MMP9, RANTES y VEGF.

Los niveles de expresión observados para cada citoquina, grupo y método de imputación se muestran en la Tabla 8.

| Citoquina | Grupos | S0 | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| **EGF** | Severo | 8.0812±2.7195 | 8.2452±2.2421 | 8.4092±1.8466 | 8.567±1.6013 | 8.4413±1.7859 |
| | Moderado | 8.9860±3.1369 | 9.1459±2.6958 | 9.3059±2.3122 | 9.4707±2.0035 | 9.2717±2.387 |
| | Leve | 8.7111±3.709 | 9.0655±2.8921 | 9.4199±2.1578 | 9.5772±1.8481 | 9.5905±1.8197 |
| | Sano | 9.376±2.6081 | 9.5396±2.0346 | 9.7032±1.5507 | 9.7255±1.4998 | 9.7718±1.3984 |
| **Fractalkina** | Severo | 10.1117±1.3185 | 10.1117±1.3185 | 10.1117±1.3185 | 10.1117±1.3185 | 10.1117±1.3185 |
| | Moderado | 8.4955±3.298 | 8.8121±2.5137 | 9.1287±1.8943 | 9.1575±1.8404 | 8.9582±2.2054 |
| | Leve | 8.3784±2.9483 | 8.6805±2.1376 | 8.9827±1.5652 | 8.9638±1.581 | 8.8539±1.852 |
| | Sano | 8.3883±3.641 | 8.8867±2.5296 | 9.3851±1.6934 | 9.2608±1.8605 | 9.1077±2.1061 |
| **IL1RA** | Severo | 12.8209±3.6158 | 13.1175±2.7658 | 13.4141±2.6744 | 13.0538±2.898 | 13.2957±2.6982 |
| | Moderado | 11.9639±2.7813 | 11.9833±2.7019 | 12.0028±2.6334 | 12.0993±2.4831 | 12.0281±2.5624 |
| | Leve | 11.5357±2.8265 | 11.655±2.4563 | 11.7742±2.394 | 11.7242±2.2931 | 11.7422±2.3069 |
| | Sano | 10.4963±2.4988 | 10.5621±2.2496 | 10.6279±2.0803 | 10.6258±2.0843 | 10.6362±2.0661 |
| **IL6** | Severo | 5.3535±2.4031 | 5.5272±2.0299 | 5.7008±1.7344 | 5.6768±1.7728 | 5.703±1.7344 |
| | Moderado | 4.4126±2.77 | 4.6984±2.3505 | 4.9843±2.0412 | 5.1153±1.7691 | 5.1819±1.6782 |
| | Leve | 3.2339±2.7102 | 3.7232±2.1542 | 4.2125±1.7084 | 4.251±1.5917 | 4.4816±1.3526 |
| | Sano | 3.2426±3.4541 | 3.9846±2.7857 | 4.7267±2.248 | 4.5391±2.3622 | 4.5012±2.4034 |
| **IL8** | Severo | 10.4147±2.5971 | 10.4147±2.5971 | 10.4147±2.5971 | 10.4147±2.5971 | 10.4147±2.5971 |
| | Moderado | 8.1309±2.6697 | 8.1723±2.5528 | 8.2137±2.4581 | 8.2346±2.4196 | 8.2235±2.4294 |
| | Leve | 7.7313±2.7634 | 7.8177±2.5342 | 7.9041±2.3404 | 7.9741±2.1959 | 7.9294±2.2791 |
| | Sano | 7.8488±1.8371 | 7.8673±1.7599 | 7.8858±1.6901 | 7.9273±1.5659 | 7.9304±1.5589 |
| **IP10** | Severo | 12.439±4.2724 | 12.7818±3.4152 | 13.1245±3.0952 | 13.2848±2.5041 | 13.1678±2.729 |
| | Moderado | 14.3602±3.0647 | 14.418±2.7955 | 14.4758±2.5378 | 14.6792±1.7947 | 14.6691±1.8217 |
| | Leve | 14.6564±2.714 | 14.6766±2.6098 | 14.6963±2.5155 | 14.7681±2.2754 | 14.7478±2.3259 |
| | Sano | 13.9309±1.6872 | 13.9309±1.6872 | 13.9309±1.6872 | 13.9309±1.6872 | 13.9309±1.6872 |
| **MMP9** | Severo | 10.3328±5.969 | 12.0053±3.0745 | 13.6778±2.3166 | 12.0974±3.0495 | 13.0153±2.5688 |
| | Moderado | 8.5311±3.8631 | 8.7069±3.4927 | 8.8827±3.1828 | 8.9537±3.0788 | 8.937±3.1041 |
| | Leve | 7.0245±4.1728 | 7.4458±3.4675 | 7.8671±2.8471 | 7.8597±2.8617 | 7.9808±2.7132 |
| | Sano | 6.7783±3.8363 | 708035±3.2952 | 7.3823±2.8034 | 7.6339±2.4605 | 7.2611±3.0092 |
| **RANTES** | Severo | 5.3782±2.0839 | 5.4981±1.7886 | 5.618±1.563 | 5.7148±1.4679 | 5.6316±1.5454 |
| | Moderado | 4.6513±2.546 | 4.939±2.0445 | 5.2266±1.6379 | 5.4016±1.5175 | 5.051±1.8783 |
| | Leve | 3.8893±2.7649 | 4.4567±2.0555 | 5.0242±1.7009 | 4.6734±1.8096 | 4.6486±2.1072 |
| | Sano | 3.3331±2.9013 | 4.4732±1.6585 | 5.6132±1.1875 | 4.4805±1.6588 | 5.0043±1.4254 |
| **VEGF** | Severo | 7.9438±4.0911 | 8.3891±3.1983 | 8.8346±2.4082 | 9.2105±1.6401 | 9.4016±1.3726 |
| | Moderado | 6.659±4.0118 | 7.3867±2.8299 | 8.1145±1.8535 | 8.2933±1.5445 | 8.0517±1.8515 |
| | Leve | 5.2904±4.5478 | 6.5435±3.0991 | 7.7966±1.7844 | 7.8784±1.6894 | 7.7049±1.8263 |
| | Sano | 3.1761±4.4157 | 5.3575±2.81 | 7.5389±1.2816 | 7.3794±1.4720 | 6.9666±1.7112 |

*Tabla 8: Descriptivos de las 9 citoquinas seleccionadas, diferenciando grupo y método.*

En la Tabla 9 se recogen los p-valores ajustados de los contrastes globales del ANOVA por citoquina y método de imputación.

| Citoquina | S0 | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|
| **EGF** | 0.327 | 0.202 | 0.0604 | 0.0353 | 0.0203 |
| **Fractalkina** | 0.259 | 0.112 | 0.0923 | 0.129 | 0.1608 |
| **IL1RA** | < 0.0001 | < 0.0001 | < 0.0001 | < 0.0001 | < 0.0001 |
| **IL6** | 0.0002 | 0.0004 | 0.001 | 0.001 | 0.0005 |
| **IL8** | < 0.0001 | < 0.0001 | < 0.0001 | < 0.0001 | < 0.0001 |
| **IP10** | 0.088 | 0.066 | 0.048 | 0.013 | 0.0099 |
| **MMP9** | < 0.0001 | < 0.0001 | < 0.0001 | < 0.0001 | < 0.0001 |
| **RANTES** | 0.00516 | 0.0593 | 0.171 | 0.0052 | 0.1016 |
| **VEGF** | 0.0003 | 0.0017 | 0.1274 | 0.0012 | 0.0001 |

*Tabla 9: P-valores ajustados por comparaciones múltiples de los contrastes globales del ANOVA por citoquina y método*

Todas las citoquinas, salvo Fractalkina, muestran diferencias significativas entre alguno de los grupos y métodos utilizados. Cualquier método de imputación detecta diferencias globales entre grupos para los niveles de IL1RA, IL6, IL8 y MMP9.

En el caso de EGF e IP10, los métodos de sustitución no detectan diferencias estadísticamente significativas. Las diferencias en RANTES son detectadas con el método se sustitución por 0 y con el método de ROS. Por último, en el caso de VEGF, el método de sustitución por el LOD es el único a partir del cual no se detecta expresión diferencial.

Para evaluar entre que grupos de severidad se encuentran las diferencias significativas en las Tablas 10-17, se recogen los resultados de la comparación por pares de las 8 moléculas que mostraron un efecto global significativo con alguno de los métodos de imputación.

En cada celda de las tablas se muestra el log2(FC), su IC del 95% y el p-valor corregido del contraste correspondiente. Además, se representan gráficamente el log2(FC) el grupo de individuos sin ninguna patología ocular (sano) vs el resto de grupos, individuos que presentan algún grado de ojo seco.

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| EGF | Leve - Sano | -0.664907 (-1.7833,0.4535) ~ 1 | -0.1813 (-1.3462,0.3980) ~ 1 | -0.2833 (-1.1811,0.181) ~ 1 | -0.1483 (-0.724,0.4274) ~ 1 | -0.1813 (-0.7407,0.3782) 0.4902 |
| | Moderado - Sano | -0.390008 (-1.3542,0.5742) ~ 1 | -0.5 (-1.2050,0.4177) ~ 1 | -0.3973 (-0.8348,0.4723) ~ 1 | -0.2548 (-0.8568,0.3472) ~ 1 | -0.5 (-1.1892,0.1891) 0.1335 |
| | Severo - Sano | -1.2948 (-2.5073,-0.0822) 0.046 | -1.3304 (-2.2603,-0.3285) 0.02 | -1.2941 (-2.0904, -0.5705) 0.038 | -1.1585 (-1.8619,-0.4551) 0.036 | -1.3305 (-2.0374,-0.6235) 0.001 |
| | Moderado - Leve | 0.2749 (-0.6417,1.1915) ~ 1 | -0.0538 (-0.7047,0.5971) ~ 1 | -0.114 (-0.9394 , 0.7114) ~ 1 | -0.1118 (-0.6355,0.4118) ~ 1 | -0.3188 (-0.876,0.2384) 0.259 |
| | Severo - Leve | -0.6298 (-1.9046,0.6449) ~ 1 | -0.9451 (-1.8488,0.0414) 0.830 | -1.0108 (-1.8361, 0.0185) 0.137 | -1.0055 (-1.7320,-0.2789) 0.046 | -1.1492 (-1.9206,-0.3778) 0.004 |
| | Severo - Moderado | -0.9048 (-2.0927,0.2832) ~ 1 | -0.8913 (-1.7815,0.0012) 0.59 | -0.8967 (-1.7206 , 0.0728) 0.24 | -0.8936 (-1.6206,-0.1666) 0.041 | -0.8304 (-1.6422,-0.0186) 0.0453 |

*Tabla 10: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina EGF.*

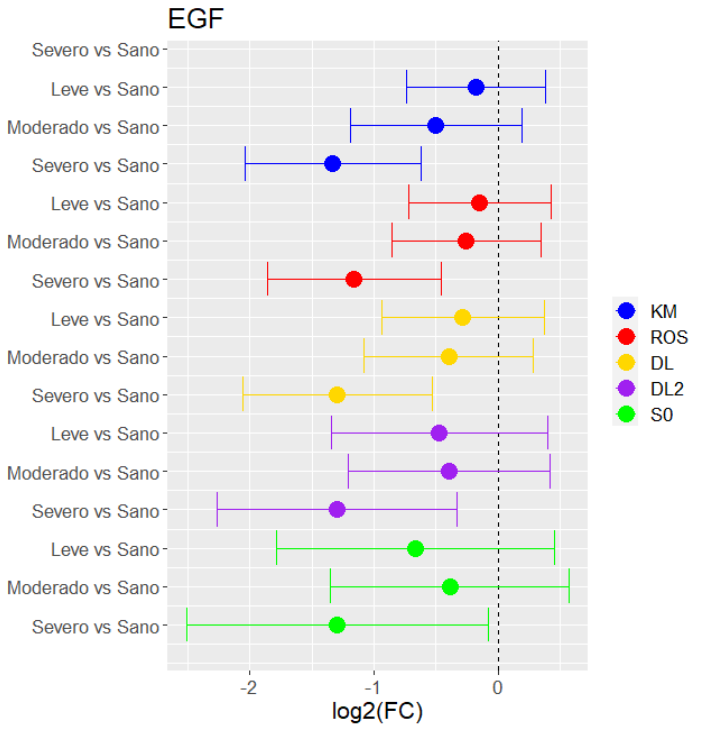



*Figura 5: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina EGF.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| IL1RA | Leve - Sano | 1.0394 (0.1056,1.9733) 0.0416 | 1.0929 (0.2722,1.9136) 0.0473 | 1.1463 (0.3595,1.9332) 0.0470 | 1.0984 (0.3341,1.8627) 0.0348 | 1.106 (0.3407,1.8713) 0.0098 |
| | Moderado - Sano | 1.4676 (0.5489,2.3863) 0.0095 | 1.4212 (0.5485,2.294), 0.0084 | 1.3749 (0.5361,2.2136) 0.0083 | 1.4736 (0.6697,2.2775) 0.0051 | 1.3919 (0.5714,2.2125) 0.0024 |
| | Severo - Sano | 2.3246 (0.9554,3.6938) < 0.0001 | 2.5554 (1.4164,3.6943) < 0.0001 | 2.7862 (1.7113,3.861) < 0.0001 | 2.4281 (1.3093,3.5469) < 0.0001 | 2.6595 (1.584,3.735) 0.0006 |
| | Moderado - Leve | 0.2436 (-0.5466,1.0338) ~ 1 | 0.2289 (-0.5203,0.9781) ~1 | 0.2286 (-0.5102,0.9674) ~ 1 | 0.2418 (-0.488,0.9715) ~ 1 | 0.2318 (-0.4883,0.9519) 0.4294 |
| | Severo - Leve | 1.6938 (0.5143,2.8733) 0.0208 | 1.6509 (0.4885,2.8134) 0.0161 | 1.6398 (0.4807,2.7989) 0.0135 | 1.63604 (0.478,2.7941) 0.0124 | 1.6177 (0.464,2.7715) 0.0157 |
| | Severo - Moderado | 1.4502 (0.2562,2.6442) 0.047 | 1.422 (0.2396,2.6044) 0.0472 | 1.4113 (0.2326,2.59) 0.0490 | 1.3943 (0.2206,2.568) 0.041 | 1.3859 (0.2143,2.5575) 0.0434 |

*Tabla 11: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina IL1RA.*

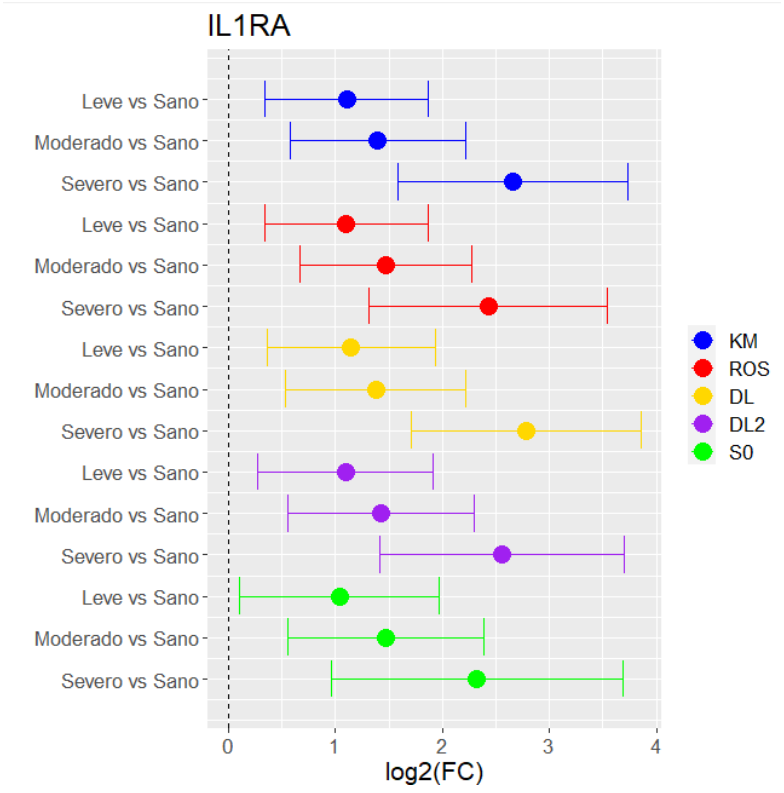



*Figura 6: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina IL1RA.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| IL6 | Leve - Sano | -0.0086 (-0.9972,0.98) ~ 1 | -0.2614 (-1.0524,0.5297) ~ 1 | -0.5142 (-1.1466,0.1183) 0.6989 | -0.2881 (-0.914,0.3378) ~ 1 | -0.0196 (-0.6096,0.5703) 0.9559 |
| | Moderado - Sano | 1.17 (0.2022,2.1379) 0.0461 | 0.7138 (-0.0911,1.5187) 0.3812 | 0.2576 (-0.4229,0.9381) ~ 1 | 0.5762 (-0.0601,1.2125) 0.4540 | 0.6807 (0.0574,1.304) 0.0477 |
| | Severo - Sano | 2.1109 (0.6767,3.5451) 0.0086 | 1.5425 (0.3745,2.7105) 0.0271 | 0.9741 (0.0191,1.9291) 0.0483 | 1.1377 (0.1406,2.1347) 0.045 | 1.2018 (0.1962,2.2074) 0.0282 |
| | Moderado - Leve | 1.1787 (0.4526,1.9048) 0.0131 | 0.9153 (0.3283,1.5024) 0.0121 | 0.7717 (0.2759,1.2675) 0.0197 | 0.9696 (0.4087,1.5305) 0.0096 | 0.7003 (0.2994,1.1013) 0.003 |
| | Severo - Leve | 2.1195 (1.0556,3.1835) 0.0028 | 1.7598 (0.9243,2.5954) 0.0017 | 1.4883 (0.7446,2.2319) 0.0019 | 1.5326 (0.5662,2.499) 0.0054 | 1.2214 (0.5079,1.9349) 0.0046 |
| | Severo - Moderado | 0.9409 (-0.1087,1.9905) 0.6691 | 0.8445 (-0.011,1.689) 0.5282 | 0.7166 (-0.0448,1.478) 0.4566 | 0.563 (-0.3944,1.5204) ~ 1 | 0.5211 (-0.209,1.2511) 0.1858 |

*Tabla 12: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina IL6.*

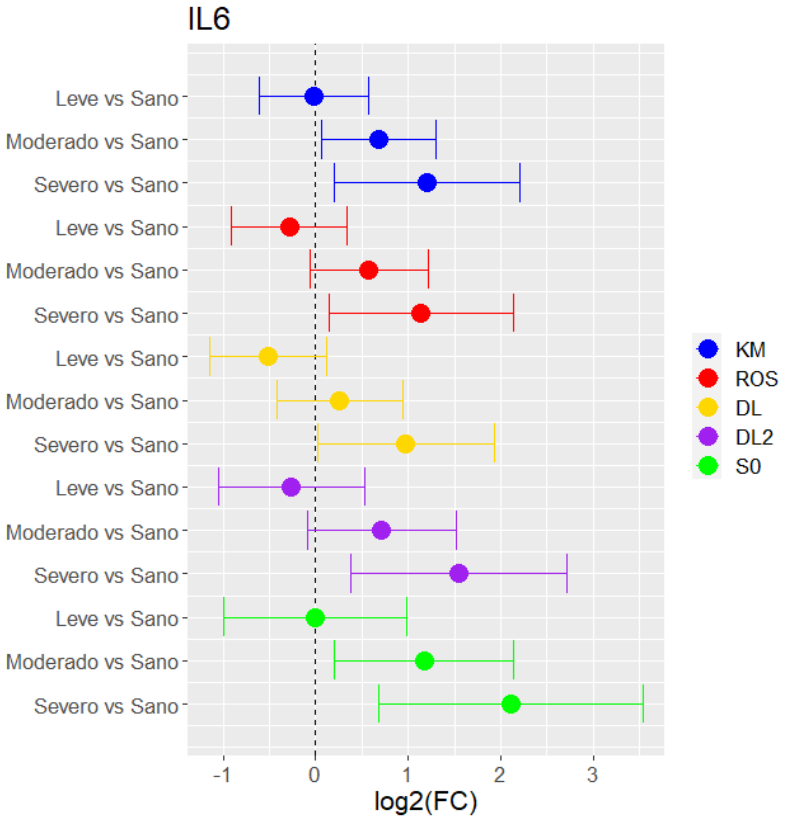



*Figura 7: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina IL6.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| IL8 | Leve - Sano | -0.1174 (-0.9417,0.7068) ~ 1 | -0.0496 (-0.8118,0.7127) ~ 1 | 0.0183 (-0.6913,0.7279) ~ 1 | 0.0468 (-0.6172,0.7108) ~ 1 | -0.001 (-0.6844,0.6825) 0.9975 |
| | Moderado - Sano | 0.2822 (-0.5076,1.0719) ~ 1 | 0.305 (-0.4504,1.0604) ~ 1 | 0.3279 (-0.3992,1.0549) ~ 1 | 0.3074 (-0.4012,1.016) ~ 1 | 0.2932 (-0.4198,1.0061) 0.5837 |
| | Severo - Sano | 2.5659 (1.5899,3.5419) 0.0001 | 2.5474 (1.5911,3.5037) < 0.0001 | 2.5289 (1.5901,3.4677) < 0.0001 | 2.4874 (1.5787,3.3962) < 0.0001 | 2.4843 (1.5772,3.3914) 0.0001 |
| | Moderado - Leve | 0.3996 (-0.3218,1.1211) ~ 1 | 0.3424 (-0.3179,1.0027) ~ 1 | 0.3095 (-0.3257,0.9448) ~ 1 | 0.2869 (-0.3495,0.9233) ~ 1 | 0.2941 (-0.3303,0.9185) 0.223 |
| | Severo - Leve | 2.6834 (1.5502,3.8165) < 0.0001 | 2.5682 (1.4627,3.6738) < 0.0001 | 2.5106 (1.4165,3.6046) < 0.0001 | 2.5035 (1.4114,3.5955) < 0.0001 | 2.4853 (1.3965,3.5741) 0.0001 |
| | Severo - Moderado | 2.2837 (1.1778,3.3896) 0.0001 | 2.2258 (1.1327,3.3188) < 0.0001 | 2.2010 (1.1122,3.2898) < 0.0001 | 2.2166 (1.1251,3.3081) < 0.0001 | 2.1912 (1.1039,3.2785) 0.0004 |

*Tabla 13: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina IL8.*

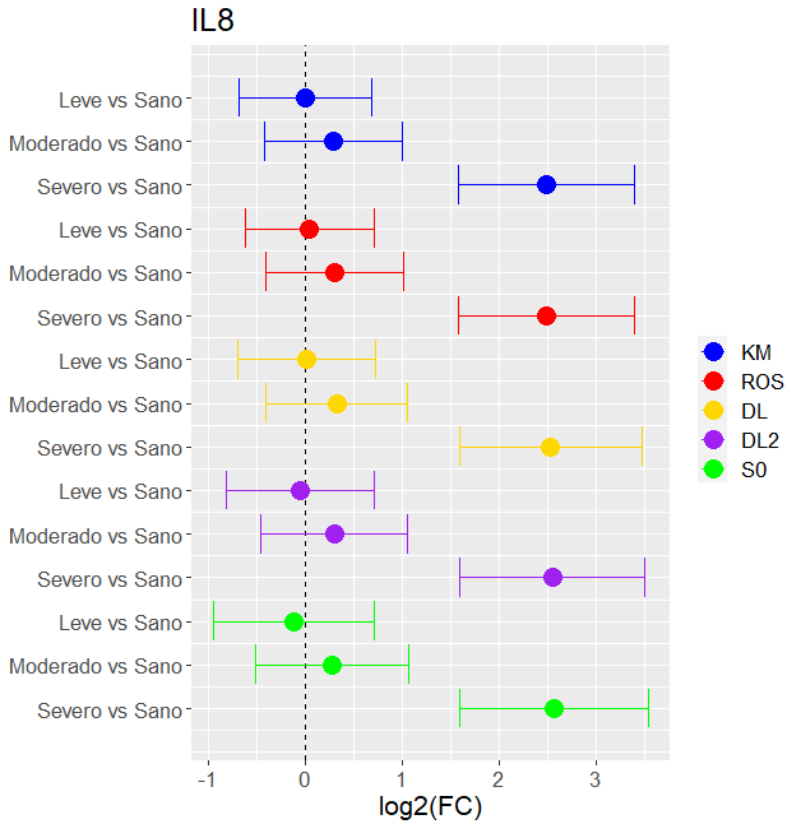



*Figura 8: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina IL8.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| IP10 | Leve - Sano | 0.7255 (-3.0883,0.1044) ~ 1 | 0.7654 (-0.1893,1.72) 0.998 | 0.7654 (-0.15595,1.6902) 0.779 | 0.8372 (-0.0125,1.687) 0.269 | 0.8169 (-0.0486,1.6823) 0.0605 |
| | Moderado - Sano | 0.4292 (-0.6688,1.5273) ~ 1 | - 0.5449 (-0.4666,1.5564) ~ 1 | 0.5449 (-0.3846,1.4744) ~ 1 | 0.7483 (0.0474,1.4491) 0.049 | 0.7382 (0.0293,1.4471) 0.0503 |
| | Severo - Sano | -1.4919 (-3.0883,0.1044) ~ 1 | -0.8064 (-2.1389,0.5261) ~ 1 | -0.8064 (-2.0439,0.4311) ~ 1 | -0.6461 (-1.7153,0.4232) ~ 1 | -0.7631 (-1.895,0.3688) 0.2494 |
| | Moderado - Leve | -0.2963 (-1.1459,0.5533) ~ 1 | -0.2416 (-1.0123,0.529) ~ 1 | -0.2205 (-0.9628,0.5218) ~ 1 | -0.0996 (-0.7099,0.5107) ~ 1 | -0.0787 (-1.1983,1.041) 0.7999 |
| | Severo - Leve | -1.66 (-3.1549,-0.165) 0.045 | -1.5964 (-2.9603,-0.2325) 0.033 | -1.5718 (-2.8895,-0.254) 0.024 | -1.4143 (-2.5226,-0.3059) 0.011 | -1.4681 (-2.6233,-0.3128) 0.0382 |
| | Severo - Moderado | -1.3637 (-2.8854,0.158) 0.163 | -1.3547 (-2.7248,0.0154) 0.114 | -1.3513 (-2.6673,0.0352) 0.077 | -1.3147 (-2.381,-0.2483) 0.02 | -1.3894 (-2.4993,-0.2795) 0.0328 |

*Tabla 14: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina IP10.*

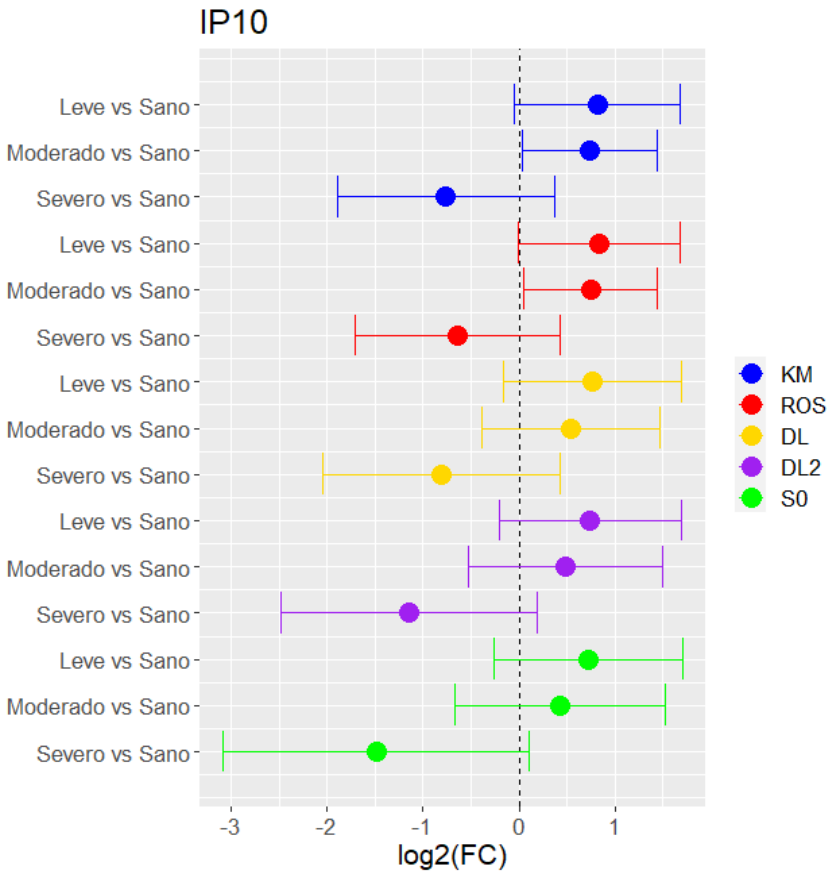



*Figura 9: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina IP10.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| MMP9 | Leve - Sano | 0.2461 (-2.0287,2.5209) ~ 1 | 0.3655 (-1.5493,2.2802) ~ 1 | 0.4848 (-1.1101,2.0796) ~ 1 | 0.2258 (-1.2967,1.7483) ~ 1 | 0.7196 (-0.8822,2.3215) 0.3847 |
| | Moderado - Sano | 1.7528 (-0.1743,3.6799) 0.41 | 1.6266 (-0.0927,3.3459) 0.33 | 1.5004 (-0.041,3.0417) 0.27 | 1.3198 (-0.1412,2.7808) 0.52 | 1.6758 (0.1372,3.2145) 0.0488 |
| | Severo - Sano | 3.5545 (0.2433,6.8657) < 0.0001 | 4.925 (2.6929,7.1571) < 0.0001 | 6.2954 (4.4696,8.1213) < 0.0001 | 4.4636 (2.587,6.3402) < 0.0001 | 5.7541 (3.7748,7.7335) 0.0001 |
| | Moderado - Leve | 1.5067 (-0.3703,3.3836) 0.49 | 1.1279 (-0.6322,2.8879) 0.58 | 1.0156 (-0.3532,2.3844) 0.78 | 1.1505 (-0.3595,2.6605) 0.53 | 0.9562 (-0.3604,2.2729) 0.1779 |
| | Severo - Leve | 3.3083 (0.5325,6.7741) < 0.0001 | 4.5595 (2.8195,6.2995) < 0.0001 | 5.8107 (4.0642,7.5572) < 0.0001 | 4.2377 (2.5377,5.9377) < 0.0001 | 5.197 (3.484,6.91) 0.0001 |
| | Severo - Moderado | 1.8017 (0.3017,3.3017) < 0.0001 | 3.2984 (2.0984,3.2984) < 0.0001 | 4.7951 (3.1769,6.4133) < 0.0001 | 3.1437 (1.5637,4.7237) < 0.0001 | 4.1408 (2.5351,5.7464) 0.0004 |

*Tabla 15: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina MMP9.*

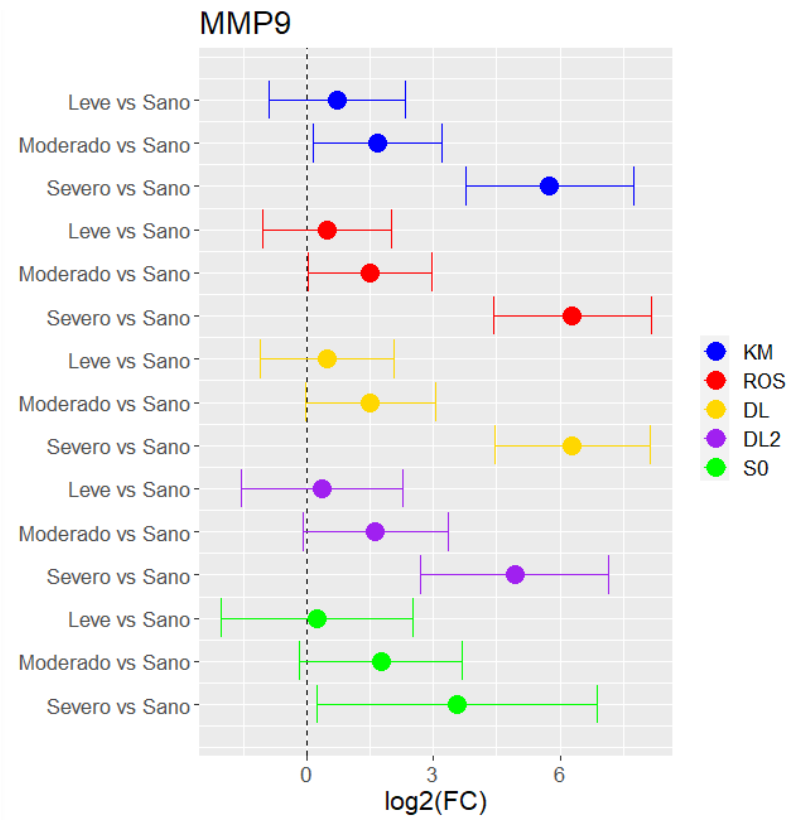



*Figura 10: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina MMP9.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| RANTES | Leve - Sano | 0.5562 (-0.5606,1.673) ~ 1 | -0.0164 (-0.7959,0.7631) ~ 1 | -0.589054 (-1.2181,0.04) 0.41 | 0.1933 (-0.5121,0.8987) ~ 1 | -0.3557 (-1.1314,0.42) 0.2881 |
| | Moderado - Sano | 1.3182 (-0.2693,2.3671) 0.079 | 0.4658 (-0.3077,1.2392) ~ 1 | -0.3866 (-0.9942,0.221) ~ 1 | 0.921 (0.3039,1.5382) 0.041 | 0.0467 (-0.6554,0.7488) 0.8821 |
| | Severo - Sano | 2.0451 (0.7353,3.3549) 0.016 | 1.0249 (0.1478,1.902) 0.037 | 0.00473 (-0.6943,0.7037) ~ 1 | 1.2342 (0.4298,2.0386) 0.029 | 0.6274 (-0.1284,1.3832) 0.2194 |
| | Moderado - Leve | 0.762 (-0.0194,1.5434) 0.309 | 0.4822 (-0.1201,1.0845) 0.597 | 0.2024 (-0.2883,0.6931) ~ 1 | 0.7277 (-0.0723,1.5277) 0.162 | 0.4024 (-0.1846,0.9894) 0.2649 |
| | Severo - Leve | 1.4889 (0.4836,2.4942) 0.046 | 1.0413 (0.2176,1.8651) 0.043 | 0.5938 (-0.1152,1.3028) 0.53 | 0.9103 (0.1415,1.6792) 0.042 | 0.983 (0.2302,1.7359) 0.045 |
| | Severo - Moderado | 0.7269 (-0.2459,1.6997) ~ 1 | 0.5591 (-0.2595,1.3778) ~ 1 | 0.3914 (-0.3078,1.0905) ~ 1 | 0.3132 (-0.3668,0.993) ~ 1 | 0.5807 (-0.1396,1.301) 0.2194 |

*Tabla 16: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina RANTES.*

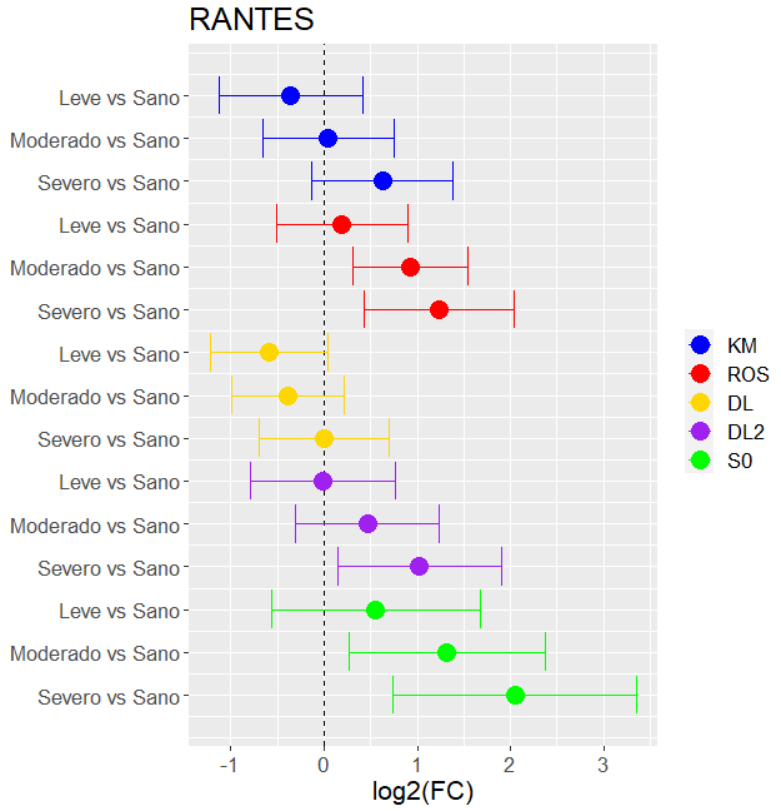



*Figura 11: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina RANTES.*

| Citoquina | Grupos | SO | SDL2 | SDL | ROS | K-M |
|---|---|---|---|---|---|---|
| VEGF | Sano - Leve | 2.1143 (0.316,3.9125) 0.037 | 1.186 (-0.0189,2.3909) 0.2951 | 0.2577 (-0.4052,0.9206) ~ 1 | 0.4991 (-0.1519,1.15) 0.5336 | 0.7384 (0.0227,1.454) 0.023 |
| | Sano - Moderado | 3.4829 (1.7381,5.2277) 0.0011 | 2.0292 (0.8433,3.2151) 0.009 | 0.5756 (-0.1305,1.2817) 0.78 | 0.9139 (0.2753,1.5525) 0.0256 | 1.0851 (0.3269,1.8433) 0.0096 |
| | Severo - Sano | 4.7677 (2.1363,7.3991) 0.0021 | 3.0317 (1.2415,4.8219) 0.0056 | 1.2957 (0.2502,2.3411) 0.031 | 1.8311 (0.9006,2.7616) 0.0018 | 2.435 (1.4518,3.4183) 0.0004 |
| | Moderado - Leve | 1.3687 (-0.0117,2.749) 0.3207 | 0.8433 (-0.1146,1.8012) 0.5001 | 0.3179 (-0.2755,0.9113) ~ 1 | 0.506 (-0.0575,1.0695) 0.6099 | 0.3467 (-0.2519,0.9454) 0.2516 |
| | Severo - Leve | 2.6535 (0.2459,5.061) 0.0495 | 1.8457 (-0.0047,3.6867) 0.1401 | 1.038 (-0.3064,2.3824) 0.2 | 1.412 (0.3844,2.4395) 0.03 | 1.6967 (0.858,2.5353) 0.0018 |
| | Severo - Moderado | 1.2848 (-1.1395,3.7091) ~ 1 | 1.0025 (-0.8588,2.8637) ~ 1 | 0.7201 (-0.6535,2.0937) 0.91 | 0.906 (-0.1197,1.9316) 0.44 | 1.3499 (0.4698,2.23) 0.0096 |

*Tabla 17: Diferencia de medias, IC y p-valor de los contrastes por pares para la citoquina VEGF.*

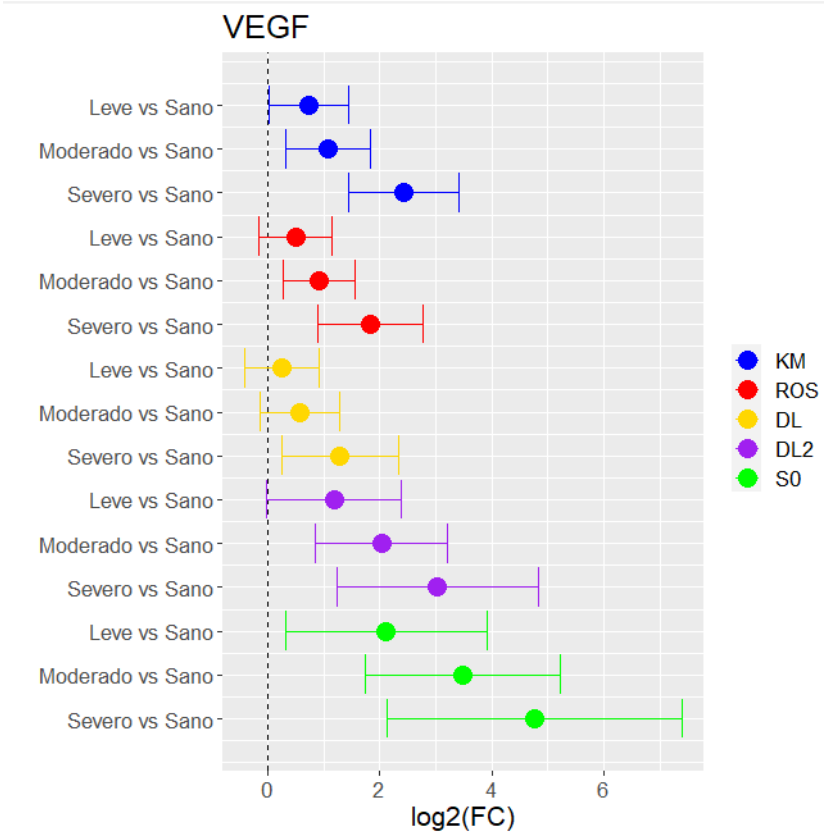



*Figura 12: Gráfico log2(FC) del grupo sano vs el resto de grupos usando la citoquina VEGF.*

Para todas las moléculas evaluadas se observa que los resultados obtenidos son coherentes con los publicados en el meta-análisis (Roda et al, 2020): los niveles de EGF e IP10 aparecen infra-expresados en los pacientes con ojo seco más severo, mientras que para el resto de moléculas se observa sobre-expresión.

En el caso de EGF (Tabla 10, Figura 5) se observan diferencias significativas entre el grupo de severos y el de sanos para cualquiera de los métodos de imputación, sin embargo, el contraste global no era significativo para los tres métodos de sustitución. Además, los métodos ROS y Kaplan-Meier detectan como significativas las diferencias entre los grupos de severos y leves.

En cuanto a IL1RA (Tabla 11, Figura 6) los cinco métodos detectan diferencias significativas entre el grupo de sanos y el resto de grupos con ojo seco, siendo mayor la expresión en aquellos grupos de mayor gravedad. La única diferencia no significativa en este caso es entre los grupos moderado y leve.

Para IL6 (Tabla 12, Figura 7) todos los métodos detectan como significativa la diferencia entre severos y sanos, y, además, los métodos de sustitución por 0 y Kaplan-Meier detectan la diferencia entre los grupos de moderados y sanos. Es llamativa, la diferencia en cuanto a la estimación de la diferencia, mientras que con el método de sustitución por 0, el log2(FC) de los severos respecto de los sanos es mayor que 2, para el método de sustitución por la mitad del LOD se estima en 1.5 y en el resto de métodos entorno a 1.

La citoquina IL8 (Tabla 13, Figura 8) apenas se ve afectada por el método de imputación, ya que en general tiene una proporción de valores ND pequeña, siendo 0 en el grupo de sanos.

Ninguno de los métodos de imputación detecta como significativa la infra-expresión que se observa en el grupo de severos para la citoquina IP10 (Tabla 14, Figura 9).

MMP9 está sobreexpresada en los grupos de pacientes de ojo seco, especialmente en el de severos (Tabla 15, Figura 10). Son llamativas las diferencias en la estimación del log2(FC) dependiendo del método de imputación utilizado, siendo menor en el caso de la sustitución por 0.

La citoquina RANTES, con un porcentaje de ND de 22.45%, es la que presenta más dependencia del método de imputación (Tabla 16, Figura 11). Los métodos que detectan más diferencias entre grupos son ROS y el de sustitución por 0, siendo significativa las diferencias entre los sanos y los pacientes de ojo seco moderados o severos. Esta molécula se caracteriza porque el grupo de sanos tiene un porcentaje de ND elevado, lo que puede explicar la variabilidad del resultado en relación al método de imputación empleado.

Por último, con la citoquina VEGF (Tabla 17, Figura 12), se detecta la diferencia entre los individuos sanos respecto de los pacientes de ojo seco moderados o

severos. Las excepciones se observan con el método de sustitución por el LOD que no detecta como significativa la diferencia entre moderados y sanos, y el de sustitución por 0 que además, detecta la diferencia entre leve y sanos como significativa.

Para todas las moléculas, se observa que la estimación de la diferencia entre grupos tiene una mayor variabilidad en el caso de los métodos de sustitución.

# 4. CONCLUSIONES

A partir del estudio de simulación:

- Cuando trabajamos con dos grupos sin expresión diferencial, los métodos que mejores resultados proporcionan son la sustitución por el LOD y Kaplan-Meier, siendo ambos métodos prácticamente idénticos con cualquier tamaño de grupos y porcentaje de ND. Por el contrario, los métodos ROS y la sustitución por 0 son las peores elecciones.

- Si uno de los grupos tiene el doble de expresión Kaplan-Meier empeora mucho y las mejores opciones pasan a ser ROS y la sustitución por el LOD cuando el porcentaje de ND es < 50%. Cuando se maneja un porcentaje de ND > 50% la mejor opción es la sustitución por el LOD.

- Por último, cuando uno de los grupos tiene cuatro u ocho veces más de expresión, el mejor método cuando el porcentaje de ND es < 50 % es ROS, seguido de cerca por la sustitución por el LOD. Si el porcentaje de ND es > 50% los mejores resultados se van a obtener con el método se sustitución por el LOD/2.

- En general, salvo para los escenarios con la proporción de valores ND más extremos, los mejores resultados se obtienen con el método ROS. Este mal comportamiento con tasa de no detección elevada tiene sentido, puesto que en estos casos, prácticamente todas las observaciones del grupo de referencia serán valores ND y serán estimados a partir de los valores observados en el grupo expresado diferencialmente, por lo que será muy difícil observar diferencias estadísticamente significativas.

A partir del conjunto de datos reales de expresión de citoquinas en lágrima:

- Los métodos de sustitución estiman las diferencias entre grupos con un error mucho mayor.

- En una situación extrema, en la que la tasa de ND es alta en uno de los grupos, aunque no lo sea en global, realizar la sustitución por un valor constante podrían provocar que se detecten como significativas diferencias inexistentes y por tanto una toma de decisiones errónea.

En general:

- Los métodos de Kaplan-Meier y ROS son capaces de detectar diferencias entre los grupos cuando el porcentaje de ND es < 50%, obteniendo resultado ligeramente mejores con el método ROS.

- El método de imputación utilizado puede tener un efecto importante en los resultados obtenidos a partir del mismo conjunto de datos, por lo que es crucial elegir el método más adecuado en cada caso.

# REFERENCIAS


1. Banta-Green, C. J., Brewer, A. J., Ort, C., Helsel, D. R., Williams, J. R., & Field, J. A. (2016). Using wastewater-based epidemiology to estimate drug consumption—Statistical analyses and data presentation. *Science of the Total Environment*, *568*, 856-863.

2. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

3. Enríquez-de-Salamanca, A., & Calonge, M. (2008). Cytokines and chemokines in immune-based ocular surface inflammation. Expert review of clinical immunology, 4(4), 457-467.

4. Genser, B., Cooper, P. J., Yazdanbakhsh, M., Barreto, M. L., & Rodrigues, L. C. (2007). A guide to modern statistical analysis of immunological data. *BMC immunology*, *8*(1), 1-15.

5. Grima-Olmedo, C., Grima-Olmedo, J., Luque-Espinar, J. A., Pardo-Iguzquiza, E., & Ramírez-Gómez, Á. (2019). Análisis de supervivencia a la detección y corrección de la contaminación en series con valores por debajo del límite de detección. *Revista internacional de contaminación ambiental*, *35*(1), 165-178.

6. Helsel, D. R. (2010). Summing nondetects: Incorporating low-level contaminants in risk assessment. *Integrated Environmental Assessment and Management*, *6*(3), 361-366.

7. Helsel, D. R. (2011). *Statistics for censored environmental data using Minitab and R* (Vol. 77). John Wiley & Sons.

8. Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data* (Vol. 1230). New York: Springer.

9. Lee L. (2020). NADA: Nondetects and Data Analysis for Environmental Data. R package version 1.6-1.1. https://CRAN.R-project.org/package=NADA

10. Mancin, M., Barco, L. & Ricci, A. (2017). Statistical methods to estimate the assigned value in presence of multiple censored results. Obtenido de:https://www.eurachem.org/images/stories/workshops/2017_10_PT/pdf/contrib/O05-Mancin.pdf

11. Quignon Santana, S., & Alfonso Sánchez, O. (2009). Principales manifestaciones oculares en la artritis reumatoide: modelos de diagnóstico y evaluación. MediSur, 7(6), 52-58.

12. Quintanilla Casas, B. (2017). Estadística en variables con censura. Aplicación a datos medioambientales. Obtenido de: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/63786/6/bquintanillacTFM0617memoria.pdf

13. Roda, M., Corazza, I., Bacchi Reggiani, M. L., Pellegrini, M., Taroni, L., Giannaccare, G., & Versura, P. (2020). Dry Eye Disease and Tear Cytokine Levels—A Meta-Analysis. *International journal of molecular sciences*, *21*(9), 3111.

14. Sadegh, M. K. (2008). A note on the regression of order statistics. *Communications in Statistics—Theory and Methods*, *37*(4), 475-480.

15. Shoari, N., & Dubé, J. S. (2018). Toward improved analysis of concentration data: embracing nondetects. *Environmental toxicology and chemistry*, *37*(3), 643-656.

16. Tamhane, M., Cabrera-Ghayouri, S., Abelian, G., & Viswanath, V. (2019). Review of biomarkers in ocular matrices: challenges and opportunities. Pharmaceutical research, 36(3), 1-35.