ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
Grado en Ingeniería de la Salud, mención Bioinformática

Análisis de citoquinas en pacientes adictos a la cocaína
Cytokine analysis in patients with cocaine use disorder

Realizado por
Rosa María Maza Quiroga
Tutorizado por
Dr. Leonardo Franco
Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA. Junio de 2017

Fecha defensa:
El Secretario del Tribunal

Agradecimientos

Gracias a todas las personas que han construido la biografía de mi vida. Yo soy, y con ellos lo he ido desarrollando. Cada uno sabe y siente, qué dosis le corresponde.

Gracias por dar, y recibir.

Namasté.

Resumen

La adicción a la cocaína es un problema de salud creciente, siendo España el país de la Unión Europea donde se produce el mayor consumo de esta sustancia. La adicción conlleva el desarrollo de comorbilidad psiquiátrica, lo que agrava la salud de los consumidores. Con el fin de facilitar el proceso de análisis de datos de pacientes con trastornos por consumo de cocaína (TCC) se ha desarrollado una aplicación informática para tratar datos que describen a pacientes reclutados en régimen ambulatorio en periodo de abstinencia. La aplicación permite el análisis de variables sociodemográficas, clínicas, de adicción y moleculares, mediante técnicas de la estadística inferencial permitiendo comparaciones con un grupo de control. Adicionalmente, la aplicación permite la utilización de técnicas pertenecientes al área de minería de datos, que potencialmente permitirían obtener nuevas conclusiones.

Para la consecución de los objetivos planteados, se ha llevado a cabo un proceso aplicando la metodología de la ingeniería del software, a fin de crear una aplicación eficiente, cumpliendo con los requisitos acordes a las necesidades de los investigadores relacionados. Una vez construída la aplicación, se realizaron las pruebas de las distintas fases con un conjunto de datos reducido, para posteriormente llevar a cabo el análisis final con el conjunto total de datos disponible.

En cuanto a los resultados obtenidos, la aplicación permitió encontrar las diferencias significativas entre ambas poblaciones (grupo de pacientes y grupo de control) en apenas minutos, reduciendo los tiempos de análisis y la complejidad de los mismos. De esta forma, fue posible considerar múltiples opciones en el análisis comparativo, concluyendo que el factor de crecimiento transformante alfa (TGFα) es un posible biomarcador relevante. Independientemente de lo resultados obtenidos hasta el momento, la automatización de la aplicación permitirá que se sigan extrayendo conclusiones a medida que el conjunto de datos crezca; sin duda ayudando a obtener un diagnóstico más adecuado, mejorando así el tratamiento y la asistencia de los pacientes.

Palabras clave: biomarcadores, cocaína, comorbilidad psiquiátrica, estadística inferencial, ingeniería del software, minería de datos, trastorno por consumo de cocaína (TCC).

Abstract

Cocaine addiction is a growing health problem worlwide, with the addition that Spain is the country within the European Union with the highest consumption of this substance. Addiction is associated with serious health problems including psychiatric comorbidity. In order to facilitate the data analysis process of patients with cocaine use disorders (CUD), a software application has been developed to study abstinence patients recruited in an outpatient treatment. The application allows the analysis of sociodemographic, clinical, addictive and molecular variables, using inferential statistics techniques allowing comparisons with a control group. Further, it is worth noting that data mining techniques have been incorporated to the application, which could potentially lead to new results and conclusions.

In order to achieve the stated objectives, a process has been carried out applying the software engineering methodology, with the aim of creating an efficient application, fulfilling all the requirements of the involved researchers. Once the application was built, the tests of the different phases were carried out with a reduced data set, to then carry out the final analysis with the total set of data available.

Regarding the obtained results, the application allowed to find the significant differences between both populations (patient and control group) in minutes, reducing the analysis times and the complexity related to the application of the several tests involved. In this way, it was possible to consider multiple options in the comparative analysis, concluding that the transforming growth factor alpha (TGFα) is a possible relevant biomarker. Regardless of the results obtained so far, automation of the application will allow further conclusions to be drawn as the data set grows; without doubt helping to obtain a more adequate diagnosis, and thus improving the treatment and the assistance of the patients.

Keywords: biomarkers, cocaine, cocaine use disorder (CUD), data mining, inferential statistics, psychiatric comorbidity, software engineering.

Índice general

Resumen vii

Índice general ix

Índice de figuras xi

Acrónimos xiii

Página

Glosario xv

1 Introducción 1
1.1. Motivación, estado del arte, objetivos y organización de la memoria 2

2 Datos 5
2.1. Participantes y reclutamiento 5
2.1.1. Criterios de exclusión e inclusión 5
2.1.2. Población con trastorno por consumo de cocaína 6
2.1.3. Población control 6
2.2. Declaración ética 7
2.3. Evaluación psiquiátrica 7
2.3.1. PRISM 7
2.3.2. Gravedad de los síntomas por consumo de cocaína 8
2.4. Selección de moléculas en plasma a analizar 9
2.5. Métodos de laboratorio 12
2.5.1. Colección y análisis de muestras de plasma 12
2.5.2. Análisis de inmunoensayo 'multiplex' 12

3 Herramientas para el análisis de los datos 14
3.1. Técnicas estadísticas 15
3.1.1. Prueba de Kolmogorov-Smirnov 15
3.1.2. Prueba de Lilliefors 15
3.1.3. Prueba de Levene 17
3.1.4. Prueba t de Student 18
3.1.5. Prueba de Wilcoxon-Mann-Whitney 19

3.1.6. Prueba χ^2 . . . 21
3.1.7. ANCOVA . . . 22
3.1.8. Medias marginales estimadas . . . 27
3.2. Algoritmos de clustering . . . 30
3.2.1. Clustering jerárquico . . . 30
3.2.2. Algoritmo de K-medias (K-means) . . . 31
3.2.3. Red de Kohonen de mapas auto-organizados (SOM networks) . . . 32
3.2.4. Índice de silueta . . . 33
3.3. Tecnología Software . . . 34
3.3.1. R y Rstudio . . . 34
3.3.2. Shiny . . . 34
3.3.3. Softwares adicionales . . . 35

4 Desarrollo de la aplicación para el análisis de datos de pacientes con TCC 36
4.1. Metodología seguida para el desarrollo de la aplicación . . . 37
4.1.1. Fases seguidas en el proceso de desarrollo de la aplicación . . . 37
4.1.2. Análisis de requisitos . . . 38
4.2. Manual de usuario . . . 40
4.2.1. Carga de datos . . . 40
4.2.2. Carga de datos . . . 41
4.2.3. Análisis de datos . . . 42

5 Resultados 47
5.1. Estudio de pacientes con TCC contrastados con sujetos control . . . 47
5.1.1. Características sociodemográficas-clínicas . . . 47
5.1.2. Impacto de la cocaína en los niveles en plasma de las citoquinas . . . 48
5.2. Estudio de pacientes con TCC diagnosticados con comorbilidad psiquiátrica 50
5.2.1. Efecto de la comorbilidad psiquiátrica en las características sociodemográficas-clínicas . . . 50
5.2.2. Impacto del diagnóstico dual en los niveles plasmáticos de citoquinas 50
5.3. Impacto de la gravedad de los síntomas por consumo de cocaína en los niveles plasmáticos moleculares en pacientes con TCC . . . 53
5.4. Aplicación de las técnicas de clustering . . . 54
5.4.1. Clustering jerárquico . . . 54
5.4.2. Clustering utilizando SOM . . . 55
5.4.3. Clustering utilizando K-means . . . 57

6 Conclusiones 60

Bibliografía **63**

Índice de figuras

3.1. Aplicación: test de Kolmogorov-Smirnov con la corrección de Lilliefors para la variable molecular Eotaxin-1 comparando pacientes con TCC y sujetos control. 16
3.2. Aplicación: distribución de los datos según la variable molecular Eotaxin-1 comparando pacientes con TCC y sujetos control. 16
3.3. Aplicación: test de Kolmogorov-Smirnov con la corrección de Lilliefors para la variable molecular Eotaxin-1, modificando el nivel de significación a 0.037 . . . 17
3.4. Aplicación: sujetos control enfrentados a pacientes CUD en la variable IMC, que es homocedástica. 18
3.5. Distribución normal. División de dos colas. 19
3.6. Aplicación: los datos presentan distribuciones normales. 20
3.7. Aplicación: prueba de Wilcoxon-Mann-Whitney: resultado significativo 21
3.8. Aplicación: prueba de Wilcoxon-Mann-Whitney con resultado no significativo 22
3.9. Aplicación: prueba χ^2 donde el estado civil guarda relación con la adicción del sujeto. 23
3.10. Aplicación: ANCOVA. Variable dependiente: IFNγ. Variables independientes: cofactores sexo, diagnóstico dual, trastorno por consumo de otra sustancia y síntomas de gravaedad; covariables: edad, IMC, edad de prueba de la droga y tiempo de abstinencia. 24
3.11. Aplicación: ANCOVA. Variable dependiente: IFNγ. Variables independientes: cofactores sexo y diagnóstico dual; covariables: edad e IMC. 25
3.12. Aplicación: análisis de normalidad de la variable dependiente Eotaxin-1 con trasformación logarítmica, comparando pacientes con TCC y sujetos control. . 26
3.13. Suma de cuadrados tipo III. 26
3.14. Aplicación: 'Cofactors means' representa la rejilla de referencia para calcular Medias marginales estimadas en este estudio de TGFα considerando el diagnóstico dual como factor principal. 29
3.15. Proceso del algoritmo K-means. 32
3.16. Estructura de un red auto-organizada (SOM network) 33

4.1. Instalación de paquetes. 40
4.2. Captura de pantalla del lanzamiento de la aplicación. 41
4.3. Captura de pantalla del proceso de carga de un archivo de datos. 41
4.4. Análisis sociodemográfico y clínico de los datos utilizados. 42
4.5. Captura de pantalla del análisis de adicción-clínico. 43

4.6. Análisis de normalidad Eotaxin-1 diferenciando la muestra con TCC según la variable de comorbilidad por otra sustancia. 44
4.7. Análisis de normalidad mediante histogramas de Eotaxin-1 diferenciando la muestra con TCC según la variable de comorbilidad por otra sustancia. 45
4.8. Análisis de covarianza de MIP-1α usando como factor principal que el sujeto tenga TCC, junto con el factor sexo y las dos covariables edad e IMC. 45
4.9. Análisis de covarianza de MIP-1α usando como factor principal que el sujeto tenga TCC, junto con el factor sexo y las dos covariables edad e IMC. Representa dos gráficas de los datos que aparecen en la figura (4.8) 46
4.10. Clustering jerárquico con valor de 3 para el parámetro de número de clusters. 46

5.1. Estudio sociodemográfico-clínico de la población total. 48
5.2. Medias marginales estimadas de las concentraciones citoquínicas en plasma para pacientes con TCC y grupo de control. 49
5.3. Características sociodemográficas y clínicas en pacientes con TCC. 51
5.4. Medias marginales estimadas de las concentraciones citoquínicas en plasma diferenciando diagnóstico dual. 52
5.5. Gravedad de los síntomas por el consumo de cocaína en pacientes con TCC. . 53
5.6. Clustering jerárquico con K=3 clusters aplicado a la población total. 54
5.7. Valores medios de las variables utilizadas en el clustering jerárquico con K=3. 55
5.8. SOM: Mapeado Estándar. 55
5.9. SOM: Mapeado Múltiple. 56
5.10. SOM: Cantidad de sujetos asignados a cada cluster. 56
5.11. SOM: Medida de calidad de los clusters. 57
5.12. Datos agrupados según k-means. 58
5.13. Resultados del clustering obtenido mediante aplicación del algoritmo de k-means. 58
5.14. Datos agrupados según k-means. 59

Acrónimos

Español

CPD: Centro Provincial de Drogodependencia.
IMC: Índice de Masa Corporal. En inglés: BMI.
RTA: Red de Trastornos Adictivos.
TCC: Trastorno por consumo de cocaína. En inglés: CUD.
TCS: trastorno por consumo de una sustancia, en este estudio se refiere a sustancias diferentes a la cocaína. En inglés SUD.

Inglés

BLUE: Best Linear Unbiased Estimator. En castellano es 'mejores estimadores lineales insesgados' pero se les conoce como BLUE.
BMI: Body Mass Index.
CUD: Cocaine use disorder.
EMM: Estimated Marginal Means, also called Predicted Marginal Means or Least Square Means. En castellano Medias Marginales Estimadas, Medias Marginales Predichas o Medias de Mínimos Cuadrados, pero se conoce como EMM, PMM o lsmeans.
SUD: Substance use disorder.
DDSI: Dual Diagnosis Screening Instrument.
DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders - 4th Edition (revised). En español Manual diagnóstico y estadístico de los trastornos mentales, 4º edición revisada.
PRISM: Psychiatric Research Interview for Substance and Mental Disorders. En español: Entrevista de investigación psiquiátrica para trastornos mentales o por consumo de sustancia.
SS: Sum of square values.

Glosario

Circulante: hace referencia a que una proteína no está anclada a ningún tejido o célula. Se trata por tanto de una proteína que se difunde mediante un fluido, como la sangre, saliva o linfa. En este estudio cuando se menciona el término "proteína circulante"se hace referencia a la sangre como medio de difusión.

Comorbilidad: coexistencia de dos patologías en una misma persona, no neceriamente en el mismo espacio temporal.

Comorbilidad por consumo de otra sustancia: coexistencia de dos patologías por consumo de sustancias. En lo que se refiere a este estudio el paciente tiene asociado el trastorno por consumo de cocaína además de otro trastorno por consumo de otra/s sustancias diferentes a la cocaína (alcohol, heroína, cannabis, benzodiacepinas, alucinógenos u otros estimulantes).

Comorbilidad psiquiátrica: también denominada comorbilidad psicopatológica. Es la coexistencia de dos comorbilidades psiquiátricas. Aclaraciones: El trastorno por consumo de sustancia se considera una comorbilidad psiquiátrica, pues la adicción es un trastorno mental; el trastorno por patología dual también se considera comorbilidad psiquiátrica. Este concepto representa una variable clínica, determinada en todos los pacientes de este estudio.

Diagnóstico dual: también denominado patología dual. Es la coexistencia de dos patologías psiquiátricas, es decir, de trastorno mental. En este estudio los pacientes al menos han tenido un trastorno psiquiátrico a lo largo de la vida (trastorno de estado de ánimo, psicótico o de personalidad (antisocial o límite) además del trastorno por consumo de cocaína).

Prevalencia: cuando se menciona este término se hace referencia a los estudios de prevalencia, que consisten en medir la prevalencia de la exposición a la cocaína y del efecto que esta exposición tiene en una muestra poblacional, en un solo momento temporal, permitiendo así estimar la magnitud y distribución de una patología en un momento concreto.

Psiquiatría: parte de la medicina que se ocupa del estudio, diagnóstico, tratamiento y prevención de las enfermedades mentales de carácter orgánico y no orgánico.

Psicopatología: estudio de las enfermedades o trastornos mentales, así como de su naturaleza y sus causas.

Régimen ambulatorio: un paciente con tratamiento bajo este régimen tiene que acudir regularmente a un centro de salud para la continuidad del tratamiento, pero no está internado en el mismo.

Capítulo 1

Introducción

Este documento detalla el trabajo fin de grado del alumno Rosa María Maza Quiroga, en el Grado en Ingeniería de la Salud, mención Bioinformática de la Universidad de Málaga. Este proyecto se titula Análisis de citoquinas en pacientes adictos a la cocaína, y con él se pretende contribuir al estudio de los factores que afectan a personas adictas a la cocaína. Se pretende principalmente identificar moléculas que guarden relación con el diagnóstico clínico de estos pacientes, pues esto permitirá realizar un adecuado tratamiento de los pacientes además de ayudar en la mejora de su pronóstico. Por otro lado, el estudio incluye un análisis de tipo etiológico del Trastorno por Consumo de Cocaína (TCC), el cual utiliza información personal del paciente. Con el fin de realizar este estudio se ha desarrollado una aplicación software novedosa e interactiva que permitirá a los investigadores de este campo poder analizar la información disponible de cada paciente, facilitando su tarea de análisis.

1.1. Motivación, estado del arte, objetivos y organización de la memoria

El consumo de cocaína es un hecho creciente en la sociedad actual. Este consumo se torna a adicción y pasa a ser un problema de salud grave de la ciudadanía. Se trata de un problema con una alta prevalencia en edades tempranas y con un fuerte impacto social, sanitario y económico. Un dato destacable es que España es el país de la Unión Europea donde más cocaína se ha consumido en los últimos años, duplicando en muchos casos la media de la comunidad (Araos et al., 2014). La adicción de cocaína conlleva una serie de complicaciones para el paciente, destacándose entre ellas la elevada prevalencia de la comorbilidad psiquiátrica[1]. Detectar los trastornos psicopatológicos asociados al consumo de cocaína al comienzo de la asistencia sanitaria es necesario para mejorar el diagnóstico y pronóstico de los pacientes. Este conjunto de pacientes está considerado de alto riesgo, pues frecuentan en mayor medida los servicios hospitalarios (Rufles, 2009), tienen mayores tasas de suicidios (Friedler et al., 2012) y responden peor a los tratamientos que los pacientes que sólo sufren el TCC (Karila et al., 2012). Pese a la gravedad de la situación son pocos los grupos de investigación dedicados al estudio de este tema tanto en Europa como en España. La gran mayoría de los grupos en el entorno español utilizan herramientas software no libres (que necesitan licencia) que en muchos casos limitan el alcance de los análisis, pues sólo pueden usar las herramientas que el software propone e incluso algunos parámetros a considerar vienen fijados. Además, en ninguno de los estudios analizados se utilizan para el estudio de los pacientes técnicas complementarias al anásis estadístico, como los métodos basados en algoritmos provenientes del campo de la inteligencia computacional, por lo que podría no estar actualmente aprovechándose toda la información disponible en los datos.

Para asistir y estudiar a los pacientes con TCC en España se ha llevado a cabo la labor de reclutar a pacientes con esta adicción. Estos pacientes se encuentran en régimen ambulatorio[2], de los cuales se toman muestras de sangre y además se les somete a una entrevista psiquiátrica hecha por especialistas. En España esta tarea la ha llevado a cabo La Red de Trastornos Adictivos (RTA)[3]. Ésta es una estructura organizativa financiada por el Instituto de Salud Carlos III del Ministerio de Sanidad y Consumo constituida por distintos grupos de investigación biomédica cuyo objetivo es investigar los orígenes de los procesos adictivos, sus manifestaciones clínicas y su impacto social con el fin último de proporcionar herramientas de intervención terapéuticas y sociales que ayuden a reducir el tremendo impacto de las adicciones y las patologías asociadas. El Investigador principal de este proyecto es Dr. Fernando Rodríguez de Fonseca. Se encarga de la coordinación de diferentes grupos en España, y en concreto del que tiene sede en Málaga (Grupo de investigación en Neuropsicofarmacología, Hospital Regional Universitario de Málaga). Este grupo ha llevado a cabo estudios sobre pacientes con TCC, incluyendo en algunos de ellos análisis

[1] Coexistencia en una misma persona de un trastorno por consumo de una sustancia, en este estudio cocaína, y de otras psicopatologías

[2] pacientes no ingresados de forma interna en un centro especializado

[3] `http://redrta.es/index.php?option=com_content&task=view&id=38&Itemid=38`

moleculares en busca de posibles biomarcadores que ayuden al diagnóstico y pronóstico clínicos (Pedraz et al., 2015). Sin embargo, hasta el momento no se habían analizado una serie de moléculas, en concreto las citoquinas, a fin de buscar posibles relaciones con la adicción a la cocaína y sus consecuencias en el paciente. La gran mayoría de los grupos de investigación llevan a cabo el estudio de los datos mediante técnicas estadísticas, buscando conclusiones a través de estudios etiológicos (construcción de modelos explicativos). Técnicas provenientes del área de aprendizaje computacional (Machine learning) son bastante desconocidas para la mayoría de los grupos de investigadores mencionados, principalmente por la falta de personal cualificado para su implementación y aplicación, y este trabajo intenta contribuir en este sentido.

Por lo anteriormente expuesto se ha creado dentro de este proyecto una aplicación bioinformática de manejo sencillo que permita de forma más directa y en un modo más visual realizar los análisis estadísticos actualemente utilizados, y a su vez aplicar métodos de agrupamiento (clustering) que permitirían un mejor análsis de los datos. Además la herramienta desarrollada es fácilmente adaptable a futuras incorporaciones tanto de diferentes tipos de datos como de herramientas de análisis. Ha sido desarrollada en el lenguaje *R* de uso libre, utilizando el entorno gráfico *Shiny*. Además de la ventajas de la utilización de un software libre, la utilización del software *R* de uso muy extendido en la investigación científica, permitirá la utilización de sus potentes métodos a investigadores no expertos en programación, acercándolos a técnicas y análisis más innovadores y potencialmente más eficaces para extraer todo el conocimiento disponible en los datos. Este trabajo fin de grado tiene como objetivo principal el análisis y estudio de biomarcadores en plasma, a través del desarrollo de un software que permita a través de ellos diferenciar a los sujetos control de los pacientes consumidores de cocaína y también, la identificación de subgrupos dentro de la población con TCC, en principio tratando de diferenciar subgrupos acorde a los trastornos psicopatológicos que padecen. A este estudio a nivel molecular le precede un análisis exploratorio de la población en el cual se estudian la características sociodemográficas, psiquiátricas, de adicción y de patrones de consumo. Esto, a su vez complementa a publicaciones anteriores y reafirma conclusiones de las mismas (Pavón et al., 2013). Por tanto la investigación se centra en enfrentar a dos muestras de datos de pacientes con y sin TCC, a fin de obtener conclusiones que permitan diferenciar a las mismas, buscando principalmente relaciones en cuanto a la concentración molecular de citoquinas en plasma que ayuden al tratamiento de los pacientes con TCC.

Para el adecuado desarrollo de este proyecto se ha desglosado la metodología en varios bloques, permitiendo así de esta forma profundizar en cada uno de ellos para sacar el máximo partido a las oportunidades que las circunstancias ofrecen. En primer lugar se ha realizado en la sección 2 un estudio de los datos disponibles y las publicaciones previas relacionadas a fin de enfocar el análisis. Entender las subpoblaciones que se considerarán para el análisis y estudiar las distribuciones que sus datos siguen en lo que a diferentes variables se refiere, es el primer paso para emprender el análisis. A continuación se incluyen también los métodos de laboratorio utilizados para la obtención de los datos. La sección 3 de esta memoria incluye todas las herramientas utilizadas para el análisis de los datos. En una primera parte se detallan los test estadísticos tradicionales que se utilizan para

el análisis comparativo entre las poblaciones analizadas. A continuación se presentan tres técnicas diferentes que pueden aplicarse para realizar un proceso de agrupamiento o clustering, mientras que al final de la sección se describen las tecnologías utilizadas para el desarrollo de la aplicación que se presenta, incluyéndose también descripciones sobre otros softwares utilizados. La sección 4 describe el proceso de desarrollo e implementación de la aplicación desarrollada para el análisis de los datos de pacientes, incluyendo en una primera parte el proceso de ingeniería de software y una segunda donde se muestra la utilización y funcionalidades de la aplicación. En el capítulo 5 se presentan los resultados obtenidos, mediante la aplicación de los diferentes tipos de análisis disponibles mediante el uso del software desarrollado. Finalmente se extraen conclusiones y se discute el trabajo a futuro que puede realizarse como continuación del presente proyecto en la sección 6.

Capítulo 2

Datos

2.1. Participantes y reclutamiento

Uno de los análisis que se lleva a cabo en este proyecto consiste en un estudio de prevalencia (también llamado estudio transversal o vertical), que se compone de un análisis estadístico y demográfico de los datos de los pacientes. Este estudio consiste en medir a la vez la prevalencia de la exposición a la cocaína y del efecto que esta exposición tiene en una muestra poblacional, en un solo momento temporal, permitiendo así estimar la magnitud y distribución de una patología en un momento concreto. Es por ello que se necesita un conjunto de sujetos en el que se diferencien dos subconjuntos, los cuales sean homogéneos en cuanto a variables sociodemográficas y difieran en incluir ó no la patología que se va a analizar en el estudio (TCC). Seleccionando los datos disponibles más actualizados hasta la fecha, el estudio está constituido por un total de 160 sujetos de raza caucásica, en el que se diferencian dos subpoblaciones: el grupo control y el grupo con TCC.

2.1.1. Criterios de exclusión e inclusión

Para poder participar en el estudio la edad del sujeto debía estar comprendida entre 18 y 65 años. Se determinaron los siguientes criterios de exclusión: (1) historia de enfermedades crónicas (ej. Enfermedades cardiovascular, respiratoria, renal, hepática, neurológica o endocrina); (2) historia de cáncer; (3) enfermedades infecciosas; (4) incapacidades cognitivas y (5) mujeres embarazadas.

Los sujetos del grupo control no podían estar siendo tratados mediante ningún tipo de medicación. Los pacientes que demandan atención por uso de cocaína en los distintos centros provinciales de drogodependencia (CPD), en su mayoría, no tienen un tratamiento farmacológico específico para paliar su adicción a cocaína, eso sí, a un alto porcentaje de ellos se les receta benzodiacepinas, antidepresivos y anticonvulsivantes para reducir estados psicológicos alterados. Dado que no existe en la actualidad un medicamento exclusivo para tratar la adicción a cocaína. Adicionalmente, se excluyeron del grupo de control suje-

tos que hubiesen tenido o tuvieran trastorno de Eje I[1] o de Eje II[2], trastorno neurológico, o trastorno por consumo de alguna sustancia a lo largo de la vida. En este estudio el consumo de tabaco no fue considerado dentro del conjunto de sustancias que provocan trastorno.

2.1.2. Población con trastorno por consumo de cocaína

Se reclutaron consumidores de cocaína por vía nasal exclusivamente. Estos consumidores pertenecían a un programa de tratamiento ambulatorio[3] por trastorno de consumo de cocaína en la provincia de Málaga (España) durante un período de 24 meses (Agosto de 2013 a Agosto de 2015). Basándose en los criterios de inclusión y exclusión, finalmente se seleccionaron 79 pacientes para el estudio. Era necesario que los pacientes consumidores de cocaína hubiesen sido diagnosticados con un 'consumo patológico' de cocaína a lo largo de su vida [abuso de cocaína y/o dependencia ('Diagnostic and Statistical Manual of Mental Disorders-4th EditionText Revision', DSM-IV-TR)], además de encontrarse en periodo de abstinencia de la droga por al menos 2 semanas.

El diagnóstico del trastorno por consumo de cocaína y otros trastornos psiquiátricos se determinó mediante entrevista psiquiátrica. Para cerciorarse de que el paciente se encontraba en periodo de abstinencia se realizó un análisis de orina semanal a cada sujeto dispuesto a participar en el estudio, en el centro de tratamiento ambulatorio de adicción a la cocaína al que cada paciente pertenecía.

Los análisis de orina para cocaína, anfetamina, opiáceos, barbitúricos, fenciclidina y cannabis se realizaron usando el 'V-Twin Drug Testing System' (Siemens AG, Erlangen, Alemania). El resultado de estos análisis sirvió para determinar los pacientes con trastorno por consumo de cocaína que al menos, llevaban 2 semanas sin consumir esta droga. Posteriormente se hicieron análisis en plasma para verificar la abstinencia en cocaína.

2.1.3. Población control

El reclutamiento de la población control se hizo teniendo especial cuidado en el sexo, edad e índice de masa corporal, para que se asemejara al grupo con TCC. En esta población se practicaron también los análisis de orina y plasma para verificar la abstinencia requerida de 2 semanasde duración, quedando finalmente 81 sujetos para el estudio.

[1] Trastornos Clínicos.

[2] Trastornos de la personalidad.

[3] aquel en el que el paciente acude regularmente a un centro de salud por razones de diagnóstico o tratamiento pero que no necesita pasar la noche allí (es decir, no queda internado).

2.2. Declaración ética

Cada participante firmó un informe de consentimiento después de haber sido informado con una descripción completa del estudio, habiéndole resuelto cualquier duda o problema. El estudio y protocolos para reclutamiento fueron aprobados por el Comité de ética del CEI Provincial de Málaga, en concordancia con el 'Ethical Principles for Medical Research Involving Human Subjects', adoptado en la Declaración de Helsinki por el 'World Medical Association (64th WMA General Assembly, Fortaleza, Brasil, Octubre 2013. Recomendación No. R (97) 5 del Comité de Ministros a los Estados Miembros sobre la Protección de Datos Médicos (1997) y la Ley de Protección de Datos de España (Ley Orgánica 15/1999 de Protección de Datos, LOPD).

2.3. Evaluación psiquiátrica

Todos los pacientes con trastorno por consumo de cocaína fueron evaluados de acuerdo al DSM-IV-TR (Diagnostic and Statistical Manual of Mental Disorders). Para ello se usó la versión en español 'Psychiatric Research Interview for Substance and Mental Disorders' (PRISM 6.0) (Hasin et al., 1996) (Torrens et al., 2004). Los sujetos de control fueron evaluados inicialmente por PRISM para detectar trastornos por uso de sustancias por la versión española del 'Dual Diagnosis Screening Instrument' (DDSI) (Mestre-Pintó et al., 2014). Todas las entrevistas fueron llevadas a cabo por psicólogos que habían recibido un certificado de formación en esta herramienta.

2.3.1. PRISM

La entrevista de diagnóstico PRISM 6.0 (Entrevista de investigación psiquiátrica para trastornos mentales o por consumo de sustancia) es el instrumento de diagnóstico en versión castellana, empleado para evaluar los trastornos que recoge el DSM-IV-TR (manual diagnóstico y estadístico de los trastornos mentales, 4º edición revisada). PRISM es una entrevista clínica semiestructurada[4] diseñada para resolver los problemas de diagnóstico en personas con un consumo elevado de sustancias y/o alcohol, evaluando en un primer módulo algunas preguntas del historial del consumo del sujeto, arrojando diagnósticos de abuso y dependencia tanto en el último año como en el anterior a la entrevista. Además, evalúa 20 trastornos del Eje I y los 2 trastornos del Eje II más prevalentes en esta población: el trastorno límite de personalidad y el trastorno antisocial de personalidad.

Los diagnósticos se realizan en dos marcos temporales de evaluación. Por un lado, la entrevista valora los trastornos actuales, es decir, los que están presentes en el último año. Por otro lado, también se evalúan trastornos previos, es decir, los definidos con anterioridad al último año. Generalmente, la prevalencia diagnóstica presentada por un

[4]Se determina de antemano cual es la información relevante que se quiere conseguir. Se hacen preguntas abiertas dando oportunidad a recibir mas matices de la respuesta, permite ir entrelazando temas, pero requiere de una gran atención por parte del investigador para poder encauzar y estirar los temas. (Actitud de escucha)

sujeto 'a lo largo de la vida', incluiría la totalidad de los diagnósticos enunciados con anterioridad observados en ambos periodos temporales, es decir, los más comunes. Una de las características más importantes de este instrumento es que permite diferenciar los trastornos mentales primarios o independientes de los trastornos inducidos por sustancias, así como los síntomas esperados del efecto de la intoxicación y la abstinencia. El criterio establecido por la entrevista PRISM para que un trastorno psicopatológico pueda ser considerado inducido por la sustancia, es que debe ocurrir en el contexto de un consumo patológico de la sustancia, en cualquiera de estas dos situaciones: a) intoxicación crónica, es decir, un consumo de 4 o más días a la semana durante un periodo de 3 o más semanas; b) atracones, durante un periodo de consumo de 3 días continuados (Hasin et al., 1996) (Torrens et al., 2004). En contraposición al consumo patológico se encuentra el consumo ocasional, cuando el consumo es menor a 4 días en la semana (a menos que la sustancia se consuma dentro de un patrón de atracones). En el presente estudio únicamente se adoptaron pacientes con trastorno por consumo patológico a lo largo de la vida, excluyendo a los sujetos diagnosticados con un consumo ocasional. A fin de diferenciar los síntomas psicopatológicos inducidos de los esperados durante la intoxicación o la abstinencia, se debe determinar si hubo un cambio brusco en el patrón de consumo . La entrevista PRISM presenta buena fiabilidad test-retest, validez y fiabilidad inter-examinadores (los valores del coeficiente Kappa oscilan entre 0,66 y 1,00) (Morgello et al., 2006).

2.3.2. Gravedad de los síntomas por consumo de cocaína

El hecho de que el criterio del manual DSM-IV-TR sea unidimensional para los trastornos por consumo de cocaína ha permitido que se utilicen para determinar la gravedad del tratamiento por cocaína, combinando los 7 criterios de dependencia (se requirieron 3 o más síntomas concurrentes en un periodo de 12 meses) y los 4 criterios de abuso (es necesario 1 síntoma), todo lo cual está de acuerdo con el criterio del manual DSM-5 (Hasin et al., 2012, 2013). En estudios anteriores se han descrito más detalles con respecto a los análisis de la gravedad sintomática de cocaína usando el criterio unidimensional del manual DSM-IV-TR (Pavón et al., 2013; Araos et al., 2015; Pedraz et al., 2015).

Los sujetos fueron inicialmente evaluados a través de la herramienta PRISM (para diferenciarlos de acuerdo a los grupos de consumo de abuso o dependencia). Posteriormente fueron sometidos a la versión española de DDSI para detectar los trastornos psiquiátricos (Robins et al., 1988; Mestre-Pintó et al., 2014). Todas las entrevistas se llevaron a cabo por psicólogos especializados, quienes habían recibido un certificado que garantiza el correcto manejo de ambas herramientas.

2.4. Selección de moléculas en plasma a analizar

Las citoquinas[5] son pequeñas proteínas segregadas por células que tienen un efecto específico sobre las interacciones y las comunicaciones entre las células, siendo su función principal la de regular el mecanismo de la inflamación. La citoquina es un nombre genérico; otros nombres incluyen linfoquina (citoquinas hechas por linfocitos), monocinas (citoquinas hechas por monocitos), quimioquinas [6] (citoquinas con actividades quimiotácticas [7]) e interleucina (citoquinas hechas por un leucocito y que actúan sobre otros leucocitos). Son producidas fundamentalmente por células inmunes como los linfocitos[8] y los macrófagos activados, aunque también pueden ser producidas por leucocitos polimorfonucleares (PMN)[9], células endoteliales[10], epiteliales, adipocitos[11], miocitos[12] y del tejido conjuntivo. Las citoquinas pueden actuar sobre las células que las segregan (acción autócrina), en las células cercanas (acción parácrina) o, en algunos casos, en células distantes (acción endocrina). Siendo su acción fundamental la regulación del mecanismo de la inflamación, las citoquinas pueden actuar en las rutas provocando la inflamación o la protección de la misma, denominándose pro-inflamatoria o anti-inflamatoria respectivamente. Esta es una distinción tradicional entre citoquinas.

Distinguir a las citoquinas por sus funciones hizo que en el pasado se las clasificara de una forma u otra, pero a medida que la investigación en este campo ha avanzado se ha descubierto que una misma citoquina puede actuar de las dos formas. Por ello en la actualidad se indica solamente si la función de la citoquina en cuestión es principalmente inflamatoria, sin distinciones. Aunque existen determinadas citoquinas a las que se les alude con el tipo de inflamación que acometen, pues ocurre en ocasiones que esta función es la única y está notoriamente definida. Además del mecanismo de inflamación, las citoquinas son proteínas que regulan otras funciones de las células que las producen sobre otros tipos celulares. Por ejemplo, son los agentes responsables de la comunicación intercelular, inducen la activación de receptores específicos de membrana, funciones de proliferación y diferenciación celular, quimiotaxis, modulación de la secreción de inmunoglobulinas y crecimiento. En un principio, esta última función fue observada en algunas proteínas y es por ello que recibieron el nombre de Factor de Crecimiento. Sucesivas investigaciones indicaron que proteínas que se creía que únicamente cumplían función de crecimiento lo que desarrollaban como papel principal era un mecanismo de inflamación; sin embargo, por comodidad se les mantuvo el nombre que se les asignó en su descubrimiento original, y es lo que le ocurre a una de las citoquinas objeto del estudio: TGFα.

[5] También denominadas citocinas

[6] También llamadas quimiocinas o «citocinas quimiotácticas» son proteínas de tamaño pequeño y bajo peso molecular (8 a 14 kDa).

[7] Sustancia que induce a determinadas células a migrar hacia el órgano diana.

[8] Tipo de glóbulo blanco.

[9] Engloba a varios tipos de glóbulo blanco: neutrófilo, eosinófilo y basófilo.

[10] Células que forman el endotelio vascular que es un epitelio plano simple (de una sola capa de células) que recubre la cara interna de los vasos sanguíneos y el corazón.

[11] Células que forman el tejido adiposo

[12] Células del tejido muscular.

En la última década, un creciente número de estudios han explorado el potente papel que las citoquinas pueden jugar en poblaciones de pacientes con trastornos mentales (Raison et al., 2006; Engler et al., 2017; Michopoulos et al., 2017; Notter et al., 2017) y/o adicción a drogas, incluyendo TCC (Irwin et al., 2007; Parikh et al., 2014; Levandowski et al., 2016). Algunos estudios han encontrado reducción en las citoquinas circulantes[13] en pacientes con ansiedad generalizada antes y después del tratamiento con la práctica de Mindfulness [14], cuando son comparados con los grupos de control (Hoge et al., 2017). Existe evidencia significativa que demuestra que ciertas citoquinas están implicadas en la iniciación y persistencia del dolor patológico al activar directamente las neuronas sensoriales nociceptivas. Cabe destacar el hecho de que las citoquinas actúan a través de receptores en la regulación de la inflamación o la embriogénesis[15]. Algunas de estas señales pueden cruzar la barrera hematoencefálica[16] y producir un estado inflamatorio que se ha relacionado con disfunciones en sistemas monoaminérgicos ascendentes (Petrulli et al., 2017). De hecho, las citoquinas y sus receptores se expresan en microglía[17], en las células inmunes presentes en el cerebro que participan en procesos como la remodelación de una poda sináptica[18] (Paolicelli et al., 2011), así como en astrocitos[19] y neuronas.

En consideración de lo anteriormente expuesto se descartaron las moléculas ya estudiadas para la población disponible en este estudio y se escogieron las citoquinas que podrían dar resultados que completaran y ampliaran análisis previos, siendo entonces las seleccionadas para este estudio:

- Eotaxin-1: CCL11 es una quimioquina que cumple la función de ser un poderoso quimiotáctico contra los eosinófilos.
- IFNγ: es una de las dos citoquinas principales relacionadas con la inmunidad. Es el principal activador de macrófagos y células citotóxicas T y NK. Esta citoquina tiene acción en la inmunidad humoral[20], induciendo la producción de IgG.
- IL-4: Interleucina 4, es la citoquina mejor caracterizada en la regulación de la respuesta inmune humoral. Se caracteriza por su acción antiinflamatoria.
- IL-8: Interleucina 8, es una de las principales citoquinas que actúan en la respuesta inespecífica. Se caracteriza por su acción proinflamatoria.
- IL-17α: Interleucina 17, se caracteriza por su actividad proinflamatoria.

[13] Apartado de Definiciones

[14] Atención plena, la práctica de la meditación.

[15] Parte de la biología que trata de la formación y desarrollo del embrión.

[16] Barrera entre los vasos sanguíneos y el sistema nervioso central. La barrera impide que muchas sustancias tóxicas la atraviesen, al tiempo que permite el pasaje de nutrientes y oxígeno.

[17] También llamadas microgliales o células de Hortega, son células que tienen capacidad fagocitaria, que forman parte del conjunto de células neurogliales del tejido nervioso.

[18] Eliminación de conexiones sinápticas entre neuronas.

[19] Son las principales y más numerosas células gliales, que son células del tejido nervioso, donde actúan en funciones auxiliares, complementando a las neuronas.

[20] Es el principal mecanismo de defensa contra los microorganismos extracelulares y sus toxinas

- MIP-1α: CCL3 es una quimioquina implicada en el estado inflamatorio agudo del reclutamiento y activación de leucocitos polimorfonucleares. Cumple la función de efecto adhesivo en linfocitos.
- TGFα: Factor de Necrosis Tumoral Alfa, es una citoquina que presenta en su nombre vestigios de que fue en su origen descubierta como factor de crecimiento, aunque después se observó que es una de las principales citoquinas que actúan en la respuesta inespecífica o inflamación.

2.5. Métodos de laboratorio

2.5.1. Colección y análisis de muestras de plasma

La extracción de sangre se llevó a cabo al mismo tiempo que se realizaba la entrevista psiquiátrica. Las muestras fueron extraídas por la mañana (de 08:00 a 10:00 am) antes de un ayuno de 8 a 12 h. Las muestras de sangre venosa se extrajeron en tubos de 10 ml K2 EDTA (BD, Franklin Lakes, NJ, USA). Para obtener el plasma, las muestras se centrifugaron a 2200 g durante 15 minutos (4 °C). Cada muestra de plasma fue individualmente analizada mediante 3 rápidas pruebas para detectar enfermedades infecciosas: VIH (Retroscreen HIV, QualPro Diagnostics-Tulip Group Ltd, Goa, India), hepatitis B (HBsAg Test, Toyo Diagnostics- Turklab Inc., Esmirna, Turquía) y hepatitis C (Flaviscreen HCV, QualPro Diagnostics-Tulip Group Ltd). Los análisis de plasma para metabolitos de cocaína (Benzoylecgonine Specific Direct ELISA Kit purchased from Immunalysis Co., Pomona, CA, USA) también sirvieron para confirmar la abstinencia en cocaína. Las muestras de plasma fueron conservadas a -80 °C hasta la realización de nuevos análisis.

2.5.2. Análisis de inmunoensayo 'multiplex'

Para este estudio se han elegido moléculas que no hayan sido analizadas en estudios previos, ó en estudios desarrollados bajo la misma temática y procedimientos (Araos et al., 2015).

La técnica Bio-Plex Suspension Array 200 (Bio-Rad Laboratories, Hercules, CA, USA) se usó para cuantificar los niveles de citoquinas en plasma con un Kit de Inmunoensayo Procarta, usando bolas de poliestireno y una disolución adecuada de una disolución Estándar en Plasma (Affymetrix-Panomics, Santa Clara, CA, USA). Este método de análisis está basado en tecnología 'Luminex' y en un panel 6-plex de citoquina humana, que fue usado simultáneamente para detectar los siguientes analitos: Quimiocina C-C motifligand 11 (CCL11, eotaxin-1), Interferón γ (IFN γ), Interleucina-4 (IL-4), Interleucina-8 (IL-8), interleukin-17 α (IL-17 α), Proteínas inflamatorias de macrófago 1 α (macrophage inflammatory proteins, MIP-1 α) and Factor de crecimiento transformador α (Transforming Growth Factor,TGF α). La medida de estos analitos en plasma se realizó siguiendo las instrucciones del fabricante.

Seguidamente (terminado el procedimiento para una placa filtrante de 96 pocillos), después de preparar 1x tampón de lavado de 10x tampón de lavado con doble cantidad de agua destilada, la placa filtrante de 96 pocillos se incubó con 150 μl de tampón de lectura durante 5 minutos a temperatura ambiente (TA). Se añadieron cincuenta microlitros de perlas conjugadas con anticuerpo a cada pocillo de la placa de filtrado. Después de añadir 1x tampón de lavado, se añadieron 25 μl de tampón plasma específico y 25 μl de patrones (rango = 0,62 pg / ml a 40 600 pg / ml) o muestras (n = 2 repeticiones) en cada pocillo. A continuación, las placas de filtro se dejaron incubando bajo agitación durante 60 minutos a temperatura ambiente (TA). La solución resultante se eliminó mediante filtración a vacío. A continuación, se lavó la placa de filtro y se añadieron 25 μl de anticuerpos

de detección de multiplex, bajo agitación durante 30 minutos a TA. Después del lavado, se añadieron 50 μl de una disolución de estreptavidina R-ficoeritrina durante otros 30 minutos de agitación seguido de un posterior lavado y la adición de 120 μl del tampón de lectura en cada pocillo. Se realizó una última incubación durante 30 minutos a TA antes de la lectura en el instrumento Luminex. Los datos recogidos (intensidad de fluorescencia media) se analizaron utilizando el Software Bio-Plex Manager 4.1 (Bio-Rad Laboratories). Los datos se expresan como pg de proteína / ml de plasma.

Capítulo 3

Herramientas para el análisis de los datos

Como se ha mencionado en el capítulo anterior, el estudio está constituido por un total de 160 sujetos en el que se diferencian dos subpoblaciones principales: el grupo control y el grupo con trastorno por consumo de cocaína. Tras las entrevistas y análisis de sangre realizados, se obtuvieron diferentes variables de medida con la que poder realizar el análisis: socio-demográficas, psiquiátricas, adicción, patrones de consumo y concentración molecular en plasma. La aplicación que se ha desarrollado analiza cada una de estas variables mediante unas técnicas estadísticas, que se exponen a continuación, y un análisis de clustering, explicado a posteriori.

3.1. Técnicas estadísticas

A continuación se van a definir las técnicas estadísticas que se han utilizado para analizar los datos de este proyecto:

3.1.1. Prueba de Kolmogorov-Smirnov

Se trata de una prueba no paramétrica[1] que determina la bondad de ajuste[2] de dos distribuciones de probabilidad entre sí. En concreto se utiliza para comprobar si una muestra sigue una distribución normal[3] al compararla con una muestra que sí es normal, fijada por el estadístico que se usa en esta prueba. Se asume la hipótesis nula de que las distribuciones (la que aportamos como estudio y la del estadístico que incluye la prueba) son iguales y normales. Si el P-valor resultante de la prueba de normalidad es inferior a un cierto nivel de significación (típicamente 0.05), es poco probable que las diferencias entre la distribución de estudio y la del estadístico se hayan producido sobre la base de un muestreo aleatorio de una población que sí era normal. Por lo tanto, la hipótesis nula de igualdad de distribuciones se rechaza y se acepta la hipótesis alternativa de que las distribuciones no son iguales, y por consiguiente la distribución de estudio no sigue una distribución normal o de Gauss. Cuando lo que se desea es verificar la normalidad de una distribución, la prueba de Lilliefors presenta algunas mejoras frente a la prueba de Kolmogorov-Smirnov. Cabe decir que el test de Shapiro-Wilk o la prueba de Anderson-Darling son alternativas más potentes a la prueba de Kolmogorov-Smirnov. Sin embargo, esta última es más sensible a los valores cercanos a la mediana que a los extremos en la distribución, y es por ello que se optó por ella, pues en el estudio que se lleva a cabo hay presencia de valores atípicos, más comúnmente conocidos como 'outliers'.

3.1.2. Prueba de Lilliefors

Como se indica en el párrafo anterior, la prueba de Lilliefors es una prueba de normalidad basada en la prueba de Kolmogorov-Smirnov. Con ella se calcula la discrepancia máxima entre la distribución de datos que se aporta y la distribución de comparación, que en este caso es una normal. En la aplicación realizada, se lleva a cabo esta prueba en varias ocasiones. Se analiza así la normalidad de diferentes variables, cada una clasificada según su procedencia de acuerdo a :

- Variables sociodemográficas: Edad e IMC.
- Variables de adicción a la cocaína: Edad de Prueba, Tiempo de Abstinencia, Criterios de Gravedad (continua).
- Variables moleculares: Eotaxin-1, IFNγ, IL-4, IL-8, IL-17α, MIP-1α y TGFα, ver figuras (3.1 y 3.2).

[1] No se conocen los parámetros de la población de la cual la muestra es extraída. Únicamente se tienen las distribuciones que se desean comparar.

[2] Lo bien que describe el modelo estadístico a un conjunto de observaciones.

[3] También conocida como Campana de Gauss.

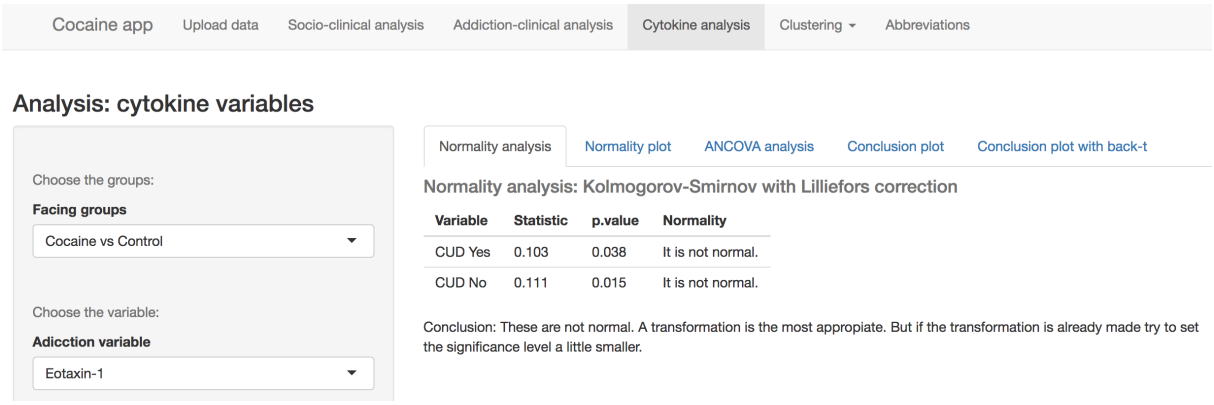


Figura 3.1: Aplicación: test de Kolmogorov-Smirnov con la corrección de Lilliefors para la variable molecular Eotaxin-1 comparando pacientes con TCC y sujetos control.

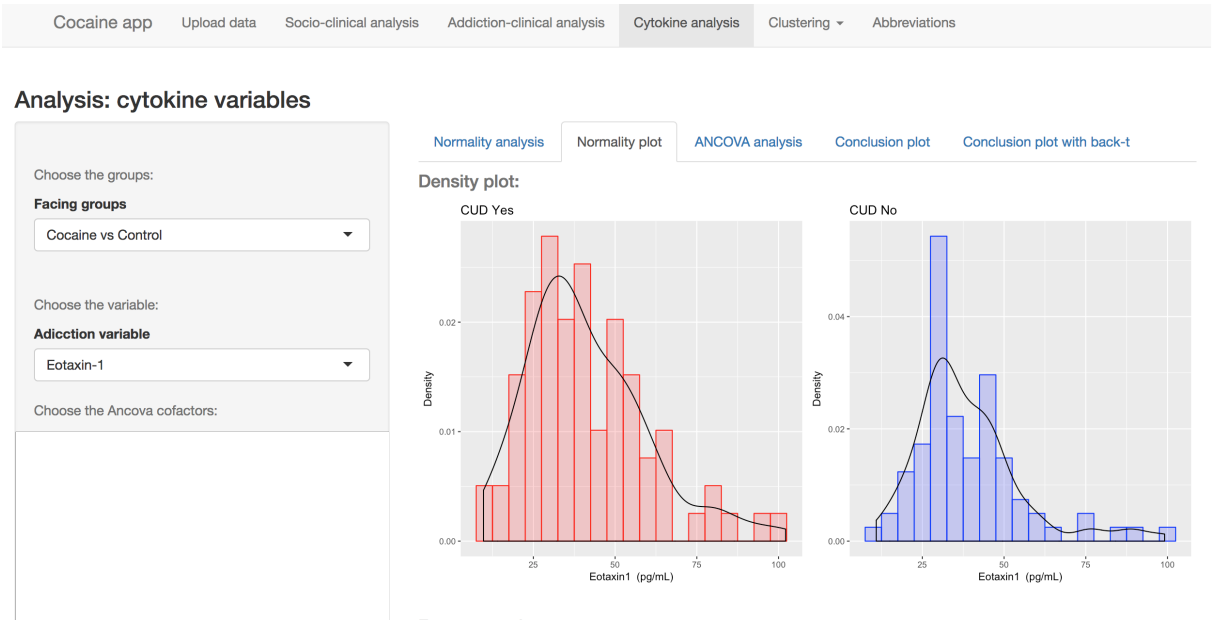


Figura 3.2: Aplicación: distribución de los datos según la variable molecular Eotaxin-1 comparando pacientes con TCC y sujetos control.

El valor se significación que suele usarse para determinar si la distribución a estudiar sigue o no una Campana de Gauss es de 0.05. Este valor es fijado por el usuario dependiendo de lo estricta que quiera llevar a cabo la prueba de Lilliefors, pudiendo aumentar el valor (sería más restrictiva) o disminuirlo. Pero la realidad, y en concreto el estudio de los datos relacionados con este proyecto, implica que, en algunas ocasiones, una distribución se

considere normal con una significación menor a 0.05. En este sentido, la aplicación desarrollada considera este aspecto incluyendo la posibilidad de modificar de forma interactiva el valor de significación, proporcionando valores de los resultados de forma inmediata. Continuando con el ejemplo anteriormente expuesto de la molécula Eotaxin-1, puede verse el cambio del tipo de distribución que siguen los pacientes diagnosticados con TCC, ver figura (3.3), al modificar el nivel de significación a 0.037.

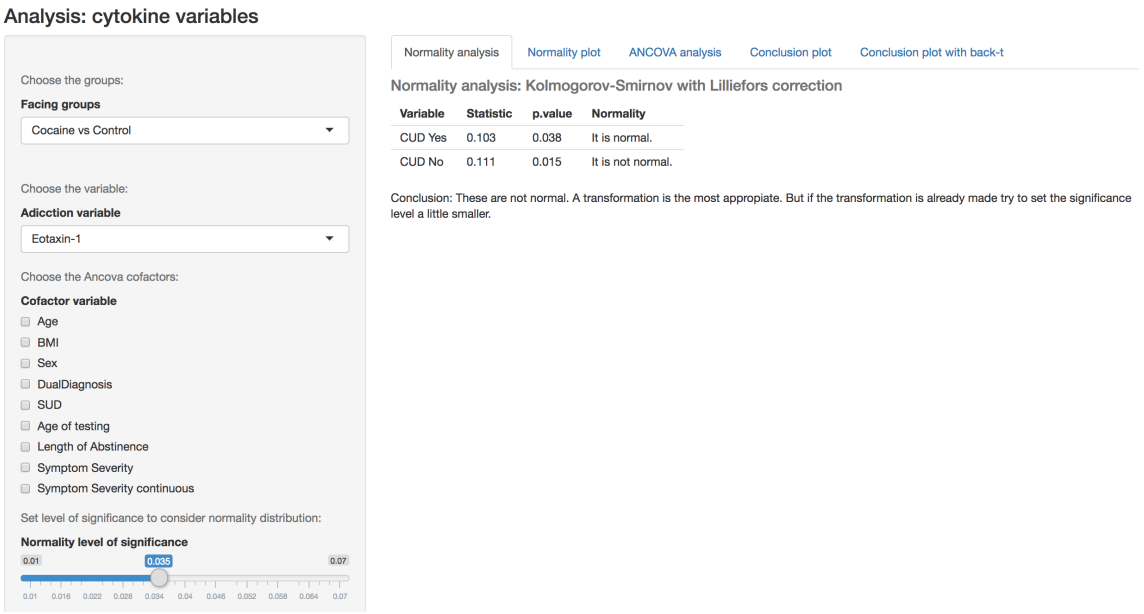


Figura 3.3: Aplicación: test de Kolmogorov-Smirnov con la corrección de Lilliefors para la variable molecular Eotaxin-1, modificando el nivel de significación a 0.037

3.1.3. Prueba de Levene

Es una prueba estadística inferencial que se usa para evaluar la homocedasticidad[4], es decir, la igualdad de las varianzas para una variable calculada para dos o más grupos. En este proyecto la población a estudiar se divide siempre en dos subpoblaciones, por lo que la prueba de Levene es para dos grupos.

Se pone a prueba la hipótesis nula de que las varianzas poblacionales son iguales. Si el P-valor resultante de la prueba de Levene es inferior a un cierto nivel de significación (típicamente 0.05), es poco probable que las diferencias obtenidas en las variaciones de la muestra se hayan producido sobre la base de un muestreo aleatorio de una población con varianzas iguales. Por lo tanto, la hipótesis nula de igualdad de varianzas se rechaza y se concluye que hay una diferencia entre las variaciones en la población. Como antes se mencionaba, el valor de significación para que dos distribuciones se consideren homocedásticas

[4]También conocida como homogeneidad

se viene fijando en 0.05. Pero nuevamente, la aplicación aporta interactividad para que este valor pueda ser modificado. La homocedasticidad es un requisito a cumplir en algunos de las pruebas que se explican a continuación, por lo que en ocasiones es necesario modificar el valor de significación. Para modificar este parámetro se ha elegido el botón de tipo deslizante, ver figura (3.4), y el resultado de que efectivamente, en el caso de enfrentar sujetos control y pacientes CUD en la variable IMC, la variable es homocedástica.

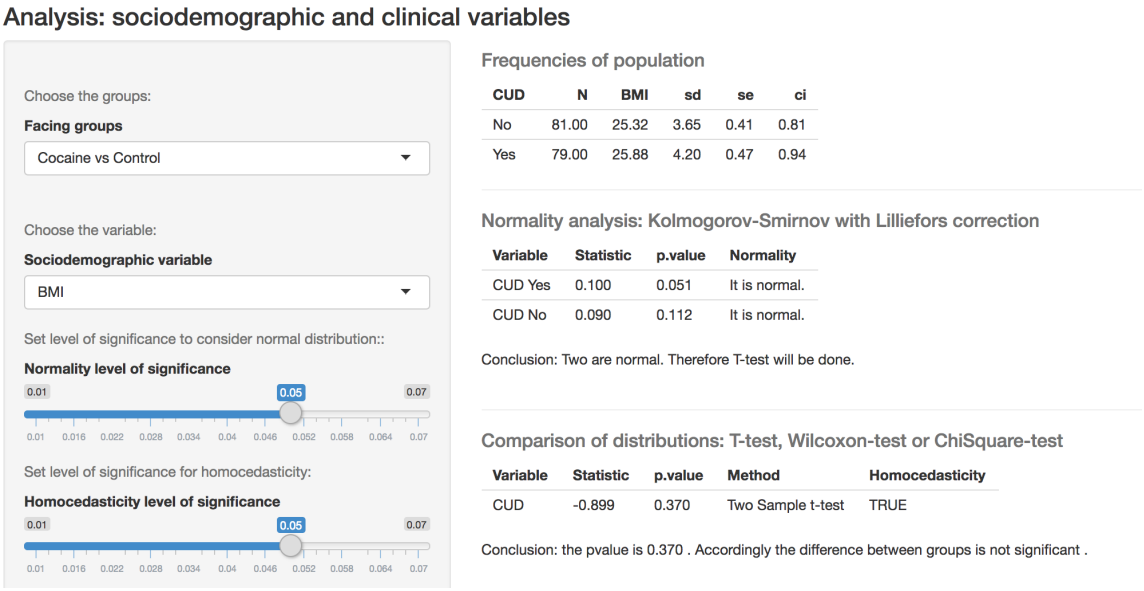


Figura 3.4: Aplicación: sujetos control enfrentados a pacientes CUD en la variable IMC, que es homocedástica.

3.1.4. Prueba t de Student

La prueba t de Student [5] es una prueba de contraste donde el estadístico utilizado tiene una distribución tipo t de Student. Se aplica cuando la población estudiada sigue una distribución normal pero el tamaño muestral es demasiado pequeño como para que el estadístico en el que está basada la inferencia esté normalmente distribuido, utilizándose una estimación de la desviación típica en lugar del valor real. La prueba t-Student se aplica con el fin de determinar la existencia, o no, de diferencias entre dos medias muestrales. Además aporta información para la construcción del intervalo de confianza, cumpliendo así el objetivo de la prueba: determinar si hay diferencias entre las medias de dos poblaciones. Esta prueba es llevada a cabo cuando se desconoce la desviación típica de una población, siendo ésta estimada a partir de los datos de una muestra. Existen varias maneras de asignar la hipótesis alternativa, y en este estudio optaremos por la hipótesis alternativa de dos colas. Esto quiere decir que la diferencia de medias que se asume en la hipótesis alternativa puede se por defecto o por exceso, por lo que al hacer la prueba t-Student

[5]También llamada prueba t-Student, o Test-T

buscamos lo que pasa en la cola inferior o en la cola superior, véase imagen 3.5. Suponiendo un nivel de significación del 5 % (o 0.05), la cola inferior sería el 2,5 % y la cola superior el otro 2,5 % (0.025 para cada cola).

Figura 3.5: Distribución normal. División de dos colas.

El valor de significación para que dos distribuciones se consideren similares en medias típicamente toma el valor de 0.05. Pero la interactividad de la aplicación permite que este valor pueda ser modificado por medio de un botón de tipo deslizante, ver figura (3.6). Efectivamente, en el caso de enfrentar pacientes con adicción a cocaína que tienen patología dual asociada o no, se comparan las medias de las dos muestras mediante el la prueba t-Student, pues ambas muestras siguen una distribución normal. La prueba indica que el resultado no es significativo, pues el valor del P-valor es mayor que el fijado en el botón tipo deslizante ($0.089 > 0.05$).

Un requisito crucial para poder aplicar esta prueba a dos muestras es que ambas tengan una distribución normal (por ello que se haga la prueba de Lilliefors con anterioridad a ésta). Una vez asegurada la normalidad de las muestras, se continúa comprobando la homocedasticidad de las mismas(prueba de Levene). Dependiendo de si las muestras son homocedásticas o no, el valor resultado de la prueba difiere. En el caso de que las muestras tengan varianzas diferentes, la varianza es estimada separadamente para ambos grupos y se usa la modificación de Welch[6] de los grados de libertad. Por el contrario, en el caso de ser las dos pruebas homocedásticas, para el cálculo del estadístico se utiliza el valor de lo que se conoce como 'varianza pooled', véase ecuación (3.1).

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{3.1}$$

3.1.5. Prueba de Wilcoxon-Mann-Whitney

La prueba de Wilcoxon-Mann-Whitney [7] es la versión no paramétrica de la prueba t de Student, pues no se puede asumir la normalidad de las muestras. Esta prueba se utiliza para comprobar la heterogeneidad de dos muestras ordinales. En este estudio se utiliza cuando se quieren analizar las medias de dos muestras. El planteamiento de partida es: las observaciones de ambos grupos son independientes. Además la variable de estudio es

[6]No es correcto usar el estadístico t para muestras independientes, por lo que se lleva a cabo el test de Welch, que usa un estadístico t diferente.

[7]También llamada prueba U de Mann-Whitney o prueba de suma de rangos Wilcoxon.

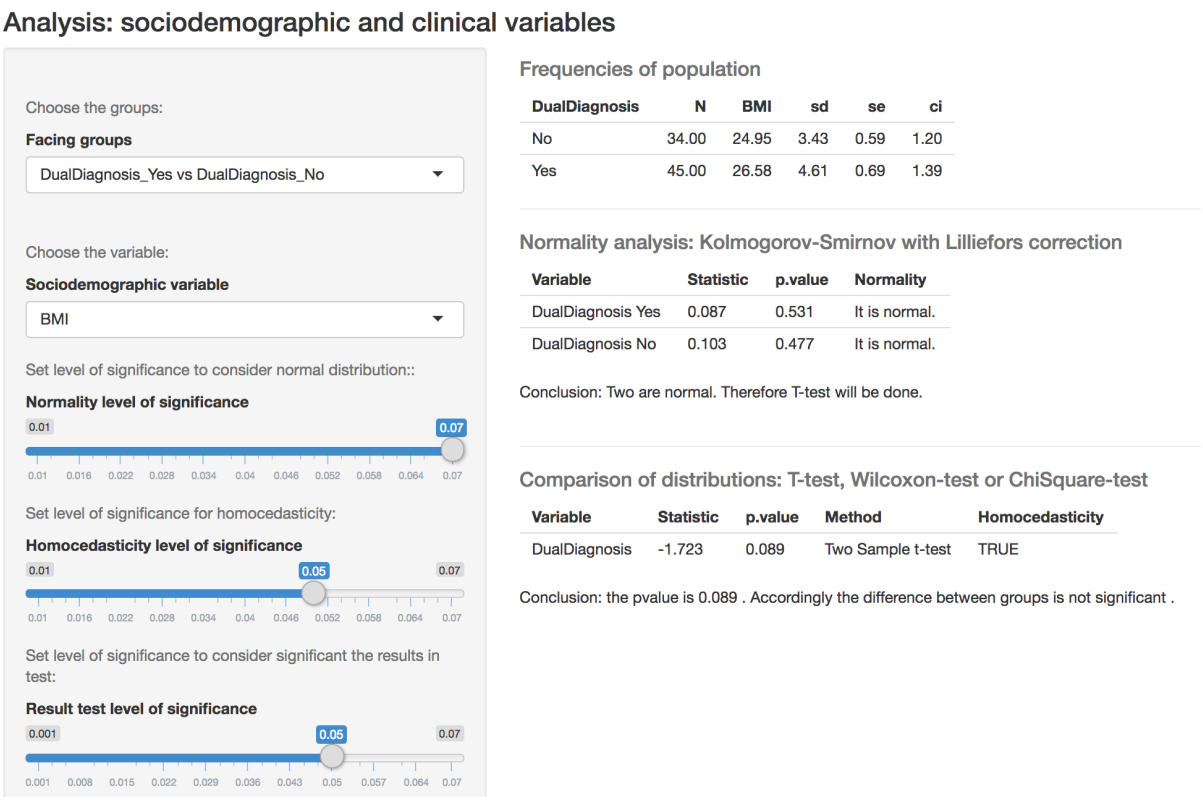


Figura 3.6: Aplicación: los datos presentan distribuciones normales.

continua y medida en escala ordinal, de intervalo o razón. En este estudio las variables que se estudian con este método variables con escala ordinal (Nivel de Gravedad continua) y de razón (Edad, IMC, Tiempo de abstinencia, Edad de prueba). Se utiliza la versión de datos no pareados.

Se pone a prueba la hipótesis nula de que las medianas poblacionales son iguales. Si el P-valor resultante de la prueba de Wilcoxon-Mann-Whitney es inferior a un cierto nivel de significación (típicamente 0.05), es poco probable que las diferencias obtenidas en las variaciones de la muestra se hayan producido sobre la base de un muestreo aleatorio de una población con medianas iguales. Por lo tanto, la hipótesis nula de igualdad de medianas se rechaza y se concluye que hay una diferencia entre las medianas en la población. El valor de significación para que dos distribuciones se consideren similares en medianas típicamente toma el valor de 0.05. Pero la interactividad de la aplicación permite que este valor pueda ser modificado,véase figuras (3.7) y (3.8). Se aprecia cómo el botón de tipo deslizante modifica la conclusión del resultado. En el caso de enfrentar pacientes con adicción a otra sustancia (además de a la cocaína) o no, se comparan las medianas de las dos muestras mediante el la prueba de Wilcoxon-Mann-Whitney, pues una de las muestras

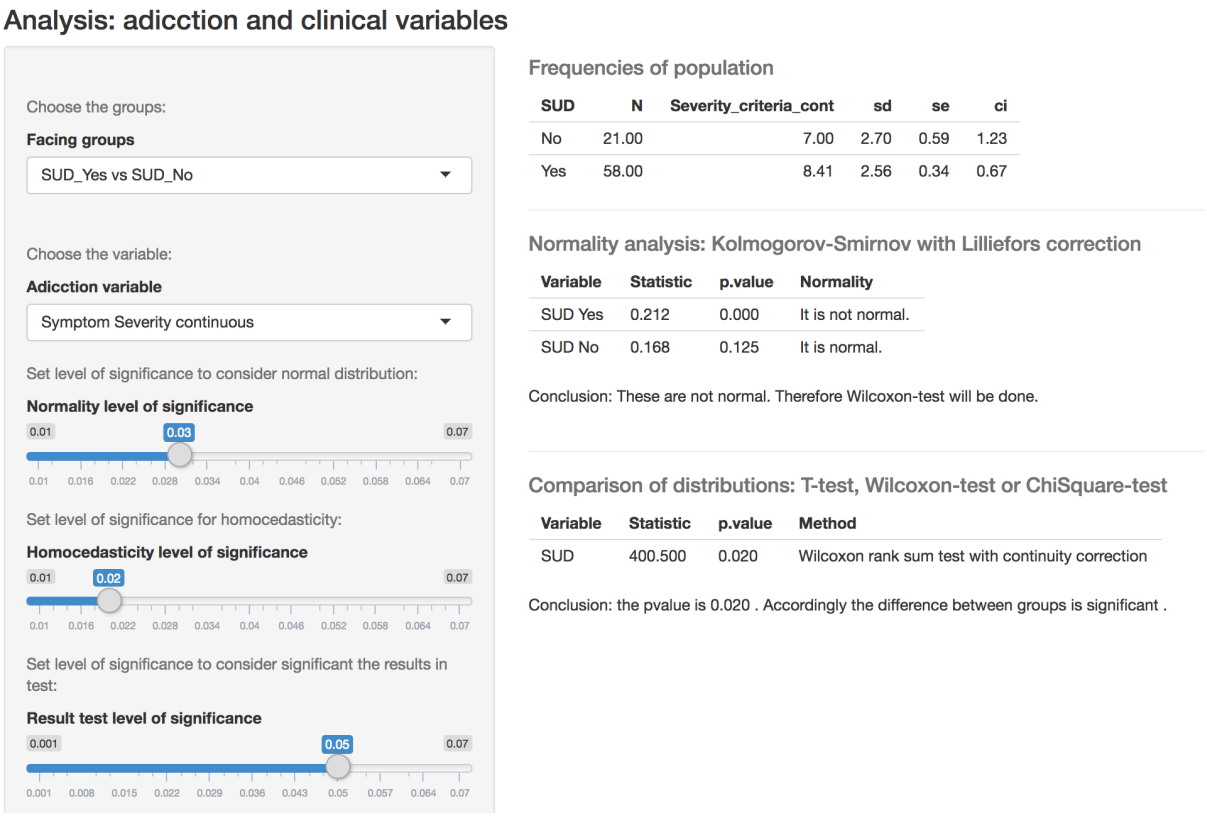


Figura 3.7: Aplicación: prueba de Wilcoxon-Mann-Whitney: resultado significativo

no sigue una distribución normal. La prueba indica que el resultado es o no significativo, dependiendo del valor fijado en el botón tipo deslizante.

3.1.6. Prueba χ^2

La prueba χ^2, llamada también ji-cuadrado o chi-square, es una prueba de independencia para determinar si el valor observado de una variable categórica depende del valor observado de otra variable, ver figura (3.9). La hipótesis nula postula que las dos variables a estudiar son independientes, indicando por tanto, que una no influye en el resultado de la otra. Si el p-valor resultante de la prueba de χ^2 y es inferior a un cierto nivel de significación (normalmente 0.05), es poco probable que las diferencias obtenidas en las variaciones de la muestra se hayan producido sobre la base de un muestreo aleatorio de una población con muestras independientes. Por lo tanto, la hipótesis nula de que las variables son independientes se rechaza y se concluye que hay una relación entre las variables en la población de estudio, por lo que las variables son dependientes. Es decir, se rechaza la hipótesis nula si los valores observados difieren mucho de los esperados.

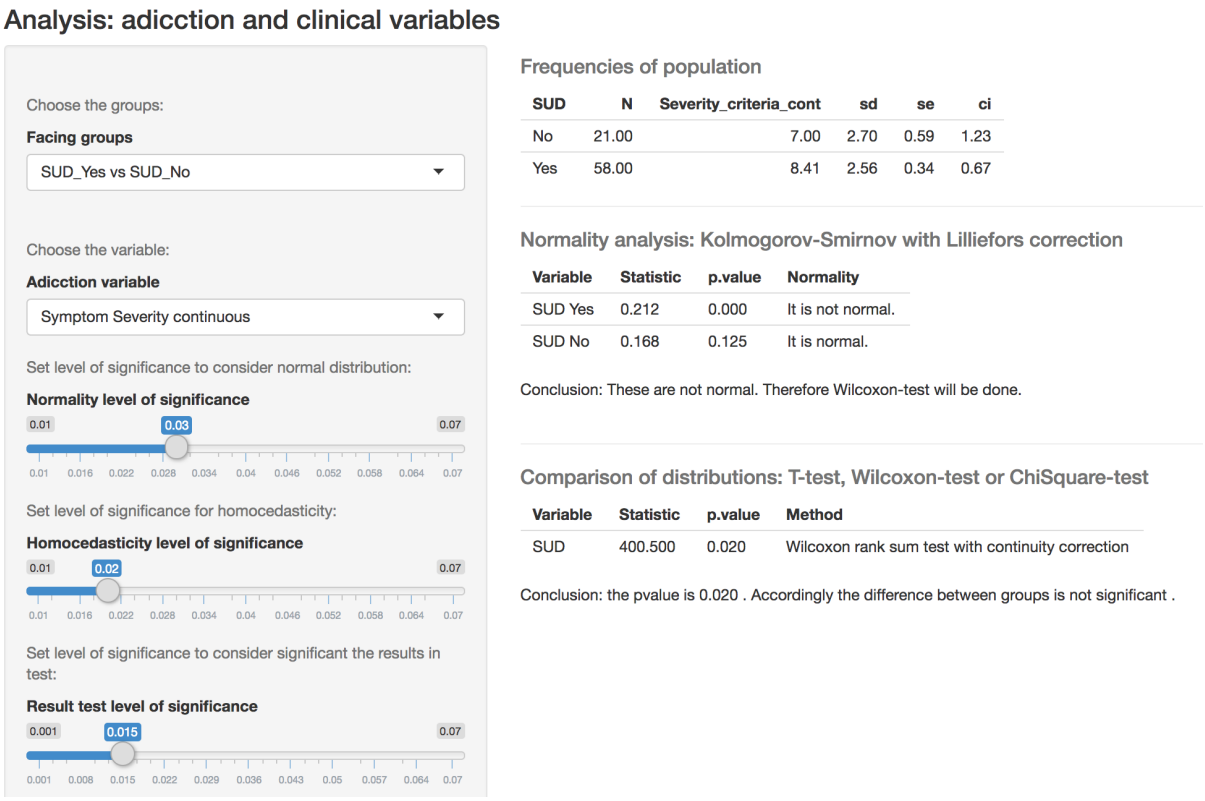


Figura 3.8: Aplicación: prueba de Wilcoxon-Mann-Whitney con resultado no significativo

3.1.7. ANCOVA

Es el acrónimo de ANálisis de COVArianza. Se trata de un modelo lineal general (MLG[8]) con una variable cuantitativa y uno o más factores. El ANCOVA es una unión del ANOVA (ANálisis de VArianza) y de la regresión lineal múltiple[9]. Por tanto se puede decir que el fundamento del ANCOVA es un ANOVA al que a la variable dependiente se le ha eliminado el efecto predicho por una o más covariables por regresión lineal múltiple. Hay que tener en cuenta que la inclusión de covariables puede aumentar la potencia estadística porque a menudo reduce la variabilidad. Se elige ANCOVA y no ANOVA porque el análisis de covarianza es apropiado para lograr dos objetivos específicos: a) eliminar cualquier error sistemático fuera del control del investigador que puede sesgar los resultados, y b) tener en cuenta las diferencias en las respuestas debidas a las características propias de los sujetos.

En el análisis de ANCOVA es posible cualquier número de factores (variables categóricas) y covariables (variables numéricas), ver figura (3.10). Lo que el investigador hace es

[8]Es una generalización de la regresión lineal ordinaria. Los MLG unifican varios modelos estadísticos, como regresión lineal, regresión logística o regresión de Poisson.

[9]Es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente y la variable independiente (si son varias se denomina múltiple)

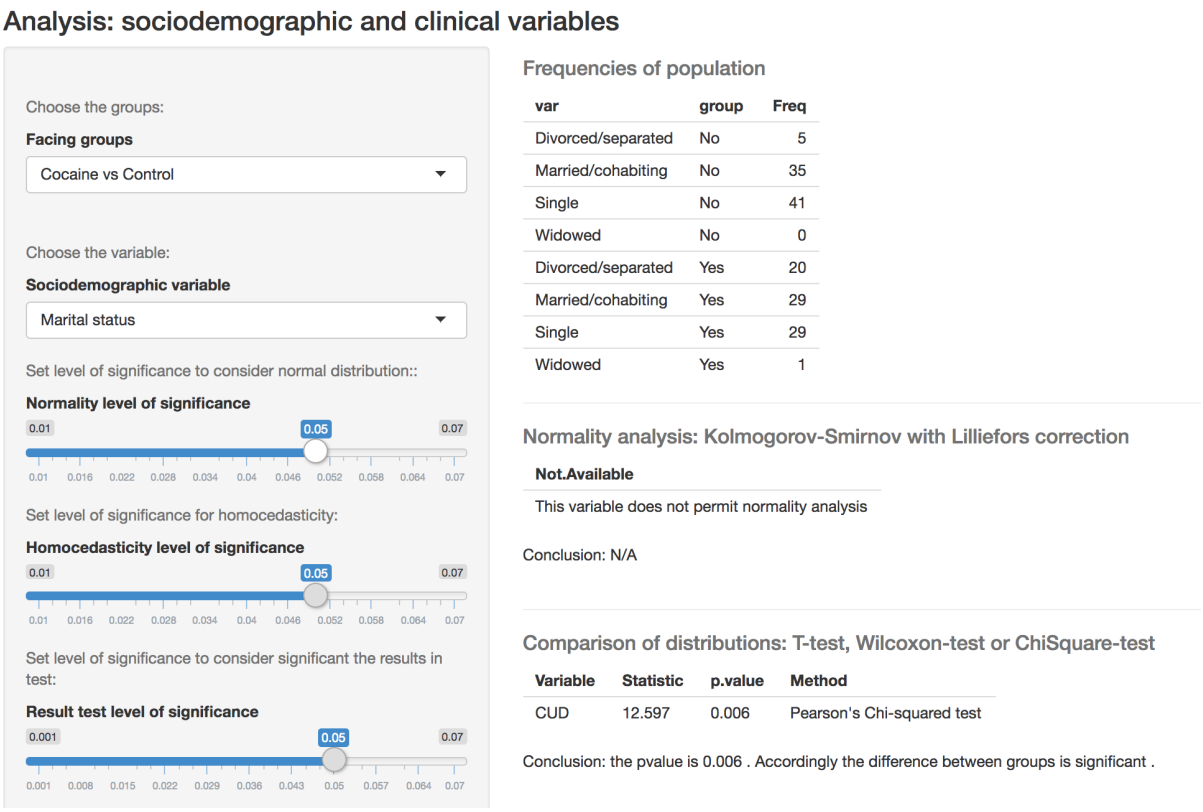


Figura 3.9: Aplicación: prueba χ^2 donde el estado civil guarda relación con la adicción del sujeto.

eliminar la heterogeneidad causada en la variable de interés, añadiendo aquellas variables independientes, ya sean factores o covariables. Una vez seleccionadas las variables y hecho el análisis, el investigador interpreta los resultados, prestando atención a aquella variable independiente que sea de su interés, y deja el resto de variables independientes como variables 'correctoras' para el análisis de ANCOVA. Las ANCOVAs utilizadas en este proyecto se han aplicado para contrastar la expresión de una citoquina en dos subpoblaciones, por lo que la variable de interés del investigador es la variable independiente categórica que actúa como factor en la ANCOVA, diferenciando a los sujetos en las dos poblaciones. Para aclarar estos conceptos se expone un ejemplo, ver figura 3.11. Para analizar si la expresión de la citoquina IFNγ depende de la población a la que pertenece, una que tiene patología dual y otra que no, se hace un análisis de ANCOVA indicando a la variable diagnóstico dual como factor. Por tanto el objetivo es comprobar si la variable diagnóstico dual tiene un efecto significativo o no en el análisis de ANCOVA (comúnmente un valor menor a 0.05). En el marco de la ventana se genera un modelo y su ANCOVA en la parte superior, donde la variable dependiente es la expresión de la citoquina, y la variable, y única en este caso, independiente es la patología dual. Esto se hace para que el investigador siempre

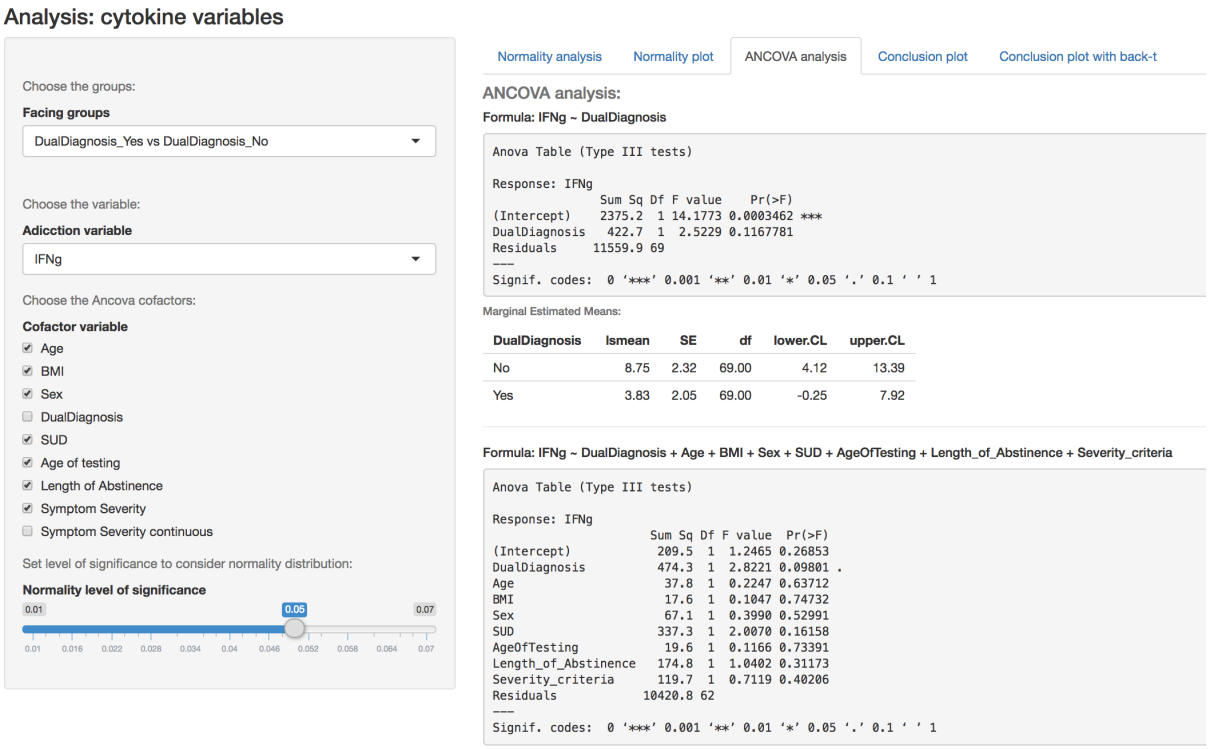


Figura 3.10: Aplicación: ANCOVA. Variable dependiente: IFNγ. Variables independientes: cofactores sexo, diagnóstico dual, trastorno por consumo de otra sustancia y síntomas de gravaedad; covariables: edad, IMC, edad de prueba de la droga y tiempo de abstinencia.

tenga en la parte superior la referencia de una ANCOVA base, con el único cofactor de interés. Cabe decir que en este proyecto, cuando se hace el análisis de ANCOVA con una sola variable independiente, ésta es siempre categórica, por lo que realmente se hace es un análisis de ANOVA, para ser políticamente correctos. El resultado de la ANCOVA, y por tanto, la conclusión que se va buscando, difiere si se consideran unas variables independientes u otras en el modelo, y por tanto más covariables y/o cofactores en la ANCOVA. Es por ello que las variables deben ser elegidas con cierto criterio y razonamiento previos, ver figura (3.11). Para el análisis que se está llevando a cabo, las variables de la Edad, Sexo e IMC del sujeto son imprescindibles para hacer el análisis ANCOVA de las moléculas que tienen lugar (referencia paper). Este ejemplo es una ANCOVA de dos vías porque incluye dos factores. Y es que el objetivo de la covarianza, y por tanto el usar ANCOVA, es eliminar cualquiera de los efectos que a) influyen solamente a una parte de los sujetos, b) varían entre los sujetos. Por tanto investigador utiliza una covarianza para extraer cualquiera de las diferencias debidas a estas variables antes de que los efectos del experimento sean calculados. En resumen, el análisis de ANCOVA en esta aplicación estudia la relación de una variable cualitativa (factor) con una variable cuantitativa (variable dependiente molecular), eliminando la influencia de una terceras variables (cuantitativas

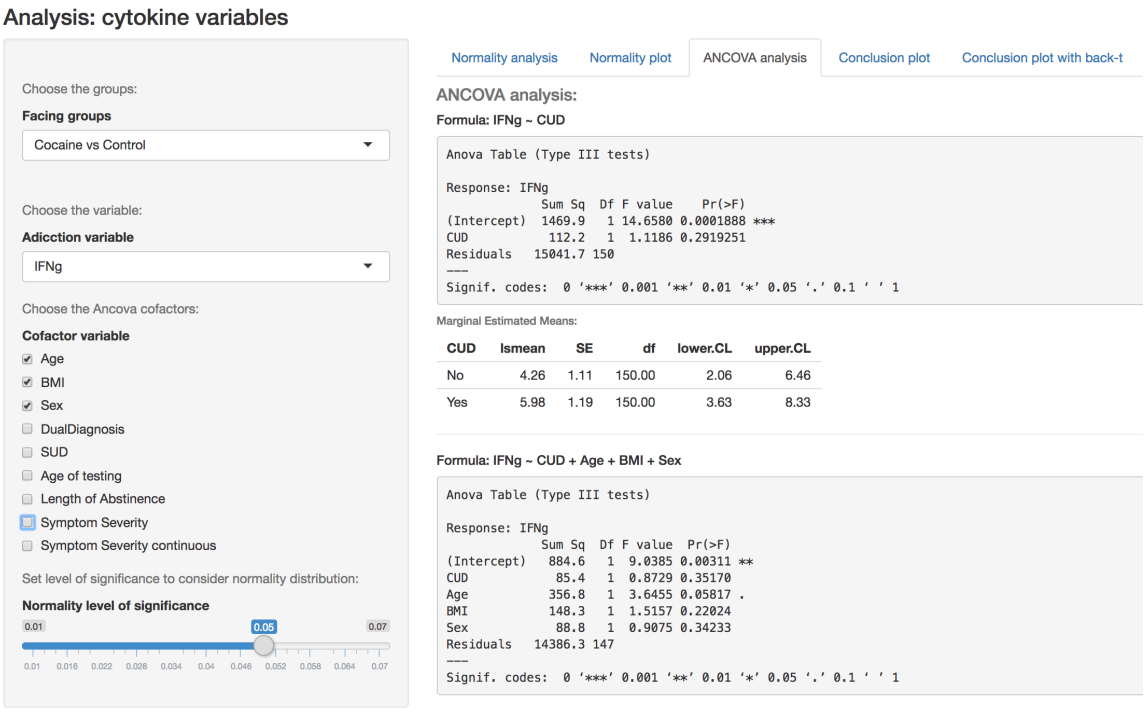


Figura 3.11: Aplicación: ANCOVA. Variable dependiente: IFNγ. Variables independientes: cofactores sexo y diagnóstico dual; covariables: edad e IMC.

y cualitativas). Para llevar a cabo el análisis de la covarianza es necesario que los datos cumplan una serie de supuestos y consideraciones. Se exponen a continuación los supouestos, los cuales se deben cumplir estrictamente para que el ANCOVA sea válido, y alguno de las consideraciones más importantes.

Supuestos: el conjunto de datos debe cumplir normalidad, homocedasticidad e independencia.

- La normalidad se evalúa con la prueba de Kolmogorov-Smirnov con la corrección de Lilliefors, ver apartado (3.1.2). En la aplicación se permite comprobar este supuesto, que es el más destacable de los tres, para evaluar la validez del ANCOVA, ver figura (3.12). Por ello se permite el tipo de trasformación logarítmica de las variables moleculares, para que de esa forman sigan, o se asemejen, a una distribución normal, ya que sin la trasformación, ninguna citoquina sigue una Campana de Gauss.

Consideraciones:

- Las covariables deben poseer una relación lineal con la variable de respuesta. De no ser así, se debe aplicar una transformación para convertir a relación en lineal.
- Debe cumplirse una regresión de homogeneidad entre grupos, donde las líneas de regresión son paralelas.

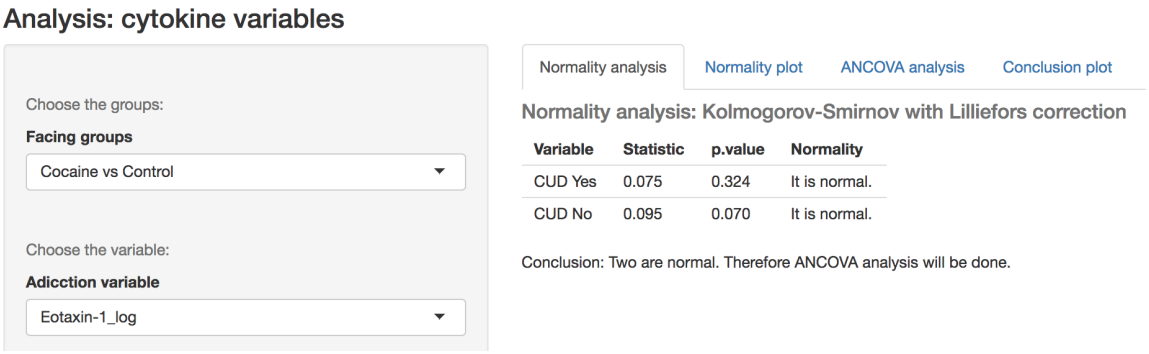


Figura 3.12: Aplicación: análisis de normalidad de la variable dependiente Eotaxin-1 con trasformación logarítmica, comparando pacientes con TCC y sujetos control.

Pero para el tipo de análisis que se lleva a cabo, el ANCOVA es suficientemente robusto como para que no cumpla todas condiciones anteriores, pudiendo incluso no cumplimentar los supuestos. En la aplicación se permite comprobar la normalidad de las variables moleculares antes de realizar sus análisis de ANCOVA, pero el nivel de significación puede ser menor a lo típicamente usado (0.05) y los resultados seguirán siendo adecuados. Pues en este tipo de estudio hay poco error cuando no se satisfacen todos los supuestos siempre y cuando el nivel de significación no sea excesivamente bajo. El tipo de suma de cuadrados que se usa en los ANCOVAs de la aplicación es el tipo III, porque es el tipo que se suele usar cuando los datos de los que se dispone para el estudio no están balanceados, i.e., existe un diferente número de observaciones para cada grupo. Se trata de una suma de cuadrados marginal porque se tiene en cuenta todos los factores en todo momento (ver figura 3.13). Sean dos factores del modelo A y B: A se ajusta para B y para A∗BA⋆B, B se ajusta para A y para A⋆BA⋆B y, por último, A⋆BA⋆B se ajusta para A y B.

Término	SS tipo III
A	$R(B, AB) - R(A, B, AB)$
B	$R(A, AB) - R(A, B, AB)$
AB	$R(A, B) - R(A, B, AB)$

Figura 3.13: Suma de cuadrados tipo III.

Este tipo de sumas de cuadrados calculan la reducción en la suma de cuadrados residual para un efecto tras haber ajustado todos los demás efectos en el modelo. Para estimar la suma de cuadrados para cada uno de los coeficientes del modelo se hace lo siguiente: al modelo completo (una vez que hemos estimado los coeficientes del mismo) se le quita una

variable y se estima la suma de cuadrados de esta variable calculando la diferencia entre la suma de cuadrados explicada del modelo completo y la suma de cuadrados explicada del modelo al que se le ha extraido dicha variable. Por tanto, las sumas de cuadrados son las que se obtienen cuando cada variable se añade la última en el modelo. Es decir, el efecto de cada variable es el que se calcula después de que todos los otros factores ya se han tenido en cuenta. Entre las ventajas de este tipo de sumas de cuadrados hay que destacar que no dependen del tamaño muestral de cada grupo y se pueden utilizar en modelos no balanceados, como se mencionaba anteriormente. Cuando no hay casillas vacías, las medias en cada subpoblación son las medias de mínimos cuadrados [10], que son los mejores estimadores lineales insesgados (BLUE) para las medias marginales del diseño.

Debe quedar claro que considerar interacciones a la hora de generar el modelo difiere de hacer interacciones en el análisis de la covarianza. En este estudio, como se ha explicado en este apartado se han considerado interacciones en el modelo, al seleccionar ANCOVA tipo III. Pero no se han estudiado las interacciones entre los factores del modelo, ver figura (3.11) (en el caso de la figura entre CUD y Sex) debido a los siguientes argumentos:

- El objetivo del estudio, en un principio, está centrado en TCC. Por tanto las otras covariables, incluido el sexo, solo se emplean para eliminar interferencias. Incluir la interacción del sexo con TCC es desviar la atención hacia una variable secundaria, tal y como se plantea inicialmente el estudio.
- Una segunda razón, más técnica quizás, es que se parte de una proporción de hombres y mujeres bastante desigual en los grupos experimentales, y los efectos de interacciones entre sexo y TCC sobre estas moléculas pueden ser erróneas. La desproporción en el número de hombres y mujeres parte de la propia naturaleza exploratoria del estudio, pues se reclutan a los pacientes posibles y la proporción de mujeres es bastante más baja.

3.1.8. Medias marginales estimadas

Las Medias marginales estimadas, también conocidas por Medidas marginales predichas o Medias por mínimos cuadrados, para un modelo lineal son simplemente predicciones - o promedios de los mismos - sobre una rejilla regular de configuraciones de predicción que es llamada rejilla de referencia. Por tanto estas medias se calculan referentes a la rejilla, la cual es establecida con anterioridad. Puesto que la rejilla de referencia es fundamental, es el punto de partida. Para cada predictor en el modelo, se define un conjunto de uno o más niveles de referencia. La cuadrícula de referencia es entonces el conjunto de todas las combinaciones de niveles de referencia. Si no se especifica explícitamente, los niveles de referencia se obtienen mediante:

- Para cada predictor que es un factor, su nivel de referencia son los únicos niveles del factor en cuestión.

[10] También denominada medias magrinales estimadas o medias marginales predichas.

- Cada predictor numérico tiene un nivel de referencia, la media sobre el conjunto de datos.

Así que la rejilla de referencia depende tanto del modelo como del conjunto de datos. En la aplicación se puede apreciar la rejilla de referencia que se usa para los análisis de ANCOVA, facilitando así la comprensión de estos conceptos, ver figura (3.14).

Dejando a la rejilla a parte, se expone un ejemplo como aclaración del concepto, ver figura (3.14). Las medias por mínimos cuadrados representan el valor medio predicho de la variable dependiente (TGFα) en un nivel especificado del factor categórico (DualDiagnosis) que es la variable independiente cuantitativa de interés, cuando todas las variables cuantitativas son fijas igual a sus medias observadas (Age e BMI) y todas las variables indicadoras para los factores categóricos (Sex) son fijados iguales a 0. Cada media es mostrada junto con su estimación del error estándar y un intervalo de confianza.

Formula: TGFa ~ DualDiagnosis + Age + BMI

```
Anova Table (Type III tests)

Response: TGFa
              Sum Sq Df F value  Pr(>F)
(Intercept)     1.52  1  0.0600 0.80895
DualDiagnosis 131.78  1  5.2107 0.03353 *
Age            36.02  1  1.4243 0.24666
BMI             6.31  1  0.2495 0.62290
Residuals     505.79 20
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Marginal Estimated Means with cofactors:

DualDiagnosis	lsmean	SE	df	lower.CL	upper.CL
No	5.57	1.61	20.00	2.21	8.93
Yes	0.72	1.36	20.00	-2.11	3.55

Cofactors means:

DualDiagnosis	Age	BMI
No	33.88	25.45
Yes	33.88	25.45

Figura 3.14: Aplicación: 'Cofactors means' representa la rejilla de referencia para calcular Medias marginales estimadas en este estudio de TGFα considerando el diagnóstico dual como factor principal.

3.2. Algoritmos de clustering

Los algoritmos de clustering son un conjunto de técnicas pertenecientes al área de minería de datos (Duda et al., 2000; Maimon and Rokach, 2005). El objetivo del clustering o agrupamiento es principalmente descriptivo, intentando descubrir nuevos conjuntos de categorías, obteniendo nuevos grupos de interés principalmente de forma intrínsica, es decir de manera no supervisada, es decir sin utilizar etiquetas que indiquen las categorías de los datos. Las técnicas de clustering intentan crear grupos de datos separando el conjunto original en subconjuntos de manera que patrones similares are agrupados conjuntamente, y de esta forma obtener una representación eficiente del conjunto que categorice la población bajo estudio. En nuestro caso, con las técnicas de clustering podemos agrupar sujetos (en nuestro estudio se trata de pacientes o controles) en regiones donde la similitud mutua es elevada, de forma que cada grupo sea homogéneo y distintos de los demás. A cada grupo se le denomina cluster, y son puntos en el espacio donde la "distancia" entre cada par de sujetos es menor que la distancia de cualquiera de ellos a otro sujeto que no pertenece al cluster. Para el cálculo de esta distancia entre sujetos pueden utilizarse diferentes medidas, siendo muy utilizadas por ejemplo el coeficiente de correlación estadística (rango de valores entre -1 y +1) o la distancia euclidiana (la raíz cuadrada de la suma de las diferencias cuadradas entre los valores de características correspondientes). A continuación se exponen los métodos de agrupamiento utilizados en la aplicación desarrollada en este proyecto. En cada apartado se expone en qué consiste el método y el por qué de su elección para este estudio.

3.2.1. Clustering jerárquico

Se ha incluido esta técnica de agrupamiento porque se ha convertido casi en un estándar dado los buenos resultados obtenidos en diversos estudios. En el clustering jerárquico los sujetos se agrupan de manera arborescente, siendo la forma común de representarlos un dendograma[11], y es la que se ha adoptado en este proyecto. Para conectar los sujetos se utiliza un criterio de enlace, que determina la distancia entre conjuntos de observaciones como una función de las distancias entre observaciones dos a dos. Existen diversos criterios, y de todos ellos se han elegido 3 posibilidades, de manera que el usuario de la aplicación pueda elegir en el momento del análisis el criterio que más se adecúe a su problema.

Sean A y B dos grupos, y d la métrica escogida, los cuatro tipos de enlace permitidos en la aplicación desarrollada son:

Agrupamiento de máximo o completo enlace: La distancia (o similitud) entre dos clusters se mide atendiendo a sus elementos más dispares, por lo que la distancia (o similitud) entre clusters viene dada por la máxima distancia (o mínima similitud) entre sus componentes, ver ecuación (3.2).

$$max\{d(a,b) : a \in A, b \in B\} \tag{3.2}$$

[11]Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos.

Agrupamiento de mínimo o simple enlace: La distancia (o similitud) entre dos clusters viene dada por la mínima distancia (o máxima similitud) entre sus componentes, véase ecuación (3.3).

$$min\{d(a,b) : a \in A, b \in B\} \tag{3.3}$$

Agrupamiento de enlace media o promedio: [12] La distancia entre dos grupos es definida a partir del promedio de todas las posibles distancias entre los miembros de cada grupo. Este método es muy robusto, porque todos los miembros de un grupo participan en la construcción de la distancia entre grupos, véase ecuación (3.4).

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b) \tag{3.4}$$

Un aspecto negativo del clustering jerárquico y de los métodos de clustering en general, es que el resultado es altamente dependiente del método de agrupamiento y la distancia empleada, lo que introduce cierta subjetividad en los resultados. Además, otro problema general de los algoritmos de clustering, es la definición del número de clusters apropiados para hacer una interpretación. Si bien existen diversas métricas que pueden ayudar en esta decisión , como el índice de silueta que se describe más abajo, es conveniente en el caso del clustering jerárquico analizar distintos valores para el número de grupos, y es por ello que la aplicación permite la elección interactiva del número de clusters directamente o a través de la elección del nivel de distancia de corte, permitiendo así analizar diferentes opciones posibles.

3.2.2. Algoritmo de K-medias (K-means)

Se trata de uno de los métodos clásicos de agrupamiento, pues es uno de los algoritmos más simples para resolver el clustering y es por ello que se ha incluido en este proyecto. El clustering es no jerárquico, generando así particiones a un solo nivel. El algoritmo de K-medias calcula, por medio de etapas sucesivas, los clusters que agrupan a los diferentes sujetos. Lo que caracteriza a este método es que el número de clusters es prefijado y para el cálculo de los mismos se usan los centroides. Para entender mejor el proceso de este algoritmo, se definen a continuación los pasos qinvolucrado (ver también la figura (3.15)).

1. Una vez prefijados el número de clusters K, el algoritmo sitúa K puntos en el espacio en el que se encuentran los n sujetos que se quieren clasificar. Estos puntos representan los centroides iniciales de los grupos.
2. Asigna cada sujeto al grupo que tiene el centroide más cercano.
3. Tras haber asignado todos los sujetos, vuelve a calcular las posiciones de los K centroides.
4. Divide el espacio en diferentes clusters, asignando a cada sujeto el grupo con el centroide más cercano.

[12] También llamado UPGMA

5. Repite los pasos 3 y 4 hasta que los centroides se mantengan estables. Esto produce una clasificación de los sujetos en grupos que permite dar una métrica entre ellos.

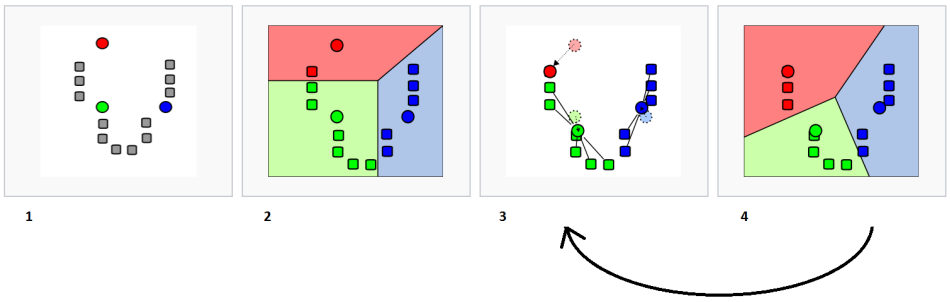


Figura 3.15: Proceso del algoritmo K-means.

El algoritmo de K-medias no alcanza en todos los casos la solución óptima, pues es muy sensible a las condiciones de partida (posición de inicial de centroides, número de clusters, etc.). Por tanto, para abarcar diferentes posibilidades y llegar a la mejor solución, la aplicación de este proyecto permite elegir de forma interactiva el número de clusters.

3.2.3. Red de Kohonen de mapas auto-organizados (SOM networks)

Un mapa auto-organizado[13] Kohonen et al. (2001) es una retícula finita (mapa bidimensional) formada por nodos que representarán a los sujetos de entrada. El entrenamiento se realiza a través de un esquema de aprendizaje competitivo, durante el cual para cada patrón de entrada (en nuestro caso datos de un sujeto) se calcula el nodo más próximo (nodo vencedor o neurona BMU[14]) mediante la distancia euclidiana a todos los vectores. Ese nodo, y sus vecinos, se adaptan para aproximarse al vector de entrada, siempre manteniendo la topología de la retícula. El mapa final, a través de sus nodos, agrupa los sujetos en grupos (clusters), representando cada nodo un cierto número de sujetos (ver figura (3.16)). Una gran ventaja de los algoritmos tipo SOM, y por eso su inclusión entre las posibilidades ofrecidas por la aplicación desarrollada, es que los nodos generados forman un representación topológica de los datos, ya que nodos cercanos (lejanos) representarán vectores similares (diferentes). Adicionalmente, la aplicación permite visualizar el número de sujetos representados por cada nodo, la calidad de cada uno de ellos y la distancia entre los mismos. Cabe mencionar que se ha demostrado que SOMs con un pequeño número de neuronas tienen un comportamiento similar al agrupamiento basado en el algoritmo de K-medias.

[13]Self-Organizing Maps (SOM) o red de Kohonen
[14]Del inglés 'Best Matching Unit'

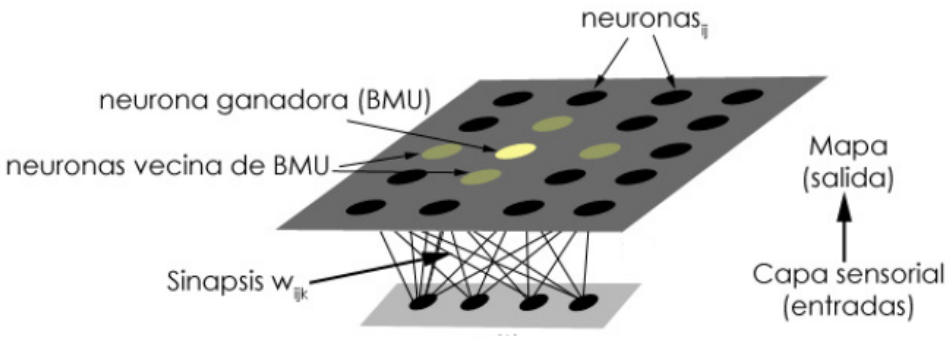


Figura 3.16: Estructura de un red auto-organizada (SOM network)

3.2.4. Índice de silueta

El índice de silueta es una métrica para evaluar el buen funcionamiento de los algoritmos de clustering, siendo la utilidad principal de este índice la identificación del número óptimo de agrupamientos. En los algoritmos de aprendizaje no supervisado, el número de clústeres puede ser un parámetro de entrada del algoritmo (K-means) o determinado automáticamente por el algoritmo (DBSCAN). En el primer caso la determinación del número óptimo de clústeres tiene que ser realizado mediante alguna medida externa al algoritmo. El coeficiente de Silueta para un conjunto de datos se define como la media del coeficiente de Silueta de cada objeto de la muestra, $s(i)$, siendo este coeficiente de Silueta para un objeto igual a:

$$si = \frac{b - a}{\text{máx}(a - b)} \tag{3.5}$$

- a es la distancia media entre el objeto y todos los otros objetos de la misma clase, y
- b es la distancia media entre el objeto y todos los otros objetos del clúster más próximo.

3.3. Tecnología Software

A continuación se van a exponer los diferentes softwares que se han usado para llevar acabo este proyecto. Se presenta una descripción general de éstos, además del justificante de por qué se han incluido en la aplicación.

3.3.1. R y Rstudio

Todos los análisis que se han llevado a cabo se han realizado con el software libre R versión 3.3.3, que utiliza el lenguaje R de programación (Verzani, 2004; Radziwill, 2015), junto con el entorno de desarrollo RStudio versión 10.0.44. Las librerías utilizadas se citan a continuación, mencionando que a la librería Shiny utilizada para el desarrollo específico de la aplicación se le reserva un apartado independiente:

- **stats:** ajustar un modelo de análisis de varianza: aov(), prueba Chi-cuadrada de Pearson para los datos del recuento chisq.test(), tablas de continjencia plana: ftable(), clustering jerárquico: hclust(), prueba t de Student: t.test(), prueba U de Mann-Whitney: wilcox.test().
- **car: (Meulders, 2013)** tablas de Anova para varios modelos estadísticos: anova().
- **kohonen:** función para crear de un super-mapa autoorganizado : supersom(), herramienta para generar dos mapas paralelos supervisados de SOM: xyf().
- **lsmeans (Lenth, 2016):** Medias marginales estimadas (o medias de mínimos cuadrados): lsmeans(), métodos para obtener los análisis de objetos ref.grid y lsmobj: contrast() .
- **visreg:** visualización de funciones de regresión: visreg()
- **nortest:** prueba de normalidad de Lilliefors (Kolmogorov-Smirnov): lillie.test().
- **lawstat (W. Hui, 2008):** pruebas de bondad de ajuste para la distribución de Laplace: levene.test()
- **Rmisc:** resumen de datos: summarySE()
- **Som (Wittek et al., 2017):** función para la generación de un mapa autoorganizado(Self-Organizing Map): som().
- **factoextra:** visualización de resultados de agrupación: fvizcluster().

3.3.2. Shiny

Shiny es un marco de aplicación web para R (Shiny, 2017). El lenguaje de programación es sencillo de escribir, por lo que no se necesitan grandes habilidades de programación en el campo de desarrollo web. Shiny es la simbiosis perfecta del lenguaje R y la interactividad web, a la que tan acostumbrados estamos con las webs modernas. Por tanto se convierte en una herramienta con un gran poder interactivo, tanto como el programador quiera darle.

3.3.3. Softwares adicionales

3.3.3.1. Graph-Pad Prism

Algunas gráficas que aparecen en este documento (gráficas de columnas indicando medias marginales estimadas y límite de confianza) se realizaron con el software privado Graph-Pad Prism version 6.01 (Graph Pad Software, San Diego, CA, USA), software elegido por continuar con el mismo formato que con el que se hicieron previas publicaciones que sirven de antesala a este proyecto.

3.3.3.2. Microsoft Word Office

Las tablas que recogen los análisis de datos en este proyecto se hicieron con el software, libre para estudiantes de la Universidad de Málaga, Microsoft Word Office 365 ProPlus versión 16.0.8067.2115.

3.3.3.3. Latex

La redacción del proyecto se ha llevado a cabo utilizando el software libre TeXstudio versión 2.12.2, usando el lenguaje LaTeX(Goossens et al., 1993). El gestor de paquetes y motor de LaTeXha sido MiKTeX versión 2.9. Además se ha utilizado el servicio de LaTeXcolaborativo ofrecido por la aplicación Overleaf[15].

[15] `https://www.overleaf.com`

Capítulo 4

Desarrollo de la aplicación para el análisis de datos de pacientes con TCC

Este proyecto surge inicialmente a partir de la necesidad de analizar, principalmente mediante técnicas estadísticas, un conjunto de datos de pacientes con TCC. A partir de esa idea, surge la propuesta de hacer algo más novedoso de manera de facilitar el análisis de los datos, creando una aplicación software que incluya además de los análisis estadísticos necesarios, otras técnicas de minería de datos como el anásis de clustering, y que posibilite una mejor visualización de los resultados. Con este fin, era por tanto necesario un estudio de las funcionalidades y alcance que la aplicación debía tener, siempre centrándonos en el usuario futuro de la aplicación. Es decir, para este proyecto se tiene como cliente al grupo de investigación Unidad de Gestión Clínica de Salud Mental, Grupo de Neuropsicofarmacología del Hospital Regional Universitario de Málaga. Los miembros de este grupo, entre los que se encuentran biólogos, médicos y psicólogos, serán los usuarios futuros de esta aplicación.

4.1. Metodología seguida para el desarrollo de la aplicación

Se ha llevado a cabo para el desarrollo de la aplicación que este proyecto propone, un análisis de Ingeniería del Software (Sommerville and Sawyer, 1997). Para el desarrollo de esta aplicación se pretendió seguir una metodología de modelo en cascada, por tratarse de un modelo secuencial en el que el proceso final es visible desde el principio. Sin embargo, al desarrollarse la aplicación con una tecnología novedosa (Shiny dentro del entorno R-Studio), y por tanto, desconocida para el alumno, se adoptó un modelo de desarrollo evolutivo. Aunque hubo un estudio previo del alcance de la herramienta con la que se iba a desarrollar el software, a medida que se entraba en profundidad en la capacidad de la herramienta se fueron desarrollando diferentes prototipos de la aplicación.

4.1.1. Fases seguidas en el proceso de desarrollo de la aplicación

1. Definición de la funcionalidad del software y las restricciones de operaciones del mismo.
   - Se concertó una cita con el cliente, en concreto, con los coordinadores principales de la Unidad de Gestión Clínica de Salud Mental de Málaga, que abarcan cada área: el Investigador principal cuya formación es médico y doctor en bioquímica, un doctor en biología y un doctor en psicología.
   - En la cita se definieron las técnicas estadísticas que debían abarcarse, así como las posibles variaciones de parámetros que éstas rigen.
   - El cliente propuso un conjunto de datos de pacientes con adicción a la cocaína en estado de abstinencia y de régimen ambulatorio, además de la muestra de control correspondiente.
   - Conjuntamente con el cliente se acordaron las variables a estudiar así como los diferentes gráficos que podrían ser de utilidad al usuario de la aplicación, pues el alumno propuso los gráficos explicativos, para que tanto el cliente actual como uno futuro pudiese comprender mejor los resultados.
2. Estudio del estado del arte (alcance del lenguaje R y Shiny).
3. Desarrollo de requisitos funcionales[1] y no funcionales[2] del software.
   - El cliente especificó los análisis estadísticos que debían llevarse a cabo con el conjunto de datos propuesto.
   - El tutor de este proyecto concretó las técnicas de clustering que debían abarcarse.
   - El alumno planificó la Ingeniería de Requisitos en función de la demanda del cliente y el tutor.

[1] Tareas que el sistema tiene que hacer.
[2] Condiciones o restricciones bajo las que se ejecuta/desarrolla el programa.

- El alumno diseñó bocetos manuales de la apariencia de la aplicación, que recreó mediante un primer prototipo que fue evaluado por el tutor del proyecto.

4. Diseño lógico de la aplicación.
   - El alumno diseñó bocetos manuales de la lógica que debía seguir la aplicación, que recreó mediante el primer prototipo que fue revisado por el tutor del proyecto para así poder continuar con el diseño de la aplicaión.
5. Implementación de la aplicación (diferentes prototipos).
   - El alumno fue desarrollando sucesivos prototipos de manera que en cada uno de ellos se mejoraba la funcionalidad, apariencia y eficiencia del anterior.
6. Validación del software (refrendación del prototipo)
   - El alumno ejecutó diversas pruebas para la comprobación correcta del software.
7. Inclusión de mejoras en el software hasta llegar al prototipo final.
   - El alumno incluyó algunas mejoras a medida que conocía el alcance de la tecnología con la que se estaba desarrollando la aplicación, el marco Shiny.
   - El tutor propuso mejoras que se incluyeron en el software.
8. Generación de resultados.
9. Análisis de los resultados.
10. Conclusiones.

4.1.2. Análisis de requisitos

4.1.2.1. Requisitos funcionales

En una primera fase, se expresan los requisitos funcionales mediante la siguiente cláusula, definiendo la acción final de la aplicación y cómo se beneficia el usuario:

- El usuario (médico, biólogo, psicólogo o futuro usuario) obtendrá resultados finales de varios tipos de análisis a un conjunto de datos que él propone.

Una vez enfocada la intención de la aplicación, se abordan de forma más específica los requisitos funcionales de la misma:

1. Importación de un archivo de datos.
2. Visualización de todo el conjunto de datos.
3. Análisis estadísticos de las variables socio-demográficas
4. Análisis estadísticos de las variables relacionadas con la adicción a la cocaína.
5. Análisis de la expresión de las moléculas.
6. Estudio de clustering: clustering jerárquico, SOM y K-means.

7. Los resultados de los análisis deben estar acompañados de conclusiones por escrito.
8. Los análisis de expresión de moléculas deben incluir figuras para la comprensión de resultados.
9. El estudio de clustering se mostrará mediante figuras y tablas, brindando información de los grupos de sujetos asignados a cada cluster.

4.1.2.2. Requisitos no funcionales

1. El lenguaje R será el elegido por el alcance del mismo en cuanto a paquetes en áreas de estadística, aprendizaje computacional y herramientas de representación gráfica.
2. Será necesario una interfaz gráfica intuitiva debido a que los usuarios actuales no poseen conocimientos informáticos avnazados. Se decide utilizar el marco Shiny para su desarrollo.
3. El archivo de datos importado debe estar en formato csv, que es el usualmente utilizado por el cliente.
4. Las técnicas estadísticas usadas para variables sociodemográficas serán también utilizadas para las variables de adicción a la cocaína.
5. El código debe ser muy legible para poder extender su capacidad en posibles versiones futuras.
6. Las diferentes funcionalidades deben estar divididas en funciones diferentes para garantizar que los análisis ya implementados se mantengan en futuras versiones, facilitando además la edición de los mismos.
7. El software se debe desarrollar de forma eficiente en cuanto al tiempo y el espacio. Para ello se debe analizar en profundidad el alcance de R y Shiny (funciones 'reactive', funciones externas, variables globales, etc.).
8. La aplicación debe lanzar mensajes de error aclaratorios cuando alguna función no se ejecuta de modo correcto.

4.2. Manual de usuario

Junto a esta memoria se adjunta el código software de la aplicación que ha resultado de este proyecto. Se presentan 3 archivos principales en lenguaje R que se detallan a continuación:

- Un archivo de instalación *installPackage.R* con extensión .R, en el cual se encuentran los paquetes que necesitan ser instalados por el usuario para poder ejecutar la aplicación.
- Dos archivos *ui.R* y *server.R* con extensión .R que conforman la aplicación.

Además se incluyen 13 archivos en lenguaje R, y cada uno de ellos implementa una función que es llamada por el código de la aplicación. Pero el usuario no necesita ejecutar estos archivos.

En el desarrollo de la aplicación se ha intentado que la misma fuera de manejo fácil e intuitivo. A continuación se describen los pasos básicos para la utilización de la misma, mostrándose una serie de capturas de pantalla que dirijan al usuario a modo de tutorial para que pueda apreciar el alcance de la aplicación.

4.2.1. Carga de datos

Dentro del entorno RStudio, se abren los 3 archivos disponibles.

El primer paso es ejecutar el archivo de instalación de paquetes: abrir el archivo *installPackage.R* en el entorno RStudio e ir pulsando el botón *Run*, que está en la barra de herramientas a la derecha, para que se ejecute cada línea de código, ver imagen (4.1). Al instalar el paquete *Shiny* en la barra de herramientas de RStudio debe salir la opción a la derecha de Run App, ver imagen (4.2).

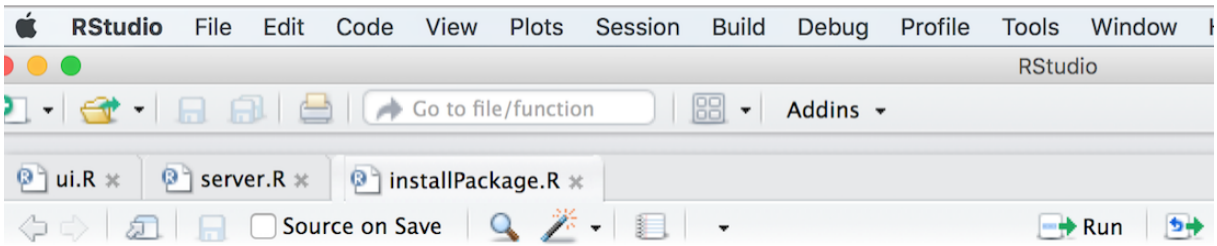


Figura 4.1: Instalación de paquetes.

A continuación se abren los archivos *ui.R* y *server.R* y se lanza la aplicación pulsando el ícono *Run App* (representado por una flecha verde) (puede hacerse en una nueva ventana o en el navegador[3]), ver imagen (4.2).

[3]Tener en cuenta que con algunos navegadores la configuración no es totalmente compatible y pueden no aparecer todas las secciones.

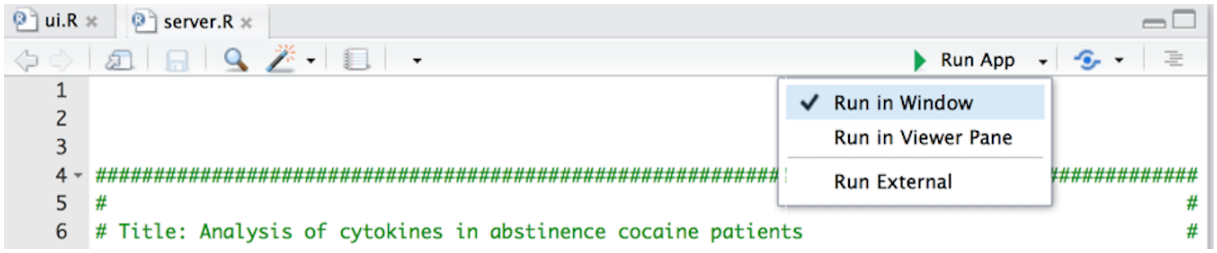


Figura 4.2: Captura de pantalla del lanzamiento de la aplicación.

4.2.2. Carga de datos

El proceso de carga de datos es muy simple, y a modo de ejemplo, el software se suministra con un conjunto de datos de prueba. Para este conjunto de datos se marca la opción de que tienen cabecera y que el separador es una coma, ver imagen (4.3). En la pestaña actual ('Upload data') hay tres secciones para visualizar la información que aportan los datos: 'Data', 'Summary' y 'About file'. En Data se encuentran todos los datos y variables, pudiendo navegar por el archivo gracias a los múltiples buscadores disponibles. Summary recoge un resumen descriptivo de todas las variables y la última sección es información de archivo.

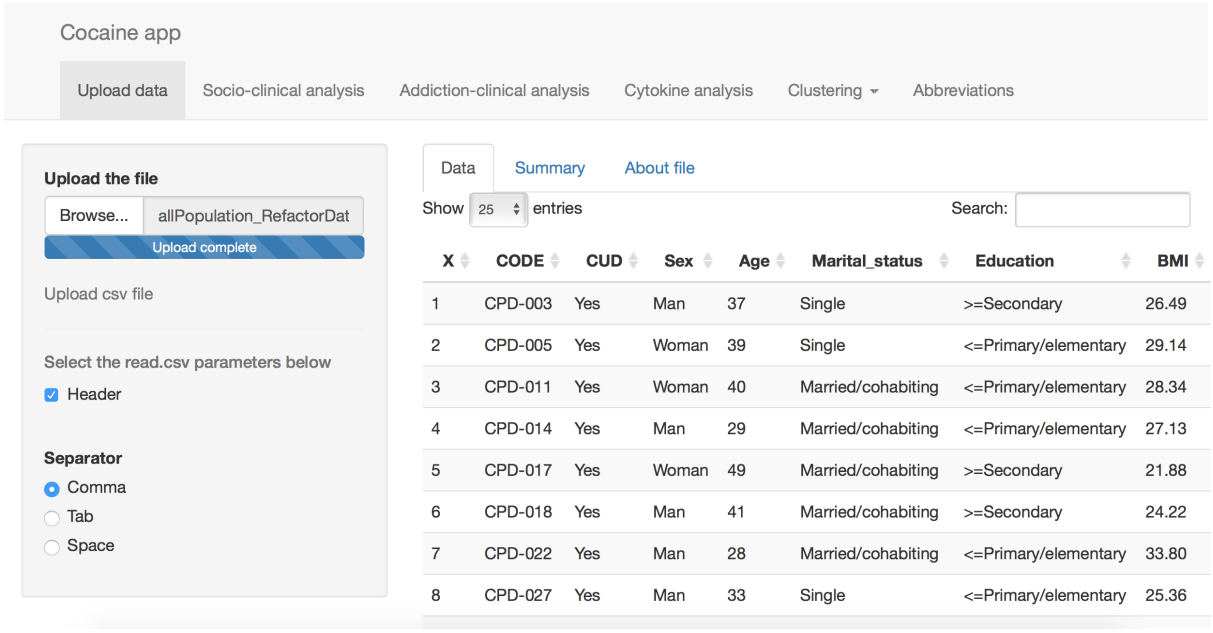


Figura 4.3: Captura de pantalla del proceso de carga de un archivo de datos.

4.2.3. Análisis de datos

La segunda pestaña permite en análisis de las variables sociodemográficas que caracterizan a todos los sujetos: edad, sexo, IMC, estado civil, nivel educacional y tratamiento psiquiátrico, ver imagen (4.4). Además se permite la selección de los grupos a comparar en los análisis (controles con respecto a los pacientes con adicción, o se enfrentan subgrupos dentro de los pacientes adictos diferenciándolos por el tipo de comorbilidad que paedecen). Una tabla de frecuencias, análisis de normalidad (para variables continuas) y uso de la estadística inferencial mediante comparación de distribuciones, son los análisis llevados a cabo en esta sección. Tanto en el apartado de normalidad como en el de comparación de distribuciones se muestran conclusiones de los resultados. En la parte izquierda se permite modificar los valores de significación para los diferentes análisis (normalidad, homocedasticidad y comparación de distribuciones). Estas variaciones influyen en el tipo de pruebas que se realizan finalmente sobre el conjunto de datos selecciondos, y además en las conclusiones de los resultados.

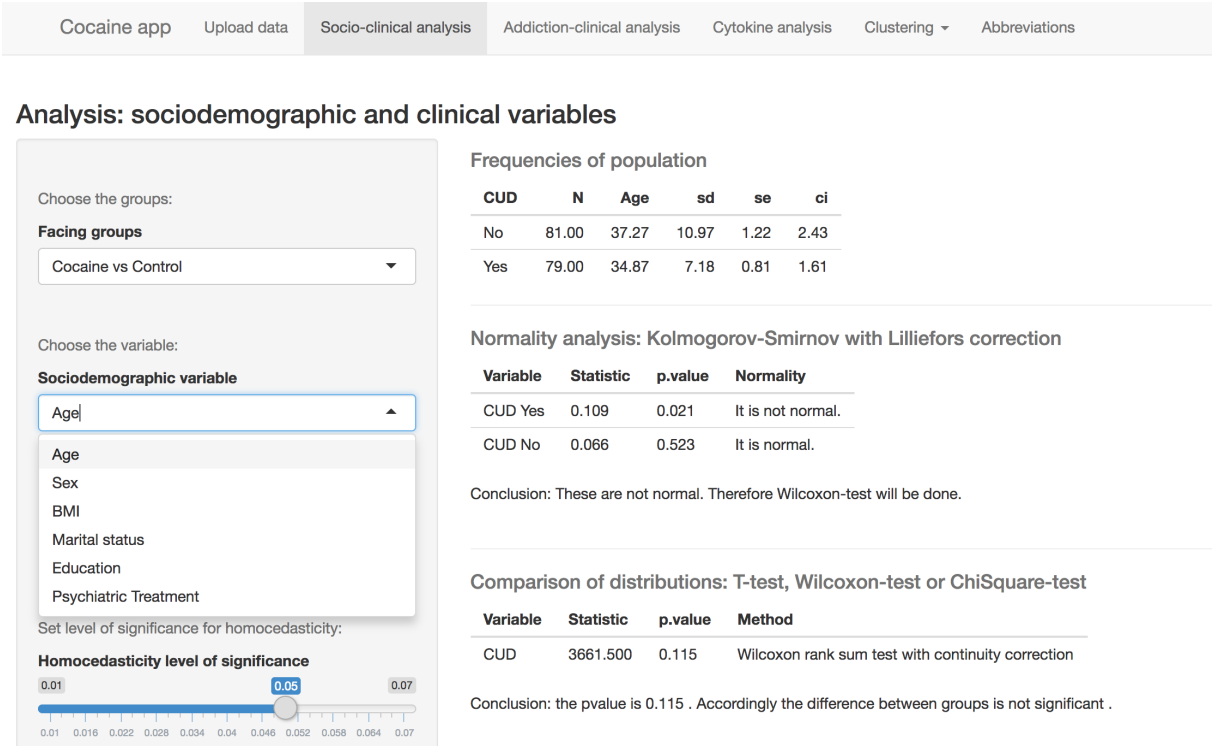


Figura 4.4: Análisis sociodemográfico y clínico de los datos utilizados.

La tercera pestaña permite en análisis de las variables de adicción que caracterizan a todos los pacientes con adicción a la cocaína: edad de prueba de la droga, tiempo de abstinencia, síntomas de la gravedad de la adicción (continua y discreta), comorbilidad por consumo de otra sustancia y diagnóstico dual, ver imagen (4.5). Al igual que en la pestaña segunda, se permite la selección de los grupos a comparar en los análisis (subgrupos dentro de los

pacientes adictos, diferenciando las muestras según el tipo de comorbilidad psiquiátrica que padecen). De nuevo, una tabla de frecuencias, análisis de normalidad (para variables continuas) y uso de la estadística inferencial mediante comparación de distribuciones, son los análisis llevados a cabo en la sección. Tanto en el apartado de normalidad como en el de comparación de distribuciones se muestran conclusiones de los resultados para que el usuario inexperto en la estadística pueda interpretar los datos. En la parte izquierda se permite modificar los valores de significación para los diferentes análisis (normalidad, homocedasticidad y comparación de distribuciones).

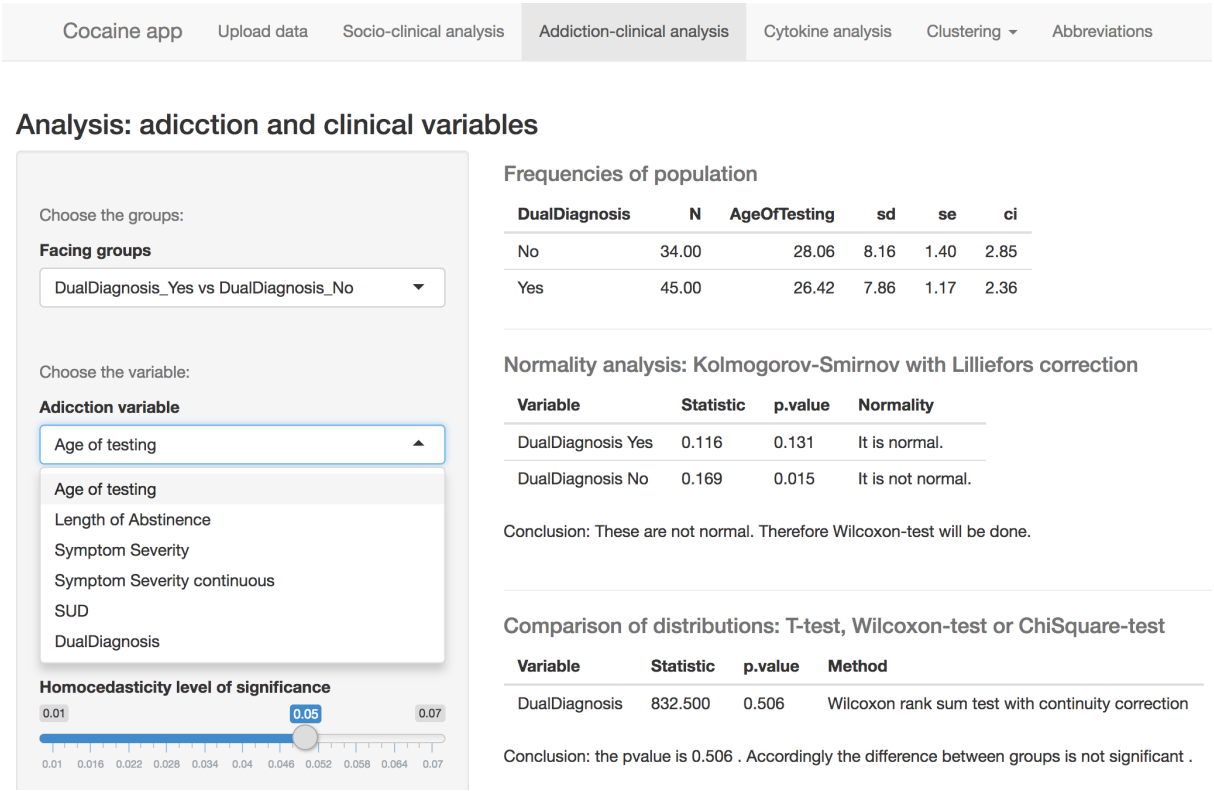


Figura 4.5: Captura de pantalla del análisis de adicción-clínico.

La cuarta sección es un poco más extensa y se dividirá en partes. En principio se muestra el análisis de normalidad de la molécula seleccionada, ver imagen (4.6). Según las conclusiones que se indican en el análisis de normalidad, puede modificarse la selección de la variable a la misma con la trasformación logarítmica, o variar el nivel de significación, pues lo que se pretende es que la variable en cuestión siga una distribución normal, o al menos, pueda considerarse que la sigue, ver apartado (3.1.7).

A continuación pueden apreciarse sucesivas figuras, muy intuitivas, que pretenden orientar al usuario en la búsqueda y comprensión de la normalidad de la molécula a analizar, ver imagen (4.7). El uso de gráficas de frecuencia, de densidad, de cajas y QQ está

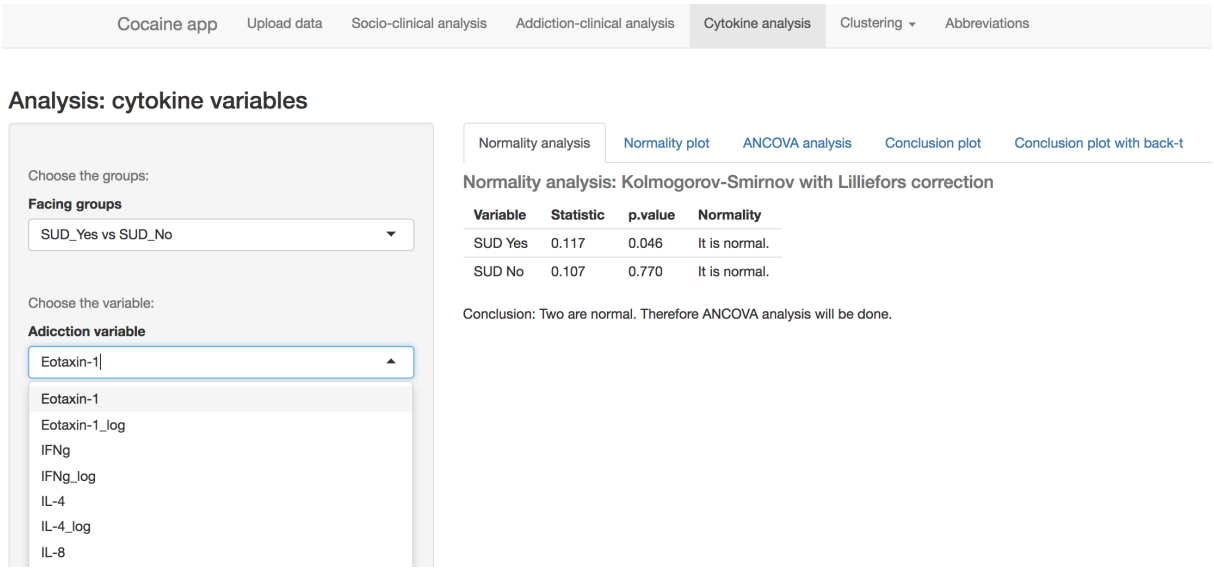


Figura 4.6: Análisis de normalidad Eotaxin-1 diferenciando la muestra con TCC según la variable de comorbilidad por otra sustancia.

muy extendido en el ámbito científico, y son necesarias para observar la cercanía de los datos de estudio a una distribución normal. Es por ello que se muestran en esta sección, además de dos gráficas de solapamiento. La pestaña de ANCOVA es el análisis de la covarianza, permitiendo elegir los cofactores (variables categóricas) y covariables (variables continuas) que formarán el modelo, ver imagen (4.8). Debajo del análisis, se muestran las Medias Marginales Estimadas y los intervalos de confianza, valores de interés para el investigador. Antes del análisis con múltiples covariables/cofactores se incluye el ANCOVA de la variable molecular y el factor de interés, para que el usuario no se desoriente a la hora de observar los resultados en el modelo más completo.

El análisis de la covarianza finaliza con gráficos de gran interés para el investigador, ver imagen (4.9). Existe la posibilidad de descarga de estas figuras, así como de apreciar los resultados gráficos de las variables trasformadas, véase de nuevo figura (4.9), y su 'back-transformation' en la pestaña que le sigue. Por último se lleva a cabo el análisis de clustering. Se permiten tres tipos de técnicas (ver imagen(4.10)), y para cada una de ellas es posible seleccionar las variables moleculares que deseen incluirse en el agrupamiento. Dependiendo del método elegido, se podrán fijar unos métodos y/o variables específicas (número de clusters, tipo de distancia, altura del cluster, etc.). En la última pestaña de la apliación, se ha añadido un apartado de acrónimos para que el usuario tenga disponible la terminología utilizada de forma inmediata.

Figura 4.7: Análisis de normalidad mediante histogramas de Eotaxin-1 diferenciando la muestra con TCC según la variable de comorbilidad por otra sustancia.

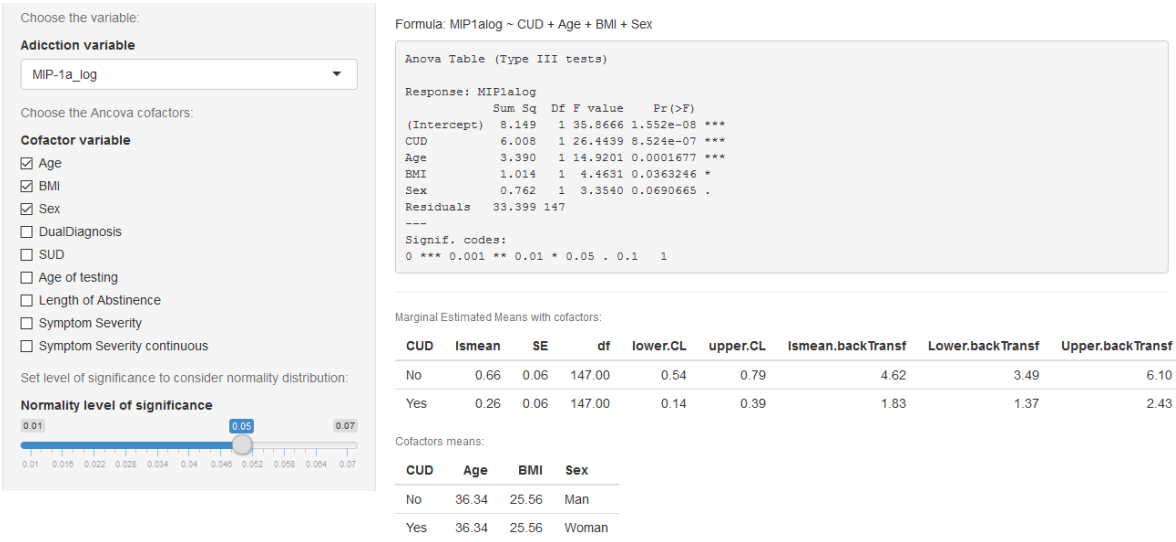


Figura 4.8: Análisis de covarianza de MIP-1α usando como factor principal que el sujeto tenga TCC, junto con el factor sexo y las dos covariables edad e IMC.

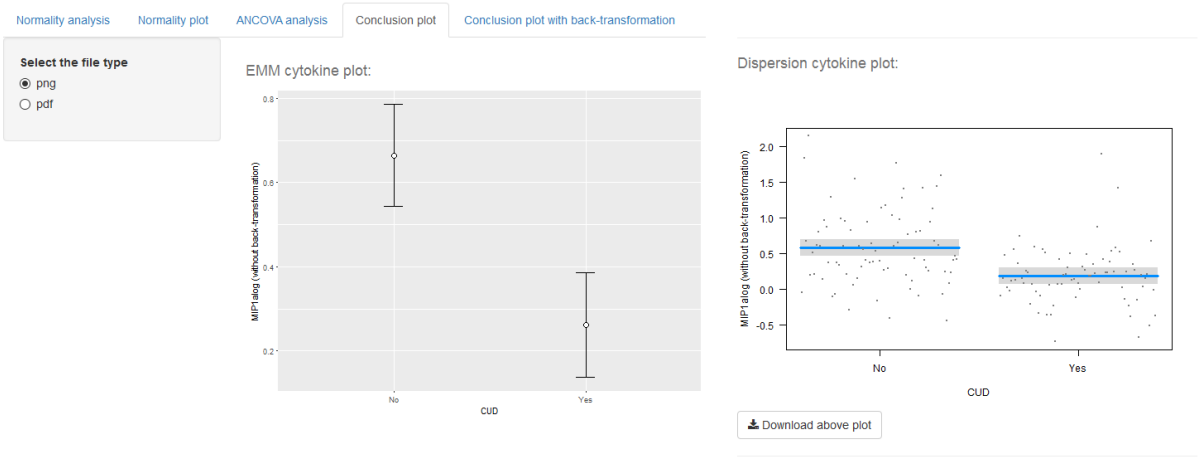
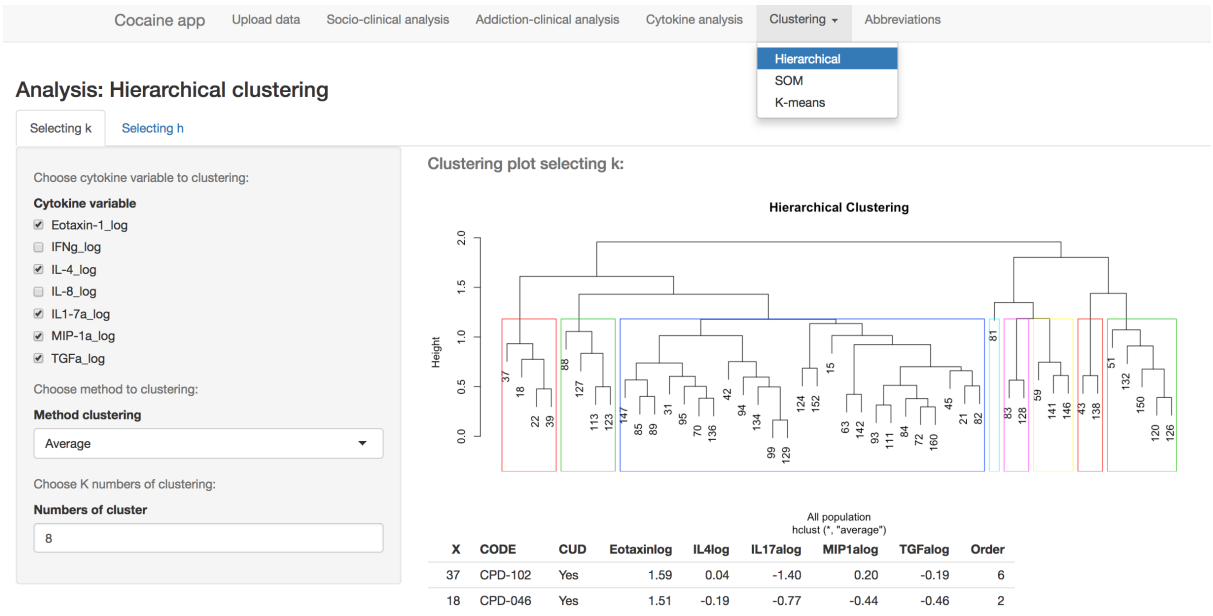


Figura 4.9: Análisis de covarianza de MIP-1α usando como factor principal que el sujeto tenga TCC, junto con el factor sexo y las dos covariables edad e IMC. Representa dos gráficas de los datos que aparecen en la figura (4.8)



X	CODE	CUD	Eotaxinlog	IL4log	IL17alog	MIP1alog	TGFalog	Order
37	CPD-102	Yes	1.59	0.04	-1.40	0.20	-0.19	6
18	CPD-046	Yes	1.51	-0.19	-0.77	-0.44	-0.46	2

Figura 4.10: Clustering jerárquico con valor de 3 para el parámetro de número de clusters.

Capítulo 5

Resultados

5.1. Estudio de pacientes con TCC contrastados con sujetos control

5.1.1. Características sociodemográficas-clínicas

Se presenta en primer lugar un estudio de las variables sociodemográficas y de patologías psiquiátricas de la población analizada. Un total de 160 sujetos cumplieron con los criterios de elegibilidad para este estudio, siendo esta población total dividida en dos subgrupos: los pacientes con adicción a cocaína en estado actual de abstinencia (N=79) y los sujetos control utilizados para el anásis de contraste (N=81). La figura (5.1) muestra los valores medios de las diferentes variables analizadas para los pacientes con TCC y grupo de control.

La media de edad del grupo con trastornos por consumo de cocaína es de 34.87 años y tiene un IMC(BMI) de 25.88. Se han encontrado diferencias significativas entre pacientes con TCC y controles en la variable que designa el estado civil ($p = 0{,}004$). Además los pacientes cocainómanos en estado de abstinencia tienen un significativo menor nivel educacional y un mayor porcentaje de tratamiento psiquiátrico a lo largo de la vida, comparado con el grupo control ($p < 0{,}001$). Un total de 56.96 % de los pacientes con TCC fueron diagnosticados con patología dual, es decir que como mínimo han sufrido un trastorno psiquiátrico a lo largo de la vida [trastorno de estado de ánimo, psicótico o de personalidad (antisocial o límite)]. Por otro lado, un 73.42 % de los pacientes del grupo de cocaína fueron diagnosticados con comorbilidad por otras sustancias (alcohol, heroína, cannabis, benzodiacepinas, alucinógenos u otros estimulantes). El tiempo de abstinencia de cocaína en estos pacientes ha sido de 133.4 días y la media de la gravedad de los síntomas del consumo de cocaína ha sido de 8.04 en la escala de criterios-TCC, todo ello medido en el momento del reclutamiento de los pacientes.

VARIABLE		TOTAL N=160		
		COCAINE N=79	CONTROL N=81	p value
Age [mean(SD)]	*Years*	34.87 (7.18)	37.27 (10.97)	0.115[a]
BMI [mean(SD)]	*Kg/m²*	25.88 (4.20)	25.32 (3.65)	0.370[b]
Sex [N(%)]	Women Men	16 (20.25) 63 (79.75)	16 (19.75) 65 (80.25)	1[c]
Marital Status [N(%)]	Single Married/Cohabiting Divorced/separated Widowed	29 (36.71) 29 (36.71) 20 (25.32) 1 (1.27)	26 (47.3) 22 (40.0) 7 (12.7) 0 (0.00)	**0.004[c]**
Education [N(%)]	≤Primary/elementary ≥Secondary	63 (79.75) 16 (20.25)	9 (11.11) 72 (88.89)	**<0.001[c]**
Psychiatric Treatment [N(%)]	No Yes	52 (65.82) 27 (34.18)	72 (88.89) 9 (11.11)	**<0.001[c]**
Dual Diagnosis [N(%)]	No Yes	34 (43.04) 45 (56.96)	-	-
Other Substance Use Disorders [N(%)]	No Yes	21 (26.58) 58 (73.42)	-	-
Length of Abstinence [mean(SD)]	*Days*	133.4 (114.53)	-	-
Cocaine Symptom Severity [mean(SD)]	*(0-11) criteria*	8.04 (2.66)	-	-

Abbreviations: BMI: body mass index
([a]) p-value from Student's-test
([b]) p-value from Wilcoxon-test
([c]) p-value from Chi-square-test

Figura 5.1: Estudio sociodemográfico-clínico de la población total.

5.1.2. Impacto de la cocaína en los niveles en plasma de las citoquinas

Las medias marginales estimadas de las moléculas Eotaxin-1, IFNγ, IL-4, IL-8, IL-17α, MIP-1α y TGFα en plasma, ver figura (5.2), se representan de acuerdo al historial de consumo de cocaína (grupos de cocaína y control). Las diferencias en concentraciones en plasma de estos factores han sido llevados acabo por análisis de ANCOVA usando el trastorno por cocaína como factor principal y la edad, sexo e IMC como variables correctoras.

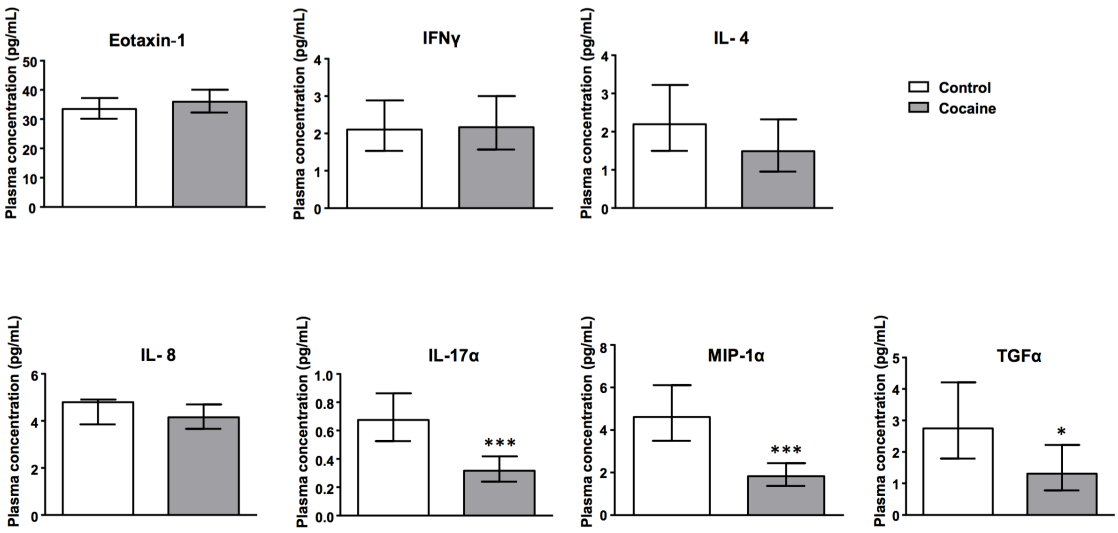


Figura 5.2: Medias marginales estimadas de las concentraciones citoquínicas en plasma para pacientes con TCC y grupo de control.

Se han encontrado efectos significativos por el consumo de cocaína en los niveles en plasma de IL-17α ($F_{1,133} = 20{,}713$, $p < 0{,}001$), MIP-1α ($F_{1,147} = 26{,}444$, $p < 0{,}001$) y TGFα ($F_{1,540} = 5{,}251$, $p < 0{,}05$), ver figura (5.2). Los pacientes diagnosticados con TCC muestran bajos niveles de expresión en estas citoquinas con respecto a los sujetos que forman el grupo control. Por el contrario, no se han encontrado diferencias contrastando las dos subpoblaciones, control y cocaína, en los niveles de las citoquinas Eotaxin-1, IFNγ, IL-4 e IL-8.

5.2. Estudio de pacientes con TCC diagnosticados con comorbilidad psiquiátrica

Los pacientes con adicción a la cocaína han sido divididos en subgrupos de acuerdo a un diagnóstico que evidencia la comorbilidad psiquiátrica. Por ello, se han utilizado dos criterios diferentes de clasificación para explorar los efectos de la comorbilidad psiquiátrica sobre los factores inflamatorios y las características sociodemográficas y de adicción:

1. Diagnóstico de trastornos por consumo de otras sustancias, o comorbilidad por otra sustancia.
2. Diagnóstico dual por trastornos de salud mental, o patología dual.

5.2.1. Efecto de la comorbilidad psiquiátrica en las características sociodemográficas-clínicas

Una primera clasificación del grupo con TCC se realizó de acuerdo al diagnóstico de trastornos por consumo de otras sustancias (TCS) [SÍ (N=58); NO (N=21)], ver figura (5.3) columnas 2, 3 y 4. La comparación entre ambos grupos muestra que no hay diferencias significativas en edad, IMC, sexo, tratamiento psiquiátrico, edad de inicio del consumo de la cocaína o tiempo de abstinencia. Sin embargo, se encontraron diferencias en la gravedad de los síntomas del consumo de cocaína ($p < 0{,}001$), pues los pacientes con TCS tienen más elevado el índice que denota la gravedad de los síntomas del consumo de cocaína que los que no tienen el trastorno por consumo de otra sustancia (8.41 vs. 7.00 criterios), véase figura (5.3). Aunque los pacientes con TCS tienen una elevada prevalencia de diagnóstico dual (63.79 %) comparado con los pacientes que no tienen el trastorno por consumo de otra sustancia, la diferencia no alcanza la significancia estadística ($p = 0{,}075$).

Una clasificación adicional se ha llevado a cabo de acuerdo a la detección del diagnóstico dual [SÍ (N=45); NO (N=34)], véase figura (5.3). No se han encontrado diferencias en IMC, sexo, tratamiento psiquiátrico, edad de inicio del consumo de la cocaína o tiempo de abstinencia. En este caso, se han observado diferencias significativas en edad ($p < 0{,}05$) y en la gravedad de los síntomas del consumo de la cocaína ($p < 0{,}001$). Por lo tanto, los pacientes con patología dual son mayores que los que no la tienen cuando acceden voluntariamente a incorporarse a este estudio (36.53 vs. 32.68 años). Los pacientes que tienen diagnóstico dual asociado sufren la adicción con más gravedad (9.20 vs. 6.50 criterios). Además, los pacientes con patología dual tienen una elevada prevalencia de comorbilidad por consumo de otra sustancia (82.22 %) frente a los pacientes que no tienen la tienen (61.76 %), pero esta diferencia no presenta significación estadística ($p = 0{,}075$).

5.2.2. Impacto del diagnóstico dual en los niveles plasmáticos de citoquinas

Las expresiones de las citoquinas Eotaxin-1, IFNγ, IL-4, IL-8, IL-17α, MIP-1α y TGFα en plasma, ver figura (5.4), de pacientes con TCC agrupados de acuerdo a la presencia

VARIABLE		COCAINE N=79					
		OTHER SUBSTANCE USE DISORDERS YES	OTHER SUBSTANCE USE DISORDERS NO	p value	DUAL DIAGNOSIS YES	DUAL DIAGNOSIS NO	p value
Participants [N (%)]		58 (73.42)	21 (26.58)	-	45 (56.96)	34 (43.04)	-
Age [mean (SD)]	*Years*	34.72 (7.15)	35.29 (7.42)	0.726[b]	36.53 (6.65)	32.68 (7.18)	**0.025**[b]
BMI [mean(SD)]	*kg/m²*	25.58 (3.41)	26.70 (5.89)	0.300[a]	26.58 (4.61)	24.95 (3.43)	0.089[a]
Sex [N (%)]	*Women* *Men*	10 (17.24) 48 (82.76)	6 (28.58) 15 (71.43)	0.430[c]	11 (24.44) 34 (75.56)	5 (14.71) 29 (85.29)	0.433[c]
Psychiatric Treatment [N(%)]	*No* *Yes*	37 (63.79) 21 (36.21)	15 (71.43) 6 (28.58)	0.716[c]	26 (57.78) 19 (42.22)	26 (76.47) 8 (23.53)	0.135[c]
Age of Cocaine Initiation [mean (SD)]	*Years*	26.48 (8.12)	28.90 (7.47)	0.236[a]	26.42 (7.86)	28.06 (8.16)	0.507[b]
Length of Abstinence [mean (SD)]	*Days*	139.50 (113.20)	116.30 (119.26)	0.287[b]	133.5 (102.68)	133.2 (130.19)	0.334[b]
Cocaine Symptom Severity [mean (SD)]	*(0-11) criteria*	8.41 (2.56)	7.00 (2.70)	**<0.001**[b]	9.20 (1.65)	6.50 (2.97)	**<0.001**[b]
Dual Diagnosis [N (%)]	*No* *Yes*	21 (36.21) 37 (63.79)	13 (61.90) 8 (38.10)	0.075[c]	-	-	-
Other Substance Use Disorders [N (%)]	*No* *Yes*	-	-	-	8 (17.78) 37 (82.22)	13 (39.24) 21 (61.76)	0.075[c]

Abbreviations: BMI: body mass index
([a]) p-value from Student's-test
([b]) p-value from Wilcoxon-test
([c]) p-value from Chi-square-test

Figura 5.3: Características sociodemográficas y clínicas en pacientes con TCC.

de patología dual es un estudio que busca ahondar en los resultados encontrados previamente, ver apartado 5.1.2. La concentración en plasma de estos marcadores ha sido analizado mediante ANCOVAs, haciendo uso del diagnóstico dual como factor principal en el estudio, e incluyendo la edad, sexo e IMC como variables de control.

Se encuentran efectos significativos en la patología dual en los niveles en plasma de TGFα ($F_{1,190} = 6{,}812$, $p < 0{,}05$). Así que los pacientes diagnosticados con patología dual mues-

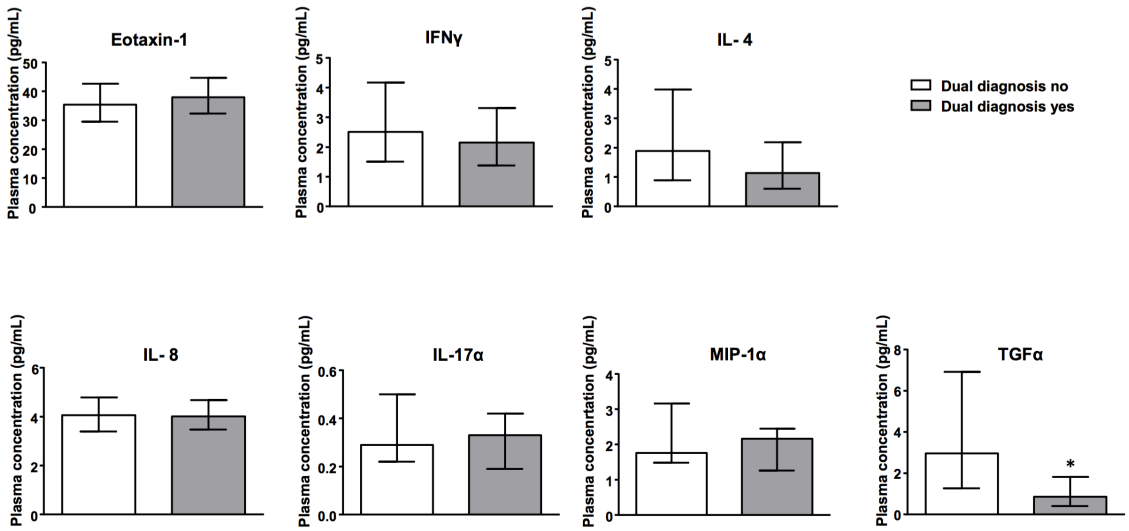


Figura 5.4: Medias marginales estimadas de las concentraciones citoquínicas en plasma diferenciando diagnóstico dual.

tran menores niveles de TGFα con respecto a los pacientes que no padecen la patología. En consecuencia, la presencia del diagnóstico dual en el grupo de cocaína incrementa la disminución de los niveles de TGFα en comparación al grupo control.

5.3. Impacto de la gravedad de los síntomas por consumo de cocaína en los niveles plasmáticos moleculares en pacientes con TCC

Debido a que se observaron diferencias significativas en la gravedad de los síntomas por el consumo de cocaína cuando los pacientes fueron agrupados de acuerdo a la comorbilidad psiquiátrica (diagnóstico dual y comorbilidad por consumo de sustancia), se exploró el impacto del grado de gravedad en las expresiones de las citoquinas. Por tanto, el grupo de consumidores de cocaína se dividió en dos subgrupos: leve/moderado [(0-8 criterios) N=38] y grave [(9-11 criterios) N=41] TCC, véase figura (5.5).

VARIABLE	CONCENTRATION (pg/mL) [mean(SD)]						
	Eotaxin-1	IFNγ	IL-4	IL-8	IL-17α	MIP-1α	TGFα
MILD/MODERATE (0-8) criteria[a] **N=38**	35.16 (1.09)	2.13 (1.26)	1.38 (1.37)	4.02 (1.08)	0.27 (1.21)	1.88 (1.19)	1.71 (1.46)
SEVERE (9-11) criteria[a] **N=41**	38.46 (1.09)	2.47 (1.25)	1.45 (1.41)	4.04 (1.08)	0.35 (1.21)	1.96 (1.19)	0.95 (1.70)
p value[b]	0.405	0.604	0.908	0.949	0.280	0.837	0.315

([a]) Based on DSM-IV-TR abuse and dependence criteria.
([b]) P-value from ANCOVA analysis. It denotes significant differences when it is <0.05

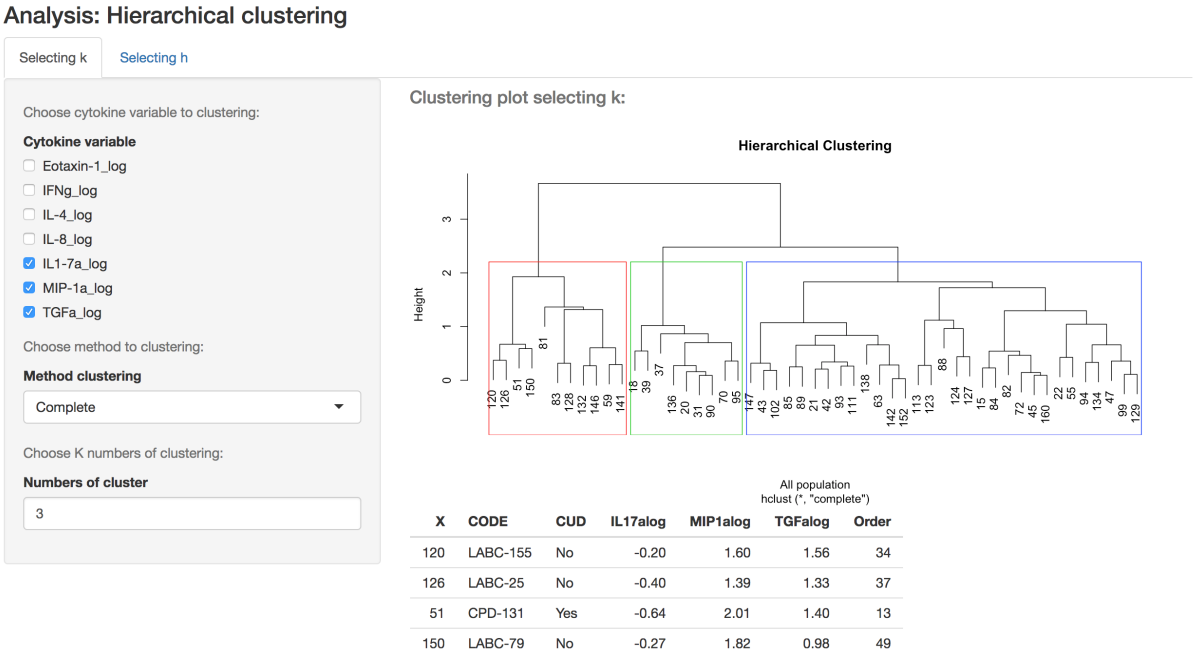
Figura 5.5: Gravedad de los síntomas por el consumo de cocaína en pacientes con TCC.

De forma similar a previos análisis, las concentraciones en plasma de estas proteínas han sido analizadas mediante ANCOVAs usando la gravedad de los síntomas (variable discreta) como factor principal, siendo controlada por la edad, sexo e IMC. Sin embargo, en la gravedad de los síntomas no se han encontrado diferencias destacables en ninguna de las expresiones moleculares entre ambos grupos, ver figura (5.5).

5.4. Aplicación de las técnicas de clustering

5.4.1. Clustering jerárquico

En el apartado (5.1.2) de resultados moleculares según el impacto de la cocaína, ver figura (5.2), se aprecia que son 3 las moléculas que indican diferencias significativas cuando se comparan las distribuciones de la muestras de pacientes con TCC respecto de los sujetos de control. Por tanto, al hacer un estudio de agrupamiento, es de esperar que al seleccionar estas tres moléculas candidatas, con su respectiva transformación logarítmica cada una de ellas, se obtuvieran resultados trascendentales, ver figura (5.6).



X	CODE	CUD	IL17alog	MIP1alog	TGFalog	Order
120	LABC-155	No	-0.20	1.60	1.56	34
126	LABC-25	No	-0.40	1.39	1.33	37
51	CPD-131	Yes	-0.64	2.01	1.40	13
150	LABC-79	No	-0.27	1.82	0.98	49

Figura 5.6: Clustering jerárquico con K=3 clusters aplicado a la población total.

Bajo el dendograma que genera la aplicación se muestra una tabla con los valores moleculares y la variable dicotómica de interés (TCC). Se esperaba que cada una de las dos ramas hijas pincipales agrupara a los sujetos de manera que el clustering jerárquico diferenciara en la primera división a los pacientes con TCC del grupo de control. Pero esto no ocurre así, y es entonces cuando se opta por fijar el número de clusters en K=3 a fin de estudiar más en profundidad los resultados obtenidos. Agrupando los valores moleculares por clusters, y calculando la media de los mismos, se obtiene que los clusters se agrupan según el valor medio molecular. Esto indica que el algoritmo está funcionando correctamente, ver figura (5.7). Sin embargo, ni siquiera en el caso en el que se seleccionan dos únicos clusters, la clasificación guarda una relación con la variable TCC.

```
> df
          Cluster1    Cluster2   Cluster3
MIP1alog 1.749514  0.04148847  0.6312403
IL17alog 0.211817 -0.75173738 -0.1590787
TGFalog  1.070054 -0.18156022  0.1779506
```

Figura 5.7: Valores medios de las variables utilizadas en el clustering jerárquico con K=3.

5.4.2. Clustering utilizando SOM

Los resultados de aplicar el clustering basado en las redes auto-organizadas de Kohonen (SOM) se muestran en las figuras (5.8), (5.9), (5.10) y (5.11). En la figura (5.8) se muestra el agrupamiento realizado por el algoritmo donde puede observarse dentro de cada neurona de representación (círculo) el valor de cada una de las variables seleccionadas (tres en el ejemplo mostrado). Dentro del agrupamiento realizado es posible identificar un grupo abajo a la izquierda con valores altos de las tres citoquinas seleccionadas, un segundo grupo arriba a la izquierda con valores altos para las variables IL-17α y MIP-1α. Adicionalmente, otro grupo que puede identificarse representa sujetos con valores bajos para todas las variables ubicado arriba la derecha de la figura. Si bien la interpretación de los resultados no es del todo directa, el valor añadido del clustering basado en SOM viene por el hecho de que ofrece una representación topológica de los datos, ya que neuronas cercanas representan sujetos con valores similares de las variables.

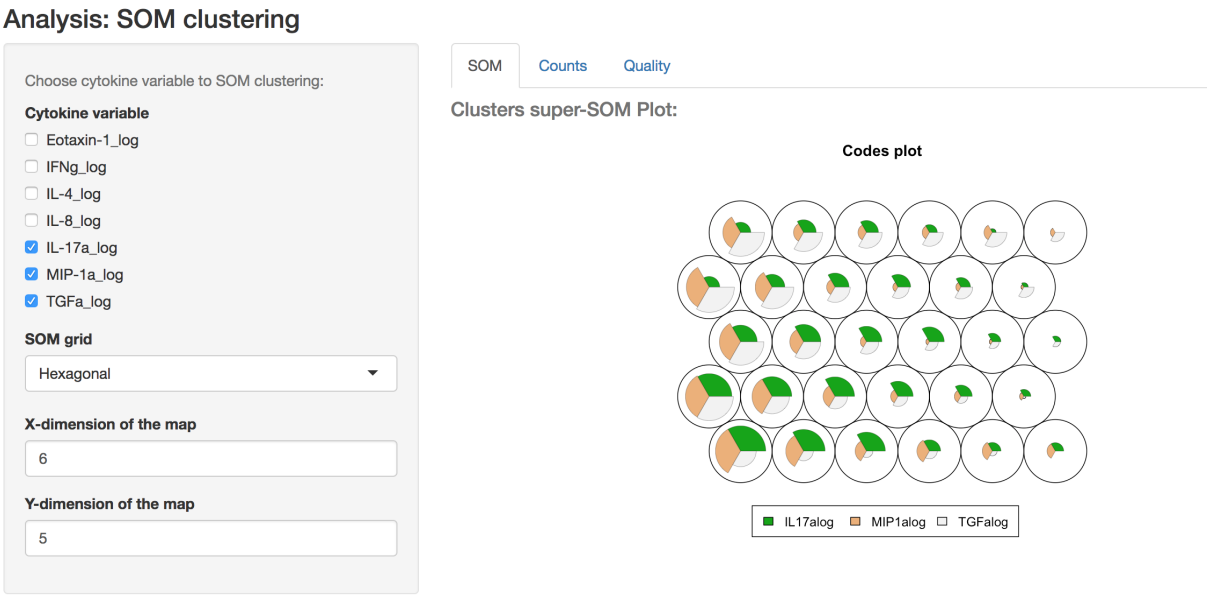


Figura 5.8: SOM: Mapeado Estándar.

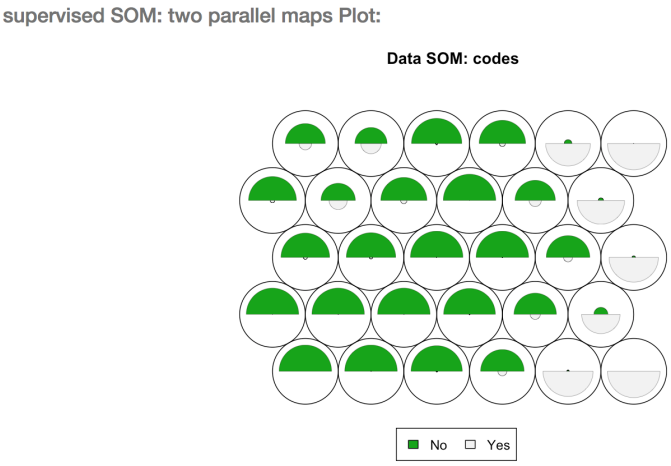


Figura 5.9: SOM: Mapeado Múltiple.

La aplicación dentro de la sección SOM genera también mapas de la cantidad de sujetos asignados a cada neurona utilizada en el agrupamiento (figura 5.10), así como de la calidad del mismo (figura 5.10), factores que ayudan en la interpretación de los resultados.

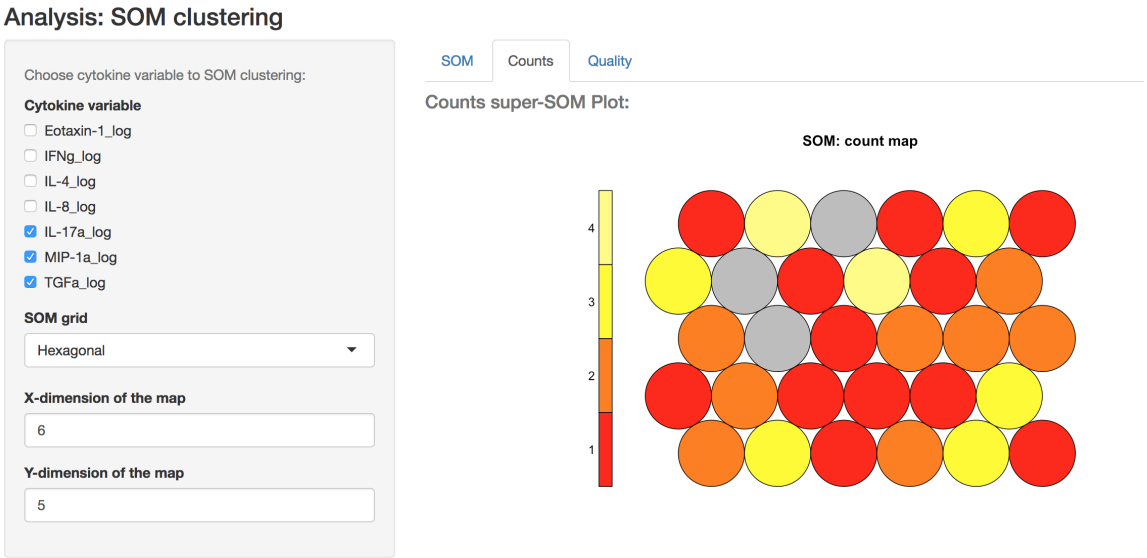


Figura 5.10: SOM: Cantidad de sujetos asignados a cada cluster.

Figura 5.11: SOM: Medida de calidad de los clusters.

5.4.3. Clustering utilizando K-means

Siendo consistentes con lo estudiado en apartados anteriores, ver apartados 5.1.2 y 5.4.1, se estudia el comportamiendo del algoritmo K-means seleccionando las 3 moléculas que, según la estadística, diferencian la muestra con TCC del grupo control. Este proyecto pretende dilucidar cómo el consumo continuado de cocaína modifica la cantidad en plasma de determinadas citoquinas. Ello trasladado al clustering que tiene lugar en esta sección implica diferenciar entre dos grupos: los consumidores en periodo de abstinencia y los controles. De esta forma es necesario comprobar si es correcto dividir la población total en dos grupos. El Índice de Silueta aplicado a las moléculas que aparentemente son capaces de cribar la población, indica que K=2 clusters es el valor óptimo, ver figura (5.12).

En cualquier caso para poder realizar una comparación con las secciones anteriores de clustering se fija el número de clusters en 3, véase figura (5.13). Los resultados del clustering, ver figura (5.14), presenta dos variable dicotómicas: TCC y diagnóstico dual. A simple vista, no se aprecia ninguna correspondencia entre los clusters obtenidos y el hecho de que el sujeto tenga un valor determinado en las variables dicotómicas. Incluso fijando los clusters en 2, como indica el Índice de Silueta de forma óptima, los resultados no indican que el agrupamiento sea coherente con las variabes que son objeto del estudio (TCC y diagnóstico dual). Podemos apreciar del agrupamiento obtenido que la componente principal 1 cuantifica en un 66.7 % las diferencias existentes entre los datos, quedando entonces pendiente por analizar las variables con mayor influencia dentro de esta componente principal.

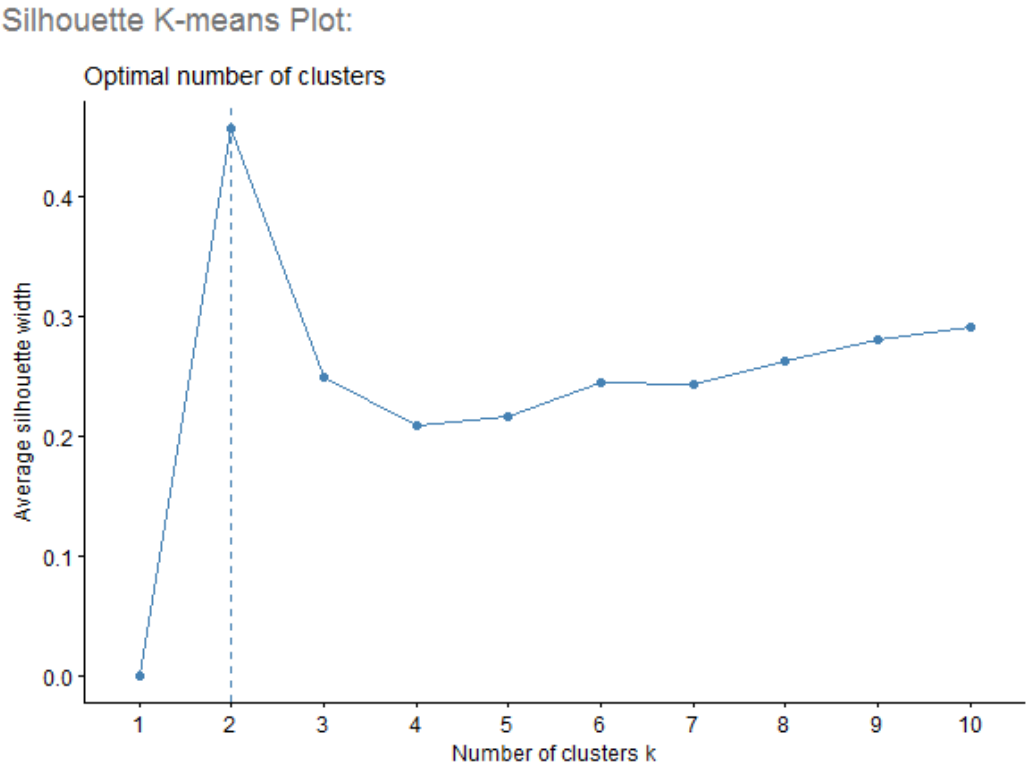


Figura 5.12: Datos agrupados según k-means.

Figura 5.13: Resultados del clustering obtenido mediante aplicación del algoritmo de k-means.

X	CODE	CUD	DualDiagnosis	Cluster
15	CPD-041	Yes	No	1
21	CPD-054	Yes	Yes	1
42	CPD-109	Yes	Yes	1
43	CPD-111	Yes	No	1
45	CPD-119	Yes	No	1
63	CPD-151	Yes	No	1
72	CPD-177	Yes	Yes	1
84	H12C-15	No	NA	1
85	H12C-16	No	NA	1
88	H12C-22	No	NA	1
89	H12C-27	No	NA	1
93	H12C-37	No	NA	1
94	H12C-42	No	NA	1
102	H12C-8	No	NA	1
111	LABC-124	No	NA	1
113	LABC-136	No	NA	1
124	LABC-18	No	NA	1
127	LABC-26	No	NA	1
134	LABC-41	No	NA	1
138	LABC-62	No	NA	1
142	LABC-67	No	NA	1
152	LABC-81	No	NA	1
160	LABC-92	No	NA	1
51	CPD-131	Yes	No	2
59	CPD-143	Yes	No	2
81	H12C-12	No	NA	2
83	H12C-14	No	NA	2
120	LABC-155	No	NA	2
123	LABC-17	No	NA	2
126	LABC-25	No	NA	2
128	LABC-27	No	NA	2
132	LABC-38	No	NA	2
141	LABC-66	No	NA	2
146	LABC-75	No	NA	2
150	LABC-79	No	NA	2
18	CPD-046	Yes	Yes	3
20	CPD-053	Yes	Yes	3
22	CPD-059	Yes	Yes	3
31	CPD-080	Yes	No	3
37	CPD-102	Yes	Yes	3
39	CPD-104	Yes	Yes	3
47	CPD-125	Yes	Yes	3
55	CPD-138	Yes	No	3
55	CPD-138	Yes	No	3
70	CPD-166	Yes	Yes	3
82	H12C-13	No	NA	3
90	H12C-31	No	NA	3
95	H12C-43	No	NA	3
99	H12C-53	No	NA	3
129	LABC-33	No	NA	3
136	LABC-50	No	NA	3
147	LABC-76	No	NA	3

Figura 5.14: Datos agrupados según k-means.

Capítulo 6

Conclusiones

Las conclusiones se van a abordar desde dos ámbitos, tal como se viene realizando hasta ahora en este proyecto. Por un lado, se afrontará el área que concierne al análisis de los datos y por otro en cuanto al software desarrollado para llevar a cabo dicho análisis.

Conclusiones sobre el estudio de pacientes con TCC

La búsqueda de biomarcadores objetivos para los trastornos mentales es un campo de interés creciente en la actualidad. Hay pruebas de que el sistema inmunológico está desempeñando un papel importante en la patogénesis de los trastornos mentales, abriendo expectativas en la búsqueda de biomarcadores de diagnóstico o dianas terapéuticas para trastornos adictivos y enfermedades psiquiátricas asociadas (Araos et al., 2015; García-Marchena et al., 2017).

En el presente estudio se ha intentado identificar posibles biomarcadores de trastorno por consumo de cocaína y/o trastorno de comorbilidad mental a través de la determinación de determinadas citoquinas y factores de crecimiento en el plasma. Los resultados obtenidos indican que los valores de las citoquinas IL-17α, MIP-1α y TGFα están alterados en los pacientes diagnosticados con TCC, aunque no se observa relación significativa con la gravedad de los síntomas de adicción. Los niveles de TGFα están además influenciados por la presencia de comorbilidad psiquiátrica, en particular en los pacientes con patología dual. Como antítesis, es posible mencionar una publicación reciente que demuestra que los pacientes con trastorno por consumo de alcohol tienen alterada la citoquina Eotaxin-1 en pacientes que presentan comorbilidad mental, en especial aquellos con trastornos de estado de ánimo y ansiedad (García-Marchena et al., 2017a). Sin embargo, no se han encontrado estos efectos en los niveles de la citoquina mencionada en pacientes con TCC con diagnóstico dual.

La disminución de los niveles plasmáticos de IL-17α, MIP-1α y TGFα son similares a los descritos para la misma población para otras citoquinas y quimioquinas como TNFα, MCP-1 y SDF-1 (Araos et al. 2015). Aunque los estudios de las consecuencias del consumo de cocaína describen efectos proinflamatorios, la disminución de los niveles de IL-17α,

MIP-1α y TGFα puede ser interpretada como un efecto compensatorio que reduce el tono inflamatorio a lo largo del periodo de abstinencia. Un fenómeno similar se describe para otras citoquinas en los mismos pacientes del estudio en régimen ambulatorio (Araos et al., 2015; Pedraz et al., 2015). Sin embargo, ninguna de las citoquinas que se analizan en el estudio presente están asociadas con los síntomas de gravedad del TCC, a diferencia de lo que se encuentra para IL-1β, CX3CL1[1] o SDF-1 como moléculas sensibles a la gravedad de los síntomas del TCC (Araos et al., 2015). Cabe exponer que los pacientes con diagnóstico dual muestran niveles bajos de TGFα con respecto a los pacientes que no lo padecen. En consecuencia, la presencia del diagnóstico dual en un gran número de pacientes en el grupo de cocaína hace notoria la disminución de los niveles de TGFα en comparación al grupo control.

Deben considerarse algunas limitaciones en el presente estudio, para mejorar hallazgos futuros que surjan de continuar este proyecto. En primer lugar, se necesitan estudios más amplios en lo que a población se refiere, ya que existe una amplia variabilidad cuando se considera la asociación de señales moleculares inmunes con los trastornos mentales. Un segundo aspecto a mejorar sería el bajo número de pacientes mujeres, siendo necesario incrementar este conjunto para tener una mejor visión general de los resultados obtenidos.

Conclusiones sobre el software desarrollado

El software desarrollado ha superado las expectativas iniciales del alumno en lo que se refiere al diseño y a la versatilidad. Conseguir una interfaz gráfica amigable era uno de los requisitos principales, destacándose que el producto final posee una apariencia profesional, permitiendo realizar los análisis requeridos en un tiempo mucho menor y de forma más organizada a como el cliente los realizaba a través del uso de una combinación de herramientas. Con el lenguaje escogido ha sido posible implementar todo el conjunto de técnicas estadísticas solicitadas por el cliente, siendo satisfactoria la elección por tanto del lenguaje R, de su entorno de desarrollo Rstudio, y de su marco de desarrollo de aplicaciones Shiny. Vale la pena destacar también que la aplicación desarrollada, además de incorporar todos los métodos estadísticos requeridos por el cliente, cuenta con un módulo de análisis de clustering, el cual permite la aplicación de tres algoritmos diferentes (Clustering jerárquico, Algoritmo de K-medias y Mapas auto-organizados), posibilitando también la visualización y exportación de los resultados. Si bien este módulo no es un requisito del cliente, consideramos que es un valor añadido para la aplicación, ya que por un lado muestra la potencia del lenguaje R, a través del uso de sus paquetes, sino que permitirá el análisis de datos desde una nueva perspectiva con la potencialidad de avanzar en el estudio de pacientes con TCC.

A pesar del exitoso resultado existen varios factores que podrían modificarse y/o ampliarse en un corto plazo para que la aplicación sea lo más fructífera posible. En primer lugar se debería añadir la funcionalidad de seleccionar y transformar logarítmica, exponencial e inversamente determinadas variables, ya que en la versión actual únicamente se incluye

[1] Fractalquina

la transformación logarítmica. Un segundo aspecto a considerar sería que las variables moleculares a estudiar no sean únicamente las analizadas en este proyecto, sino que el usuario decida las moléculas que desea incluir en el estudio. En tercer lugar, existe la necesidad de añadir análisis psicológicos, de forma que en la misma aplicación se puedan comprobar y analizar todos los aspectos que los especialistas en psicología consideran necesarios para los pacientes con TCC. En cuarto lugar deberían añadirse algunas opciones matemáticas adicionales en lo que concierne a la definición de parámetros de las funciones. Por ejemplo, permitir hacer ANCOVAS tipo I, tipo II o tipo III según el criterio del usuario acorde al conjunto de datos en cuestión, permitiendo también la comprobación de todos los supuestos y consideraciones necesarios para poder llevar a cabo un análisis ANCOVA. Como quinto aspecto se propone ampliar la aplicación en técnicas de Aprendizaje Computacional, siendo necesario un estudio exploratorio previo a fin de definir las técnicas más adecuadas. Finalmente se está analizando dar acceso público a la aplicación a través de repositorios en la nube utilizando una licencia del tipo "creative commons", de tal forma que un mayor número de usuarios puedan acceder a ella.

Bibliografía

P. Araos, E. Vergara-Moragues, M. Pedraz, F.J. Pavón, R. Campos Cloute, M. Calado, J.J. Ruiz, N. García-Marchena, I. Gornemann, M. Torrens, and F. Rodríguez de Fonseca. Psychopathological comorbidity in cocaine users in outpatient treatment. Adicciones, 26(1):15–26, 2014.

P. Araos, M. Pedraz, A. Serrano, M. Lucena, V. Barrios, N. García-Marchena, R. Cloute, J.J. Ruiz, P. Romero, J. Suárez, E. Baixeras, R. de la Torre, J. Montesinos, C. Guerri, M. Rodríguez-Arias, J. Miñarro, R. Martínez-Riera, M. Torrens, J.A. Chowen, J. Argente, B.J. Mason, F.J. Pavón, and F. Rodríguez de Fonseca. Plasma profile of pro-inflammatory cytokines and chemokines in cocaine users under outpatient treatment: influence of cocaine symptom severity and psychiatric co-morbidity. Adicction Biology, 20(4):756–72, 2015.

R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley and Sons, New York, 2000.

H. Engler, P. Brendt, J. Wischermann, A. Wegner, R. Röhling, T. Schoemberg, U. Meyer, R. Gold, J. Peters, S. Benson, and M. Schedlowski. Selective increase of cerebrospinal fluid il-6 during experimental systemic inflammation in humans: association with depressive symptoms. Molecular Psychiatry, Epub ahead of print, 2017.

K.K. Friedler, N. Kim, D.G. Kondo, and P.F. Renshaw. Cocaine use in the past year is associated with altitude of residence. Journal of Addiction Medicine, 6(2):166–71, 2012.

N. García-Marchena, P. Araos, L V. Barrios, Sánchez-Marín, J.A. Chowen, M. Pedraz, E. Castilla-Ortega, P. Romero-Sanchiz, G. Ponce, F A. Gavito, Rodríguez de Fonseca, and F.J. Pavón. Plasma chemokines in patients with alcohol use disorders: Association of ccl11 (eotaxin-1) with psychiatric comorbidity. Front Psychiatry, 7(214), 2017.

M. Goossens, F. Mittelbach, and A. Samarin. The LaTeX Companion. Addison-Wesley, Reading, Massachusetts, 1993.

D.S. Hasin, K.D. Trautman, G.M. Miele, S. Samet, M. Smith, and J. Endicott. Psychiatric research interview for substance and mental disorders (prism): reliability for substance abusers. American Journal Psychiatry, 153(9):1195–201, 1996.

D.S. Hasin, M.C. Fenton, C. Beseler, J.Y. Park, and M.M. Wall. Analyses related to the development of dsm-5 criteria for substance use related disorders: 2. proposed dsm-5 criteria for alcohol, cannabis, cocaine and heroin disorders in 663 substance abuse patients. Drug and Alcohol Dependence, 122(1-2):28–37, 2012.

D.S. Hasin, C.P. O'Brien, M. Auriacombe, G. Borges, K. Bucholz, A. Budney, W.M. Compton, T. Crowley, W. Ling, N.M. Petry, M. Schuckit, and B.F. Grant. Dsm-5 criteria for substance use disorders: recommendations and rationale. American Journal Psychiatry, 170(8):834–51, 2013.

E.A. Hoge, E. Bui, S.A. Palitz, N.R. Schwarz, M.E. Owens, J.M. Johnston, M.H. Pollack, and N.M. Simon. The effect of mindfulness meditation training on biological acute stress responses in generalized anxiety disorder. Psychiatry Research, 1781(16):30847–2, 2017.

M.R. Irwin, L. Olmos, M. Wang, E.M. Valladares, S.J. Motivala, T. Fong, T. Newton, A. Butch, R. Olmstead, and S.W. Cole. Cocaine dependence and acute cocaine induce decreases of monocyte proinflammatory cytokine expression across the diurnal period: autonomic mechanisms. Journal of Pharmacology and Experimental Therapeutics, 320 (2):507–15, 2007.

L. Karila, A. Petit, W. Lowenstein, and M. Reynaud. Diagnosis and consequences of cocaine addiction. Current Medical Chemistry., 19(33):5612–8, 2012.

T. Kohonen, M.R. Schroeder, and T.S. Huang, editors. Self-Organizing Maps. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001. ISBN 3540679219.

R.V. Lenth. Least-squares means: The r package lsmeans. Journal of Statistical Software, 69(1):33, 2016.

M.L. Levandowski, T.W. Viola, C.H. Prado, A. Wieck, M.E. Bauer, E. Brietzke, and R. Grassi-Oliveira. Distinct behavioral and immunoendocrine parameters during crack cocaine abstinence in women reporting childhood abuse and neglect. Drug and Alcohol Dependence, 167:140–8, 2016.

O. Maimon and L. Rokach. Data Mining and Knowledge Discovery Handbook. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387244352, 9780387244358.

J.I. Mestre-Pintó, A. Domingo-Salvany, R. Martín-Santos, M. Torrens, and PsyCoBarcelona Group. Dual diagnosis screening interview to identify psychiatric comorbidity in substance users: development and validation of a brief instrument. European Addiction Research, 20(1):41–8, 2014.

M. Meulders. An r package for probabilistic latent feature analysis of two-way two-mode frequencies. Journal of Statistical Software, 54(14):29, 2013.

V. Michopoulos, A. Powers, C.F. Gillespie, K.J. Ressler, and T. Jovanovic. Inflammation in fear- and anxiety-based disorders: Ptsd, gad, and beyond. Neuropsychopharmacology, 42(1):254–270, 2017.

S. Morgello, C.E. Holzer, E. Ryan, C. Young, M. Naseer, S.A. Castellon, A.B. Frol, J.H. Atkinson, B.B. Gelman, I. Grant, and E.J. Singer. Interrater reliability of the psychiatric research interview for substance and mental disorders in an hiv-infected cohort: experience of the national neuroaids tissue consortium. International Journal of Methods in Psychiatry Research, 15(3):131–8, 2006.

T. Notter, J.M. Coughlin, T. Gschwind, U. Weber-Stadlbauer, Y. Wang, M. Kassiou, A.C. Vernon, D. Benke, M.G. Pomper, A. Sawa, and U. Meyer. Translational evaluation of translocator protein as a marker of neuroinflammation in schizophrenia. Molecular Psychiatry, Epub ahead of print, 2017.

R.C. Paolicelli, G. Bolasco, F. Pagani, L. Maggi, M. Scianni, P. Panzanelli, M. Giustetto, T.A. Ferreira, E. Guiducci, L. Dumas, D. Ragozzino, and C.T. Gross. Synaptic pruning by microglia is necessary for normal brain development. Science, 333(6048):1456–8, 2011.

N. Parikh, W. Dampier, R. Feng, S.R. Passic, W. Zhong, B. Frantz, B. Blakey, B. Aiamkitsumrit, V. Pirrone, M.R. Nonnemacher, J.M. Jacobson, and B. Wigdahl. Cocaine alters cytokine profiles in hiv-1-infected african american individuals in the drexelmed hiv/aids genetic analysis cohort. Journal of Acquired Immune Deficiency Syndromes, 66(3):256–64, 2014.

F.J. Pavón, P. Araos, A. Pastor, M. Calado, M. Pedraz, R. Campos-Cloute, J.J. Ruiz, A. Serrano, E. Blanco, P. Rivera, J. Suárez, M. Romero-Cuevas, M. Pujadas, E. Vergara-Moragues, I. Gornemann, M. Torrens, R. de la Torre, and F. Rodríguez de Fonseca. Evaluation of plasma-free endocannabinoids and their congeners in abstinent cocaine addicts seeking outpatient treatment: impact of psychiatric co-morbidity. Addiction Biology., 18(6):955–69, 2013.

M. Pedraz, A.I. Martín-Velasco, N. García-Marchena, P. Araos, A. Serrano, P. Romero-Sanchiz, J. Suárez, E. Castilla-Ortega, V. Barrios, R. Campos-Cloute, J.J. Ruiz, M. Torrens, J.A. Chowen, J. Argente, R. de la Torre, L.J. Santín, M.A. Villanúa, F. Rodríguez de Fonseca, and F.J. Pavón FJ. Plasma concentrations of bdnf and igf-1 in abstinent cocaine users with high prevalence of substance use disorders: relationship to psychiatric comorbidity. Plos One, 10(3):756–72, 2015.

J.R. Petrulli, B. Kalish, N.B. Nabulsi, Y. Huang, J. Hannestad, and E.D. Morris. Systemic inflammation enhances stimulant-induced striatal dopamine elevation. Translational Psychiatry, 7(3):–, 2017.

N. M. Radziwill. Statistics (The easier way) with R. Lapis Lucera, San Francisco, California, 2015.

C.L. Raison, L. Capuron, and A.H. Miller. Cytokines sing the blues: inflammation and the pathogenesis of depression. Trends in immunology, 27(1):24–31, 2006.

L.N. Robins, J. Wing, H.U. Wittchen, J.E. Helzer, T.F. Babor, J. Burke, A. Farmer, A. Jablenski, R. Pickens, D.A. Regier, and et al. The composite international diagnostic interview. an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Plos One, 45(12):1069–77, 1988.

G. Rufles. Emergency care for young people who have used cocaine. Emergency Nurse., 17(8):34–6, 2009.

Shiny. Shiny tutorial and articles. `https://shiny.rstudio.com`, 2017. Accessed: 2017-06-28.

I. Sommerville and P. Sawyer. Requirements Engineering A good practice guide. John Wiley and Sons, New Jersey, 1997.

M. Torrens, D. Serrano, M. Astals, G. Pérez-Domínguez, and R. Martín-Santos. Diagnosing comorbid psychiatric disorders in substance abusers: validity of the spanish versions of the psychiatric research interview for substance and mental disorders and the structured clinical interview for dsm-iv. American Journal Psychiatry, 161(7):1231–7, 2004.

J. Verzani. Using R for Introductory Statistics. Chapman & Hall, London, United Kingdom, 2004.

J.L. Gastwirth W. Hui, Y.R. Gel. lawstat: An r package for law, public policy and biostatistics. Journal of Statistical Software, 28(3):33, 2008.

P. Wittek, S.C. Gao, I.S. Lim, and L. Zhao. somoclu: An efficient parallel library for self-organizing maps. Journal of Statistical Software, 78(9):21., 2017.