A bioinformatics approach to the study of comorbidity. Insight into mental disorders.

Alba Gutiérrez-Sacristán

TESI DOCTORAL UPF / 2017

DIRECTOR DE LA TESI

Dra. Laura Inés Furlong

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra Barcelona

We received support from ISCIII-FEDER (PI13/00082, CP10/00524, CPII16/00026), IMI-JU under grants agreements no. 115191 (Open PHACTS)], no. 115372 (EMIF), no. 115735 (iPiE), resources of which are composed of financial contribution from the EU-FP7 (FP7/2007- 2013) and EFPIA companies in kind contribution, and the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. AGS acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the "María de Maeztu" Programme for Units of Excellence in R&D (MDM-2014-0370).

Para mamá y papá.

Para ti abuelo Helios, mi inspiración.

Acknowledgements

Un paso más hacia adelante, un pasito más que sola hubiese sido imposible dar. A todas esas personas que me han ayudado a levantarme, a seguir hacia delante y no tirar la toalla. **Lo mejor de mí, sin lugar a duda, es la gente que me rodea.** Porque hay trabajos que solo llevan un autor, pero que guardan dentro de ellos a todos los que lo han hecho posible. GRACIAS, gràcies, eskerrik asko, thanks, merci, grazie, ευχαριστώ, kiitoksia, spasibo, köszönöm, 谢谢, ممنون! Sabéis que podría escribir una tesis solo con los agradecimientos, y dedicaros al menos un capítulo a cada uno de vosotros, pero intentaré resumir toda mi gratitud en unas pocas páginas.

En primer lugar me gustaría agradecer esta tesis a Laura I. Furlong, gracias por darme la oportunidad de vivir esta aventura que ha sido el doctorado. Gracias por no dejar que abandonase nunca, por ayudarme a encontrar la motivación cuando parecía perdida, por apoyarme en los altibajos de estos años. Por darme la oportunidad de vivir y descubrir la ciencia aquí y en Boston, ayudándome a sacar lo mejor de mí misma. Gracias Laura por ser parte fundamental de esta aventura.

Thanks to Dr. Paul Avillach for giving me the opportunity to work in his lab in the amazing Harvard Medical School. It has been one of the best experiences of my life, six months of intense work, where I have learned about science and life. Thanks Paul for all

your advice. Each meeting with you was a unique opportunity to learn. I admire you and your work. Merci beaucoup.

Un especial agradecimiento al Dr. Ferrán Sanz, por sus grandes consejos, por saber escuchar, por sacar siempre un momento para resolver esas dudas, sobre todo en estadística! Y por confiar y pensar en mí para esas clases de R, enseñar siempre fue mi gran sueño, gracias Ferrán y Manuel Pastor por hacerlo realidad. Sin duda, una gran experiencia que me llevo conmigo de estos años. Moltes gràcies.

Y ahora permitidme que eche la vista atrás. Y es que, a lo largo de la vida, muchas son las personas que nos acompañan, que nos ayudan a crecer, como profesionales y como personas. Tantos y tantos a los que agradecer. Y es que hay personas que dejan una huella imborrable en nosotros, y que significan mucho más de lo que ellos mismos podrían imaginar. Por esos profesores que ya desde el colegio, me transmitieron no solo conocimientos, sino lo que es más importante, la pasión por lo que uno hace. Especialmente a ti Pepa, porque con los años te has convertido en mucho más que una profe de inglés, una amiga, gracias por ofrecerme tu mano en mis primeros pasos, y por no dejarme caer, tú ya sabes a que me refiero! A Chema, porque aunque el dibujo técnico nunca fue lo mío, me enseñaste que todo se puede hacer, y que hay muchas perspectivas de ver las cosas, la isométrica, la caballera... Gracias por seguir acompañándome años después de dejar el instituto! Y mil gracias a Joan Bertrán, por darme mi

primera oportunidad de trabajar en la bioinformática. Gracias a tí descubrí que aunque pipetas y microscopios no fuese lo mío, había algo más ahí fuera. ¡Nunca pierdas esa pasión con la que enseñas! Gracias. Y gracias por abrirme las puertas del BSC! Mi primera gran aventura con la bioinformática, terminales, ssh, mare nostrum… ¿cómo haber sobrevivido a esos meses sin todas y cada una de las personas tan maravillosas que allí conocí? Gracias Pau, por ayudarme esos primeros meses, por tu paciencia y ayuda, nunca lo olvidaré. Y a tí Ramón, por darme la oportunidad de ver qué era la bioinformática, y por apoyarme y confiar en mí trabajo. Sin duda, uno de los grandes culpables de que a día de hoy la bioinformática sea mi pasión.

A los compañeros de Sant Pau, Laura, Miquel, Helena, Andrey, gracias. Mi Lau, gracias por tantos momentos compartidos, por nuestros conciertos y Sonoramas, Helena, gracias por estar siempre ahí, dispuesta a escucharme, animarme, a leerte una tesis y lo que hiciese falta! ¡Aquí llega la Dra. Aranda! jaajaj Andrey, nunca podré agradecerte lo suficiente todo tu apoyo en Boston, bonita casualidad volvernos a encontrar al otro lado del charco! ¡Gracias!

Y es que, ¿cuántas horas al día pasamos con esas personas que se hacen llamar colegas de trabajo? Compañeros de cafés, de risas, de locuras, de momentos de estrés y deadlines! Rodny, Gabriel, Luis, Martina, todo un placer. Alfons, Miguel, mis ITs preferidos, gracias por estar siempre dispuestos a echar una mano, dos y las que hiciesen falta, instalándome paquetes, porque está claro que yo

siempre R que R. Y en estos agradecimientos no podían faltar ellas, Carina, Maria! Personas realmente excepcionales, grandes profesionales y mejores personas que te hacen la vida más sencilla. ¡Carina, gracias por estar siempre ahí, por saltar conmigo de alegría, y sentarte a mi lado y darme un abrazo cuando más lo necesitaba, porque Aranda y sus bodegas siempre tendrán las puertas abiertas para ti! ¡María, mi finlandesa preferida, gracias por tu paciencia ayudándome con mi inglés, corrigiendo mis artículos, cartas, emails… gracias por tantos y tantos buenos momentos! ¡Nos vemos por el sur!

Desde que llegué al IBI hace ya unos años me hicisteis sentir una más, y convertisteis el PRBB en mi segunda casa. Algunos se marcharon a descubrir nuevos mundos, Solène, merci beaucoup por enseñarme tanto durante los meses que compartimos juntas! Y gracias por ser parte imprescindible de ese primer paper y gran proyecto que es PsyGeNET. Emilio gracias por tantos buenos momentos y por tu apoyo en estos últimos meses de estrés, horroros! Y que decirte a ti, Alexia, una vez más, GRACIAS por todo. ¡He aprendido mucho de ti y contigo! ¡Gracias por estar siempre dispuesta a echarme una mano!

Y ahora les toca a ellos, mis dos pilares fundamentales en estos años, colegas, confidentes, amigos, Àlex, Núria, ¡GRACIAS! Nurieta, qué te voy a contar que tú no sepas ya. Una de las personas con el corazón más grande que he conocido, gracias por esos abrazos, por tantos y tantos momentos, por nuestros cafés, nuestro

ribera del Duero y ese vermut de Reus. ¡Gracias por convertirte en esa amiga loca que todo el mundo debería tener! Este mundo no es lo suficientemente grande como para que no volvamos a encontrarnos! Y que decirte a ti Àlex… no me imagino estos años sin ti, que suerte que la vida te pusiera en mi camino! El mejor colega y amigo que uno puede tener, siempre dispuesto a ayudar, siempre dispuesto a escuchar, siempre con una sonrisa, único, indescriptible. Si algo bueno me ha dado estos años es haberte conocido. ¡Gracias por hacerme reír cada día! Por no dejarme nunca sola y por no dejar que tirase la toalla en mis varios intentos por abandonar. Y gracias por cada uno de los momentos que me has regalado, tú y esa persona tan maravillosa que tienes como pareja, Eleonora. Gracias a los dos de corazón.

And thanks a lot to all the great people of the Avillach's lab! Tom, Jason, Andre, Alex, Ranjay, Li Li, Cartik…guys, you are awesome! It was a pleasure to share these six months with you! Work, parties and some swim race! Gabe, THANK YOU for everything! Life wouldn't be the same without those who lead you to do things that you never even thought about it, like swimming a mile! ☺ I will never forget it! Cartik, Miss Rcupcake mermaid misses you a lot! I would like to thanks specially Carlos, Niloofar and Romain. Carlos, gracias por ese par de meses juntos, porque echar lavadero contigo es lo más! ¡Nos vemos pronto! No mamees! Niloofar, my Iranian friend, you are so special! Thanks to all the moments we share together! I hope things will change soon and we could visit Spain together! ممنون Take care, my friend! And Romain, my French

friend, I have no words to thank you all you did for me during my Boston experience. Thank you for your support, for all your help and advice. And for all the moments I shared with your incredible family! Thanks because you made me feel like at home! Merci beaucoup.

A mis bioinformáticos preferidos, Ricard, Marina, Carles, Merx, Oihane, Fran. ¡Gracias por tantos momentos compartidos! Porque nadie mejor que vosotros sabe lo que cuesta una tesis. Carles, por fin lo conseguimos, nuestro artículo publicado. Gracias por enseñarme tanto, porque R que R, al final lo logramos. Gracias por acompañarme en parte de este camino. Merx, Oihane, que deciros que no sepáis. Empezar la mañana con un "buenos diaaas!" por hangouts era el chute de energía necesario para comerse el mundo. Gracias por estar a mi lado estos años, que lo que unió el master de bioinformática no lo separe nadie. ¡Oihane, prepara la hamaca, que allá voy! Merx, gracias a ti y a tu familia, por hacerme sentir como en casa, por hacerme sonreír y ser mi pequeña de las dudas infinitas… Fran, gracias, siempre es un placer charlar, discutir y debatir contigo. Gracias por estar ahí, por ayudarme más de lo que quizás tú mismo imaginas. ¡Gracias a todos! #bioinformaticsdrama

Y ahora les toca a ellos, a los de siempre. A mis arandinos, Dani, Carlos y Marian, gracias. Por demostrarme que no hay distancia que esté lejos, que aunque los años pasen, volver a casa siempre es un placer cuando sabes que personas tan maravillosas te están esperando. Gracias Marian, porque aunque haga años de esa

despedida entre lloros y abrazos, haces que cada vez que hablemos sea como si el tiempo no pasara y la distancia no existiera, gracias por estar siempre ahí! Y especialmente gracias a ti Dani, nunca podré agradecerte todo lo que has hecho por mí, sin ti hoy no estaría escribiendo esta tesis. Por creer en mí cuando ni yo lo hacía, por estar tan cerca aun estando lejos, por hacerme ver el lado bueno de las cosas, por estar siempre a mi lado y recordarme que valía más de lo que yo misma pensaba. Por estar conmigo cuando la montaña rusa estaba abajo y juntos lograr que subiese. ¡GRACIAS AMIGO!

A mi David, porque no hay nada mejor de una vuelta a casa que saber que tú estarás allí para darnos uno de nuestros achuchones, gracias. Y a ti Alfon, porque viajar es mi terapia preferida, gracias por tantas aventuras vividas y rincones descubiertos juntos. ¡Y los que nos quedan! Lore, gracias por estar ahí a nuestra manera, porque nadie dijo que fuese fácil estar sin estar. Mi Cris, mi rubia, la niña de mis ojos. Gracias por no soltarme nunca de la mano. Por recordarme en tantos momentos, que nunca estaré sola. Qué por muchos años que hayan pasado desde que jugábamos en la Glorieta, la amistad que se allí se forjó no hay nada ni nadie que pueda romperlo. ¡GRACIAS! A mis segovianas, Ana y Raquel, porque no toda distancia es ausencia ni todo silencio olvido. Gracias por estar siempre a mi lado. Y no podía faltar ella, Marta, porque no hace tanto que nos conocemos, pero parece que siempre hubieses estado ahí. Gracias por tus mensajes de apoyo, por nuestros audios interminables, y por tantos momentos juntas. A ti y a toda tu familia, que ya sois parte de la mía, gracias!

Y es que no sería la misma sin todas esas personas que han pasado por mi vida y han dejado un pedacito de ellas conmigo, esas con las que has compartido cañas después de las clases, noches de café pre-examen, risas, lloros, esos primeros pasos fuera del nido. David, Elena, ¡gracias! Luis Ángel, gracias por seguir ahí, por las cenas improvisadas, las noches arreglando el mundo, y poniendo el contrapunto a tus grandes proyectos que estoy segura harás realidad. Guillem, gracias por contar conmigo para ese maravilloso proyecto que es Biocloud y, juntos hacer realidad ese sueño que ha sido crear nuestra Start-up! ¡Moltes gràcies i molta sort en el teu futur!

Y ahora te toca a ti, más que una compañera, más que una amiga, mi relación estrecha, mi otro yo. Con una mirada puedo decirte mucho más que en mil palabras, Lier, el destino nos hizo amigas, pero la vida nos convirtió en hermanas. ¡Nunca podré agradecerte todo lo que has hecho por mí estos años! Parte de lo que soy, te lo debo a ti. Gracias por tu apoyo, porque siempre estas dispuesta a escucharme, por no dejarme nunca sola, por recordarme siempre que al final todo estará bien, y si no lo está es que todavía no ha llegado el final. Y por no dudar ni un segundo en cruzar el charco para darnos un abrazo, porque no puedo imaginar mi vida sin ti a mi lado. ¡GRACIAS! ¡Te quiero mucho! Esta tesis lleva tu nombre, no lo olvides. Y gracias a Raúl y Montse, por incluirme en el álbum familiar, por hacerme sentir una más, por vuestro apoyo y cariño. Gracias a los tres por convertiros en parte de mi familia.

A ti Alberto, por apoyarme cada día, por no dejar que la distancia fuese un impedimento para sentirte cerca. Por escucharme, por animarme y hacerme reír. Por ayudarme a ver las cosas desde otra perspectiva y a no correr más de la cuenta, paso a paso pero siempre hacia delante. Porque has sido y eres un apoyo fundamental para mí. GRACIAS feo ;)

Y ahora permitirme, que les dedique más que unas líneas a ellos, a mi familia. Que decir que no sepáis, sacristanes y monaguillos, TODOS y cada uno de vosotros....gracias por confiar en mí más que yo misma, por estar siempre a mi lado, por esos mensajes de apoyo que tanto han significado para mí. A mis primos, especialmente a David y Miriam, gracias por ser mucho más que eso, por ser amigos, confidentes, ese apoyo incondicional. A mis tíos, especialmente a ti Jose, por esa confianza en mí, por recordarme cada día durante estos últimos meses que la vida está para disfrutarla. A mis abuelos, Concha y Julián. Cuanto hubieses disfrutado y cuanto hubieses llorado en un día como hoy, ¡casi tanto como en mi graduación! Y a mis segundos padres, mis abuelos Helios y Andrea. Abuela, gracias por ser mi segunda madre, por esos abrazos interminables, por esas miradas cómplices, por creer en mí y estar siempre a mi lado. Os quiero mucho.

A ellos, a los que lo han dado todo por mí, para los que no ha habido distancias, papá, mamá, sin vosotros jamás hubiera llegado hasta aquí. Gracias por darme raíces y alas, por vuestro apoyo incondicional, porque no hay distancia que esté lejos. Esta tesis es

tan vuestra como mía. No podré agradeceros nunca todo lo que habéis hecho y hacéis por mí. Por confiar en mí cuando ni yo lo hacía, por tirar de mí hacia adelante, por estar siempre a mi lado, a las buenas y a las no tan buenas. Por recordarme siempre que el que quiere puede y lo consigue. Sois lo más valioso que tengo en mi vida. GRACIAS.

Y quisiera acabar estos agradecimientos con él, porque se merece mucha más que una frase. Quien le conoció lo sabe, la mejor persona que ha habido, quien me inspiró a dedicar mi vida a la ciencia, al saber, al investigar, mi abuelo Helios. Abuelo, sé que hoy te sentirías muy orgulloso y que nada te gustaría más que poder sentarte hoy a mi lado y juntos releer esta tesis, porque hubieses sido capaz de aprender inglés para poderla leer por ti mismo y no me cabe duda de que hubieses sido mi mejor crítico. Porque tú eras capaz de todo. Por luchar hasta el final, por darnos a todos una gran lección de lo que eran las ganas de vivir. Porque has sido, eres y serás mi guía, mi luz en los días oscuros, porque no hay día en que no piense en ti. Porque estás presente en todo lo que hago. Porque yo de mayor quiero ser como tú. Te echo de menos. ¡Esta tesis es para ti!

Abstract

Clinical and epidemiological studies show that comorbidity, the coexistence of disorders in a patient, has a great impact on the evolution of the health status of patients. Therefore, comorbidity analysis is key to identify new preventive and therapeutic strategies, walking through a more personalized medicine. In order to harness the power of the increasing volume of available health information in the era of big data, this thesis presents the development of new tools and resources for the identification of comorbidity patterns, based on the clinical and molecular information. The `comoRbidity` package and the `psygenet2r` one presented in this thesis provide an adequately complete and comprehensive analysis of comorbidities and in particular, offer the users the possibility to design their own comorbidity study according to their needs and specifications. Moreover, due to the significant role that plays the molecular information in interpreting the cause of disease comorbidities and the lack of resources to collect that information in the specific area of mental disorders, a new manual curated database, PsyGeNET, focus on gene-disease association has also been developed.

In summary, all the tools developed in this thesis, available to the scientific community and already applied to several studies in the biomedical field, are of immense practical value for the comorbidity analysis and can aid to transform clinical information in a form of knowledge that can be analyzed, interpreted by researchers and applied leading overall, to more personalized medicine.

Resumen

Estudios clínicos y epidemiológicos muestran que la comorbilidad, la coexistencia de varias enfermedades en un mismo paciente, tiene un gran impacto en la evolución de su estado de salud. Por lo tanto, el análisis de comorbilidades es clave para identificar nuevas estrategias preventivas y terapéuticas, trabajando hacia una medicina más personalizada. Con el fin de aprovechar el potencial del creciente volumen de información de salud disponible en la época del "big data", esta tesis presenta el desarrollo de nuevas herramientas y recursos para la identificación de patrones de comorbilidad, basados en la información clínica y molecular. Las herramientas `comoRbidity` y `psygenet2r` presentados en esta tesis permiten analizar las comorbilidades de forma amplia y completa, y en particular, ofrecen a los usuarios la posibilidad de diseñar su propio estudio de comorbilidad según sus necesidades y especificaciones. Por otra parte, debido al importante papel que juega la información molecular en la interpretación de la causa de comorbilidades y la falta de recursos para recopilar esta información en el área específica de los trastornos mentales, una nueva base de datos, PsyGeNET, se ha desarrollado centrada en las asociaciones gen-enfermedad. En resumen, todas las herramientas desarrolladas en esta tesis, disponibles en el dominio público y aplicadas ya en estudios del campo biomédico, son de gran valor práctico para el análisis de la comorbilidad y puede ayudar a transformar la información clínica en conocimiento que puede ser analizado, interpretado por los investigadores y aplicado para lograr una práctica de la medicina más personalizada.

Preface

"*Don't limit your challenges, challenge your limits!*" This quote summarizes the last years in my working life. During this Ph.D., every day has been a great challenge, and the main goal has been continuing growing without losing motivation.

As a bioinformatician with a specific interest in biomedical research, during the last years, I have explored a wide range of bioinformatic approaches, from molecular dynamics and systems biology to network medicine. In 2013 I joined the Integrative Biomedical Informatics (IBI) Group in order to conduct research towards a Ph.D. degree in bioinformatics, under the direction of Dr. Laura I. Furlong. In the undertaken Ph.D. work, my research has mainly focused on the development of bioinformatic approaches to analyze disease comorbidities. Developing a system medicine approach for the exploration of the relationships between diseases at different levels will provide key information to a better understanding of the biologic mechanisms of comorbidities.

At the end of this thesis, and after having exploited clinical data from patient record databases to identify comorbid diseases, I started to analyze comorbidities from the genomic point of view, exploring the molecular and cellular mechanism that substantiate comorbidities, more specifically for psychiatric disorders comorbidities, gaining a broad insight into the disease comorbidities.

This thesis is organized as follows: the challenging task of analyzing clinical health record data and the comorbidity concept will be introduced in Chapter 1. Furthermore, the need of developing new strategies for the analysis of clinical data as well as the necessity of having curated molecular information related to mental disorders will be discussed, and current tools in biomedical field will be described. In Chapter 2, the motivation and objectives of this thesis will be presented. The general methodology of the comorbidity study will be described in Chapter 3, introducing the need for new tools for analyzing clinical health records and comorbidities from the molecular perspective. Complementary information, applications, and results of the comorbidity software and PsyGeNET database will be presented in Chapter 4. In Chapter 5 a discussion of the work conducted in this thesis, together with limitations and future perspectives will be provided. Conclusions will be drawn in Chapter 6. Finally, publications and contributions to conference and workshops will be listed in the Appendixes.

Table of contents

List of figures

List of table

1 Introduction

"Science knows no country, because
knowledge belongs to humanity and
is the torch which illuminates the world."
Louis Pasteur (1822-1895)

1.1 Introduction to the concept of comorbidity

Mr. A. visits his general practitioner as he has presented cough and fever for a few days. According to what the doctor was taught in medical school, pharyngitis, pneumonia, chronic cardiac disease and pulmonary cancer are some of the candidate diseases associated with these symptoms. Furthermore, a clinical examination and certain laboratory tests can assist in making the right diagnosis (pharyngitis in this case), and then he can prescribe a particular treatment for this specific condition. This could be a naïve conception of how medicine works.

In fact, most diseases have symptoms that reflect several disorders occurring at the same time in the same patient. Mr. A may suffer, at the same time, a chronic cardiac disease and an acute pharyngitis, while both can be responsible for the symptom of cough. A significant number of questions arise: "*Is pulmonary cancer more likely because of the chronic cardiac disease, even if the patient has pharyngitis*? And if so, *is there a need for specific tests in this particular patient with a cough*? *Which drugs for the treatment of pharyngitis are not harmful to a patient with a chronic cardiac disease*? *Can the chronic cardiac disease modify the results of the laboratory tests*? ...

Patients usually do not present only one disorder, but rather the co-occurrence of a bunch of diseases, such that the latter is the rule rather than the exception. Alvan R. Feinstein was the first to

emphasize this problem in his seminal paper in 1970, in which he also coined a term for co-existing disorders: '*comorbidity*' [1].

1.1.1 The definition of Comorbidity: origin and discrepancies

Feinstein defined comorbidity as "*any distinct clinical entity that has co-existed or that may occur during the clinical course of a patient who has the index disease under study*" [1].

Since then, the concept has evolved in different directions, becoming a matter of concern in clinical care. Nowadays, we can find several definitions of comorbidity and there is no consensus since there is not a uniform methodology for the study of comorbidity.

A literature review of the multiple existing definitions of the term comorbidity was carried out by taking into account 141 articles. This review was done using the MEDLINE database looking for those publications between 2016 and 2017 that contained the term "comorbidity" or "multimorbidity" in the title and the abstract and had the full-text article publicly available.

Many authors, assuming that the meaning of the concept was widely understood, used it without providing any definition [2], and when it was provided, the exact meaning or definition varied from one author to another. More than ten different definitions can be found related to disease comorbidity (Table 1).

Table 1 General disease comorbidity definitions found in the literature. The comorbidity definitions are ordered from the most recent one to the first one given by Feinstein in 1970. For each definition, the table contains the author and the publication date. These definitions are some of the most frequently used in the current comorbidity-related publications.

Definition	***Author and date***
Coexistence of two or more long-term conditions in one patient / different systems of the body	Lawson et al., 2013 [3]
Coexistence of multiple illnesses of different types	Starfield & Kinder 2011 [4]
Co-occurrence of multiple medical conditions within one person without any reference to an index condition	Bayliss et al., 2008 [5]
Case where an individual suffers from two or more disease conditions at the same time / within a given period	Marengoni et al., 2008 [6]
Coexistence of three or more clinical conditions	Cesari et al., 2006 [7]
Coincidence of two or more diseases in a patient	Milkus & Saag, 2001 [8]
Co-occurrence of multiple chronic or acute diseases and medical conditions within one person	van de Akker et al., 2001 [9]
Co-existence of two or more chronic conditions, where one is not necessarily more central than the others	van de Akker et al., 1996 [10]
Co-occurrence of multiple (often chronic) diseases or medical conditions within one person	McGee et al., 1996 [11]
Coexistence of two or more chronic diseases in the same individual	Schellevis et al., 1993 [12]
Co-occurrence of several chronic conditions simultaneously	Verbrugge et al., 1989 [13]
Any distinct clinical entity that has co-existed or that may occur during the clinical course of a patient who has the index disease under study	Feinstein, 1970 [1]

The definitions mainly differ in: (i) considering or not an index disease, (ii) the type of disease accounted for (e.g., chronic

diseases), (iii) considering diseases of different kinds or affecting various organs and (iv) the time window between conditions.

The most frequently used definitions are "*coexistence of two or more chronic diseases in the same individual*" and "*coincidence of two or more conditions in a patient.*" The Feinstein and Bayliss et al. definitions differ in considering or not an index condition. On the other hand, Milkus et al. and van den Akker et al. have a different approach in considering or not only chronic disorders. Nevertheless, a common theme in all these definitions is the co-occurrence of multiple diseases in one patient [10], [14].

1.1.2 Alternatives to the concept of comorbidity

Apart from comorbidity, other terms such as 'multimorbidity,' 'morbidity burden' or 'patient complexity' (Figure 1) are used in the same context [15].

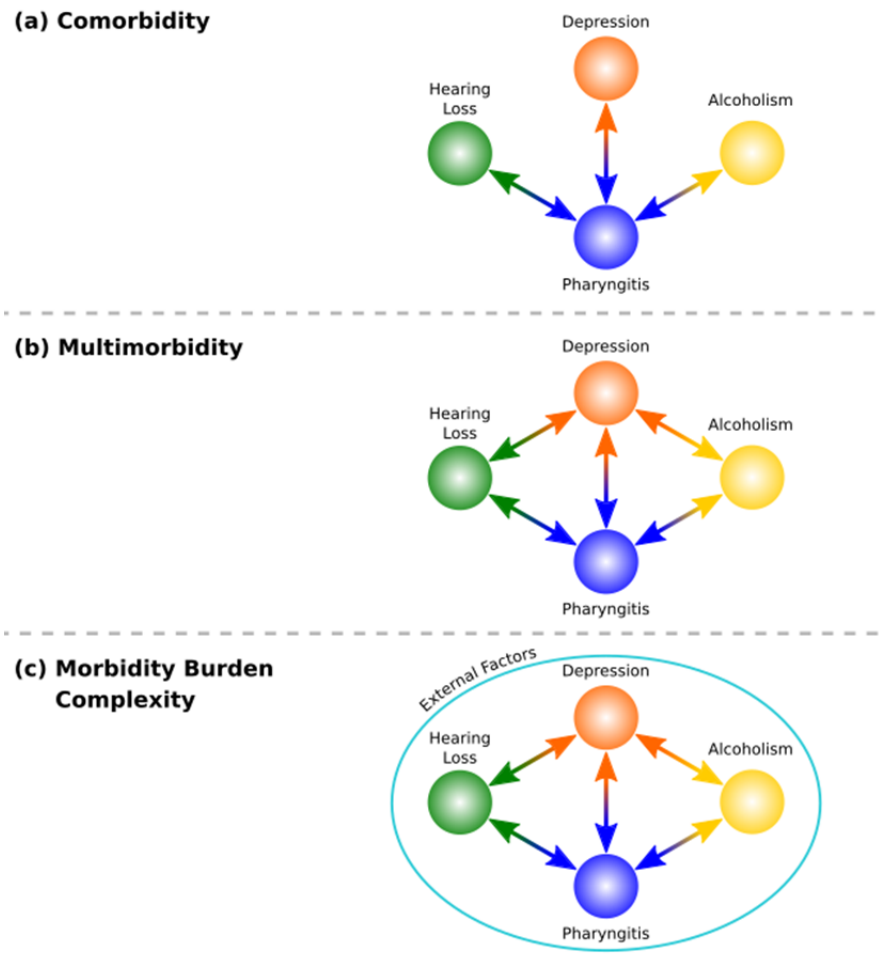

Figure 1 Comorbidity, multimorbidity, morbidity burden and complexity in a patient diagnosed with depression. a) For comorbidity, depression is the index disease and all other diseases are considered concomitant, b) In the multimorbidity concept, the patient is of staple concern and all diseases are of equal importance with interactions between each other, c) In morbidity burden and complexity, in addition to the impact of the distinct conditions, factors like age, gender, social factors, educational level or race, among others, are considered.

The second most frequent term in the literature referring to the co-occurrence of two or more diseases is ‘multimorbidity’ (Figure 2). Likewise, there is a lack of consensus about its definition and different attempts have also tried to unify them unsuccessfully. In

2014, the International Research Community of Multimorbidity [16], asked the scientific community the next question, "*which definition do you think should be used for multimorbidity?*", being *"multiple co-occurring chronic or long-term diseases or conditions, none considered as index disease,*" the preferred definition (69%) between research participants all over the world.

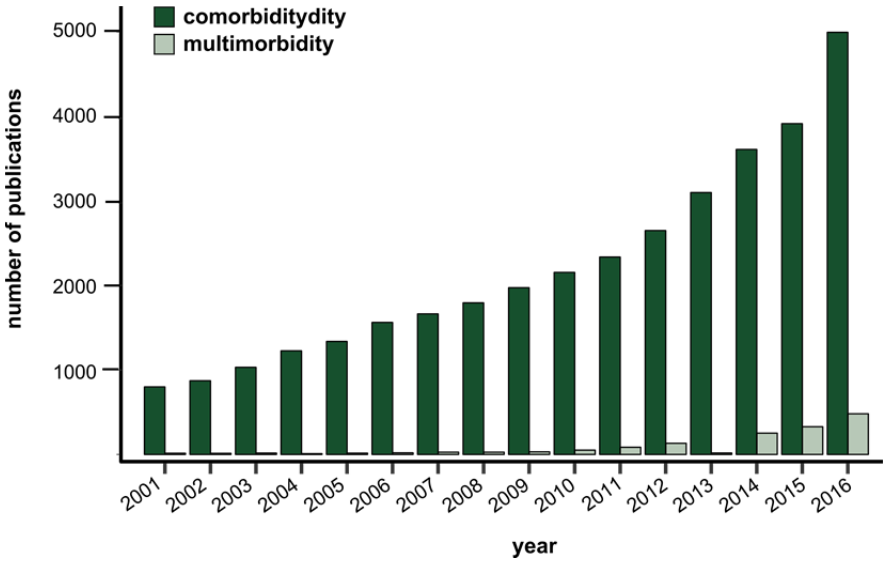

Figure 2 Number of Medline publications containing the term "comorbidity" or "multimorbidity" in the title or the abstract, from 2001 to December 2016.

"Comorbidity" is more often used when referring to the presence of multiple diseases about an index disease, while "multimorbidity" is usually employed in the context of multiple coexistent diseases without designation of an index one. The concept and definition to be chosen for a specific study depend on the particular objective of the study.

1.2 The impact of comorbidity on public health

The study of comorbidity in patients, usually, requires the analysis of a large amount of medical data in order to have sufficient statistical power. Fortunately, the era of big data has made possible this kind of studies. Moreover, comorbidity constitutes an essential aspect in the emerging field of personalized medicine [17], [18].

1.2.1 Why is important to study comorbidity?

The study of comorbidity has potential implications for fundamental and clinical research, as well as, in clinical practice [1], [19]–[21]. Determining the prevalence and causes of disease comorbidity contributes to:

1. **A better understanding of the etiology of disease:** genetic, biological, lifestyle and environmental factors might explain the co-occurrence of illnesses. For example, research in the comorbidity of diabetes with obesity [22], has led to the identification of a genetic variant associated with both, obesity and diabetes in the liver and adipose tissues [23], helping to substantiate a new hypothesis about the etiology of both disorders.
2. **Definition of new disease subtypes:** the comorbidity analysis is a potential approach to establish new disease subtypes due to the wide range of comorbidity patterns among patients [24]. For instance, in the case of obsessive-compulsive disorder (OCD), four subgroups of patients were identified based on the co-occurrence of OCD with different

diseases [25], enabling a more personalized management of patients.

3. **Determining more effective and safer treatments:** disease comorbidities are a fundamental aspect to consider for efficiency and effectiveness of pharmaceutical therapies [21]. For example, certain comorbidity studies have shown a higher co-occurrence of migraine, asthma and hypertension [14]. Although the best treatment for patients with migraine and hypertension are beta blockers, these can cause bronchoconstriction in patients with asthma [14]. Therefore, characterizing patients according to their comorbidity patterns [24] would contribute to the improvement of drug prescription.
4. **Enabling preventive medicine:** comorbidity studies may provide important opportunities for prevention [26]. Understanding the causes, nature, and mechanism of disease co-occurrence would lead to better prevention strategies, including early recognition of concomitant conditions. Distinct factors can lead to disease comorbidity: (i) those cases in which suffering one disease predisposes to the development of another one - like obesity and hypertension [27], where the 70% of the risk for hypertension can be related to obesity- and (ii) those comorbidities that arise as a consequence of sharing risk factors, such as suicide and substance use disorders (SUD) [26]. In each case, the prevention strategy applied should be different; in the former case, focusing on the causal disease – obesity - and

in the second case, addressing the risk factors – depressing mood, emotional problems [28] –in an integrated fashion.

In summary, assessing the disease comorbidity patterns of patients will allow determining the most appropriate treatment for each particular case, considering the patient as a whole entity and not treating each disorder separately.

1.2.2 Comorbidity causes

The presence of one or more diseases in a patient will likely affect health-related quality of life, mortality, health-care cost and possibly, treatment effectiveness [15]. Even if comorbidity can occur by chance without any causal correlation between the co-occurring diseases, more often than not, disorders occur together because they share underlying factors.

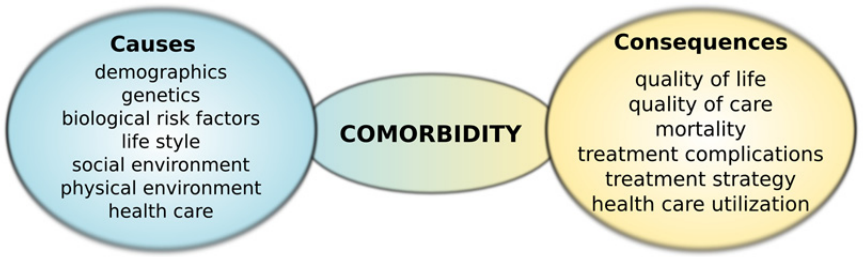

Figure 3 Comorbidity causes and consequences. Adapted from Gijsen et al. [29]

Several hypotheses try to explain the origin of comorbidities, such as common external factors that increase the risk of both diseases, or common genetic background [30], [31]. According to the

comorbidity the causes can be different. In particular, these can be differentiated between:

- **External factors:** it has been suggested that common environmental, lifestyle and socioeconomic factors increase the risk of developing certain disease comorbidities [32]–[34]. For example, diet, smoking habits or alcohol intake are highly related to the comorbidity of diabetes and obesity [35].
- **Treatment:** drug side effects constitute an undesirable outcome of medical care. In some cases, the treatment of one disorder can be responsible for the appearance of a second disease [29]. Sernyak and colleagues reported that diabetes could emerge as a side effect of atypical neuroleptics medications - clozapine, olanzapine, and quetiapine - used for the treatment of schizophrenia [36].
- **Genetic and biological factors:** for many disorders, germline as well as *de novo* mutations can contribute to disease. Alteration in one gene can be associated with one illness, such as schizophrenia, and at the same time be responsible for other disorders. Recent studies support that autism and schizophrenia co-occur more frequently than it would be expected by chance alone [37]. Although the exact reason remains unclear, in their study, Carroll & Owen proposed a biological explanation for both conditions. According to this study, the NRXN1 gene deletions, encoding the pre-synaptic protein neurexin 1, is associated with both, autism and schizophrenia [38].

1.2.3 Comorbidity prevalence

Recent studies indicate that comorbidity is a common issue especially in the elderly population [21], [39]. Nevertheless, although the prevalence of comorbidity is higher among older people, it is not only limited to this population group [40]. The prevalence differs depending on the population under study (the general population, hospitalized patients, geographic region, etc.) and the definition of comorbidity [21].

According to several studies, comorbidity is reported to range between 35% and 80% in the generally ill population [14], [39], [41], and in elderly, this fluctuates from 49% to 99% [42]. In particular, in Spain, the prevalence of comorbidity in patients older than 20 years old is 30%, and this increases by 60% for people older than 65 years old [43], [44]. Similar figures were reported for other European countries, like Germany [45].

1.3 Mental health: a worldwide problem

Although associated with a high impact on morbidity and mortality [46], [47], it was not until the beginnings of the 90s when mental health and substance use disorders started to be a staple concern in global health [46]. According to Whiteford et al. [48] in 2010, psychiatric and behavioral disorders – that comprise conditions such as depression, dysthymia, anxiety, schizophrenia, or drug and alcohol use disorders – were responsible for 7.4% of the global

burden of diseases, being depression one of the most disabling disorders.

Furthermore, as a result of the demographics changes (e.g., the increase of the life span, the population growth and the lack of access to health care for certain population groups) it is predicted that in the following years the burden of mental health and substance-use disorders will increase [49]–[51].

1.3.1 Comorbidity in mental diseases

Almost 75% percent of the patients suffering a mental health disorder, present an additional disorder of the same type, and almost half of them present two or more [52], [53] (e.g., according to the US National Comorbidity Survey 51% of patients diagnosed with major depression also presented anxiety disorder as comorbidity [54]). In particular, those patients with a mental health diagnosis have an odd of 2.7 of suffering a substance abuse disorder compared to those without a mental health diagnosis.

However, comorbidity in mental health is not restricted to diseases of the same class. Mental health disorders have been associated with a broad range of diseases. Several studies estimate that more than half of the adults suffering from mental disorders have at least one additional medical condition, and around 30% of patients with a medical condition present a psychiatric disorder [55]–[57]. Several studies show that people suffering mental health conditions, such as

depression, are more likely to have several physical comorbidities (e.g., cardiovascular disease, hypertension, diabetes [58], cancer [59], osteoporosis [60], arthritis [61] and asthma [62], [63]) than their non-mental disordered counterparts [64].

Although substantial evidence supports a strong association between several mental and substance-use disorders, as well as, between mental illness and physical-illness conditions, the nature of these relationships is not well characterized and remains to be investigated.

1.3.2 A molecular perspective on mental health disorders

Diagnosis of mental illnesses, in contrast to most of the diseases, is performed by assessing a cluster of subjective symptoms and can be made from observations of behavior by the clinician, the patient or their relatives [65]–[67]. The diagnosis is assigned to those individuals who show at least a determined number of symptoms and behaviors, during a given period of time [67]. These symptoms and behaviors tend to overlap among disorders. For instance, the occurrence of psychotic symptoms such as hallucinations, mood changes, alterations in speech, behavior and sleep can indicate a diagnosis of either schizophrenia or bipolar disorder, as outlined by O`Donovan [67].

Nowadays there are no genetic or biomarker tests with the sufficient predictive power for psychiatric disease diagnosis [68], a fact that hinders the diagnosis of mental health diseases [67]. However, the

high heritability of neuropsychiatric disorders (46%) [69]–[71] makes the study of the genetic architecture of these diseases a promising path towards a better understanding of their etiology and the identification of disease biomarkers. By studying the function of particular disease genes and how alterations in these genes are related to different symptoms and disease manifestations, better stratification of patients can be achieved, together with a better understanding of the disease co-occurrence and comorbidity.

The genetics of mental health disorders has received significant attention during the past few years and associations between hundreds of variants and neuropsychiatric disorders have been reported [72]. Due to the large-scale technologies available nowadays, it is possible to collect genomic information from large cohorts and connect the disease risk to the gene function. While microarray studies allow identifying gene expression or structural anomalies among other alterations – copy number variations (CNVs), as well as, genomic rearrangements – genome wide association studies (GWAS) have been used to detect loci associated with mental disorders status [72]. The majority of the disease-associated genetic variation lies in noncoding regions, being enriched in regulatory elements, such as promoters and enhancers that control gene expression and splicing [72].

Research in the genetic architecture of psychiatric diseases shows that in most of the cases, disease susceptibility is not only the consequence of an alteration in a single gene but rather the result of

complex interaction between multiple genes, as well as, environmental influences [68], [72], [73]. Furthermore, the same genetic alteration can be responsible for several disorders. O'Donovan and Owen highlight the cross-disorder effects observed from the genetic studies [67]. The International Schizophrenia Consortium (ISC) showed hundreds of alleles that increase the risk of distinct mental disorders such as schizophrenia, bipolar disorder and major depression, among others. These findings support the hypothesis that comorbidities in mental disorders could be explained by a shared genetic risk.

The National Institute of Mental Health (NIMH) has made significant efforts for the development of genomic resources and the support of gene discovery in the mental health area, leading to the creation of repositories for psychiatric genetics data [74] including gene expression and functional genomic elements across a range of psychiatric disorders. Some of the resources developed include: (i) the Whole-Genome Sequencing Consortium for Psychiatric Disorders (WGSPD), (ii) the Autism Sequencing Consortium (ASC; https://genome.emory.edu/ASC/), (iii) the Bipolar Sequencing Consortium (BSC), (iv) the PsychEncode (http://psychencode.org/) [75] and (v) the Psychiatric GWAS Consortium [76] (PGC; https://www.med.unc.edu/pgc). Nevertheless, this data is not always open access, thus, limiting its exploitation and re-use [74].

In the area of mental disorders, to the best of our knowledge, there exist five databases (listed in Table 2). All of them are associated

with specific diseases. For example, SGZR, SZGene and SZDB [77] are focused exclusively to schizophrenia. Furthermore, only one of them is up to date, namely, SZDB [77], which contains information about genes associated with schizophrenia (including genetic data, gene expression data, network-based, brain eQTL data; ENCODE data and SNP function annotation information). On the other side, the SLEP and BDGene [78] databases, that leverage information about several disorders, are obsolete and not up to date. Thus, there is a need for databases that collect, harmonize and offer the scientific community accurate data about psychiatric diseases and their genes.

A rich source of information that has been growing in the recent years is the scientific literature. Moreover, in most cases, this data is publicly available. One of the most frequently-used sources of scientific publications in life science is MEDLINE, comprising more than 27 million of publications. However, its constant growth is at the same time an obstacle for the analysis of such a high volume of data. For instance, more than 3,000 articles are published in biomedical journals per day [79]. As a result, a significant part of the research results generated nowadays is locked in the literature and cannot be transferred to knowledge repositories in a timely fashion.

Automated text-processing systems, also known as text mining tools, can help to unlock the information hidden in the publications by identifying, extracting and structuring the biomedical literature

in a more efficient and cost-effective way. However, text mining for biomedical text has to deal with: (i) specialized terminology and complex name conventions, (ii) large amount of synonyms – a name can refer to different meanings – (e.g., "Neutrophil gelatinase-associated lipocalin", "NGAL" and "LCN2" are synonyms to the Lipocalin-2 gene;), (iii) high use of abbreviations and acronyms and (iv) ambiguity of terms (e.g., "AD" can be the acronym of "Alzheimer disease" as well as the synonym of "APP" gene). All these together make the task of detecting biomedical entities a challenge.

Text mining cannot ensure the quality of the extracted data and experts in the domain are required to curate the information extracted from the literature before transferring it into databases or downstream analysis. Nevertheless, the manual curation process is hard to scale up to cope with the high number of publications available nowadays. As a result, there is a gap between the data present in manually curated sources and the information available in the scientific literature [80], [81]. The manual curation process is very work-intensive and time-consuming. It requires an enormous human effort, involving (i) the recruitment of people, (ii) the time needed to train the experts and to perform the actual process of curation or (iii) the need for curation guidelines, as well as, curation tools. In this regard, text mining methods represent a valuable tool not only for extracting the information hidden in the literature, but also for supporting the work to be performed by curation teams, making easier and less time consuming the task [82].

Exploring the vast amount of biomedical publications in the field of mental disorders [83] and extracting relevant genetic information, is of crucial importance for resources that aim at providing the most recent findings in the area of psychiatric disorders.

Table 2 Databases contain genomic information about mental health disorders. For each database, the disorder on which it is focused is shown. Moreover, the article publication is provided in the reference column, as well as the last update of the database. Each one of the databases is focused on a particular disorder or set of disorders, being schizophrenia the most prevalent. The genomic information contained in each one of them goes from gene association to pathways or miRNA among others.

Database	*Disorders*	*Database information*	*Reference*	*Last update*
SLEP	ADHD, autism, bipolar disorder, eating disorders, major depression, nicotine dependence, and schizophrenia	Genome Browser	[84]	24 April 2013
BDgene	Bipolar disorders	SNP, CNV, gene, pathway of Bipolar Disorder (BD); Overlapping genetic factors between BD and SZ/MDD	[78]	31 March 2016
SGZR	Schizophrenia	Genes associated with schizophrenia obtained from different studies analysis: Association, linkage, expression, literature, GO_Annotation, gene Network, KEGG, pathway, miRNA Target	[77], [85]	25 May 2009
SZGene	Schizophrenia	Genetic association studies	[86]	23 December 2011
SZDB	Schizophrenia	Genetic data, gene expression data, network-based, brain eQTL data, ENCODE data, SNP function annotation information	[87]	16 March 2017

1.4 Data for comorbidity studies: the big data era

The analysis of comorbidities is a major challenge in healthcare systems and fundamental for the study, treatment and prevention of diseases. The availability of big data provides unique opportunities to conduct experiments that were impossible to perform only ten years ago, bringing out new challenges [88], [89]. Currently, it is possible to identify disease comorbidity from the analysis of clinical records, which can be defined as "clinical comorbidity analysis", and shed light on the mechanisms underlying comorbidities by exploring omics data available in public repositories, in a "molecular comorbidity analysis".

1.4.1 Clinical data

The computerization of clinical data has been recognized as an unprecedented opportunity for research as a byproduct of healthcare [90]. It enables easier access to data, thus, providing a window into the patient's life and serves as the backdrop for evaluating, for example, the effectiveness of clinical interventions.

Clinical data can be collected throughout the hospitalization of a patient or as part of a clinical trial, over time and across care settings [91]. According to the source of the clinical data, six staple types can be distinguished: (i) electronic health records (EHRs), (ii) administrative data, (iii) claims data, (iv) patient/disease registries, (v) health surveys and (vi) clinical trials data [92], [93]. An example of the structure of clinical data is presented in Figure 4.

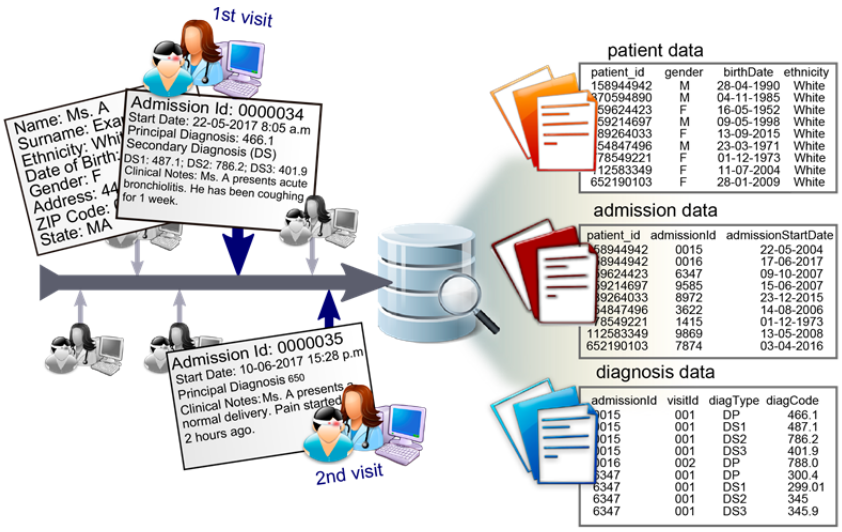

Figure 4 Example of an entry of a patient in a clinical database. In addition to basic demographic and personal data that is recorded in the system, each time the patient visits a health care provider, a new entry is added. Admission and diagnosis data is saved for each visit. Other information, such as laboratory tests, prescribed medication or clinical notes in free text are usually added to the records.

Recent studies have demonstrated the usefulness of the analysis of clinical data for discovering or confirming outcome correlations, finding subcategories of disease, and identifying adverse drug effects [94]–[96]. Advantages of the use of clinical data include (i) the accessibility to large amounts of data over time, (ii) cost-effectiveness, (iii) time reduction compared to traditional research for obtaining the data and (iv) the availability of longitudinal clinical information to carry out genetic studies [90].

However, this kind of data is primarily designed for the routine clinical care and not for research purpose and thus, it has certain limitations (Table 3).

Table 3 Challenges of using clinical data in research. Adapted from Cowie et al. [97]

Problem	***Example***
Data quality and validation	Coding errors Inaccurate information Selecting measurements of interest
Complete data capture	Clinical endpoints Death
Heterogeneity among systems	Lack of flexible architecture Lack of common data fields, definitions Difficulty with data mapping Incomplete data Missing fields of interest, relevant for some diseases but not others Inability to link systems

When using clinical data, several aspects should be considered. It contains sensitive private information; therefore, it is subject to confidentiality. Moreover, it provides a wealth of clinical information in different formats and languages and coding systems, (i.e., as structured and/or free text) [90]. Furthermore, there is a bias due to the fact that the data is generated for administrative purposes and differences arise in the manner that concepts are defined. With respect to the quality of data, it can be influenced by the low coverage, as well as, by missing data and the unavoidable incompleteness of records (Table 3) [97].

As outlined in [98], [99], the quality of electronic health record data is highly variable, regarding accuracy and completeness. Related to sensitivity [100], [101], there was a huge variability for the same clinical concepts between multiple institutions. The completeness of data, such as blood pressure recordings, ranged between 0.1% to 51% [101].

In summary, the use of this data still appears more as a burden than as a support tool. Thus, several questions arise: *are health systems able to process this information? Are we able to interpret this big data?*

1.4.2 Molecular data

Clinical data by itself is not sufficient in order to have a complete understanding of the mechanisms of disorders. Combining this large source of information with molecular data is crucial for promoting the biomedical research in the area of precision medicine. Recent studies have demonstrated that many, apparently, different diseases share common molecular mechanisms [102].

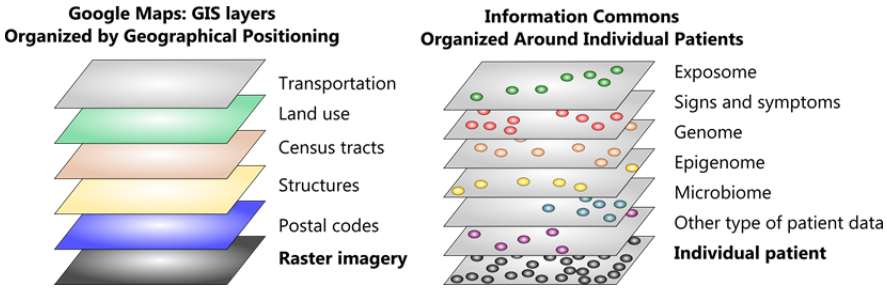

Figure 5 Like Google Maps, the information commons would consist of multiple layers of data that together provide insights that could not be gained from any of the layers alone. Figure adapted from National Research Council [103].

The National Research Council (US) (2011) presented a comprehensive overview of the data integration for the generation of a knowledge network of diseases, illustrated in (Figure 5). In the same way that google maps application combines multiple layers of data, an integrated perception of an individual patient can be

obtained using all available layers of information (e.g., exposome, signs and symptoms, genome, proteome, etc.) instead of separately processing each layer. It was only in the past few years that access to both clinical and genomic data of patients was obtained, such that there is a significant ongoing investigation on how to manage such massive and complex datasets.

In the field of comorbidities, on which this work is focused, recent studies have provided important insights into the etiology of comorbid diseases by exploring their shared genes [104]. A better understanding of the molecular mechanisms of disease comorbidities can be achieved through an integrated analysis of disease genes in the context of biological networks and pathways. Furthermore, investigating the regulation of gene expression by environmental factors by means of epigenomic marks and the role of the microbiome in diseases can shed light towards this direction.

Ideally, having access to clinical and genomic information of the same patient would enable to perform more accurate and reliable studies. Nevertheless, the number of databases that include both, clinical and molecular data of individual patients is limited and uncommon. One example of databases collecting all information about the patients is the Simon Simplex Collection (SSC) for the study of autism disorders [105]. This database has been developed in order to associate clinical, genomic, and neurobiological data [105] in autism patients. Another example is "Informatics for Integrating Biology and the Bedside" system (i2b2) [106]. The i2b2

has been installed in several hospitals and institutions with the purpose of connecting clinical to molecular and cellular data [107]. However, as mentioned earlier, such databases are rare.

Effective implementation of genomic data remains a challenge for the healthcare system [108]. The development of tools for storage, processing and integration of genome wide sequence data and clinical data systems together is required for implementing these systems [109].

Thus, clinical data systems are in widespread use but nowadays have very limited genomic capabilities. Consequently, public databases containing genomic information about disorders are used to overcome this limitation of the patient's data. Distinct databases can be found according to the layer of information under study, such as, (i) genome databases, (ii) gene expression databases, (iii) disease databases, as well as, (iv) protein's sequence, structure and modeling sources or (v) metabolic databases, among others. As claim by the *NAR* online Molecular Biology Database Collection, in 2016 the number of databases reached 1,685 [110]. How to integrate clinical and molecular data is a non-trivial task, mainly due to lack of homogeneity encountered in the vast number of databases. The source of information of each database remains different (e.g., patients, techniques), and each one of them uses specific vocabulary and standards. Consequently, it is challenging to integrate all the disease information in a unique database.

1.5 Existing tools for comorbidity analysis

To the best of our knowledge, only a few software tools have been developed for the analysis of disease comorbidities and are listed in Table 4.

In the R programming language, the comoR [111] package and medicalRisk [112] are the unique tools publicly available. Both of them assess disease comorbidity based on the ICD9-CM codification. The `comoR` package makes use of only two comorbidity metrics – relative risk and ϕ correlation (these metrics will be described in detail in Section 3.5)– using as input the US Medicare claims database [113], while, the `medicalRisk` applies other, yet very specialized measurements (such as, the Charlson Comorbidity Index and the Elixhauser comorbidity map, among others). In addition, medicalRisk accepts only input data in ICD-9 or ICD-10. Likewise, the Elixhauser comorbidity [114], developed in SAS (which is/is not publicly available) has the same limitation with respect to input data. Finally, the Network Regularized Cox [115] tool (developed in both, R and Matlab) is another comorbidity analysis software, which, however, is only focused on cancer (Table 4).

Regarding the assessment of comorbidities based on shared molecular components (e.g., genes, proteins) – which can be defined as "molecular comorbidities" –, the comoR package was the only one to consider disease co-occurrence based on genes – obtained from OMIM [116] – and pathways data – obtained from

KEGG [117] -. However, comoR package and its analogous Cytoscape plugin are not currently available.

In summary, the currently available tools for comorbidity analysis are focused on the analysis of clinical data, and none of them allows investigating the comorbidity from a molecular perspective. In this regard, there is a lack of tools to study disease comorbidities, which could be applicable to any kind of clinical data, and at the same time to allow investigating both the genetic and molecular mechanisms of comorbidity.

Table 4 Comorbidity tools. For each tool, the programming language and the comorbidity measures that can be estimated with it are summarized in this table. Finally, the availability is provided.

Software	***Programming language***	***Comorbidity measures***	***Availability***
medicalRisk	R package	-Charlson Comorbidity Index -Elixhauser Comorbidity classification -Revised Cardiac Risk Index -Risk Stratification Index	Available
Elixhauser Comorbidity Software, Version 3.7	SAS computer Programme	-29 Elixhauser comorbidity Metrics	Available
Network regularized Cox	Matlab and R	Cox proportional hazard model	Available
CytoCom	Java (Cytoscape plugin)	-Relative Risk -ϕ-correlation	Not available
comoR	R package	-Relative Risk -ϕ-correlation -Associated genes, pathway	Not available

2 Objectives

“If there is a good will, there is a great way.”

William Shakespeare (1564 - 1616)

The study of comorbidities is currently a major priority due to the impact of comorbidity on life expectancy, quality of life and healthcare cost. The availability of clinical data in the last years, gathered during routine medical care, has opened the opportunity to discover disease correlations and comorbidity patterns. This has raised the need for the development of analytical tools for the identification of disease comorbidities and the study of their underlying genetic basis.

Psychiatric disorders are characterized by a high prevalence of disease comorbidities. On the other hand, they also present a high heritability, and research in the last years has generated a large amount of data on the potential genetic contribution to the disease risk. However, this information is scattered around different repositories, thus making it difficult for the researcher to access and analyze the data due to the lack of publicly available resources and analytical tools.

In this context, the general objective of this thesis is to develop new resources and software tools for the assessment of disease comorbidities, including the investigation of their underlying genetic basis, with a special focus on psychiatric diseases.

The specific objectives are the following:

1. To develop a methodology for the exploitation of clinical data from patient record databases in order to identify comorbid

diseases, detect comorbidity patterns, and investigate their genetic causes.

2. To implement this methodology through a publicly available tool, `comoRbidity`, that encourages the re-use of clinical data and transforms this information into knowledge that can be analyzed, interpreted and applied by doctors in order to improve the health outcome of patients.

3. To apply the `comoRbidity` tool to different databases and population registries, with the purpose of analyzing comorbidity patterns in disorders with a high impact in the population, such as psychiatric disorders.

4. To develop a gene-disease database, PsyGeNET, by collecting information on mental disorder genes, and make it available to the scientific community.

5. To develop a methodology that combines text mining with curation by experts, for the development of the database (PsyGeNET) and its regular update.

6. To develop tools to extract, visualize and analyze molecular information of mental illness. In particular, to examine comorbidities from the genomic point of view, by exploring the molecular mechanism that substantiates mental disorder comorbidities.

3 Designing a comorbidity study

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

Ronald Fisher (1890 - 1962)

The analysis of comorbidity is a major challenge in healthcare systems and has implications for the study, treatment and prevention of diseases. Before starting a study, several decisions must be taken (Figure 6).

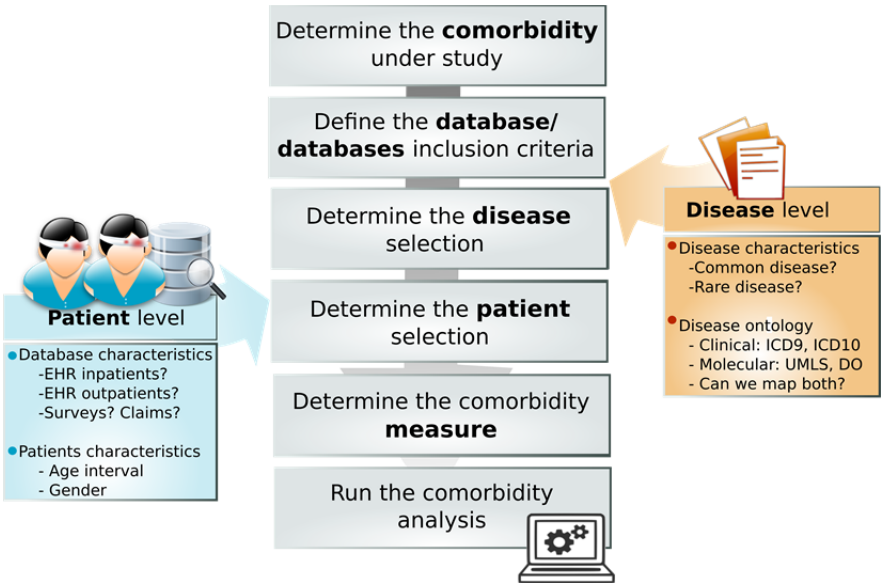

Figure 6 Comorbidity analysis conceptual decisions: (i) determine the comorbidity under study and the comorbidity definition, (ii) define the inclusion criteria of the database, considering the origin of the data; (iii) establish the disease selection; (iv) define the patient selection; (v) according to the disorders that will be analyze and population characteristics, determine the comorbidity measurements and finally (iv) look for the best comorbidity software for the study.

3.1 Comorbidity definition

The first step is to select the comorbidities of interest (e.g., autism and epilepsy comorbidity or cancer and Alzheimer comorbidity). Although the steps to follow in both cases remain the same, there will be slight differences between both studies and as introduced in Section 1.1.1, the comorbidity definition could change according to

this selection [118]. Moreover, the time range between both disorders to be considered as comorbidity is also intrinsically related to the diseases analyzed, [15] (e.g., if the diseases are chronic, temporality could be not considered).

3.2 Database inclusion criteria

Once the comorbidity has been determined, the next step is to consider the database inclusion criteria.

- From the **clinical comorbidity analysis perspective**, it should be examined what the inclusion criterion of the patients in the database is. It should be determined if the database is focused on a population subgroup (e.g., pediatric patients, cancer patients), as well as, if the data comes from specialized hospitals or departments (e.g., mental hospital, emergency department, primary care) is key to the study. For example, in autism and epilepsy comorbidity analysis, a pediatric database is required, as autism is diagnosed in the childhood. On the other hand, when analyzing the prostate cancer comorbidities, medium age and old men patients' data is required.
- From the **molecular comorbidity analysis perspective**, different questions raised around the inclusion criteria of the database. *Which diseases are included in the database? Is it a particular database focused on a specific disease, is it a general database? Which disease standards are used? Which kind of data is included in the database?*

The researcher should be conscious as to the kind of data he/she is analyzing. The comorbidity results can vary significantly from one to the other, and before performing any analysis, the bias due to the origin of the data has to be considered, as well as, the advantages and limitations of each type of data. In summary, in both cases, the first question to be asked is the following: "*Can we perform the comorbidity analysis we are interested in using the database that we have access to?"*

3.3 Disease selection features

Afterwards, the disease selection process has to be performed. From the clinical perspective, it is important to know the diagnostic criteria that have been followed since they will be strongly related to the number of patients identified. From the molecular perspective, it is fundamental to know how the diseases have been defined in the database. This will be intrinsically related, for instance, to the number of genes that are found to be associated with the disease. Furthermore, if the design of the comorbidity study includes both, clinical and molecular analysis, it should be taken into account that diseases can be provided in different controlled vocabulary or standards, and even some databases lack standards, which hinder the mapping process.

The problem arises when trying to look for the corresponding disorders in distinct standards. *Which disease standards are used in each case? What is the most accurate methodology to look for the*

equivalent code in another standard? Can we do it automatically or do we need an expert? From one hand, in the clinical data, the *International Classification of Diseases (ICD)* is usually used [119]. It is a medical classification standard or vocabulary to determine diseases, injuries, signs and symptoms and other health-related conditions. It is especially designed for the clinical purpose and it contains detailed disease information (e.g., ICD9-CM comprises more than 14,000 different codes). On the other hand, the way in which disorders are defined in the molecular database is entirely different. One of the standards used is the *UMLS (Unified Medical Language System)* [120]. UMLS is a broad terminological source that integrates distinct biomedical vocabularies and ontologies in a single resource. An example of the standard differences is shown in Table 5.

Table 5 Autism spectrum disorder according to ICD9-CM and UMLS standards. The disease code and the disease description for each case are provided in the table.

Autism Spectrum Disorder in ICD9-CM codification		***Autism Spectrum Disorder in UMLS codification***	
Code	**Disease description**	**Code**	**Disease Description**
299	Pervasive developmental disorders	C1510586	Autism Spectrum Disorders
299.0	Autistic disorder	C1854416	Macrocephaly/Autism Syndrome
299.00	Autistic disorder, current or active state	C3275438	Autism, Susceptibility To, X-Linked 5
299.01	Autistic disorder, residual state	C3550875	Autism, Susceptibility To, X-Linked 6
299.8	Other specified pervasive developmental disorders	C1845539	Autism, X-Linked, Susceptibility To, 2
299.9	Unspecified pervasive developmental disorder	C1845540	Autism, X-Linked, Susceptibility To, 1

When conducting a clinical and molecular analysis, an accurate methodology to find the equivalent code in both standards is needed. Sometimes, expert curation is required to perform the mapping between different terminologies.

3.4 Patients selection features

Equally important is to determine which patients will be included in the analysis. As explained before, patient selection will depend on the comorbidity under study. Several questions come up when thinking about patient's selection criteria: *What is the best range of patients' age for the study? Should we include males and females? Could be of interest to perform the study separately, and then compare the results or are the differences in age and gender not relevant?*

3.5 Comorbidity metrics

Measuring comorbidity is an aspect of research that is receiving increasing attention in the literature [20]. To evaluate the correlation starting from disease co-occurrence, we need to estimate the strength of the comorbidity risk. There are more than thirteen different methods to measure disease co-occurrence, although some of them are unique to a set of conditions [121]. Note that the validity of each method relies on the population group in which it is measured [122].

Between those indexes developed for particular disorders or conditions the Charlson and Elixhauser indices are the most widely used in clinical research [20]. Both of them were designed for main health issues and predict mortality using International Classification of Disease (ICD) diagnosis codes [114], [123]. Other indexes are focused on chronic disorders as well as treatments, such as the Chronic Disease Score (CDS) [124].

Table 6 Summary of comorbidity measurements. Adapted from Sarfati et al. [125] and Huntley et al. [122].

Measure Name	*Author (year)*	*Purpose*
Charlson	Charlson et al., 1987	To predict 1-year mortality among patients admitted to hospital.
Elixhauser	Elixhauser et al., 1998	To measure comorbidity using administrative data.
CDS/RxRisk	Von Korff et al., 1992; Clark et al., 1995	To develop a stable measure of chronic disease status using routine pharmacy data.
Cumulative Index Illness Rating Scale (CIRS)	Linn (1968)	To assess the medical burden of chronic illness.
Adjusted Clinical Groups (ACG) System	Weiner (1991)	Originally devised to predict morbidity burden and use of health care resources.
Kaplan Index	Kaplan and Feinstein, 1974	A measure of comorbidity among diabetic patients.
ICED	Greenfield et al., 1993	To measure the impact of comorbidity and physical functioning.
Satariano	Satariano and Ragland, 1994	To assess comorbidity in breast cancer patients.
NCI Comorbidity Index	Klabunde et al., 2000 and 2007	To measure comorbidity among cancer patients using administrative data.

All the previous comorbidity measurements (Table 6) are defined according to determined characteristics like disease severity, mortality and survival, but when the primary goal is identifying general comorbidity patterns, other statistical indexes can be applied. Specifically, some of these alternative estimators that are used nowadays in comorbidity studies are: the relative risk (RR) [126], the odds ratio, the comorbidity score [126], the ϕ-correlation [111] or the Fisher test [91], corrected by the Benjamini-Hochberg false discovery rate method [127].

- ***Relative Risk (RR):*** the relative risk expresses the relationship between the rate of incidence of a disease among the patients exposed and those patients that are not exposed to a certain risk factor. Here the risk factor is another disease. The RR is estimated as the fraction of the number of patients diagnosed with both diseases and random expectation based on disease prevalence. The RR of observing a pair of diseases A and B affect the same patient is given by Roque et al. [126].

$$RR_{ij} = \frac{C_{ij}N}{P_i P_j} \quad (1)$$

where *Cij* is the number of patients affected by both diseases, *N* is the total number of patients in the population and *Pi* and *Pj* are the prevalences of diseases *i* and *j*. The RR value moves from zero to infinity. A value of 1means that the risk is the same for patients exposed to the factor of risk than for those patients that are not exposed; when the RR is

greater than one the patients exposed to the factor of risk are more likely to suffer the disease. Finally, if RR is lower than 1 the patients exposed to the factor of risk are less likely to suffer the disease.

- ***Odds ratio:*** The odds ratio represents the increased chance that someone suffering disease A will have the comorbid disorder B. It shows the extent to which suffering a disorder increases the risk of developing another illness or disorder. The odds ratio is derived from a comparison of rates of the illness among individuals who do and do not exhibit the factor of interest. A statistically significant odds ratio (significantly different from 1.00 at the .05 level) indicates an appreciable risk associated with a particular factor. For example, an odds ratio of 2.00 means a doubled risk of the appearance of the disorder.
- ***Comorbidity score:*** this score is defined in Roque et al. [126] as follows:

  $$comorbidity\ score = \log_2\left(\frac{observed+1}{expected+1}\right)\ \ expected = \frac{n_A n_B}{N} \quad (2)$$

  where *observed* stands for the number of disease-disease associations (disease A and disease B), and *expected* is estimated based on the occurrence of each disease (number of patients diagnosed with disease A, n_A, multiplied by the number of patients diagnosed with the comorbid disorder B, n_B , and divided by the total number of patients, *N*). Since the logarithm is applied, a comorbidity score of 1.0 means that the *observed* comorbidities are two-fold higher than (approximately) than *expected*.

- ***ϕ-correlation:*** the Pearson's correlation for binary variables measures the robustness of the comorbidity association [111]. The ϕ-correlation, which is Pearson's correlation for binary variables, can be expressed mathematically as:

$$\phi_{AB} = \frac{C_{AB}N - P_A P_B}{\sqrt{P_A P_B (N-P_A)(N-P_B)}} \quad (3)$$

Where N is the total number of patients in the population, P_A and P_B are incidences/prevalence's of diseases A and B respectively. C_{AB} is the number of patients that have been diagnosed with both diseases A and B, and P_A P_B is the random expectation based on disease prevalence. The Pearson correlation coefficient can take a range of values from +1 to -1. A value of 0 indicates that there is no correlation between the two diseases; a value greater than 0 indicates a positive correlation between the two diseases and a value less than 0 indicates a negative correlation.

- ***Fisher test:*** A Fisher exact test for each pair of diseases is performed to assess the null hypothesis of independence between the two conditions [91]. The Fisher exact test is applied to estimate the p-value for each pair of diseases. Four groups of patients are defined in order to perform the statistical testing: patients suffering disease A and disease B, patients suffering disease A but not disease B, patients suffering disease B but not disease A and patients not suffering disease A nor disease B. Then the Benjamini-

Hochberg false discovery rate method [127] is used in the ranked list to correct for multiple testing.

Note that the comorbidity measures are not entirely independent of each other and they can underestimate or overestimate the relation between disorders. For instance, while relative risk overestimates rare diseases comorbidities, the ϕ underestimates the comorbidity among rare and common conditions [126].

3.6 Summary

All the points discussed in the previous sub-sections stress the importance of a proper description of the definitions, procedures, and standards used to facilitate the correct understanding and comparison of the data from different studies. Note that regarding the diseases under study, how the disorders are defined, the demographic characteristics of the patients, the results of the analysis can change, and a precise description of the criteria followed is fundamental to understand and interpret the results of the study.

4 Applications and results

"Nothing in life is to be feared; it is only to be understood. Now is the time to understand more, so that we may fear less."

Marie Curie (1867-1934)

4.1 comoRbidity: an R package to analyze disease comorbidities

Clinical databases contain a large amount of information about patient history. Using a limited number of data types, such as the age, gender, and the patient's diagnosis, it provides us an opportunity to improve patient outcomes through research and the development of clinical decision support tools in the personalized medicine. A clinical data analysis system is fundamental to understand and predict outcomes from past patient data. `comoRbidity` software process massive amounts of healthcare data to identify comorbidity (coexistence of diseases in one patient) patterns in patient level data according to age interval and gender population. Moreover, the sex ratio parameter, that allows seeing if the disease co-occurrence is equally likely for both genders or not, as well as the temporal direction analysis is assessed for the comorbidities identified in the analysis can be estimated. A particular focus is made on the results visualization, providing a variety of representation formats, such as networks, heatmaps or bar plots.

Gutiérrez-Sacristán A, Bravo À, Giannoula A, Mayer MA, Sanz F, Furlong LI. comoRbidity: an R package for the systematic analysis of disease comorbidities. Bioinformatics. 2018; 34(18): 3228-3230. DOI: 10.1093/bioinformatics/bty315

4.2 Molecular and clinical diseasome of comorbidities in exacerbated COPD patients

The frequent occurrence of comorbidities in patients with chronic obstructive pulmonary disease (COPD) suggests that they may share pathobiological processes and/or risk factors. To explore these possibilities we compared the clinical diseasome and the molecular diseasome of 5447 COPD patients hospitalized because of an exacerbation of the disease. The clinical diseasome is a network representation of the relationships between diseases, in which diseases are connected if they co-occur more than expected at random; in the molecular diseasome, conditions are linked if they share associated genes or interaction between proteins. The results showed that about half of the disease pairs identified in the clinical diseasome had a biological counterpart in the molecular diseasome, particularly those related to inflammation and vascular tone regulation. Interestingly, the clinical diseasome of these patients appears independent of age, cumulative smoking exposure or severity of airflow limitation. These results support the existence of shared molecular mechanisms among comorbidities in COPD.

Faner R, Gutiérrez-Sacristán A, Castro-Acosta A, Grosdidier S, Gan W, Sánchez-Mayor M, et al. Molecular and clinical diseasome of comorbidities in exacerbated COPD patients. Eur Respir J. 2015; 46(4):1001–10. DOI: 10.1183/13993003.00763-2015

4.3 Using Electronic Health Records to Assess Depression and Cancer Comorbidities

Comorbidities are an important concern in oncology since they can affect the choice and effectiveness of treatment and the diagnosis of depression in cancer patients has an important impact on the quality of life of these patients. Although there is no consensus about a particular relationship of depression with specific cancer types, some authors have proposed that depression constitutes a risk factor for cancer. The objective of this study is to identify the presence of comorbidities in a massive EHR system, between depression and the 10 most common cancers in women and men and to determine if there is a preferred temporal ordering in the co-occurrence of these diseases. All the cancers studied showed a significant co-occurrence with depression, in particular twice more frequent than what could be expected by chance. A preferred directionality was identified between some of the comorbid diseases, such as breast cancer followed by depression, and depression followed by either stomach cancer, colorectal cancer or lung cancer.

Mayer MA, Gutierrez-Sacristan A, Leis A, De La Peña S, Sanz F, Furlong LI. Using Electronic Health Records to Assess Depression and Cancer Comorbidities. Stud Health Technol Inform. 2017; 235:236–40. DOI: 10.3233/978-1-61499-753-5-236

4.4 PsyGeNET: a knowledge platform on psychiatric disorders and their genes

Comorbidity is the norm among common mental illness, as more than 50% of affected people meet criteria for multiple diseases. The coexistence of mood and substance use disorders (SUD) is attracting growing interest in the scientific community because of its high prevalence rates and its association with a greater severity of illness and rate of recurrence of both disorders. In particular, alcohol and cocaine dependencies are frequently associated with depression. Several mechanisms have been proposed to explain the coexistence of diseases in one patient, being the genetic origin one of them. With the objective of having a curated gene-disease resource focus on psychiatric disorders, PsyGeNET was developed. PsyGeNET is a new resource that integrates information on mental illness and their genes, offering exploratory tools for the analysis of gene-disease associations. Due to its special focus on psychiatric diseases, comprehensiveness, and high-quality database, PsyGeNET represents a valuable resource for the discussion of the molecular underpinning of mental disorders and their comorbidities

Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. Bioinformatics. 201; 31(18):3075–7. DOI: 10.1093/bioinformatics/btv301

4.5 Text mining and expert curation to develop a database of psychiatric diseases and their genes

Mental disorders constitute one of the leading causes of disability worldwide. The difficulty in accessing up-to-date, relevant genotype-phenotype information has hampered the application of this wealth of knowledge to translational research and clinical practice in order to improve diagnosis and treatment of psychiatric patients. PsyGeNET contains up-to-date information on genes associated with mood disorders (depression, bipolar disorder), psychosis (schizophrenia) and substance use disorders (alcohol, cannabis, and cocaine use disorders, substance-induced depressive disorder and psychoses). The PsyGeNET database has been developed by extracting gene-disease associations from the literature with the text mining tool BeFree, followed by process of curation by a team of 22 domain experts. A web-based annotation tool supported the curation process. Due to its special focus on psychiatric diseases and comprehensiveness, PsyGeNET represents a valuable resource for the analysis of the molecular underpinning of mental disorders and their comorbidities.

Gutiérrez-Sacristán A, Bravo À, Portero-Tresserra M, Valverde O, Armario A, Blanco-Gandía MC, et al. Text mining and expert curation to develop a database on psychiatric diseases and their genes. Database (Oxford). 2017 Jan 1;2017. DOI: 10.1093/ database/bax043

4.6 psygenet2r: a R/Bioconductor package for the analysis of psychiatric disease genes

PsyGeNET (Psychiatric disorders Gene association NETwork) is a database developed for the exploratory study of mental health disorders and their associated genes. The `psygenet2r` package (https://bioconductor.org/packages/release/bioc/html/psygenet2r.html) implements several functions for exploring and analyzing PsyGeNET data in a clear and meaningful way and allows performing comorbidity analysis based on shared genes. `psygenet2r` contains a variety of functions for leveraging PsyGeNET using the powerful visualization and statistical capabilities of the R environment. `psygenet2r` eases the exploration of gene-disease associations from different perspectives. It offers different types of visualization, such as heatmaps and networks. The `psygenet2r` package expedites the integration of PsyGeNET data with other R packages and allows the development of complex bioinformatic workflows.

Gutiérrez-Sacristán A, Hernández-Ferrer C, González JR, Furlong LI. psygenet2r: a R/Bioconductor package for the analysis of psychiatric disease genes. Bioinformatics. 2017 Dec 15;33(24):4004–6. DOI: 10.1093/bioinformatics/btx506

5 Discussion

"If we knew what it was we were doing,

*it would not be called research, would it**?"***

Albert Einstein (1879-1955)

5.1 Overview

The fundamental motivation behind this work resulted from the increasingly growing demand, in the recent years, for a more personalized medicine in order to improve the health outcome of patients. As presented in the Introduction, the population-based comorbidity studies can contribute towards this direction. Nowadays, the amount of data in biomedical research is increasing dramatically [128], [129]. Undoubtedly, there is an urgent need for new efficient and publicly available tools that facilitate the collection of this data and turn it into knowledge [130], [131]. However, despite the great effort made in helping researchers and medical doctors to have a better understanding of human health and disorders, how to harness and interpret the promising biomedical big data remains a great challenge.

In this regard, this Ph.D. thesis has addressed the detection of comorbidity patterns, both at the clinical and molecular levels. Towards this objective, novel software tools and resources for comorbidity analysis, which are already publicly available, have been developed and presented in this Ph.D. thesis. The main outcomes described throughout this manuscript, that is, the novel tools developed for the analysis of comorbidities (`comoRbidity` and `psygenet2r`) and the new manually curated database for psychiatric disorders (PysGeNET), will critically be discussed in the following sections.

5.2 Patient's data as the new big data, challenges and perspectives in the context of comorbidities

More data has been generated in the past decades than in the whole human history [128]. Nowadays, researchers and clinicians have access to this vast amount of information, such as electronic health records of patients, comprising, for example, diagnostics as well as laboratory test results or clinical notes [129]. Nevertheless, as the volume of data is growing, so does the necessity of tools to process and create value from it, addressing the present and future health care needs [128].

The idea of predicting what kind of diseases will affect a patient in the future is decades old. However, actual advances in the field of disease comorbidities, both at the clinical and genomic level, have been only recently made possible, due to the massive availability of patients health data. Nonetheless, when considering patient data as the new big data, several challenges must be resolved before making extensive its use for data discovery, including data protection, quality and devising ways to integrate over time and space the records for a particular person [130].

With these challenges in mind and based on the key features and limitations of the pre-existing comorbidity tools (Section 1.5), the `comoRbidity` software for exploiting clinical and molecular information for comorbidity analysis (Section 4.1) has been developed. ComoRbidity has been applied to different studies: (i) the analysis of comorbidities between cancer and depression based

on electronic health records from general hospitalization (section 4.3), (ii) the comorbidity analysis on chronic obstructive pulmonary disease (COPD) in hospitalized patients from clinical audits (section 4.2) and (iii) the comorbidity analysis in pregnant women.

Herein it is important to mention that the results of these studies actively support the usefulness and validity of the `comoRbidity` tool. Known comorbidities supported by the literature, as well as, new disease correlations were found. For instance, Mayer et al. found comorbidities between depression and the most common cancers, as expected, while novel results were found regarding comorbidity directionality [132]. When comorbidity patterns were analyzed in pregnant women, expected comorbidities such as anemia and infections of genitourinary tract or premature labor and tobacco use disorder were identified. Finally, in the COPD analysis, one of the three novel findings of Faner et al. was the suggestion about shared molecular mechanisms across comorbidities [133] when comparing the clinical and molecular comorbidities.

The `comoRbidity` package enables the user to select the dataset to use for the comorbidity study under consideration (e.g., his/her hospital data, a cohort study, public datasets) and execute it locally, without the need to move the data to an external server, thereby, guaranteeing data protection standards. Moreover, as there exist different comorbidiy definitions and plenty of comorbidity metrics, the `comoRbidity` package allows the user to define the comorbidity term and select the best comorbidity metrics according

to the nature of the analysis [20]. Additionally, the package performs a molecular comorbidity analysis based on a public database by gathering gene-disease information data. The `psygenet2r` package, also presented in this thesis, permits the analysis of comorbidities in psychiatric disorders based on shared genes. A comparison of the main features of some of the available comorbidity tools is presented in (Table 7).

However, the `comoRbidity` package has certain limitations that are due to the large quantity of data under analysis. In particular, (i) a high computational power is required to perform the analysis and (ii) not all possible confounding variables are considered. Regarding the speed of the analysis, although some strategies such as parallelization or Python scripts have been implemented in the package, high computational power is required when analyzing large datasets. On the other hand, related to the influence of confounder variables, the `comoRbidity` package allows analyzing the comorbidities according to the patients' age and gender. However other modifying variables such as demographic, socioeconomic or treatment features have not yet been implemented. In specific experimental designs, if these variables are not taken into account, they could lead to unrealistic or irrelevant outcomes [9].

In summary, the development of systems, such as the `comoRbidity` package that support clinical decision-making, represents a huge step towards a more personalized medicine.

Table 7 Comorbidity software currently available.

Comorbidity tool	*Programming language*	*Clinical comorbidity*	*Molecular comorbidity*	*Comorbidity measurements*
`comoRbidity`	R package	Based on user data Any clinical diagnosis	Shared genes Based on DisGeNET	relative risk ϕ-correlation comorbidity score fisher test corrected Jaccard Index
`psygenet2r`	R package	-	Shared genes Based on PsyGeNET	Jaccard Index
`Elixhauser Comorbidity`	SAS computer programs	Hospital discharge records ICD-9-CM and ICD10-CM	-	29 Elixhauser Comorbidity Measures
`medicalRisk`	R package	Determine comorbidity and medical risk status of a given patient ICD-9-CM codes	-	Charlson Index Elixhauser Comorbidity Revised Cardiac Risk Index Risk Stratification Index
Network regularised Cox	Matlab and R	Diseasome associations Cancer disease comorbidities Survival of cancer	Seven Microarray data Cancer gene expression OMIM and GAD dbs	Cox proportional hazard model

5.3 Biocuration as a crucial process for accurate validation of gene-disease associations

The molecular understanding of disease comorbidities plays a main role in precision medicine, as it allows a precise identification of disease drivers and therefore, a more targeted therapy. However, information on gene-disease associations is typically dispersed, and there is a lack of curated databases containing molecular information about diseases. Hence, a significant effort during this Ph.D. thesis has been dedicated to the development of a manually curated database on genes associated to mental disorders.

The development of a curated database by a group of experts requires an enormous effort. For this reason, this thesis was focused on a particular group of disorders that usually present a significant number of comorbidities, which is that of mental illnesses. In this regard, a methodology was presented (section 4.5) for the curation of gene-disease association data in mental disorders. As a result of this process, a new manually curated resource, PsyGeNET, was developed.

The strategy adopted for the development of the PsyGeNET database was to combine text-mining with expert curation. As introduced in section 1.3.2, the scientific literature contains a vast amount of information. Hence, in order to be able to extract information locked in publications and to facilitate the curation process, text mining approaches are becoming essential. In our case, the BeFree text-mining system [134], based on a supervised

learning approach for the gene-disease association identification, was applied.

The whole curation process represented a huge challenge, as it required (i) the recruitment of a team of 23 domain experts, (ii) the development of curation guidelines describing all the details involved in the annotation task and (iii) a user-friendly curation tool to support the experts. The whole process took around three months to be completed. A training process in order for the curators to follow the guidelines and apply the annotation tool was fundamental, such that feedback could be received and discrepancies could be identified. For instance, one of the main reasons of disagreement between the experts was the vast diversity of studies covered by the papers – GWAS, sequencing studies, animal models –, which requires the corresponding variety of expertise between the curators. An inter-annotator agreement value, in order to assess the consensus achieved between annotators, was established. It was required to be higher than 60% in all the different steps of the curation process.

The PsyGeNET database was developed considering only those gene-disease associations for which agreement between curators was found. Furthermore, PsyGeNET includes not only positive associations between genes and diseases. A significant percentage of gene-disease associations (~30%) are supported by at least one negative evidence, which is a publication that states that the gene is not associated to a disease. This stresses the importance of

collecting negative associations from the literature in databases. This information has been considered to rank the associations in PsyGeNET, and an evidence index has been developed to show the positive and negative findings for each association. In addition, collecting this data is relevant for the development of corpus for the training of text mining systems able to identify negative gene-disease associations from the literature. Moreover, annotated corpus – defined as a set of documents with labeled information (e.g., genes, diseases, drugs, relationship between entities) – is key for improving and evaluating text mining methodologies. The PsyGeNET corpus, consisting of the sentences curated by the experts, is available in the PsyGeNET web.

PsyGeNET is a publicly available database that currently covers 3,771 associations, between 1,549 genes and 117 diseases. Compared with other curated databases, such as CTD human, Human Phenotype Ontology (HPO), BDgene and SZDB, it represents a valuable source regarding gene coverage in each psychiatric category included in the second release of PysGeNET.

Although there is an overlap between the knowledge found in distinct databases (e.g., around five hundred gene-disease associations present in PsyGeNET are also found in other specific mental disorder data sources, like SZDB and BDgene database) (Figure 7), PsyGeNET contains information that is not available in other sources. A total of 1244 gene-disease associations are only present in PsyGeNET.

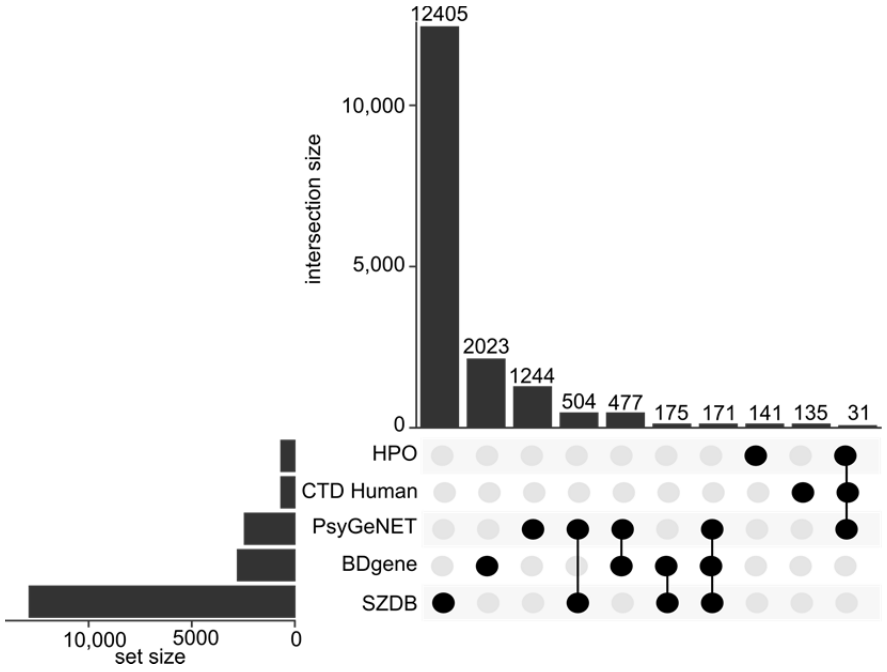

Figure 7 Gene-disease association overlap between different psychiatric curated resources. The horizontal barplot represents the set size of each database, while the vertical one shows the intersection size. The black filled circles joined by a black line represent the sources between which we find the overlap if it exists; otherwise, a gray circle is shown. This graphic has been done using the UpSetR [135]

The first release of the PsyGeNET database was launched in August 2014, and the current one, PsyGeNET v.02 in September 2016. Both versions of the database have been published in peer-reviewed journals and since it was created, it has attracted more than 7,900 users from all over the world and 1,500 users during the last year.

Moreover, the database was enriched with an intuitive web browser and analysis tools to establish the PsyGeNET platform as a comprehensive knowledge platform. Additionally, the `psygenet2r` package, which allows exploring and visualizing PysGeNET data, was released as a Bioconductor package. The

`psygenet2r` package has been used by more than 1,000 users and downloaded more than 2,000 times since it was launched in March 2016, according to Bioconductor [136]. More than ten publications make use of PsyGeNET for the analysis of psychiatric disorders with different applications. PsyGeNET is also used in the H2020 MedBioinformatics project and indexed in several bioinformatic tools registries (such as Omic tools https://omictools.com/).

In summary, PsyGeNET allows gaining insight into the genetics of psychiatric conditions, providing a comprehensive catalog of gene-disease associations, including data not provided by other similar databases. The suite of tools offered is aimed at fostering the study of the molecular and biological mechanisms behind psychiatric disorders and their comorbidities.

5.4 Future Perspectives

In this thesis, it was shown that the `comoRbidity` package can be used to analyze comorbidity patterns at the clinical and molecular level. Furthermore, the importance of a curated database, such as PsyGeNET, was demonstrated for studying the molecular basis of mental health disorders. However, there is still room for improvement in both areas.

Throughout this Ph.D. thesis, it has become evident that in order to gain insight into the causes of disease comorbidities, a promising approach appears to be connecting the clinical with molecular data [107]. Personalizing the treatment of patients, by taking into account individual risks and variations in the treatment response, has been a goal of modern medicine for a long time [137]. The idea of using genomics to further this vision has become widespread [108], [138], aided by the plummeting cost of DNA sequencing. However, as pointed out in the introduction (section 1.4.2), there are currently only a few databases that include both, clinical and molecular data of individual patients. Undoubtedly, this is an ideal dataset for conducting a comorbidity study.

In order to overcome this data limitation and to provide a complete analysis including clinical and molecular data, the use of molecular public databases is needed. The `comoRbidity` package includes both types of analyses, although they run independently. In this thesis, it was shown that the `comoRbidity` package could analyze the clinical comorbidities using the user data while the molecular

one is performed based on the DisGeNET database, which integrates gene-disease association data from distinct sources. However, as explained in section 1.4.2, the process of linking clinical data with molecular data gathered in public health databases, still remains a challenge, mainly due to the lack of homogeneity and the use of specific standards in each database, which difficult the integration task.

Future directions in this regard include the implementation of a curation process to map between disorders in a clinical and molecular level that cope with the mapping between standards. A promising technique towards this direction could be the application of semantic similarity approaches that would facilitate the curation task, by prioritizing similar terms.

On the other side, from the perspective of mental health genetics, several different improvements could be made. Apart from adding other mental health disorders into the database, curating additional information from studies that report the presence or absence of a gene-disease association would help to better understand psychiatric disorders. During the curation process, several concerns rose as fundamental sources of divergence between curators, such as: (i) the challenge in determining whether or not animal models studies capture well the disease pathophysiology under investigation, (ii) the consideration of studies focused on pharmacogenomics or the response to drug treatments as part of the evidence for a gene-disease association or (iii) the assessment of the statistical

significance threshold in certain publications. In this regard, future work involves revisiting the annotation guidelines to make clear the previous curation concerns risen.

Moreover, as explained before, the manual curation process requires a great effort and it is highly time-consuming. For this reason, alternative strategies could be explored. Implementing a periodical text mining analysis for those novel publications together with an automatic email sender to the publication author could help maintaining the database up to date. This could be achieved by validating the text mining results by the main author of the publication.

As it has been discussed throughout the thesis, many challenges exist in the "big data" era, where biology has acquired the capacity to systematically compile clinical and molecular data at a scale that was unimaginable 20 years ago. A wide range of opportunities for improved healthcare is awaiting, representing an unprecedented opportunity for the biomedical science and the clinical communities to work together and transfer their knowledge.

6 Conclusions

Now, this is not the end.

It is not even the beginning of the end.

But it is, perhaps, the end of the beginning

Sir Winston Churchill (1874 - 1965)

(1) A publicly available tool, called `comoRbidity`, has been developed for the exploitation of clinical data and the identification of comorbidity patterns, making possible the formulation of hypothesis on the etiology of disease comorbidities.

(2) The `comoRbidity` R package has been applied to different studies, enabling researchers to analyze their own clinical data for highly prevalent disorders, such as depression, cancer or COPD, among others.

(3) We developed an approach to distill knowledge from the literature by automatic text mining tools coupled to curation by experts in order to enable the development and maintenance of knowledge resources.

(4) We designed a protocol that includes training the curators and iteratively improving both the tools and annotation guidelines. It was shown to be successful in incorporating new information into the database.

(5) A high-quality database, PsyGeNET (Psychiatric disorders Gene association NETwork) (http://www.psygenet.org/), for the exploratory analysis of mental diseases and their associated genes, has been developed with the obtained curated data.

(6) The R tool `psygenet2r` has been developed for querying, visualizing and analyzing PsyGeNET, thereby providing a unique opportunity to gain insight into the molecular basis of mental disorders. In particular, `psygenet2r` allows performing comorbidity studies in psychiatric disorders at the molecular level.

7 Appendix 1: Publications included in the results section

- A. Gutiérrez-Sacristán, À. Bravo, A. Giannoula, M. A. Mayer, F. Sanz, and L. I. Furlong, "comoRbidity: an R package to analyze disease comorbidities" *Bioinformatics*, Under review

- A. Gutiérrez-Sacristán, C. Hernandez-Ferrer, J. R. Gonzalez,, & L. I. Furlong, "psygenet2r: a R/Bioconductor package for the analysis of psychiatric disease genes." *Bioinformatics,* btx506, Ag. 2017. https://doi.org/10.1093/bioinformatics/btx506

- A. Gutiérrez-Sacristán, À. Bravo, M. Portero-Tresserra, O. Valverde, A. Armario, M. C. Blanco-Gandía, A. Farré, L. Fernández-Ibarrondo, F. Fonseca, J. Giraldo, A. Leis, A. Mané, M. A. Mayer, S. Montagud-Romero, R. Nadal, J. Ortiz, F. J. Pavon, E. J. Perez, M. Rodríguez-Arias, A. Serrano, M. Torrens, V. Warnault, F. Sanz, and L. I. Furlong, "Text mining and expert curation to develop a database on psychiatric diseases and their genes," *Database*, vol. 2017, no. 1, Jan. 2017.

- M. A. Mayer, A. Gutiérrez-Sacristán, A. Leis, S. De La Peña, F. Sanz, and L. I. Furlong, "Using Electronic Health Records to Assess Depression and Cancer Comorbidities.,"

in *Studies in health technology and informatics*, 2017, vol. 235, pp. 236–240.

- R. Faner, A. Gutiérrez-Sacristán, A. Castro-Acosta, S. Grosdidier, W. Gan, M. Sánchez-Mayor, J. L. Lopez-Campos, F. Pozo-Rodriguez, F. Sanz, D. Mannino, L. I. Furlong, and A. Agusti, "Molecular and clinical diseasome of comorbidities in exacerbated COPD patients.," *Eur. Respir. J.*, vol. 46, no. 4, pp. 1001–10, Oct. 2015.

- A. Gutiérrez-Sacristán, S. Grosdidier, O. Valverde, M. Torrens, À. Bravo, J. Piñero, F. Sanz, and L. I. Furlong, "PsyGeNET: a knowledge platform on psychiatric disorders and their genes.," *Bioinformatics*, vol. 31, no. 18, pp. 3075–7, Sep. 2015.

8 Appendix 2: Other publications

- A. Gutiérrez-Sacristán, R. Guedj, F. G. Korodi, J. Stedman, L. I. Furlong, CJ. Patel, IS Kohane and P. Avillach, "Rcupcake: an R package for querying and analyzing biomedical data through the BD2K PICSURE RESTful API" *Bioinformatics*, Under review, Sept. 2017.

- A. Giannoula , A. Gutiérrez-Sacristán, À. Bravo, F. Sanz, and L. I. Furlong, "Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study" *Scientific Report*, Under review, Ag. 2017.

- P. Gomez-Rubio, V. Rosato, M. Márquez, C. Bosetti, E. Molina-Montes, M. Rava, J. Piñero, C. W. Michalski, A. Farré, X. Molero, M. Löhr, L. Ilzarbe, J. Perea, W. Greenhalf, M. O'Rorke, A. Tardón, T. Gress, V. M. Barberá, T. Crnogorac-Jurcevic, L. Muñoz-Bellvís, E. Domínguez-Muñoz, A. Gutiérrez-Sacristán, J. Balsells, E. Costello, C. Guillén-Ponce, J. Huang, M. Iglesias, J. Kleeff, B. Kong, J. Mora, L. Murray, D. O'Driscoll, P. Peláez, I. Poves, R. T. Lawlor, A. Carrato, M. Hidalgo, A. Scarpa, L. Sharp, L. I. Furlong, F. X. Real, C. La Vecchia, N. Malats, and PanGenEU Study Investigators, "A systems approach identifies time-dependent associations of multimorbidities

with pancreatic cancer risk," *Ann. Oncol.*, vol. 28, no. 7, pp. 1618–1624, Jul. 2017.

- J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D833-D839, Oct. 2016.

- M. A. Mayer, L. I. Furlong , P. Torre, I. Planas, F. Cots, E. Izquierdo, J. Portabella, J. Rovira, A. Gutiérrez-Sacristán, and F. Sanz, "Reuse of EHRs to Support Clinical Research in a Hospital of Reference.," in *MIE*, 2015, pp. 224–226.

9 Appendix 3: Contribution in conferences and workshops

- A. Giannoula, A. Gutiérrez-Sacristán, À. Bravo, F. Sanz & L. I. Furlong, "Extraction of temporal associations from patient disease trajectories for a population-based comorbidity study". Poster session; 13th [BC]2 - *the Basel Computational Biology Conference*, Basel, Switzerland; 13-15 September, 2017

- A. Gutiérrez-Sacristán, R. Guedj, E. Centeno, P. Avillach & L. I. Furlong, "Visualization of comorbidities in time series HER: application to pregnancy." Poster session to *Translational Bioinformatics Wellcome Genome Campus* Hinxton, Cambridge, UK; 12-13 June 2017

- A. Gutiérrez-Sacristán, À. Bravo, M. Portero-Tresserra, O. Valverde, A. Armario, M. C. Blanco-Gandía, A. Farré, L. Fernández-Ibarrondo, F. Fonseca, J. Giraldo, A. Leis, A. Mané, M. A. Mayer, S. Montagud-Romero, R. Nadal, J. Ortiz, F. J. Pavon, E. J. Perez, M. Rodríguez-Arias, A. Serrano, M. Torrens, V. Warnault, F. Sanz & L. I. Furlong, "PsyGeNET: a knowledge resource on psychiatric diseases and their genes." Poster session to *Translational Bioinformatics Wellcome Genome Campus Hinxton*, Cambridge, UK; 12-13 June 2017

- A. Gutiérrez-Sacristán, À. Bravo, O. Valverde, M. Torrens, F. Sanz & L. I. Furlong, "Leveraging text mining, expert curation and data integration to develop a database on psychiatric diseases and their genes." Oral participation in the *XIII Symposium on Bioinformatics (JBI2016)*. Valencia, Spain; 10-13 May 2016

- A. Gutiérrez-Sacristán & L. I. Furlong, "comoRbidity: An R package to analyze comorbidities from clinical data" Poster session to *III Symposium on Bioinformatics (JBI2016)*. Valencia, Spain; 10-13 May 2016

- A. Gutiérrez-Sacristán, À. Bravo, O. Valverde, M. Torrens, F. Sanz & L. I. Furlong, "Leveraging text mining, expert curation and data integration to develop a database on psychiatric diseases and their genes." Oral participation in the *Ninth International Biocuration Conference*. Geneva, Switzerland; 10-14 April 2016

- A. Gutiérrez-Sacristán, S. Grosdidier, O. Valverde, M. Torrens, À. Bravo, J. Piñero, F. Sanz & L. I. Furlong, "PsyGeNET: a curated resource on associations between genes and psychiatric disorders." Poster session to *XII Symposium on Bioinformatics*, Sevilla, Spain; 21-24 September 2014

- A. Gutiérrez-Sacristán, S. Grosdidier, M. A. Mayer, F. Sanz & L. I. Furlong, "A network medicine approach to explore comorbidity patterns in Catalonia." Oral participation to a workshop on *Interdisciplinary Signaling*. Visegrad, Hungary; 21-24 July 2014

10 Bibliography

[1] A. R. Feinstein, “The pre-therapeutic classification of co-morbidity in chronic disease,” *J. Chronic Dis.*, vol. 23, no. 7, pp. 455–468, Dec. 1970.

[2] J. Almirall, M. Fortin, and M. Fortin, “The coexistence of terms to describe the presence of multiple concurrent diseases.,” *J. Comorbidity*, vol. 3, no. 1, pp. 4–9, Oct. 2013.

[3] K. D. Lawson, S. W. Mercer, S. Wyke, E. Grieve, B. Guthrie, G. C. Watt, and E. A. Fenwick, “Double trouble: the impact of multimorbidity and deprivation on preference-weighted health related quality of life a cross sectional analysis of the Scottish Health Survey.,” *Int. J. Equity Health*, vol. 12, no. 1, p. 67, Aug. 2013.

[4] B. Starfield and K. Kinder, “Multimorbidity and its measurement.,” *Health Policy*, vol. 103, no. 1, pp. 3–8, Nov. 2011.

[5] E. A. Bayliss, A. E. Edwards, J. F. Steiner, and D. S. Main, “Processes of care desired by elderly patients with multimorbidities,” *Fam. Pract.*, vol. 25, no. 4, pp. 287–293, Aug. 2008.

[6] A. Marengoni, B. Winblad, A. Karp, and L. Fratiglioni, “Prevalence of Chronic Diseases and Multimorbidity Among the Elderly Population in Sweden,” *Am. J. Public Health*, vol. 98, no. 7, pp. 1198–1200, Jul. 2008.

[7] M. Cesari, G. Onder, A. Russo, V. Zamboni, C. Barillaro, L. Ferrucci, M. Pahor, R. Bernabei, and F. Landi, “Comorbidity and Physical Function: Results from the Aging and Longevity Study in the Sirente Geographic Area (ilSIRENTE Study),” *Gerontology*, vol. 52, no. 1, pp. 24–32, Jan. 2006.

[8] T. R. Mikuls and K. G. Saag, “Comorbidity in rheumatoid arthritis.,” *Rheum. Dis. Clin. North Am.*, vol. 27, no. 2, pp. 283–303, May 2001.

[9] M. van den Akker, F. Buntinx, S. Roos, and J. A. Knottnerus, “Problems in determining occurrence rates of

multimorbidity.," *J. Clin. Epidemiol.*, vol. 54, no. 7, pp. 675–679, Jul. 2001.

[10] M. van den Akker, F. Buntinx, and J. A. Knottnerus, "Comorbidity or multimorbidity. What's in a name? A review of literature.," *Eur. J. Gen. Pract.*, vol. 2, no. 2, pp. 65–70, Jan. 1996.

[11] D. McGee, R. Cooper, Y. Liao, and R. Durazo-Arvizu, "Patterns of comorbidity and mortality risk in blacks and whites.," *Ann. Epidemiol.*, vol. 6, no. 5, pp. 381–5, Sep. 1996.

[12] F. G. Schellevis, J. van der Velden, E. van de Lisdonk, J. T. van Eijk, and C. van Weel, "Comorbidity of chronic diseases in general practice.," *J. Clin. Epidemiol.*, vol. 46, no. 5, pp. 469–73, May 1993.

[13] L. M. Verbrugge, J. M. Lepkowski, and Y. Imanaka, "Comorbidity and its impact on disability.," *Milbank Q.*, vol. 67, no. 3–4, pp. 450–84, 1989.

[14] V. Bonavita and R. De Simone, "Towards a definition of comorbidity in the light of clinical complexity," *Neurol. Sci.*, vol. 29, no. S1, pp. 99–102, May 2008.

[15] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: implications for understanding health and health services.," *Ann. Fam. Med.*, vol. 7, no. 4, pp. 357–63, Jan. 2009.

[16] International Research Community on Multimorbidity, "Looking for a consensus for a definition of multimorbidity: the results." [Online]. Available: http://crmcspl-blog.recherche.usherbrooke.ca/?p=927. [Accessed: 14-May-2017].

[17] R. Tabarés-Seisdedos and J. L. Rubenstein, "Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders.," *Nat. Rev. Neurosci.*, vol. 14, no. 4, pp. 293–304, Mar. 2013.

[18] K. Ibáñez, C. Boullosa, R. Tabarés-Seisdedos, A. Baudot,

and A. Valencia, “Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses.,” *PLoS Genet.*, vol. 10, no. 2, p. e1004173, Feb. 2014.

[19] M. Jakovljević, “Psychopharmacotherapy and comorbidity: conceptual and epistemiological issues, dilemmas and controversies.,” *Psychiatr. Danub.*, vol. 21, no. 3, pp. 333–40, Sep. 2009.

[20] V. de Groot, H. Beckerman, G. J. Lankhorst, and L. M. Bouter, “How to measure comorbidity. a critical review of available methods.,” *J. Clin. Epidemiol.*, vol. 56, no. 3, pp. 221–9, Mar. 2003.

[21] M. Jakovljević and L. Ostojić, “Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other.,” *Psychiatr. Danub.*, vol. 25 Suppl 1, pp. 18–28, Jun. 2013.

[22] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási, “The implications of human metabolic network topology for disease comorbidity.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 29, pp. 9880–5, Jul. 2008.

[23] Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, A. Leonardson, L. W. Castellini, S. Wang, M.-F. Champy, B. Zhang, V. Emilsson, S. Doss, A. Ghazalpour, S. Horvath, T. A. Drake, A. J. Lusis, and E. E. Schadt, “Variations in DNA elucidate molecular networks that cause disease,” *Nature*, vol. 452, no. 7186, pp. 429–435, Mar. 2008.

[24] B. Starfield, “Threads and yarns: weaving the tapestry of comorbidity.,” *Ann. Fam. Med.*, vol. 4, no. 2, pp. 101–3, Mar. 2006.

[25] D. McKay, J. S. Abramowitz, J. E. Calamari, M. Kyrios, A. Radomsky, D. Sookman, S. Taylor, and S. Wilhelm, “A critical evaluation of obsessive–compulsive disorder subtypes: Symptoms versus mechanisms,” *Clin. Psychol. Rev.*, vol. 24, no. 3, pp. 283–313, Jul. 2004.

[26] G. Andrews, E. Atkinson, A. Baker, E. Brewin, M. Dadds, L. Degenhardt, K. Gournay, W. Hall, C. Isssakidis, D. Kavanagh, M. Lynskey, L. Manns, K. Mueser, H. Proudfoot, T. Slade, and M. Teesson, *Comorbid mental disorders and substance use disorders: epidemiology, prevention and treatment*. Sidney: National Drug and Alcohol Research Centre, University of New South Wales, 2003.

[27] W. B. Kannel, N. Brand, J. J. Skinner, T. R. Dawber, and P. M. McNamara, "The relation of adiposity to blood pressure and development of hypertension. The Framingham study.," *Ann. Intern. Med.*, vol. 67, no. 1, pp. 48–59, Jul. 1967.

[28] T. Dragisic, A. Dickov, V. Dickov, and V. Mijatovic, "Drug Addiction as Risk for Suicide Attempts.," *Mater. Sociomed.*, vol. 27, no. 3, pp. 188–91, Jun. 2015.

[29] R. Gijsen, N. Hoeymans, F. G. Schellevis, D. Ruwaard, W. A. Satariano, and G. A. M. van den Bos, "Causes and consequences of comorbidity," *J. Clin. Epidemiol.*, vol. 54, no. 7, pp. 661–674, Jul. 2001.

[30] C. Caron and M. Rutter, "Comorbidity in child psychopathology: concepts, issues and research strategies.," *J. Child Psychol. Psychiatry.*, vol. 32, no. 7, pp. 1063–80, Nov. 1991.

[31] K. T. Mueser, R. E. Drake, and M. A. Wallach, "Dual diagnosis: A review of etiological theories.," *Addict. Behav.*, vol. 23, no. 6, pp. 717–734, Nov. 1998.

[32] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease.," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[33] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to CARE: a collaborative engine for practical disease prediction," *Data Min. Knowl. Discov.*, vol. 20, no. 3, pp. 388–415, May 2010.

[34] J. P. Mackenbach, I. Stirbu, A.-J. R. Roskam, M. M. Schaap, G. Menvielle, M. Leinsalu, A. E. Kunst, and European Union Working Group on Socioeconomic Inequalities in Health,

"Socioeconomic Inequalities in Health in 22 European Countries," *N. Engl. J. Med.*, vol. 358, no. 23, pp. 2468–2481, Jun. 2008.

[35] A. Astrup, "Healthy lifestyles in Europe: prevention of obesity and type II diabetes by diet and physical activity.," *Public Health Nutr.*, vol. 4, no. 2B, pp. 499–515, Apr. 2001.

[36] M. J. Sernyak, D. L. Leslie, R. D. Alarcon, M. F. Losonczy, and R. Rosenheck, "Association of Diabetes Mellitus With Use of Atypical Neuroleptics in the Treatment of Schizophrenia," *Am. J. Psychiatry*, vol. 159, no. 4, pp. 561–566, Apr. 2002.

[37] S. J. Wood, "Autism and schizophrenia: one, two or many disorders?," *Br. J. Psychiatry*, vol. 210, no. 4, pp. 241–242, Apr. 2017.

[38] L. S. Carroll and M. J. Owen, "Genetic overlap between autism, schizophrenia and bipolar disorder.," *Genome Med.*, vol. 1, no. 10, p. 102, Oct. 2009.

[39] A. W. Taylor, K. Price, T. K. Gill, R. Adams, R. Pilkington, N. Carrangis, Z. Shi, and D. Wilson, "Multimorbidity - not just an older person's issue. Results from an Australian biomedical study," *BMC Public Health*, vol. 10, no. 1, p. 718, Dec. 2010.

[40] A. Marengoni, E. von Strauss, D. Rizzuto, B. Winblad, and L. Fratiglioni, "The impact of chronic multimorbidity and disability on functional decline and survival in elderly persons. A community-based, longitudinal study," *J. Intern. Med.*, vol. 265, no. 2, pp. 288–295, Feb. 2009.

[41] J. E. Mezzich and I. M. Salloum, "Clinical complexity and person-centered integrative diagnosis.," *World Psychiatry*, vol. 7, no. 1, pp. 1–2, Feb. 2008.

[42] M. Fortin, L. Lapointe, C. Hudon, and A. Vanasse, "Multimorbidity is common to family practice: is it commonly researched?," *Can. Fam. Physician*, vol. 51, no. 2, pp. 244–5, Feb. 2005.

[43] N. Garin, B. Olaya, M. V. Moneta, M. Miret, A. Lobo, J. L. Ayuso-Mateos, and J. M. Haro, "Impact of Multimorbidity on Disability and Quality of Life in the Spanish Older Population," *PLoS One*, vol. 9, no. 11, p. e111498, Nov. 2014.

[44] S. S. Lim, T. Vos, A. D. Flaxman, G. Danaei, K. Shibuya, H. Adair-Rohani, M. Amann, H. R. Anderson, K. G. Andrews, M. Aryee, C. Atkinson, L. J. Bacchus, A. N. Bahalim, K. Balakrishnan, J. Balmes, S. Barker-Collo, A. Baxter, M. L. Bell, J. D. Blore, F. Blyth, C. Bonner, G. Borges, R. Bourne, M. Boussinesq, M. Brauer, P. Brooks, N. G. Bruce, B. Brunekreef, C. Bryan-Hancock, C. Bucello, R. Buchbinder, F. Bull, R. T. Burnett, T. E. Byers, B. Calabria, J. Carapetis, E. Carnahan, Z. Chafe, F. Charlson, H. Chen, J. S. Chen, A. T.-A. Cheng, J. C. Child, A. Cohen, K. E. Colson, B. C. Cowie, S. Darby, S. Darling, A. Davis, L. Degenhardt, F. Dentener, D. C. Des Jarlais, K. Devries, M. Dherani, E. L. Ding, E. R. Dorsey, T. Driscoll, K. Edmond, S. E. Ali, R. E. Engell, P. J. Erwin, S. Fahimi, G. Falder, F. Farzadfar, A. Ferrari, M. M. Finucane, S. Flaxman, F. G. R. Fowkes, G. Freedman, M. K. Freeman, E. Gakidou, S. Ghosh, E. Giovannucci, G. Gmel, K. Graham, R. Grainger, B. Grant, D. Gunnell, H. R. Gutierrez, W. Hall, H. W. Hoek, A. Hogan, H. D. Hosgood, D. Hoy, H. Hu, B. J. Hubbell, S. J. Hutchings, S. E. Ibeanusi, G. L. Jacklyn, R. Jasrasaria, J. B. Jonas, H. Kan, J. A. Kanis, N. Kassebaum, N. Kawakami, Y.-H. Khang, S. Khatibzadeh, J.-P. Khoo, C. Kok, F. Laden, R. Lalloo, Q. Lan, T. Lathlean, J. L. Leasher, J. Leigh, Y. Li, J. K. Lin, S. E. Lipshultz, S. London, R. Lozano, Y. Lu, J. Mak, R. Malekzadeh, L. Mallinger, W. Marcenes, L. March, R. Marks, R. Martin, P. McGale, J. McGrath, S. Mehta, G. A. Mensah, T. R. Merriman, R. Micha, C. Michaud, V. Mishra, K. Mohd Hanafiah, A. A. Mokdad, L. Morawska, D. Mozaffarian, T. Murphy, M. Naghavi, B. Neal, P. K. Nelson, J. M. Nolla, R. Norman, C. Olives, S. B. Omer, J. Orchard, R. Osborne, B. Ostro, A. Page, K. D. Pandey, C. D. H. Parry, E. Passmore, J. Patra, N. Pearce, P. M. Pelizzari, M. Petzold, M. R. Phillips, D. Pope, C. A. Pope, J. Powles, M. Rao, H. Razavi, E. A. Rehfuess, J. T. Rehm, B. Ritz, F. P. Rivara, T. Roberts, C. Robinson, J. A. Rodriguez-Portales, I. Romieu,

R. Room, L. C. Rosenfeld, A. Roy, L. Rushton, J. A. Salomon, U. Sampson, L. Sanchez-Riera, E. Sanman, A. Sapkota, S. Seedat, P. Shi, K. Shield, R. Shivakoti, G. M. Singh, D. A. Sleet, E. Smith, K. R. Smith, N. J. C. Stapelberg, K. Steenland, H. Stöckl, L. J. Stovner, K. Straif, L. Straney, G. D. Thurston, J. H. Tran, R. Van Dingenen, A. van Donkelaar, J. L. Veerman, L. Vijayakumar, R. Weintraub, M. M. Weissman, R. A. White, H. Whiteford, S. T. Wiersma, J. D. Wilkinson, H. C. Williams, W. Williams, N. Wilson, A. D. Woolf, P. Yip, J. M. Zielinski, A. D. Lopez, C. J. L. Murray, M. Ezzati, M. A. AlMazroa, and Z. A. Memish, "A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010.," *Lancet (London, England)*, vol. 380, no. 9859, pp. 2224–60, Dec. 2012.

[45] S. Afshar, P. J. Roderick, P. Kowal, B. D. Dimitrov, and A. G. Hill, "Global Patterns of Multimorbidity: A Comparison of 28 Countries Using the World Health Surveys," in *Applied Demography and Public Health in the 21st Century*, Springer International Publishing, 2017, pp. 381–402.

[46] H. A. Whiteford, L. Degenhardt, J. Rehm, A. J. Baxter, A. J. Ferrari, H. E. Erskine, F. J. Charlson, R. E. Norman, A. D. Flaxman, N. Johns, R. Burstein, C. J. Murray, and T. Vos, "Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010.," *Lancet*, vol. 382, no. 9904, pp. 1575–1586, Nov. 2013.

[47] C. J. L. Murray and A. D. Lopez, "Measuring the Global Burden of Disease," *N. Engl. J. Med.*, vol. 369, no. 5, pp. 448–457, Aug. 2013.

[48] H. A. Whiteford, A. J. Ferrari, L. Degenhardt, V. Feigin, and T. Vos, "The Global Burden of Mental, Neurological and Substance Use Disorders: An Analysis from the Global Burden of Disease Study 2010.," *PLoS One*, vol. 10, no. 2, p. e0116820, Feb. 2015.

[49] G. B. of D. S. 2013 Global Burden of Disease Study 2013 Collaborators, “Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.,” *Lancet*, vol. 386, no. 9995, pp. 743–800, Aug. 2015.

[50] F. Baingana, M. al’Absi, A. E. Becker, and B. Pringle, “Global research challenges and opportunities for mental health and substance-use disorders,” *Nature*, vol. 527, no. 7578, pp. S172–S177, Nov. 2015.

[51] N. Sartorius, “Comorbidity of mental and physical disorders: A major challenge for medicine in the 21 st century,” *Eur. Psychiatry*, vol. 41, p. S9, Apr. 2017.

[52] R. C. Kessler, W. T. Chiu, O. Demler, E. E. Walters, and E. E. Walters, “Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication.,” *Arch. Gen. Psychiatry*, vol. 62, no. 6, pp. 617–627, Jun. 2005.

[53] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, “Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication.,” *Arch. Gen. Psychiatry*, vol. 62, no. 6, pp. 593–602, Jun. 2005.

[54] R. C. Kessler, K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H. U. Wittchen, and K. S. Kendler, “Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey.,” *Arch. Gen. Psychiatry*, vol. 51, no. 1, pp. 8–19, Jan. 1994.

[55] B. G. Druss and E. R. Walker, “Mental disorders and medical comorbidity.,” *Synth. Proj. Res. Synth. Rep.*, no. 21, pp. 1–26, Feb. 2011.

[56] M. A. Nowels and L. M. VanderWielen, “Comorbidity indices: a call for the integration of physical and mental health,” *Prim. Health Care Res. Dev.*, pp. 1–3, 2017.

[57] “Mental health horizons,” *Nat. Med.*, vol. 22, no. 11, pp. 1213–1213, Nov. 2016.

[58] R. J. Anderson, K. E. Freedland, R. E. Clouse, and P. J. Lustman, “The prevalence of comorbid depression in adults with diabetes: a meta-analysis.,” *Diabetes Care*, vol. 24, no. 6, pp. 1069–78, Jun. 2001.

[59] M. Härter, H. Baumeister, K. Reuter, F. Jacobi, M. Höfler, J. Bengel, and H.-U. Wittchen, “Increased 12-month prevalence rates of mental disorders in patients with chronic somatic diseases.,” *Psychother. Psychosom.*, vol. 76, no. 6, pp. 354–60, Jan. 2007.

[60] L. J. Williams, J. A. Pasco, F. N. Jacka, M. J. Henry, S. Dodd, and M. Berk, “Depression and bone metabolism. A review.,” *Psychother. Psychosom.*, vol. 78, no. 1, pp. 16–25, Jan. 2009.

[61] F. Matcham, L. Rayner, S. Steer, and M. Hotopf, “The prevalence of depression in rheumatoid arthritis: a systematic review and meta-analysis.,” *Rheumatology (Oxford).*, vol. 52, no. 12, pp. 2136–48, Dec. 2013.

[62] M.-H. Chen, T.-P. Su, Y.-S. Chen, J.-W. Hsu, K.-L. Huang, W.-H. Chang, T.-J. Chen, and Y.-M. Bai, “Higher risk of developing major depression and bipolar disorder in later life among adolescents with asthma: A nationwide prospective study,” *J. Psychiatr. Res.*, vol. 49, pp. 25–30, 2014.

[63] T. W. Strine, A. H. Mokdad, L. S. Balluz, O. Gonzalez, R. Crider, J. T. Berry, and K. Kroenke, “Depression and Anxiety in the United States: Findings From the 2006 Behavioral Risk Factor Surveillance System,” *Psychiatr. Serv.*, vol. 59, no. 12, pp. 1383–1390, Dec. 2008.

[64] J. Sokal, E. Messias, F. B. Dickerson, J. Kreyenbuhl, C. H. Brown, R. W. Goldberg, and L. B. Dixon, “Comorbidity of medical illnesses among adults with serious mental illness who are receiving community psychiatric services.,” *J. Nerv. Ment. Dis.*, vol. 192, no. 6, pp. 421–7, Jun. 2004.

[65] J. C. Wakefield, “The concept of mental disorder: diagnostic

implications of the harmful dysfunction analysis.," *World Psychiatry*, vol. 6, no. 3, pp. 149–56, Oct. 2007.

[66] T. R. Insel and B. N. Cuthbert, "Brain disorders? Precisely.," *Science.*, vol. 348, no. 6234, pp. 499–500, May 2015.

[67] M. C. O'Donovan and M. J. Owen, "The implications of the shared genetics of psychiatric disorders.," *Nat. Med.*, vol. 22, no. 11, pp. 1214–1219, Oct. 2016.

[68] B. M. Cohen, D. R, H. X, and N. BM, "Embracing Complexity in Psychiatric Diagnosis, Treatment, and Research," *JAMA Psychiatry*, vol. 73, no. 12, p. 1211, Dec. 2016.

[69] T. J. C. Polderman, B. Benyamin, C. A. de Leeuw, P. F. Sullivan, A. van Bochoven, P. M. Visscher, and D. Posthuma, "Meta-analysis of the heritability of human traits based on fifty years of twin studies.," *Nat. Genet.*, vol. 47, no. 7, pp. 702–9, Jul. 2015.

[70] C. A. Prescott and K. S. Kendler, "Genetic and Environmental Contributions to Alcohol Abuse and Dependence in a Population-Based Sample of Male Twins," *Am. J. Psychiatry*, vol. 156, no. 1, pp. 34–40, Jan. 1999.

[71] M. A. Schuckit, H. J. Edenberg, J. Kalmijn, L. Flury, T. L. Smith, T. Reich, L. Bierut, A. Goate, and T. Foroud, "A genome-wide search for genes that relate to a low level of response to alcohol.," *Alcohol. Clin. Exp. Res.*, vol. 25, no. 3, pp. 323–9, Mar. 2001.

[72] M. J. Gandal, V. Leppa, H. Won, N. N. Parikshak, and D. H. Geschwind, "The road to precision psychiatry: translating genetics into disease mechanisms," *Nat. Neurosci.*, vol. 19, no. 11, pp. 1397–1407, Oct. 2016.

[73] M. L. Santoro, P. N. Moretti, R. Pellegrino, A. Gadelha, V. C. Abílio, M. A. F. Hayashi, S. I. Belangero, and H. Hakonarson, "A current snapshot of common genomic variants contribution in psychiatric disorders," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 171, no. 8, pp. 997–1005, Dec. 2016.

[74] G. Senthil, T. Dutka, L. Bingaman, and T. Lehner, “Genomic resources for the study of neuropsychiatric disorders,” *Mol. Psychiatry*, Mar. 2017.

[75] S. Akbarian, C. Liu, J. A. Knowles, F. M. Vaccarino, P. J. Farnham, G. E. Crawford, A. E. Jaffe, D. Pinto, S. Dracheva, D. H. Geschwind, J. Mill, A. C. Nairn, A. Abyzov, S. Pochareddy, S. Prabhakar, S. Weissman, P. F. Sullivan, M. W. State, Z. Weng, M. A. Peters, K. P. White, M. B. Gerstein, A. Amiri, C. Armoskus, A. E. Ashley-Koch, T. Bae, A. Beckel-Mitchener, B. P. Berman, G. A. Coetzee, G. Coppola, N. Francoeur, M. Fromer, R. Gao, K. Grennan, J. Herstein, D. H. Kavanagh, N. A. Ivanov, Y. Jiang, R. R. Kitchen, A. Kozlenkov, M. Kundakovic, M. Li, Z. Li, S. Liu, L. M. Mangravite, E. Mattei, E. Markenscoff-Papadimitriou, F. C. P. Navarro, N. North, L. Omberg, D. Panchision, N. Parikshak, J. Poschmann, A. J. Price, M. Purcaro, T. E. Reddy, P. Roussos, S. Schreiner, S. Scuderi, R. Sebra, M. Shibata, A. W. Shieh, M. Skarica, W. Sun, V. Swarup, A. Thomas, J. Tsuji, H. van Bakel, D. Wang, Y. Wang, K. Wang, D. M. Werling, A. J. Willsey, H. Witt, H. Won, C. C. Y. Wong, G. A. Wray, E. Y. Wu, X. Xu, L. Yao, G. Senthil, T. Lehner, P. Sklar, N. Sestan, and N. Sestan, “The PsychENCODE project.,” *Nat. Neurosci.*, vol. 18, no. 12, pp. 1707–1712, Nov. 2015.

[76] P. F. Sullivan, “The psychiatric GWAS consortium: big science comes to psychiatry.,” *Neuron*, vol. 68, no. 2, pp. 182–6, Oct. 2010.

[77] P. Jia, G. Han, J. Zhao, P. Lu, and Z. Zhao, “SZGR 2.0: a one-stop shop of schizophrenia candidate genes.,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D915–D924, Jan. 2017.

[78] S.-H. Chang, L. Gao, Z. Li, W.-N. Zhang, Y. Du, and J. Wang, “BDgene: a genetic database for bipolar disorder and its overlap with schizophrenia and major depressive disorder.,” *Biol. Psychiatry*, vol. 74, no. 10, pp. 727–33, Nov. 2013.

[79] C.-C. Huang, Z. Lu, and MaoY, “Community challenges in biomedical text mining over 10 years: success, failure and the

future," *Brief. Bioinform.*, vol. 17, no. 1, pp. 132–144, Jan. 2016.

[80] A. Koike, Y. Niwa, and T. Takagi, "Automatic extraction of gene/protein biological functions from biomedical text," *Bioinformatics*, vol. 21, no. 7, pp. 1227–1236, Apr. 2005.

[81] S. Pakhomov, B. T. McInnes, J. Lamba, Y. Liu, G. B. Melton, Y. Ghodke, N. Bhise, V. Lamba, and A. K. Birnbaum, "Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies," *J. Biomed. Inform.*, vol. 45, no. 5, pp. 862–869, Oct. 2012.

[82] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress," *Brief. Bioinform.*, vol. 8, no. 5, pp. 358–375, Jun. 2007.

[83] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, "Text mining of 15 million full-text scientific articles," *bioRxiv*, 2017.

[84] T. Konneker, T. Barnes, H. Furberg, M. Losh, C. M. Bulik, and P. F. Sullivan, "A searchable database of genetic evidence for psychiatric disorders," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 147B, no. 6, pp. 671–675, Sep. 2008.

[85] P. Jia, J. Sun, A. Y. Guo, and Z. Zhao, "SZGR: a comprehensive schizophrenia gene resource.," *Mol. Psychiatry*, vol. 15, no. 5, pp. 453–62, May 2010.

[86] N. C. Allen, S. Bagade, M. B. McQueen, J. P. A. Ioannidis, F. K. Kavvoura, M. J. Khoury, R. E. Tanzi, and L. Bertram, "Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database," *Nat. Genet.*, vol. 40, no. 7, pp. 827–834, Jul. 2008.

[87] Y. Wu, Y.-G. Yao, and X.-J. Luo, "SZDB: A Database for Schizophrenia Genetic Research," *Schizophr. Bull.*, vol. 43, no. 2, pp. 459–471, Mar. 2017.

[88] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, G. De Moor, and D. Kalra, "Electronic health records: new opportunities for clinical research.," *J. Intern. Med.*, vol. 274, no. 6, pp. 547–560, Dec. 2013.

[89] C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng, "Big Data Bioinformatics," *J. Cell. Physiol.*, vol. 229, no. 12, pp. 1896–1900, Dec. 2014.

[90] W.-Q. Wei and J. C. Denny, "Extracting research-quality phenotypes from electronic health records to support precision medicine," *Genome Med.*, vol. 7, no. 1, p. 41, Dec. 2015.

[91] S. Silow-Carroll, J. N. Edwards, and D. Rodin, "Using electronic health records to improve quality and efficiency: the experiences of leading hospitals.," *Issue Brief (Commonw. Fund)*, vol. 17, pp. 1–40, Jul. 2012.

[92] D. P. Jutte, L. L. Roos, and M. D. Brownell, "Administrative Record Linkage as a Tool for Public Health Research," *Annu. Rev. Public Health*, vol. 32, no. 1, pp. 91–108, Apr. 2011.

[93] M. S. M. van Mourik, P. J. van Duijn, K. G. M. Moons, M. J. M. Bonten, and G. M. Lee, "Accuracy of administrative data for surveillance of healthcare-associated infections: a systematic review," *BMJ Open*, vol. 5, no. 8, p. e008424, Aug. 2015.

[94] F. Doshi-Velez, Y. Ge, and I. Kohane, "Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis," *Pediatrics*, vol. 133, no. 1, pp. e54–e63, Jan. 2014.

[95] K. P. Liao, T. Cai, G. K. Savova, S. N. Murphy, E. W. Karlson, A. N. Ananthakrishnan, V. S. Gainer, S. Y. Shaw, Z. Xia, P. Szolovits, S. Churchill, and I. Kohane, "Development of phenotype algorithms using electronic medical records and incorporating natural language processing.," *BMJ*, vol. 350, p. h1885, Apr. 2015.

[96] D. M. Roden, H. Xu, J. C. Denny, and R. A. Wilke,

"Electronic Medical Records as a Tool in Clinical Pharmacology: Opportunities and Challenges," *Clin. Pharmacol. Ther.*, vol. 91, no. 6, pp. 1083–1086, Jun. 2012.

[97] M. R. Cowie, J. I. Blomster, L. H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, A. Michel, S. Ong, J. P. Pell, M. R. Southworth, W. G. Stough, M. Thoenes, F. Zannad, and A. Zalewski, "Electronic health records to facilitate clinical research," *Clin. Res. Cardiol.*, vol. 106, no. 1, pp. 1–9, Jan. 2017.

[98] W. R. Hogan, M. M. Wagner, van der S. E, W. D, T. CK, and F. J, "Accuracy of Data in Computer-based Patient Records," *J. Am. Med. Informatics Assoc.*, vol. 4, no. 5, pp. 342–355, Sep. 1997.

[99] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 144–51, Jan. 2013.

[100] K. Thiru, A. Hassey, and F. Sullivan, "Systematic review of scope and quality of electronic patient record data in primary care," *BMJ*, vol. 326, no. 7398, pp. 1070–0, May 2003.

[101] K. S. Chan, J. B. Fowles, and J. P. Weiner, "Review: Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature," *Med. Care Res. Rev.*, vol. 67, no. 5, pp. 503–527, Oct. 2010.

[102] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis," *Bioinformatics*, vol. 31, no. 4, pp. 608–609, Feb. 2015.

[103] National Research Council (U.S.). Committee on A Framework for Developing a New Taxonomy of Disease. and National Research Council (U.S.). Board on Life Sciences., *Toward precision medicine : building a knowledge network for biomedical research and a new taxonomy of disease*. .

[104] J. X. Hu, C. E. Thomas, and S. Brunak, "Network biology

concepts in complex disease comorbidities.," *Nat. Rev. Genet.*, vol. 17, no. 10, pp. 615–629, Aug. 2016.

[105] G. D. Fischbach and C. Lord, "The Simons Simplex Collection: a resource for identification of autism genetic risk factors.," *Neuron*, vol. 68, no. 2, pp. 192–5, Oct. 2010.

[106] S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).," *J. Am. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 124–30, Mar. 2010.

[107] R. B. Altman, "Translational bioinformatics: linking the molecular world to the clinical world.," *Clin. Pharmacol. Ther.*, vol. 91, no. 6, pp. 994–1000, Jun. 2012.

[108] H. Burton, T. Cole, and A. M. Lucassen, "Genomic medicine: challenges and opportunities for physicians.," *Clin. Med.*, vol. 12, no. 5, pp. 416–9, Oct. 2012.

[109] A. G. Ury, "Storing and interpreting genomic information in widely deployed electronic health record systems.," *Genet. Med.*, vol. 15, no. 10, pp. 779–85, Oct. 2013.

[110] D. J. Rigden, X. M. Fernández-Suárez, and M. Y. Galperin, "The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection.," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1-6, Jan. 2016.

[111] M. A. Moni and P. Liò, "comoR: a software for disease comorbidity risk assessment.," *J. Clin. Bioinforma.*, vol. 4, no. 1, p. 8, 2014.

[112] P. McCormick, "medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data.," *Cran R project*, Jan-2016. [Online]. Available: https://cran.r-project.org/web/packages/medicalrisk/vignettes/medicalrisk.html. [Accessed: 10-Sep-2017].

[113] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, "A dynamic network approach for the study of human phenotypes.," *PLoS Comput. Biol.*, vol. 5, no. 4, p.

e1000353, Apr. 2009.

[114] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data.," *Med. Care*, vol. 36, no. 1, pp. 8–27, Jan. 1998.

[115] H. Xu, M. A. Moni, and P. Liò, "Network regularised Cox regression and multiplex network models to predict disease comorbidities and survival of cancer," *Comput. Biol. Chem.*, vol. 59, no. Part B, pp. 15–31, Dec. 2015.

[116] A. Hamosh, A. F. Scott, J. S. Amberger, C. a Bocchini, and V. a McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.," *Nucleic Acids Res.*, vol. 33, no. suppl_1, pp. D514–D517, Jan. 2005.

[117] M. Kanehisa, "The KEGG database.," in *Novartis Foundation symposium*, 2002, vol. 247, pp. 91–103.

[118] M. van den Akker, F. Buntinx, J. F. Metsemakers, and J. A. Knottnerus, "Marginal impact of psychosocial factors on multimorbidity: results of an explorative nested case-control study.," *Soc. Sci. Med.*, vol. 50, no. 11, pp. 1679–93, Jun. 2000.

[119] "WHO | International Classification of Diseases," *WHO*, 2017. [Online]. Available: http://www.who.int/classifications/icd/en/. [Accessed: 04-Sep-2017].

[120] "Unified Medical Language System (UMLS)." [Online]. Available: https://www.nlm.nih.gov/research/umls/. [Accessed: 04-Sep-2017].

[121] J. M. Quail, L. M. Lix, B. A. Osman, and G. F. Teare, "Comparing comorbidity measures for predicting mortality and hospitalization in three population-based cohorts.," *BMC Health Serv. Res.*, vol. 11, no. 1, p. 146, Jun. 2011.

[122] A. L. Huntley, R. Johnson, S. Purdy, J. M. Valderas, and C. Salisbury, "Measures of multimorbidity and morbidity burden for use in primary care and community settings: a

systematic review and guide.," *Ann. Fam. Med.*, vol. 10, no. 2, pp. 134–41, Mar. 2012.

[123] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.," *J. Chronic Dis.*, vol. 40, no. 5, pp. 373–83, 1987.

[124] M. Von Korff, E. H. Wagner, and K. Saunders, "A chronic disease score from automated pharmacy data.," *J. Clin. Epidemiol.*, vol. 45, no. 2, pp. 197–203, Feb. 1992.

[125] D. Sarfati, A. M. Jette, G. Zahner, E. Guadagnoli, R. A. Silliman, A. Trentham-Dietz, and et al., "Review of methods used to measure comorbidity in cancer populations: no gold standard exists.," *J. Clin. Epidemiol.*, vol. 65, no. 9, pp. 924–33, Sep. 2012.

[126] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen, and S. Brunak, "Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts.," *PLoS Comput. Biol.*, vol. 7, no. 8, p. e1002141, Aug. 2011.

[127] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, "Controlling the false discovery rate in behavior genetics research.," *Behav. Brain Res.*, vol. 125, no. 1–2, pp. 279–84, Nov. 2001.

[128] S. Munevar, "Unlocking Big Data for better health.," *Nat. Biotechnol.*, vol. 35, no. 7, pp. 684–686, Jul. 2017.

[129] A. J. Butte, "Big data opens a window onto wellness.," *Nat. Biotechnol.*, vol. 35, no. 8, pp. 720–721, Aug. 2017.

[130] I. S. Kohane, "Using electronic health records to drive discovery in disease genomics," *Nat. Rev. Genet.*, vol. 12, no. 6, pp. 417–428, Jun. 2011.

[131] M. May, "Life science technologies: Big biological impacts from big data.," *Science.*, vol. 344, no. 6189, pp. 1298–1300,

Jun. 2014.

[132] M. A. Mayer, A. Gutiérrez-Sacristán, A. Leis, S. De La Peña, F. Sanz, and L. I. Furlong, "Using Electronic Health Records to Assess Depression and Cancer Comorbidities.," in *Studies in health technology and informatics*, 2017, vol. 235, pp. 236–240.

[133] R. Faner, A. Gutiérrez-Sacristán, A. Castro-Acosta, S. Grosdidier, W. Gan, M. Sánchez-Mayor, J. L. Lopez-Campos, F. Pozo-Rodriguez, F. Sanz, D. Mannino, L. I. Furlong, and A. Agusti, "Molecular and clinical diseasome of comorbidities in exacerbated COPD patients.," *Eur. Respir. J.*, vol. 46, no. 4, pp. 1001–10, Oct. 2015.

[134] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research.," *BMC Bioinformatics*, vol. 16, no. 1, p. 55, Feb. 2015.

[135] J. R. Conway, A. Lex, and N. Gehlenborg, "UpSetR: an R package for the visualization of intersecting sets and their properties," *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, Sep. 2017.

[136] "Bioconductor - psygenet2r." [Online]. Available: https://www.bioconductor.org/packages/release/bioc/html/psygenet2r.html. [Accessed: 05-Sep-2017].

[137] R. Conti, D. L. Veenstra, K. Armstrong, L. J. Lesko, and S. D. Grosse, "Personalized Medicine and Genomics: Challenges and Opportunities in Assessing Effectiveness, Cost-Effectiveness, and Future Research Priorities," *Med. Decis. Mak.*, vol. 30, no. 3, pp. 328–340, May 2010.

[138] M. J. Khoury, M. Gwinn, P. W. Yoon, N. Dowling, C. A. Moore, and L. Bradley, "The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?," *Genet. Med.*, vol. 9, no. 10, pp. 665–74, Oct. 2007.