Bioinformatics, 34(18), 2018, 3228–3230
doi: 10.1093/bioinformatics/bty315
Advance Access Publication Date: 20 April 2018
Applications Note

Data and text mining

comoRbidity: an R package for the systematic analysis of disease comorbidities

Alba Gutiérrez-Sacristán[1,2], Àlex Bravo[1,3], Alexia Giannoula[1], Miguel A. Mayer[1], Ferran Sanz[1] and Laura I. Furlong[1,*]

[1]Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences (DCEXS), Hospital del Mar Medical Research Institute (IMIM), Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain, [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA and [3]Large-Scale Text Understanding Systems Lab, TALN Research Group, Department of Information and Communication Technologies (DTIC), Universitat Pompeu Fabra, Barcelona 08018, Spain

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on September 1, 2017; revised on March 23, 2018; editorial decision on April 18, 2018; accepted on April 19, 2018

Abstract

Motivation: The study of comorbidities is a major priority due to their impact on life expectancy, quality of life and healthcare cost. The availability of electronic health records (EHRs) for data mining offers the opportunity to discover disease associations and comorbidity patterns from the clinical history of patients gathered during routine medical care. This opens the need for analytical tools for detection of disease comorbidities, including the investigation of their underlying genetic basis.

Results: We present comoRbidity, an R package aimed at providing a systematic and comprehensive analysis of disease comorbidities from both the clinical and molecular perspectives. comoRbidity leverages from (i) user provided clinical data from EHR databases (the clinical comorbidity analysis) and (ii) genotype-phenotype information of the diseases under study (the molecular comorbidity analysis) for a comprehensive analysis of disease comorbidities. The clinical comorbidity analysis enables identifying significant disease comorbidities from clinical data, including sex and age stratification and temporal directionality analyses, while the molecular comorbidity analysis supports the generation of hypothesis on the underlying mechanisms of the disease comorbidities by exploring shared genes among disorders. The open-source comoRbidity package is a software tool aimed at expediting the integrative analysis of disease comorbidities by incorporating several analytical and visualization functions.

Availability and implementation: https://bitbucket.org/ibi_group/comorbidity

Contact: laura.furlong@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The co-existence of two or more diseases in the same patient, also known as comorbidity (van den Akker *et al.*, 1996; Valderas *et al.*, 2009) is a matter of public health concern as it has important consequences both for patients and the healthcare system (Gijsen *et al.*, 2001; Valderas *et al.*, 2009). According to several studies, the prevalence of comorbidity varies between ∼20% and ∼90% (Bonavita and De Simone, 2008; Fortin *et al.*, 2005; Mezzich and Salloum, 2008; Marengoni *et al.*, 2011). This variation is due to the population under study, as well as other characteristics of the study design, such as the definition of comorbidity (van den Akker *et al.*, 1996; Valderas *et al.*, 2009). Although the prevalence of comorbidity increases with age, it is not limited to the elderly population (Doshi-Velez *et al.*, 2014; Jakovljević and Ostojić, 2013; Marengoni *et al.*, 2011;

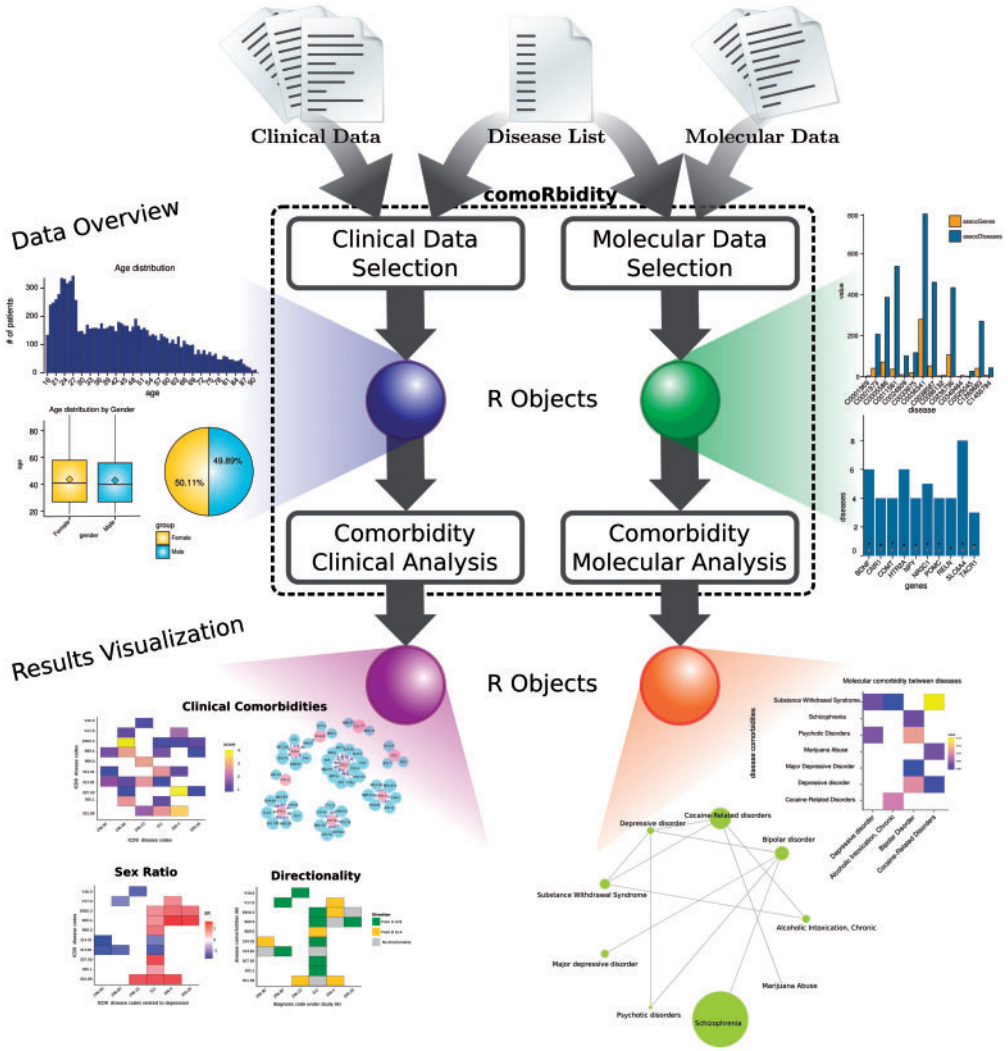

Fig. 1. Overview of comoRbidity

Taylor *et al.*, 2010). The availability of electronic health records (EHR) for data mining offers the opportunity to discover disease associations and comorbidity patterns from the clinical history of patients gathered during routine medical care (Bagley *et al.*, 2016; Backenroth *et al.*, 2016; Holmes *et al.*, 2011).

In recent years, there has been a growing interest in the re-use of clinic data for research (Jensen *et al.*, 2012). In this context, the availability of tools that enable the analysis of clinical data in a reproducible manner and in a secure environment is key. The development of analytical tools to identify comorbidity patterns from clinical data will enable: (i) the estimation of the prevalence of comorbidities in particular populations, (ii) the stratification of patients according to their comorbidities and (iii) the development of decision support systems in the clinical setting.

In this paper, we introduce comoRbidity, an R package aimed at providing a comprehensive analysis of disease comorbidities from both the clinical and molecular perspectives. comoRbidity leverages from clinical data obtained from EHR databases or health registries (the clinical comorbidity analysis), and from genotype-phenotype information of the diseases under study (the molecular comorbidity analysis) from DisGeNET (Piñero *et al.*, 2017), or provided by the user.

2 Design and implementation

comoRbidity aims at expediting the analysis of disease comorbidities by providing several analytical functions and different visualization options to analyze clinical data provided by the user. comoRbidity is based on standard CRAN and Bioconductor classes allowing for full flexibility and integration with other R packages. It runs under Linux, Windows and Mac operating systems.

The R CRAN package parallel (R, 2014) is used to speed up the comorbidity estimation by adjusting the cores according to the user requirements. comoRbidity contains 14 R functions (see Supplementary Table S1 for details) used to process clinical and molecular data to perform the disease comorbidity analysis and visualize the results. The package includes a dataset of artificially generated clinical data (http://www.emrbots.org/) to illustrate the functionalities of the package.

The software implements two types of independent analysis, the clinical comorbidity analysis and the molecular comorbidity analysis. An overview of the workflow of data analysis provided by the package is shown in Figure 1. Each analysis includes three sequential steps:

i. *Data Selection*: From the user's input data, comoRbidity provides an overview of the data, including a demographic analysis

based on age and sex, the number of genes associated to the diseases under study, or the number of diseases sharing genes.

ii. *Data Analysis*: The comorbidity analysis is performed, based on different parameters set by the user.

iii. *Results Visualization*: The package offers different options for the visualization of the results.

3 Related work

To the best of our knowledge, only few tools have been developed for the analysis of disease comorbidities, namely comoR (Moni *et al.*, 2014), CytoCom (Moni *et al.*, 2015) and medicalRisk (McCormick, 2016). In the R environment, the comoR package (Moni *et al.*, 2014) computes statistically significant associations among diseases based on the US Medicare claims database (Hidalgo *et al.*, 2009) along with several molecular and phenotypic association metrics. The same authors developed CytoCom2 (Moni *et al.*, 2015), a Cytoscape App to visualize and query their disease comorbidity networks (Hidalgo *et al.*, 2009). The medicalRisk R package (McCormick, 2016), can be used to obtain medical risk status from large datasets with diseases encoded in ICD-9-CM, based on mortality predictors such as the Charlson Comorbidity Index and the Elixhauser comorbidity map. Compared to these tools, the main advantage of comoRbidity is the possibility to analyze the user's own clinical data in his/her private workstation, avoiding any privacy issues concerning the sharing of patient data. In addition, comoRbidity allows any classification to encode diseases, and provides different statistics and functions for assessing comorbidity between disorders. Finally, it allows exploring the genetic basis of disease comorbidities by the analysis of gene-disease association data from DisGeNET (Piñero *et al.*, 2017), or from gene-disease association data provided by the user.

4 Conclusions

The comoRbidity package is a novel, publicly available tool for the processing of healthcare data to identify comorbidity patterns enabling their analysis in a user-friendly and reproducible manner. More importantly, it permits the user to provide its own clinical data, which can be analyzed locally in a secure environment. comoRbidity supports any classification system used to identify diseases and/or phenotypes. In addition, it permits full flexibility to the user in the definition of comorbidity regarding the temporal window considered, the diseases of interest and the use of primary or secondary diagnoses in the analysis, among other aspects. Several analytical and visualization functions are provided including metrics to assess disease associations and their temporal directionality. In addition, it allows performing a molecular analysis of the comorbidities even if no genomic data of the patient is available, by using publicly available information on gene-disease associations, making possible the formulation of hypothesis regarding the etiology of disease comorbidities.

Funding

The authors received support from ISCIII-FEDER (PI13/00082, CP10/00524, CPII16/00026), IMI-JU under grants agreements 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 (ElixirExcelerate). 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).

Conflict of Interest: none declared.

References

Backenroth,D. *et al.* (2016) Using rich data on comorbidities in case-control study design with electronic health record data improves control of confounding in the detection of adverse drug reactions. *PLoS One*, **11**, e0164304.

Bagley,S.C. *et al.* (2016) Constraints on biological mechanism from disease comorbidity using electronic medical records and database of genetic variants. *PLoS Comput. Biol.*, **12**, e1004885–e1004818.

Bonavita,V. and De Simone,R. (2008) Towards a definition of comorbidity in the light of clinical complexity. *Neurol. Sci.*, **29**, 99–102.

Doshi-Velez,F. *et al.* (2014) Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, **133**, e54–e63.

Fortin,M. *et al.* (2005) Multimorbidity is common to family practice: is it commonly researched? *Can. Fam. Phys.*, **51**, 244–245.

Gijsen,R. *et al.* (2001) Causes and consequences of comorbidity: a review. *J. Clin. Epidemiol.*, **54**, 661–674.

Hidalgo,C.A. *et al.* (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.*, **5**, e1000353.

Holmes,A.B. *et al.* (2011) Discovering disease associations by integrating electronic clinical data and medical literature. *PLoS One*, **6**, e21132.

Jakovljević,M. and Ostojić,L. (2013) Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other. *Psychiatr. Danub.*, **25** (Suppl. 1), 18–28.

Jensen,P. *et al.* (2012) Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.*, **13**, 395–405.

Marengoni,A. *et al.* (2011) Aging with multimorbidity: a systematic review of the literature. *Ageing Res. Rev.*, **10**, 430–439.

McCormick,P. (2016) medicalrisk: medical risk and comorbidity tools for ICD9-CM data. R package version 1.2. https://CRAN.R-project.org/package=medicalrisk.

Mezzich,J.E. and Salloum,I.M. (2008) Clinical complexity and person-centered integrative diagnosis. *World Psychiatry*, **7**, 1–2.

Moni,M. *et al.* (2014) comoR: a software for disease comorbidity risk assessment. *J. Clin. Bioinform.*, **4**, 8.

Moni,M.A. *et al.* (2015) CytoCom: a cytoscape app to visualize, query and analyse disease comorbidity networks. *Bioinformatics*, **31**, 969–971.

Piñero,J. *et al.* (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.

Taylor,A.W. *et al.* (2010) Multimorbidity–not just an older person's issue. Results from an Australian biomedical study. *BMC Public Health*, **10**, 718.

Valderas,J.M. *et al.* (2009) Defining comorbidity: implications for understanding health and health services. *Ann. Fam. Med.*, **7**, 357–363.

van den Akker,M. *et al.* (1996) Comorbidity or multimorbidity. *Eur. J. Gen. Pract.*, **2**, 65–70.