Building co-morbidity networks via Bayesian network reconstruction

Guillermo Romero Moreno, Valerio Restocchi, and Jacques Fleuriot

University of Edinburgh, EH8 9AB UK,
Guillermo.RomeroMoreno@ed.ac.uk,

1 Introduction

Patients that simultaneously suffer multiple long-term health conditions pose a problem to current healthcare systems, as these are configured for individual conditions and overlook their interaction [1]. However, more than 65% of the population above 65 year old suffers from two or more simultaneous long–term conditions, and this number is expected to increase as populations are progressively ageing [1]. Hence, the study of co-morbidity has grown in importance in the recent years, as finding the drivers of their onset —environmental or genetic— and the most common combinations may help their prevention or tailor the medical pathways to treat them in a more effective manner. A particularly useful approach to represent co-morbidity data is via networks, where nodes correspond to conditions and edges represent their relations[1]. This representation naturally enables the analysis of the role of single long–term conditions in the context of disease progression, or to detect clusters of conditions that are related [2]. However, determining the degree of co-morbidity and deciding on their relevance is not trivial. Different measures have been used in the literature for this purpose (e.g. relative risks, ϕ-correlations, cosine index, etc), but they tend to be biased against certain conditions, and determining thresholds for pruning the network remains an arbitrary process [2].

Here, we adapt a Bayesian network reconstruction method developed by Young et al. [3] to tackle the problems of co-morbidity network construction. This methodology allows us to infer the network of relevant associations between conditions without any bias or need to choose an arbitrary threshold. Instead, relevant associations are determined via a statistically sound method that accounts for the noise and uncertainty in the data. Furthermore, the Bayesian framework retrieves the whole posterior probability distribution, allowing us to factor in —or analyse independently— different hypotheses of models that generated the data. The framework also provides ample flexibility in its modelling, easily allowing to incorporate extra mechanisms to the basic assumptions of morbidity generation, while also admitting the injection of domain knowledge in the model's priors.

2 Model and methods

The Bayesian network reconstruction method assumes that a dataset of co-morbidity observations X_{ij} can be explained by a generative model that is composed by a network

[1] These networks are also referred to as *Phenotypic Disease Networks*

of 'true associations' A_{ij} and a set of parameters θ. The method samples from the posterior distribution $P(A_{ij},\theta|X_{ij})$, given a data model $P(X_{ij}|A_{ij},\theta)$, a network prior $P(A_{ij}|\theta)$, and parameter priors $P(\theta)$: $P(A_{ij},\theta|X_{ij}) \propto P(X_{ij}|A_{ij},\theta)P(A_{ij}|\theta)P(\theta)$.

As data model, we assume that a single patient may have developed a specific combination of conditions c_{ij} with a probability that depends on the independent probabilities that each of the conditions appears (σ_i and σ_j, respectively), and an excess probability $r(A_{ij})$ of their co-occurrence that depends on the association network A_{ij}: $P(c_{ij}) = \sigma_i\sigma_j[1+r(A_{ij})]$. We assume that associations may be absent ($A_{ij}=0$), weak ($A_{ij}=1$), or strong ($A_{ij}=2$), with

$$r(A_{ij}) = \begin{cases} r_2 \text{ if } A_{ij}=2 \\ r_1 \text{ if } A_{ij}=1 \\ 0 \text{ if } A_{ij}=0 \end{cases}, \tag{1}$$

where $r_1, r_2 \in \mathbb{R}^+$ and $r_2 > r_1$. The probability of observing X_{ij} co-occurrences is then given by the binomial distribution $P(X_{ij}|A_{ij},\theta) = \binom{N}{X_{ij}} P(c_{ij})^{X_{ij}}[1-P(c_{ij})]^{N-X_{ij}}$, where N is the number of patients in the dataset.

3 Results

We have applied the above network reconstruction method to a cross-sectional dataset from the Primary Care Clinical Unit at the University of Aberdeen, with primary care information from 1.7M patients of the Scottish population in 2007. The inferred association network A_{ij} and independent probabilities σ_i can be seen in Fig. 1a, with around 21% of the morbidity pairs having a strong association, and 34% of them having a weak association. Fig. 1b shows how the association strength from our model —computed as $S_{ij} = \sum A_{ij}P(A_{ij})$— compares to the *relative risk* and the *ϕ-correlation* in each of the co-morbidity pairs, showing significant disparities. It is particularly noticeable that some of the co-morbidity pairs for which the network reconstruction method infers a strong association ($S_{ij} \approx 2$) are given very low scores by the *relative risk* ($RR < 5$) and *ϕ-correlation* ($\phi < 0.05$) measures. These generally correspond to conditions with high prevalence —in the case of the *relative risk*— and combinations of high and low prevalence conditions —in the case of *ϕ-correlations*— both defects that are known in the literature [2]. Fig 1c shows the correlation between the prevalence of the different conditions in the dataset and our inferred independent probability σ_i of condition appearance. The strong correlation present corroborates the validity of the inferred parameters σ_i.

References

1. K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.
2. B. Fotouhi, N. Momeni, M. A. Riolo, and D. L. Buckeridge, "Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data," *Applied Network Science*, vol. 3, p. 46, dec 2018.
3. J.-G. Young, G. T. Cantwell, and M. E. J. Newman, "Bayesian inference of network structure from unreliable data," *Journal of Complex Networks*, vol. 8, pp. 1–26, mar 2021.

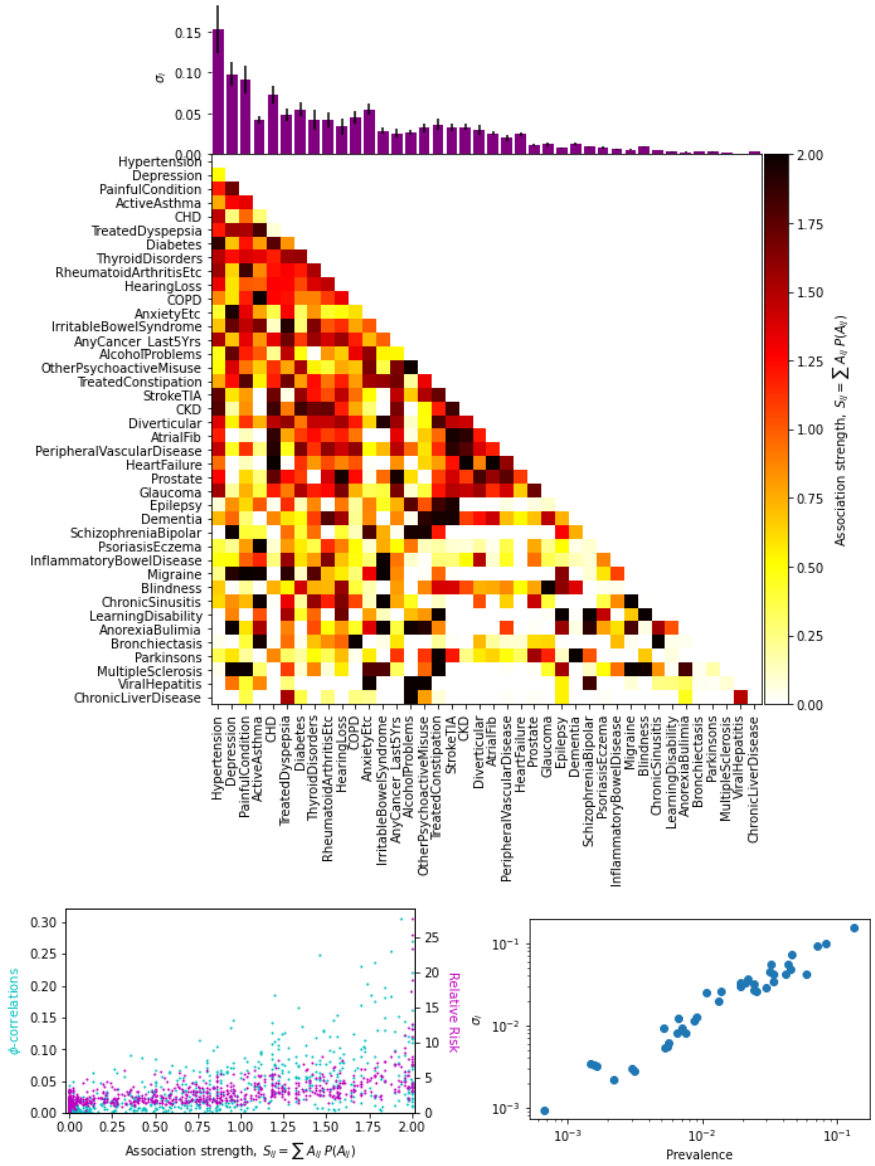


Fig. 1. a) Inferred association strengths between pairs of morbidities (only half of the matrix is represented for clarity, the matrix is symmetric). White, red, and black cells represent no, weak, or strong associations, respectively, with high certainty. Yellow and brown cells respectively represent uncertainty between no or weak association, or weak or strong associations. On the top, the independent probabilities of morbitidity appearance σ_i. b) Correlations between the inferred connections and the *relative risk* and ϕ-*correlation* measures of the same co-morbidity pairs. c) Correlation between morbidity prevalences in the dataset and the inferred probabilities of their appearance σ_i.