JMLR: Workshop and Conference Proceedings vol 52, 310-320, 2016 PGM 2016

Joint Bayesian Modelling of Internal Dependencies and Relevant Multimorbidities of a Heterogeneous Disease

Péter Marx MARXP@MIT.BME.HU
András Millinghoffer MILLI@MIT.BME.HU
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Gabriella Juhász GABRIELLA.JUHASZ@MANCHESTER.AC.UK
MTA-SE Neuropsychopharmacology and Neurochemistry Research Group
Hungarian Academy of Sciences, Semmelweis University
Budapest, Hungary
Neuroscience and Psychiatry Unit, School of Community Based Medicine
Faculty of Medical and Human Sciences
The University of Manchester and Manchester Academic Health Sciences Centre
Manchester, United Kingdom

Péter Antal ANTAL@MIT.BME.HU
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Abstract

A heterogeneous target disease represented by multiple descriptors and disease subtypes frequently has a rich internal dependency structure. The identification of comorbidities and particularly the multimorbidities of such diseases requires very large sample size as relevant comorbidities may form complex interactions. We demonstrate this phenomena by applying a Bayesian probabilistic graphical model on a large-scale medical datasets UK Biobank (117,392 samples), specifically by showing that in this case the posterior landscape of multimorbidities is still flat. As a potential solution, we evaluate a Bayesian method, which provides a hierarchic, multivariate characterization of strongly relevant morbidities and a Bayesian, systems-based score for exploring interactions for a heterogeneous disease. It explores complete sets of strongly relevant comorbidities using full multivariate representation for the internal dependencies within the target disease. We used depression as target, a heterogeneous disease in the UK Biobank dataset. Results are compared against scenarios using a univariate and an independent, multivariate representation of the target medical condition, specifically investigating multitarget interaction posteriors and its approximations.

Keywords: Multitarget; multimorbidity; Bayesian networks; comorbidity.

1. Introduction

The shift of focus from single diseases towards multiple diseases permeates medicine and biomedical research, fueled by shared pathways and overlapping molecular mechanisms of diseases (Menche et al., 2015). Although the network view of disorders is confounded with well-known terminological and nosological problems related to definitions of single diseases and social-environmental factors, the diseasome became an essential crossroad of the molecular and epidemiological levels, further supported by the availability of large health data sets, such as UK Biobank (Sudlow et al., 2015).

The large sample size opens up new possibilities beyond currently prevailing statistical comorbidity methods (Smith et al., 2014).

Utilizing large health datasets and the disease network approach, we focus on the exploration of multivariate patterns of jointly comorbid diseases, i.e. multimorbidity patterns, for a heterogeneous target medical condition. This is still an unresolved problem as the application of multivariate methods with univariate response/outcome is frequently hindered by the availability of an appropriate disease descriptor for the target medical condition. For these heterogeneous targets, multivariate response variables can be used, but current solutions are limited or their scope is different (Van Der Gaag et al., 2006; Klami et al., 2013). As approximations for heterogeneous targets, univariate or constrained multitarget approaches can be applied. However, the use of a single target variable, either by selecting a dominant descriptor best representing the target medical condition or constructing a mega-variable approximating the target medical condition, can have diverse negative consequences on the exploration of relevant factors and their interactions, such as loss of power or artificial, higher-order interaction of factors. The use of multiple separated analyses for each target can have substantial negative effects as well, because it can overestimate the relevance of predictors if intermittent dependencies between the targets are neglected (e.g. by using the others systematically for corrections), but an even more serious effect can be the inability of detecting interactions relevant for the target medical condition, e.g. those interactions, which effects different targets.

In principle, approaches based on probabilistic graphical models (PGMs) are ideal candidates both for the overall exploration of the dependencies among all the undiscriminated predictor-response variables (Madigan et al., 1996; Friedman and Koller, 2003; Schadt et al., 2005; Yeung et al., 2014) and for the exploration of relevant variables (predictors) for a given medical condition (outcome) (Yeung et al., 2005; Antal et al., 2006; Verzilli et al., 2006). Indeed, a method using a special class of PGMs in the Bayesian statistical framework, the Bayesian network-based multilevel analysis of relevance (BN-BMLA) supports the Bayesian exploration of multivariate patterns of relevance for multiply represented targets, but the effect of such group representation of the target medical condition was not investigated yet (Antal et al., 2008).

In this paper, we demonstrate that high-order interactions, specifically for complex heterogeneous target diseases still cannot be identified with high certainty from recent large-scale medical datasets such as UK Biobank. In such problems, explicit modelling of remaining uncertainty is necessary but scalable complexity of model properties of the multimorbid networks can provide a solution. Especially, we present a systematic comparison of three scenarios for the multimorbidity analysis of a heterogeneous disease: (1) if only a single, preferably dominant variable is used to represent the target medical condition in a simplified, frequently used approach, i.e. weaker descriptor variables or disease subtypes and categories are omitted (shortened as univariate target: UT), (2) if a group representation is used for the target medical condition, but interdependences among them are neglected, i.e. they are treated as independent targets (shortened as independent targets: ITs), (3) if a group representation is used for the target medical condition and the interdependencies among them are fully modeled using Bayesian networks (shortened as multitarget: MT). Furthermore, we investigate the presence of higher-order interactions for a heterogeneous disease that cannot be captured in the first two scenarios, termed as "multitarget interactions". These real-world analyses explores the multimorbidities of depression using a UK Biobank data set, which contains 117,392 samples and 110 variables. The target medical condition of depression are represented with 7 variables. The paper is organized as follows. In Section 2 we introduce basic concepts of our approach to rele-

vance analysis. In Section 3, we describe the data and method with detailed settings applied in the evaluation. In Section 4 and 5, we discuss the results and summarize conclusions.

The notation is as follows: target variables (corresponding to depression descriptors and subtypes) are denoted with $\boldsymbol{Y}$ ($\boldsymbol{Y} = \{Y_1, \ldots, Y_m\}$), explanatory variables are denoted with $\boldsymbol{X}$ ($\boldsymbol{X} = \{X_1, \ldots, X_n\}$) and V ($V = \boldsymbol{X} \cup \boldsymbol{Y}$) denotes all the variables. The data set with N samples is denoted by D_N.

2. Approach

The exploration of the multivariate-multivariate dependency relations between explanatory variables and the group representation of the target medical condition using multiple target variables requires the overall, systems-based modelling of the global dependency/independency structure of the variables. PGMs, especially BNs, provide a fundamental tool to represent structural properties of the joint distribution or the underlying causal mechanisms. Furthermore, the Bayesian statistical framework offers many advantages for this analysis, such as prior incorporation and posterior post-processing, thus we focus on a Bayesian, systems-based methodology, which performs Bayesian model averaging over BN structures representing multivariate-multivariate dependency relations between explanatory variables and the target variables. The Bayesian Network-based Bayesian Multilevel Analysis of relevance (BN-BMLA) was proposed to focused on special subgraphs in the neighbourhood of target variables (Antal et al., 2006, 2008). In this work we investigate the effect of using multiple target variables with Bayesian model averaging over their potential dependencies as the group representation of the target medical condition to explore relevant variables. Necessary concepts are as follows (for detailed discussion and references, see (Antal et al., 2008, 2014)).

The probabilistic definition of Markov Blanket and Markov Boundary became fundamental concepts in data analysis, in the feature subset selection problem, to exceed limitations of pairwise and predictive approaches ((Pearl, 1988; Koller and Sahami, 1996; Aliferis et al., 2010)).

Definition 1 *A set of variables $\mathbf{X}' \subseteq \boldsymbol{V}$ is called a* Markov blanket set (MBS) *of Y with respect to the distribution $p(\boldsymbol{V})$, if $(Y \perp\!\!\!\perp V \setminus \mathbf{X}' | \mathbf{X}')_p$, where $\perp\!\!\!\perp$ denotes conditional independence. A minimal Markov blanket is called* Markov boundary *(Pearl, 1988).*

The connections between BNs and Markov Boundaries are provided by the following theorem:

Theorem 1 *For a distribution $p(\boldsymbol{V})$ defined by the Bayesian network (G, θ) the variables $\mathrm{bd}(Y, G)$ form a (not necessarily unique or minimal) Markov blanket of Y, where $\mathrm{bd}(Y, G)$ denotes the set of parents, children and the children's other parents for Y. If $p(\boldsymbol{V})$ is a positive distribution and DAG G is Markov compatible with $p(\boldsymbol{V})$, then $\mathrm{bd}(Y, G)$ is the unique, minimal Markov blanket (called Markov boundary) (Pearl, 1988).*

In stable distributions, $\mathrm{bd}(Y, G)$ also identifies the so called strongly relevant variables (Tsamardinos and Aliferis, 2003). Because of the Bayesian framework at the level of BN parameters, we will treat $\mathrm{bd}(Y, G)$ as Markov Boundary and its members as strongly relevant (for details, see Antal et al., 2008, 2014). In this case, the Markov Boundary set is the union of the external variables, if there are multiple target variables $\boldsymbol{Y}$:

$$\mathrm{MBS}(\mathbf{Y}) = \Big(\bigcup_{Y_i \in \mathbf{Y}} \mathrm{MBS}(Y_i) \Big) \setminus \mathbf{Y}. \tag{1}$$

The induced pairwise, symmetric relation $\mathrm{MBM}(Y, X_i, G)$ with respect to G between Y and X_i is called *Markov Blanket Membership* (MBM) (although in the Bayesian framework the stricter Markov Boundary Membership could be used as well):

$$\mathrm{MBM}(Y, X_i, G) \Leftrightarrow X_i \in \mathrm{bd}(Y, G) \tag{2}$$

However, the linear number of MBM features cannot represent the multivariate aspects of relevant factors for a target medical condition, on the contrary, the exponential number of Markov Boundary sets characterize the joint relevance of factors, but they are intractable computationally, statistically and their visualization is hard. The concept of k-ary Markov Boundary subsets, focusing on k sized sets of *sub-relevant*variables, and the analogous concept of *sup-relevant* sets were introduced to support a constrained multivariate analysis (Antal et al., 2008, 2014).

Definition 2 *For a distribution $p(\boldsymbol{V})$ with Markov Boundary set* mbs*, a set of variables $\boldsymbol{s}$ is called sub-relevant if it is a k-ary Markov Boundary subset (k-subMBS), i.e. $|\boldsymbol{s}| = k$ and $\boldsymbol{s} \subseteq \mathrm{mbs}$. A set of variables $\boldsymbol{s}$ is called sup-relevant if it is a k-ary Markov Boundary superset (k-subMBS), i.e. $|\boldsymbol{s}| = k$ and $\mathrm{mbs} \subseteq \boldsymbol{s}$.*

A k-subMBS and a k-supMBS denotes a necessary and a sufficient set of variables respectively: a k-subMBS set s_{sub} contains some strongly relevant variables, in contrast the complement of a k-supMBS set s^c_{sup} contains variables that are not strongly relevant. The posterior probability of the sub-relevance of a subset s is denoted as follows:

$$\underline{p}(s|D_N) = p(\mathrm{MBS}(Y, G) = s|D_N) + \sum_{s' : s \subset s'} p(\mathrm{MBS}(Y, G) = s'|D_N), \tag{3}$$

where the first term is the exact MBS posterior of s and subsequent terms, behind the summation sign, are the MBS posteriors of each proper superset of s. The posterior probability of sup-relevance $\bar{p}(s|D_N)$ can be defined analogously. Both visualization and post-processing can exploit that subsets of relevant variables form a lattice with operations intersection and union, where the minimum and maximum are the empty and complete sets.

3. Methods

We used the following dataset and settings below together with several approximations.

3.1 Data

We used a subset of the UK Biobank disease data with 117,392 participants who filled out the mental state questionnaire. Disorders were recorded during an interview with trained professionals. Altogether 526 different diseases appeared in the data set. We assigned multilevel ICD-10 codes to these disorders. For this analysis we collected those disorders which fall in one of the following ICD-10 categories: diabetes mellitus (E08-E13), diseases of the circulatory system (I00-I99), diseases of the musculoskeletal system and connective tissue (M00-M99), diseases of the nervous system (G00-G99), mental and behavioural disorders (F01-F99), metabolic disorders (E70-E88) and irritable bowel syndrome. Furthermore, we filtered diseases with frequency below 0.1% prevalence resulting in a subset of 107 diseases together with sex and age (discretized into 3 bins with

thresholds 52 and 61). We defined obesity over 30 kg/m^2 Body Mass Index. Smith et al. (2013) calculated three depression subcategories based on severity and recurrence. Specifically depression occurring once, moderate recurrent depression and severe recurrent depression.

In the first scenario we used the interview-based depression from the UK Biobank as target. In the second scenario we used the following 7 descriptors and subtypes as a group representation of depression: interview-based depression, depression occurring once, recurrent moderate depression, recurrent severe depression, post-natal depression, nervous breakdown and mania/bipolar disorder/manic depression.

3.2 Settings

We applied a Markov Chain Monte Carlo simulation to compute the MBM and MBS posterior probabilities (Antal et al., 2008). To evaluate the three scenarios we performed nine different simulations as follows. Beside the query with multiple targets we ran a simulation for all target variables with including the other targets. Each simulation used the following parameters: 10 parallel runs with 500.000 burn-in steps and additional 2.000.000 simulation steps. Each run started with five chains in the Metropolis-coupled MCMC with swapping and different temperature (Antal et al., 2008). Assuming complete, discrete data set D_N, multinomial sampling and Dirichlet parameter priors ($\mathcal{P}(\Theta|G)$), we used the Cooper-Herskovits (BD_{CH}) prior with 1 as virtual sample size per cell ($vss = 1$) and uniform structure prior ($\mathcal{P}(G)$) (Cooper and Herskovits, 1992; Heckerman et al., 1995). To be able to model higher-order interactions of multimorbidities we allowed a relatively large number of possible parents (max. 6) which motivated the use of BD_{CH} instead of BD_{eu} as the latter may cause anomalies in case of low sample numbers in specific parental configurations (Hullám and Antal, 2013).

3.3 Approximations of MBS Posteriors

For a detailed comparison of the scenarios we applied several approximations. The posterior probability of MBM can be used to estimate the posterior of Markov Blanket Sets in two steps. First, we can approximate the multitarget MBM posteriors (see Eq. 4) and using these values we can calculate the approximated MBS probabilities (Eq. 5). We followed this approach to estimate the multitarget MBS using the multiple independent target BN-BMLA runs and for the univariate target scenario

$$P(\mathrm{MBM}(X_i, \mathbf{Y})|D_N) \approx 1 - \prod_j \left(1 - P(\mathrm{MBM}(X_i, Y_j)|D_N)\right), \tag{4}$$

where $\mathbf{Y}$ is the set of the disorders forming the target medical condition and X_i represents the factors in the data while D_N is the data set with N cases. To simplify the equation we omit the notation of the data D_N in the following equations. It is straightforward to approximate the MBS probability using the approximated MBM probabilities

$$P(\mathrm{MBS}_i(\mathbf{Y})) \approx \prod_{X_i \in \mathrm{MBS}_i} P(\mathrm{MBM}(X_i, \mathbf{Y})) * \prod_{X_i \notin \mathrm{MBS}_i} \left(1 - P(\mathrm{MBM}(X_i, \mathbf{Y}))\right). \tag{5}$$

However, this estimation is biased because using the pairwise MBM descriptors we cannot catch the interactions between the factors. Approximating the multitarget MBS posteriors by utilizing the single-target MBSs we can handle the interactions and redundancy between the variables

$$P(\mathrm{MBS}_k(\mathbf{Y})) \approx \sum_{\{I_i\}_{i=1}^n : \cup \mathrm{MBS}_{I_i}(Y_i) = \mathrm{MBS}_k} \prod_{i=1}^{n} P(\mathrm{MBS}_{I_i}(Y_i)). \tag{6}$$

We estimate the posterior of a given MBS (MBS_k) by searching for a unitarget MBS for each target variable which union equals with MBS_k. We sum up for all possible combinations resulting in MBS_k the product of the posteriors of the unitarget MB sets.

4. Results and Discussion

To demonstrate the advantages of multitarget BN-BMLA method, first we compare our results for the three scenarios: univariate target (UT), multiple independent targets (ITs) and the multitarget case (MT). Next, we present the possibilities using constrained Markov Blanket sets (k-subMBSs) to search for strongly relevant factors. Finally, we describe the pairwise and higher order interactions between comorbid disorders and discuss its biomedical relevance.

4.1 Posteriors for comorbidities and multimorbid sets

Markov Blanket Membership can represent a pairwise connection between the target group and the other disorders. By using the approximation in Eq. 4 we compared the unitarget and the independent targets cases to the multitarget MBM posteriors (see Fig.1A). The multitarget approximation follows accurately the multitarget MBM posteriors with only minor differences present (e.g. heart attack/myocardial infarction). In case of the single target MBM many differences can be detected (e.g. type 2 diabetes, osteopenia, osteoporosis). The connection between depression and these disorders is mostly underestimated compared to the multitarget case as expected, since the other members of the disease group can have different comorbidities. In summary, the error of the univariate and multiple independent targets estimations are as follows: the UT method provides precise estimate for 68% of the variables (Δ is less than 0.05 for factors with higher than 0.05 MBM posteriors). The ITs estimations are better in all of these cases and less than 0.05 for the relevant variable except for osteopenia.

We also investigated the structural interaction between relevant variables, so we examined the Markov Blanket sets for the three scenarios. In the first to cases (UT, ITs) we approximated the multitarget MBS with MBM (Eq. 5) and MBS (Eq. 6). Besides, we used the multitarget MBM to approximate the multitarget MBS as well. Figure 1B shows the ordered multitarget MBSs and the approximations. The highest multitarget MBS posterior is 0.262 followed by steep drop. No MBS have high posterior probability (> 0.5) and a few sets have medium posteriors between 0.05 and 0.5. It can be clearly seen, that MBM estimates (triangles) performed worse than MBS approximations. This implies that there are many interaction between variables which MBM cannot handle. However, MBS approximations gave better results, the estimations of the multitarget MBS posteriors were still not accurate.

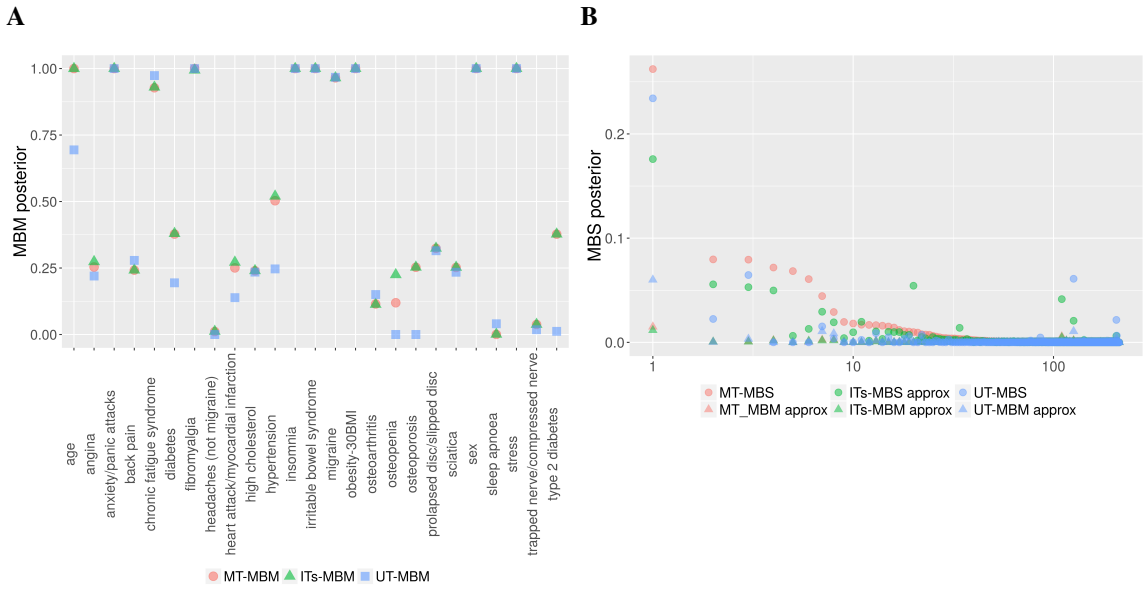

Figure 1: Approximations of MBM and MBS showing the three different scenarios. **A** shows the difference between the MT MBM the ITs MBM and UT MBM posteriors. Disorders with higher than 0.01 posterior in either case are shown. **B** presents the multitarget MBS probabilities and its different estimations. Red shows the calculated multitarget MBS and MBM by the BN-BMLA approach. Green presents the approximation of the multitarget MBS based on multiple unitarget MBMs and MBSs. Blue shows the unitarget MBM-based approximation together with the unitarget MBS posterior of the multitarget MB sets.

To summarize, these results for the multitarget MBS and its poor approximations indicates that the performance of the three scenarios should be investigated at intermediate levels bridging Markov Blanket Memberships and Markov Blanket Sets.

4.2 Limits of data in multivariate relevance analysis

The k-subMBS concept in the Bayesian framework provides a unique approach to understand the power of the data to explore strongly relevant comorbid disorder subsets. Figure 2B shows the multitarget MBS posteriors with several k-subMBSs. 1-subMBS (red dashed line) which equals the MBM posteriors predicts the number of comorbid conditions for the target group but MBMs cannot include the interactions between the factors. As expected, in case of $k = 1, 2$ a couple of variables and variable pairs have a high posterior followed by a steep fall in the posteriors probabilities. The high number of 3- and 5-subMBSs with high posteriors implies that the group representation of depression has multiple strongly relevant comorbidities which corresponds to our current knowledge (Smith et al., 2014).

The sub-MBS and sup-MBS posteriors characterize the power of the data (see Fig. 2A). Using only depression as a target we have the highest posterior probability at the intersection point (about 0.5) of the sub and sup curves. The multitarget curves have a moderate (0.375) intersection posterior with higher set size than depression. The multiple independent targets MBS approximation have

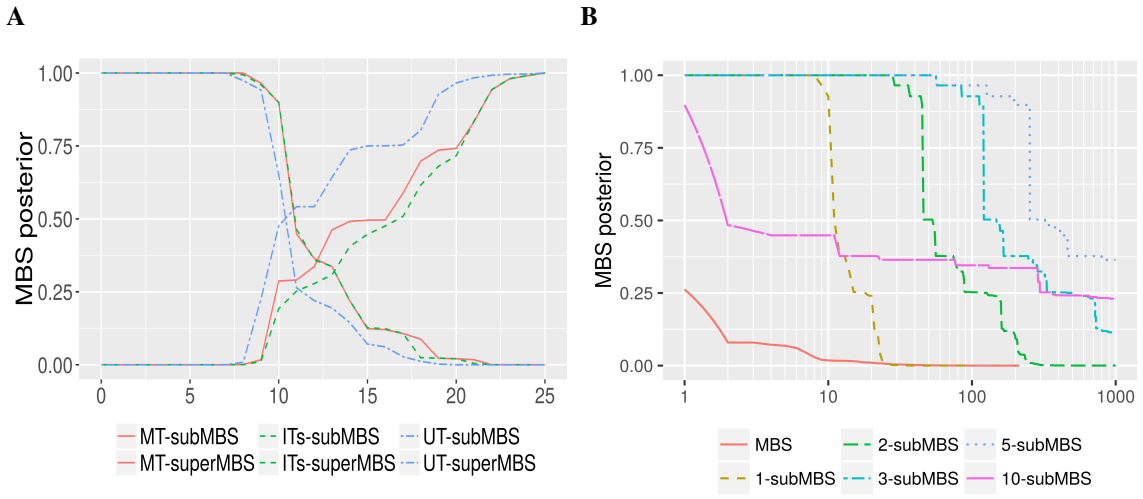

Figure 2: Sub- and supMBS curves. **A** shows the curves for the posteriors for the sub- and sup relevancies in the three scenarios (MT, ITs, UT). **B** presents the curves of the k-MBS posteriors together with MBS posteriors in the MT scenario.

the highest set size and the lowest posterior. As expected the univariate case can filter out irrelevant factors with higher confidence and can form a relevant subset with higher belief. On the other hand the multitarget case includes more factors with high posteriors in the MBS.

4.3 Pairwise and higher-order interactions in morbidities

To investigate the presence of interactions between the factors we examined an interaction-redundancy score (IRS see equation 7) based on posterior decomposition (Antal et al., 2008)

$$IRS(X_1; X_2) = \log \frac{p(\{X_1, X_2\} \subseteq \mathrm{MBS}(\mathbf{Y}))}{p(\mathrm{MBM}(\mathbf{Y}, X_1, G))p(\mathrm{MBM}(\mathbf{Y}, X_2, G))}. \tag{7}$$

Figure 3 and Table 1 shows these interactions together with the redundancies between the disorders. In the case of UT method, 11 interactions gave significantly different estimates (the ratio of IRS scores outside 0.95-1.05 interval) whereas the ITs scenario gives an even worse approximation for 38% in that sense. An interesting interaction which can be detected in the multitarget case but not present using the individual targets is the heart attack/myocardial infarction and osteoarthritis.

5. Conclusion

Complex, heterogeneous diseases pose a critical challenge in multiple domains, such as in network medicine to find comorbidities and in genetic association research to find relevant variants, genes and pathways. This challenge is further complicated by the necessity of finding higher-order interactions, e.g. multimorbidities, whose pattern of presence and absence synergistically influences various aspects of the heterogeneous medical condition. Indeed, it is currently unknown what is the presence, frequency and significance of the introduced multitarget interactions, whose defining

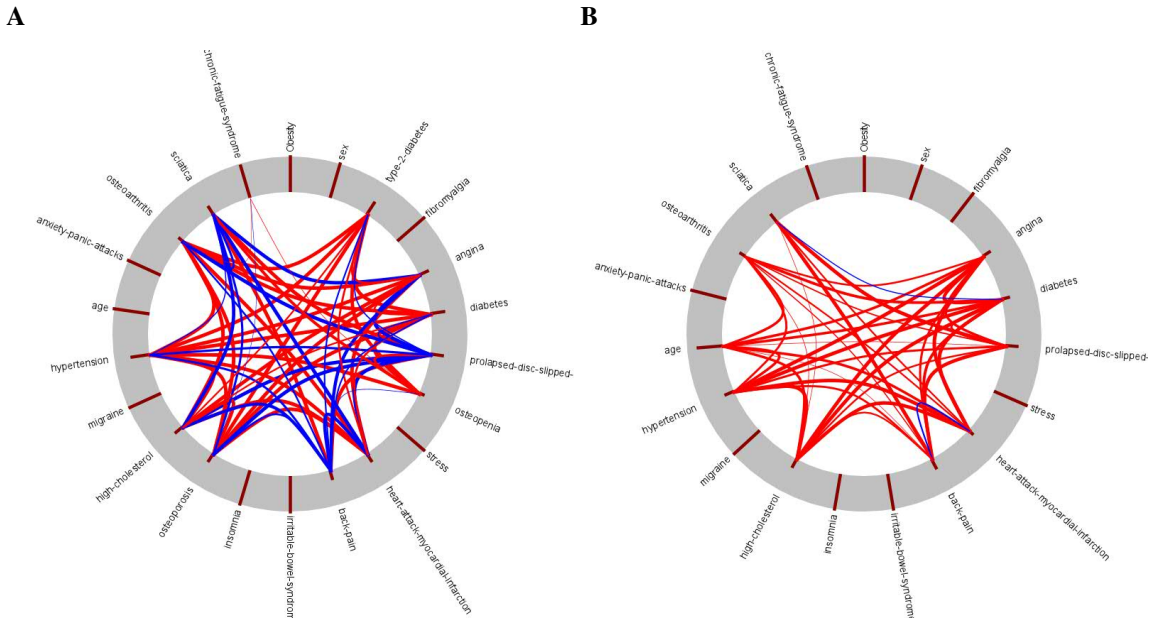

Figure 3: The interaction diagrams of the multitarget MBS (**A**) and the unitarget depression excluding the other target variables (**B**). Red line shows positive interactions while blue lines represents redundancy between the connected variables. The length of the brown lines in the grey circle shows the MBM posterior of the given variable.

characteristics is their systems-based relevance for a heterogeneous disease, consequently, they cannot be identified by simplified univariate target analysis or based on separated analyses of multiple, independently treated targets.

We confronted these challenges of exploring relevant factors for a heterogeneous disease through the problem of exploring multimorbidities for depression. Results are compared using three scenarios with (1) the univariate, (2) the independent-multivariate and (3) the group representation for the target heterogeneous disease. These investigations confirmed the limitations of the first two scenarios: brittle and constrained disease definitions lead to decreased statistical power and to asymptotic limitations as well. The univariate approach poorly approximates the posteriors of strong relevance for 32% comorbidities and it falsely quantifies 25% of the significant pairwise interacting comorbidities. The performance of the second scenario, which assumes independence of the 7 variables representing subtypes and aspects of depression, is improved, but it still cannot cope with the estimation for the exploration of interacting comorbidities. From the point of view of multitarget interactions, this result shows that such interactions are not just existing, but frequent, which indicates that systems-based methods are needed for the detailed exploration of the multivariate-multivariate relevance relations.

The above results confirm that large-scale datasets are still limited for non-ambiguous identification of high-order interactions of complex heterogeneous target diseases. As a solution, we evaluated the applicability of Bayesian probabilistic graphical models, where PGMs offer a rich toolset for representing relevance and interaction with varying complexity and the Bayesian statistical approach provides an explicit and processable representation for the confirmation of the data

Table 1: Observed posteriors of 2-MBSs (P), their MBM-based approximations ($\hat{P}$) and the interaction score (IRS) are shown for the three scenarios (MT, UT, ITs) and their respective ratios. Pairwise interactions with 2-MBS posteriors above 0.05, IRS_{MT} outside the 0.95-1.05 and the $\frac{IRS_{MT}}{IRS_{ITs}}$ is above 1.05 are shown.

	MT			ITs			UT					
Joint relevant comorbidities for depression	P	$\hat{P}$	IRS	P	$\hat{P}$	IRS	P	$\hat{P}$	IRS	$\frac{IRS_{MT}}{IRS_{ITs}}$	$\frac{IRS_{MT}}{IRS_{UT}}$	$\frac{IRS_{ITs}}{IRS_{UT}}$
osteoporosis osteoarthritis	0.091	0.029	3.151	0.024	0.028	0.844	0.001	0.006	0.234	3.732	13.477	3.611
heart attack osteoarthritis	0.113	0.029	3.927	0.044	0.031	1.455	0.023	0.006	3.684	2.698	1.066	0.395
angina osteoarthritis	0.114	0.029	3.905	0.045	0.031	1.452	0.023	0.006	3.693	2.691	1.057	0.393
diabetes osteopenia	0.119	0.045	2.648	0.085	0.085	0.997	0.000	0.000	0.185	2.656	14.332	5.397
type 2 diabetes osteopenia	0.119	0.045	2.646	0.085	0.085	0.999	0.000	0.000	0.363	2.647	7.287	2.752
osteoarthritis high cholesterol	0.114	0.028	4.133	0.042	0.027	1.576	0.023	0.006	3.905	2.622	1.058	0.404
type 2 diabetes osteoarthritis	0.091	0.043	2.106	0.035	0.042	0.830	0.001	0.009	0.155	2.538	13.577	5.351
diabetes osteoarthritis	0.091	0.043	2.108	0.036	0.043	0.831	0.001	0.009	0.158	2.536	13.359	5.268

through posteriors. Results show that the hypothesized multitarget interactions exist and they can be explored in the diseasome using large-scale health data sets using Bayesian model averaging over the dependency structures of the group representatives and the potential factors.

Acknowledgments

This research has been conducted using the UK Biobank Resource. This work has been supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (P. Antal), National Development Agency (KTIA_NAP_13-1-2013-0001), OTKA 112915, Hungarian Brain Research Program - Grant No. KTIA_13_NAP-A-II/14, by the Hungarian Academy of Sciences and the Hungarian Brain Research Program - Grant No. KTIA_NAP_13-2-2015-0001 (MTA-SE-NAP B Genetic Brain Imaging Migraine Research Group), by the MTA-SE Neuropsychopharmacology and Neurochemistry Research Group, Hungarian Academy of Sciences, Semmelweis University, and by the Manchester Academic Health Science Centre, University of Manchester.

References

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification. *Journal of Machine Learning Research*, 11:171–284, 2010.

P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex Bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

P. Antal, A. Millinghoffer, G. Hullám, C. Szalai, and A. Falus. A Bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 4:74–89, 2008.

P. Antal, A. Millinghoffer, G. Hullám, G. Hajós, P. Sárközy, C. Szalai, A. Gézsi, and A. Falus. *Bayesian, systems-based, multilevel analysis of biomarkers of complex phenotypes: from interpretation to decisions*. Oxford University Press, 2014.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.

D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

G. Hullám and P. Antal. The effect of parameter priors on bayesian relevance and effect size measures. *Periodica Polytechnica. Electrical Engineering and Computer Science*, 57(2):35, 2013.

A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *The Journal of Machine Learning Research*, 14(1):965–1003, 2013.

D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25: 2493–2520, 1996.

J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.

E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. a. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach

to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, 2005.

D. J. Smith, B. I. Nicholl, B. Cullen, D. Martin, Z. Ul-Haq, J. Evans, J. M. Gill, B. Roberts, J. Gallacher, D. Mackay, et al. Prevalence and characteristics of probable major depression and bipolar disorder within uk biobank: cross-sectional study of 172,751 participants. *PLoS One*, 8(11): e75362, 2013.

D. J. Smith, H. Court, G. McLean, D. Martin, J. L. Martin, B. Guthrie, J. Gunn, and S. W. Mercer. Depression and Multimorbidity. *The Journal of Clinical Psychiatry*, (November):1202–1208, 2014.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):1–10, 2015.

I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy, filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.

L. C. Van Der Gaag, P. R. De Waal, et al. Multi-dimensional bayesian network classifiers. In *Probabilistic graphical models*, pages 107–114. Citeseer, 2006.

C. J. Verzilli, N. Stallard, and J. C. Whittaker. Bayesian graphical models for genomewide association studies. *American journal of human genetics*, 79(1):100–12, jul 2006.

K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10):2394–2402, 2005.

K. Y. Yeung, C. Fraley, W. C. Young, R. Bumgarner, and A. E. Raftery. Bayesian model averaging methods and r package for gene network construction. In *Big Data Analytic Technology For Bioinformatics and Health Informatics (KDDBHI), workshop at the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.