Exploring Disease Interactions Using Markov Networks

Jan Van Haaren and **Jesse Davis**
KU Leuven
{jan.vanhaaren,jesse.davis}@cs.kuleuven.be

Martijn Lappenschaar and **Arjen Hommersom**
Radboud University Nijmegen
{mlappens, arjenh}@cs.ru.nl

Abstract

Network medicine is an emerging paradigm for studying the co-occurrence between diseases. While diseases are often interlinked through complex patterns, most of the existing work in this area has focused on studying pairwise relationships between diseases. In this paper, we use a state-of-the-art Markov network learning method to learn interactions between musculoskeletal disorders and cardiovascular diseases and compare this to pairwise approaches. Our experimental results confirm that the sophisticated structure learner produces more accurate models, which can help reveal interesting patterns in the co-occurrence of diseases.

Introduction

Given the complexity of the human body, a disease is rarely a consequence of an abnormality in a single part of the system, e.g., a single gene. Therefore, *network medicine* has been proposed as a systematic tool to study human diseases and identify molecular and genetic pathways (Barabási, Gulbahce, and Loscalzo 2011). One type of disease networks are *phenotypic disease networks*, where one can link disease pairs on the basis of observed comorbidity.

Resulting networks typically illustrate the impact and complexity of disease associations, and in which ways diseases can be grouped and clustered. It has been shown that such networks can capture disease progression, i.e., patients tend to develop diseases in the network vicinity of diseases they already had (Hidalgo et al. 2009). Moreover, it has been shown that phenotypic similarity correlates positively with the molecular signatures of two linked diseases (van Driel et al. 2006). Therefore, advances in this direction are essential for identifying novel biological pathways.

Commonly used measures in medicine to describe associations are the relative risk (RR) and the ϕ-correlation coefficient (Pearson's correlation coefficient for binary variables). These measures have been used to investigate cancer metastasis patterns in a network-based manner, where edges in a constructed network were added based on high enough strength of RR or ϕ-correlation (Chen et al. 2009). The resulting network represents the dependencies between the nodes involved.

In more recent years, there has been a shift from simple comorbidities to multimorbidity, i.e., the co-occurrence of two but often more than two multiple chronic or acute diseases and medical conditions within a person. The introduction of this term indicates a shift of interest from a given index condition to individuals who suffer multiple disorders (e.g., (Barnett et al. 2012)). While the previously mentioned measures give some insight into this problem, they are geared to pairwise comparisons between diseases and are unsuitable for discovering more complex patterns.

In this paper, we propose to learn more complex patterns in disease interactions using Markov network structure learning, where the conditional independencies are learned from clinical data. Markov networks are often represented as a log-linear model, which means that structure learning can be posed as a feature induction problem.

Traditionally, structure learning is addressed through standard search based techniques (e.g., (Della Pietra, Della Pietra, and Lafferty 1997; Davis and Domingos 2010)). Algorithms that follow this strategy use the current feature set to construct a set of candidate features. After evaluating each feature, the highest scoring feature is added to the model.

Alternatively, a set of local models can be learned and then combined into a global model (e.g., (Ravikumar, Wainwright, and Lafferty 2010; Lowd and Davis 2010)). Algorithms that follow this strategy consider each variable in turn and build a model to predict this variable given the remaining variables. Each predictive model is then transformed into a set of features, each of which is included in the final, global model. One of the successful approaches of this strategy employed L1 logistic regression as the local model (Ravikumar, Wainwright, and Lafferty 2010).

GSSL, a more recent approach, combines aspects of both procedures (Van Haaren and Davis 2012). In the feature generation phase, the algorithm proceeds in a data-driven, bottom-up fashion to explore the space of candidate features. In the feature selection phase, the algorithm performs weight learning only once to select the best features, and it follows the philosophy of local model based approaches that try to minimize the computational expense of weight learning.

In this paper, we focus on learning Markov network structures of rheumatic and cardiovascular related disorders. There is accumulating evidence for an increased cardiovascular burden in inflammatory arthritis (Nielen et al. 2012).

We believe that computational tools can help to better understand such relations. We learned the structure using the RR, ϕ-correlation, L1 approach, and GSSL algorithm. We compare the results in terms of pseudo-log-likehood and we study some of the learned complex features in more detail.

Preliminaries

Traditional pairwise comparisons

Let N_i be the number of patients with disease D_i, N_j the number of patients with disease D_j, N_{ij} the number of patients with both diseases D_i and D_j, and N the total number of patients. The *relative risk* of observing a pair of diseases D_i and D_j affecting the same patient is given by:

$$RR_{ij} = \frac{N_{ij}N}{N_iN_j} = \frac{P(d_i, d_j)}{P(d_i)P(d_j)}. \quad (1)$$

The *ϕ-correlation coefficient* is given by:

$$\phi_{ij} = \frac{N_{ij}N - N_iN_j}{\sqrt{N_iN_j(N - N_i)(N - N_j)}}. \quad (2)$$

Markov networks

Representation A Markov network is a model for compactly representing the joint distribution of a set of variables $X = (X_1, X_2, \ldots, X_v)$ (Della Pietra, Della Pietra, and Lafferty 1997). It consists of an undirected graph G and a set of potential functions ϕ_k. The graph has a node for each variable, and the model has a potential function for each clique in the graph. The joint distribution represented is:

$$P(X = x) = \frac{1}{Z}\prod_k \phi_k(x_{\{k\}}) \quad (3)$$

where $x_{\{k\}}$ is the state of the kth clique (i.e., the state of the variables in that clique), and Z is a normalization constant. Markov networks are often conveniently represented as log-linear models, with each clique potential replaced by an exponentiated weighted sum of features of the state:

$$P(X = x) = \frac{1}{Z}\exp\left(\sum_j w_j f_j(x)\right). \quad (4)$$

A feature $f_j(x)$ may be any real-valued function of the state. For discrete data, a feature typically is a conjunction of tests of the form $X_i = x_i$, where X_i is a variable and x_i is a value of that variable. We say that a feature matches an example if it is true for that example.

Weight Learning The weight learning task is, given a set of features and data, to learn the weight associated with each feature. Weights are assigned that optimize a given objective function. A natural choice is to learn weights that maximize the log-likelihood of the training data, which is a convex function of the weights in a Markov network, and thus the weights can be learned through an iterative optimization technique. However, this optimization typically requires evaluating the log-likelihood and its gradient in each iteration. This is typically intractable to compute exactly due to the partition function.

Consequently, many existing approaches optimize the pseudo-log-likelihood (Besag 1975) (PLL), which can also be learned via convex optimization and is much more efficient to compute. The PLL is defined as:

$$\log P_w^{\bullet}(X = x) = \sum_{j=1}^{V}\sum_{i=1}^{|D|}\log P_w(X_{i,j} = x_{i,j}|MB_x(X_{i,j})) \quad (5)$$

where V is the number of variables, $|D|$ is the number of examples, $x_{i,j}$ is the value of the jth variable of the ith example, and $MB_x(X_{i,j})$ is the state of $X_{i,j}$'s Markov blanket in the data.

Structure Learning The structure learning task is, given data, to learn both the features and their weights. Local model based structure learning algorithms try to discover the Markov blanket of each variable X_i by building a model to predict its value given the remaining variables. Finally, all features are added to the model and their weights are learned globally using any standard weight learning algorithm.

Ravikumar et al.'s (2010) algorithm employs L1 logistic regression as the local model. In the limit of infinite data, consistency is guaranteed such that X_i is in X_j's Markov blanket iff X_j is in X_i's Markov blanket. In practice, this is often not the case and there are two methods to decide which edges to include in the network. One includes an edge if either X_i is in X_j's Markov blanket or X_j is in X_i's Markov blanket. The other includes an edge if both X_i is in X_j's Markov blanket and X_j is in X_i's Markov blanket.

A weakness of this algorithm is that it only constructs pairwise features. GSSL learns features that include more than two variables. GSSL involves two phases. Given a set of training examples, the first step generates a large number of non-unique features. It begins by converting the training set into the initial feature set. The procedure then repeatedly selects a feature at random and generalizes it by dropping a random number of variable-value tests in the feature. The procedure stops after generating a user-specified maximum number of non-unique features. The second step receives the generated features and a lower bound, $thres$, on the number of times each feature was proposed during feature generation. The algorithm loops through the feature set and discards all features that were generated fewer than $thres$ and then learns the weights for each feature using L1 optimization. This reduces the number of features in the model by forcing many weights to be zero.

Experiments

In this section, we compare using RR, ϕ-correlation, Ravikumar et al.'s L1 approach, and GSSL for Markov network structure learning on a real-world clinical database. The evaluation consists of two parts. In this section, we investigate the predictive performance of the learned models using several criteria. In the following section, we analyze

the learned features to see if any interesting correlations among diseases were discovered.

We will now describe the data and our experimental methodology, and present our experimental results.

Data

The data used for analysis were obtained from the Netherlands Information Network of General Practice (LINH). All Dutch inhabitants are obligated to register with a general practice, and the LINH registry contains information of routinely recorded data about all patients from approximately 90 general practices. Our analysis includes longitudinal data from 78,436 patients, aged over 35 years. The definition of a "chronic disorder" given by (O'Halloran, Miller, and Britt 2004), which is based on the international classification of primary care (ICPC) codes, was used to determine whether patients had a cardiovascular disease (ICPC category K) or a musculoskeletal disorder (ICPC category L).

For our experiments, two datasets were constructed. The first contains information about all 78,436 patients, whereas the second only contains information about the 19,322 patients aged 65 or above. Hence, the latter dataset is a subset of the former. Both datasets contain 38 binary variables, of which 28 are related to cardiovascular diseases, and 14 to musculoskeletal diseases.[1] Each variable indicates whether a certain disease was observed for a particular patient.

Methodology

To evaluate the predictive performance of the learned models, we performed ten-fold cross-validation. We used six folds for learning, three folds for parameter tuning, and one fold for testing. We iterated over the training folds such that each fold served six times for learning, three times for parameter tuning, and one time for testing.

We used publicly available implementations for all methods. We learned weights that optimize the PLL for computational tractability. To allow for a fair comparison, we applied the same weight learning procedure for all approaches. Additionally, we also tried using the RR and ϕ-correlation coefficients as feature weights in the models produced by these two approaches. We tried standard deviations of 0.1, 0.5, 1, 2, and 5 for the Gaussian weight prior, and combined these with L1 norm weights of 1, 5, 10, 25, 50, 100, and 200, leading to 35 different configurations. We selected the model with the best tune set PLL.

We consider the following feature learning approaches:

- **GSSL best:** We selected the 50, 100, 250, 500, 1000, 2500, 5000, 10000, and 20000 most frequently generated features by GSSL.
- **GSSL simple:** To avoid overfitting on the training data, we selected the simplest model, in terms of average feature length and number of features, whose tune set PLL was within 0.01 of the best model.
- **L1:** We used Ravikumar et al.'s approach to generate the features. We tried both edge selection strategies.
- **RR:** We selected the 10, 25, 50, 100, 200, 300, 400, 500, 600, and 700 best ranked features.
- **Phi:** We selected the 10, 25, 50, 100, 200, 300, 400, 500, 600, and 700 best ranked features.
- **RRCC:** We applied the same feature selection strategy as RR but used the correlation coefficients as weights.
- **PhiCC:** We applied the same feature selection strategy as Phi but used the correlation coefficients as weights.

In this evaluation, we look at two quality measures for the learned models. First, we look at the test set pseudo-log-likelihood. We report pseudo-log-likelihood because, as alluded to in the discussion on weight learning, computing the likelihood is generally intractable in Markov networks. In essence, pseudo-log-likelihood measures how good a model's predictions would be with perfect information about an individual's disease state.

Second, we explore the interaction between cardiovascular and musculoskeletal disorders. This aims to measure, for example, what is the risk on a cardiovascular disease such as heart failure, given that the patient has a disorder in the other group such as osteoporosis or osteoarthritis. We consider six scenarios that are relevant from a medical point of view:

1. Predict the marginal probability for *heart failure*, *cerebrovascular accident*, and *transient ischaemic attack*, given a patient's state for all other disorders.
2. Predict the marginal probability for *osteoarthritis NOS*, *hip osteoarthritis*, *knee osteoarthritis*, and *osteoporosis*, given a patient's state for all other disorders.
3. Predict the marginal probability for each *musculoskeletal* disorder, given a patient's state for one *cardiovascular* disorder.
4. Predict the marginal probability for each *cardiovascular* disorder, given a patient's state for one *musculoskeletal* disorder.
5. Predict the marginal probability for each *musculoskeletal* disorder, given a patient's state for a pair of *cardiovascular* disorders.
6. Predict the marginal probability for each *cardiovascular* disorder, given a patient's state for a pair of *musculoskeletal* disorders.

The first two experiments are used to study the predictive performance of comorbidity on cardiovascular and musculoskeletal diseases. In the remaining experiments, we study how well the networks perform to predict a disease from one group given one or two diseases in the other group. Such information is useful as it elucidates the impact of one disease group on the other. For example, the burden of cardiovascular diseases in rheumatoid arthritis patients is of medical interest (Nielen et al. 2012).

For these results, we report the average per example conditional marginal log-likelihood (CMLL) of the queries on

[1] In principle all the learning algorithms can work with non-binary variables. It is also always possible to convert any non-binary variable into a binary one.

the test data. The CMLL is defined as:

$$\text{CMLL}(X = x) = \sum_{i \in Q} \log P(X_i = x_i | E) \qquad (6)$$

where Q is the query set and E is the evidence set. These are defined as described above. To compute these probabilities, we use the Gibbs sampler from the Libra package.[2] We used 10 Markov chain Monte Carlo chains of 10,000 samples with a burn-in of 1,000 samples.

Results

As shown in Table 1, both GSSL best and GSSL simple outperform the other algorithms in terms of test set pseudo-log-likelihood. Since RRCC and PhiCC are nowhere near the best score, we omit them from further analysis.

Table 2 presents the conditional marginal log-likelihoods for all six scenarios and all five approaches on the full dataset. GSSL's models yield the highest predictive accuracy in scenarios 1, 2, and 4. However, it does a bit worse than its competitors in scenarios 5 and 6, and to a lesser extent in scenario 3. Scenarios 5 and 6 pose a significantly harder inference problem for more complex models, which probably explains this. Also, pairwise features might be more important in these scenarios, whereas GSSL focuses on more complex features, as shown in Table 4. As a result, the GSSL models are probably slightly overfitted on the training data despite using an L1 penalty in the weight learning phase.

Table 3 presents the conditional marginal log-likelihoods for all six scenarios and all five approaches on the 65+ dataset. GSSL's models yield the highest predictive accuracy in scenarios 1, 2, and 6, and do only slightly worse than the other approaches in the other three scenarios. This is not surprising as GSSL learns much simpler models for this dataset, resulting in an average feature length of 2.62 for the best model and 2.27 for the simple model. As a result, GSSL's models are very similar to those produced by the other approaches.

Interpretation of the features

From a medical point of view, the question is how to interpret these complex features. To obtain some insight into this question, we investigated the diseases which often occur in these (complex) features.

Diseases in complex features

Focusing on features with three diseases, we find that the following diseases occur most often in features: hypertension uncomplicated (13.7% of the features), musculoskeletal disorders NOS (12.8%), lumbar hernia (12.2%), angina pectoris (11%), and hypertension complicated (11%). We hypothesize that these features are often an indication of (confounding) age and life-style factors. Hypertension, lumbar hernia, angina pectoris are associated with factors such as obesity and smoking. The factor musculoskeletal disorders NOS includes various less prevalent disorders, which makes it difficult to interpret. Some of these may be explained by increasing age and others by a sedentary life-style. This factor, however, requires a more in-depth investigation that is beyond the scope of this paper.

[2] http://libra.cs.uoregon.edu

Mediation effect of hypertension

To study complex features in some more detail, we investigated how the RR is altered in the presence/absence of the most common disease in these features, uncomplicated hypertension (H). For this, we computed the so-called *conditional* RR:

$$RR_{ij}^{k} = \frac{P(d_i, d_j \mid H = k)}{P(d_i \mid H = k) P(d_j \mid H = k)} \qquad (7)$$

for all possible disease pairs d_i and d_j, for the cases that (uncomplicated) hypertension is present ($k = \textit{true}$) or absent ($k = \textit{false}$). Then for each pair of diseases, we say that d_i and d_j are *mediated* by hypertension if the RR_{ij}^{true} is significantly different (higher or lower) from RR_{ij}^{false} ($p = 0.01$). We find that in 75% of the cases, pairs of diseases are mediated by hypertension: in 70% of the cases the RR significantly increases in the presence of hypertension and in 5% of the cases the RR significantly decreases in the presence of hypertension. If we examine the first 10.000 features that were generated by GSSL, in 95.0% of those features that contain uncomplicated hypertension, a pair of diseases is indeed mediated by hypertension. Likewise, for the patients of 65 years and older, 60% of the possible disease pairs show a significantly higher RR (and 2% a significant lower RR) when hypertension is present. For these patients, in 89.9% of the features that contain uncomplicated hypertension, a pair of diseases is indeed mediated by hypertension. Hence, the complex features that were generated can, for a large part, be explained by the fact that hypertension changes the relative risk of two diseases. As a well-known life-style factor, it is likely that hypertension indicates some influence from life-style in the relationship between diseases. In the older patients, differences between patients are less explained by the presence of hypertension, which also is a plausible result, as hypertension becomes more prevalent with age.

Graphical comparison

If we zoom into the learned models with respect to a subset of variables, i.e., inflammatory arthritis, osteoarthritis of the knee and hip, heart failure, heart murmur, heart tumor, myocardial infarction, and pulmonary heart disease, we obtain the models shown in Figure 1. One can see that, although it performs equally well, the model obtained by the GSSL method is less dense than the one obtained by the L1 model (which is almost maximal). Apparently, data explained by pairwise dependencies, can also be captured in more complex features, yielding a more comprehensive graph.

For example, in the GSSL model, the diseases within the cardiovascular cluster are only connected through heart failure. From a medical point of view this makes sense; heart tumor, myocardial infarction, heart murmur (if related to valvular heart disease), and pulmonary heart disease, will eventually all lead to heart failure (McMurray et al. 2012). Direct relations are indeed less known from the literature.

	GSSL best	GSSL simple	L1	RR	Phi	RRCC	PhiCC
Full dataset	-57.637	**-55.598**	-59.076	-60.743	-60.555	-784.583	-261.766
65+ dataset	-70.058	**-67.545**	-70.291	-70.481	-70.461	-964.741	-263.233

Table 1: GSSL outperforms L1, RR, and ϕ-correlation in terms of test set pseudo-log-likelihood. The best result is in bold.

	GSSL best	GSSL simple	L1	RR	ϕ
Scenario 1	-0.773	**-0.771**	-0.784	-0.784	-0.784
Scenario 2	-0.773	**-0.771**	-0.784	-0.784	-0.784
Scenario 3	-2.745	-2.711	-2.702	-2.700	**-2.683**
Scenario 4	-11.103	**-1.618**	-1.639	-1.636	-1.659
Scenario 5	-4.357	-3.512	-3.416	**-3.415**	-3.432
Scenario 6	-47.225	-3.608	-3.268	-3.277	**-3.246**

Table 2: Conditional marginal log-likelihood in all scenarios for the full dataset. The best result for each scenario is in bold.

	GSSL best	GSSL simple	L1	RR	ϕ
Scenario 1	**-1.110**	-1.111	-1.117	-1.116	-1.117
Scenario 2	**-1.110**	-1.111	-1.117	-1.116	-1.117
Scenario 3	-2.199	-2.201	-2.165	-2.165	**-2.150**
Scenario 4	-2.439	-2.426	-2.425	**-2.422**	-2.428
Scenario 5	-3.445	-3.427	**-3.420**	-3.425	-3.426
Scenario 6	-3.627	**-3.606**	**-3.606**	-3.611	**-3.606**

Table 3: Conditional marginal log-likelihood in all scenarios for the 65+ dataset. The best result for each scenario is in bold.

	Length 2	Length 3	Length 4	Length 5	Length 6	Length 7	Length 8	Length 9	Average
GSSL best	614	2641	1463	120	14	12	3	1	3.20
GSSL simple	532	1351	248	2	0	0	0	0	2.87
L1	635	0	0	0	0	0	0	0	2.00
RR	636	0	0	0	0	0	0	0	2.00
Phi	614	0	0	0	0	0	0	0	2.00

Table 4: Feature length distribution and average feature length for the full dataset.

Furthermore, there are also fewer interactions between the musculoskeletal and cardiovascular clusters in the GSSL model. Some cardiovascular diseases, e.g., myocardial infarction, are indeed found to be associated with inflammatory arthritis and osteoarthritis (Nielen et al. 2012). However, the GSSL model suggests that this is not true for heart tumors and pulmonary heart diseases.

Conclusions

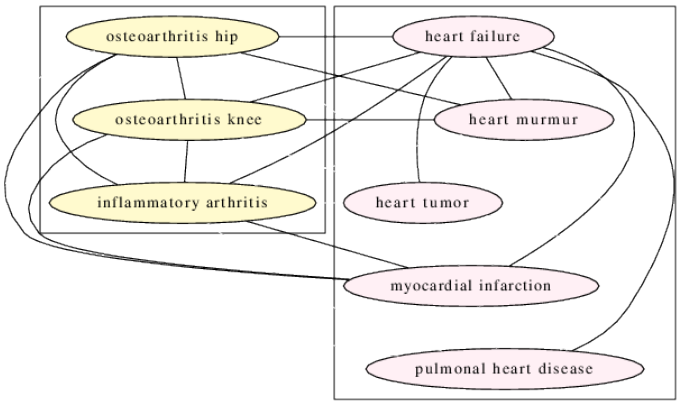
In this paper we analyzed data about patients that have multiple different diseases, which is an emerging area of work. Most existing work focuses on co-occurrence of diseases. Our primary contribution is to investigate the possibility of using Markov networks as vehicle to explore relationships involving more than two diseases. Using a real-world dataset, we found that going beyond disease occurrence can result in more accurate learned models. Additionally, we manually inspected some of the learned complex features (i.e., those that involve more than two diseases) and found that they make sense from a medical perspective. In conclusion, this seems to be a promising approach to exploring multimorbidity that requires further investigation.

Acknowledgments. JVH is supported by the Agency for Innovation by Science and Technology in Flanders (IWT). JD is partially supported by the research fund KU Leuven (CREA/11/015 and OT/11/051), and EU FP7 Marie Curie Career Integration Grant (#294068). ML is supported by the Netherlands Organisation for Health Research and Development (ZonMw) (#300020009). AH is supported by a VENI project of the Netherlands Organisation for Scientific Research (#639.021.918).

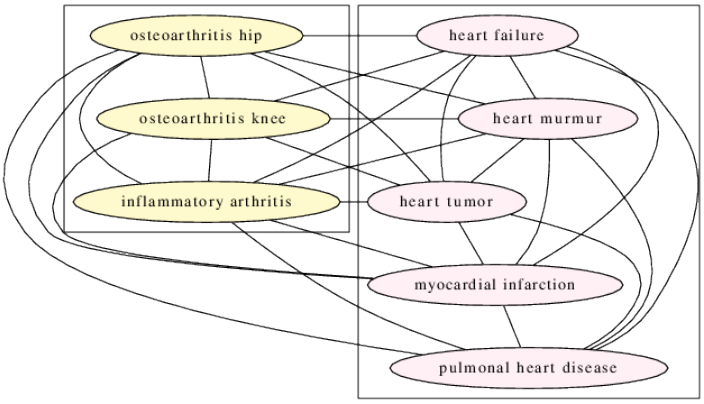
References

Barabási, A. L.; Gulbahce, N.; and Loscalzo, J. 2011. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12(1):56–68.

Barnett, K.; Mercer, S.; Norbury, M.; Watt, G.; Wyke, S.; and Guthrie, B. 2012. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 380:37–43.



(a.) Part of the GSSL model



(b.) Part of the L1 model

Figure 1: Subgraph of models learned by GSSL and L1.

Besag, J. 1975. Statistical Analysis of Non-Lattice Data. *The Statistician* 24:179–195.

Chen, L.; Blumm, N.; Christakis, N.; Barabasi, A.; and Deisboeck, T. 2009. Cancer metastasis networks and the prediction of progression patterns. *British journal of Cancer* 101:749–758.

Davis, J., and Domingos, P. 2010. Bottom-Up Learning of Markov Network Structure. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*. ACM Press.

Della Pietra, S.; Della Pietra, V.; and Lafferty, J. 1997. Inducing Features of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19:380–392.

Hidalgo, C. A.; Blumm, N.; Barabasi, A. L.; and Christakis, N. A. 2009. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* 5(4):e1000353.

Lowd, D., and Davis, J. 2010. Learning Markov Network Structure with Decision Trees. In *Proceedings of the Tenth IEEE International Conference on Data Mining*.

McMurray, J.; Adamapoulos, S.; Anker, S.; and et al. 2012. European society of cardiology guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal* 33:1787–1847.

Nielen, M.; van Sijl, A.; Peters, M.; Verheij, R.; Schellevis, F.; and Nurmohamed, M. 2012. Cardiovascular disease prevalence in patients with inflammatory arthritis, diabetes mellitus and osteoarthritis: a cross-sectional study in primary care. *BMC Musculoskeletal Disorders* 13:150–155.

O'Halloran, J.; Miller, G.; and Britt, H. 2004. Defining chronic conditions for primary care with icpc-2. *Familiy Practice* 21:381–386.

Ravikumar, P.; Wainwright, M. J.; and Lafferty, J. 2010. High-Dimensional Ising Model Selection using L1-Regularized Logistic Regression. *Annals of Statistics* 38(3):1287–1319.

van Driel, M. A.; Bruggeman, J.; Vriend, G.; Brunner, H. G.; and Leunissen, J. A. M. 2006. A text-mining analysis of the human phenome. *European Journal of Human Genetics* 14(5):535–542.

Van Haaren, J., and Davis, J. 2012. Markov Network Structure Learning: A Randomized Feature Generation Approach. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.