Modeling Inter-practice Variation of Disease Interactions using Multilevel Bayesian Networks

Martijn Lappenschaar[1], Arjen Hommersom[1],
Stefan Visscher[2], and Peter J.F. Lucas[1]

[1] Radboud University Nijmegen
Institute for Computing and Information Sciences
{mlappens,arjenh,peterl}@cs.ru.nl
[2] NIVEL (Netherlands Institute for Health Services Research)
S.Visscher@nivel.nl

Abstract. Multimorbidity is becoming a significant health-care problem for western societies, especially within the elderly. Since medical knowledge is mostly organized around single diseases, it is unlikely that the elderly patient with multiple diseases receives appropriate treatment. To get a grip on complex interactions, we aim to model domains using hierarchies, for example, patient characteristics, pathophysiology, symptomatology and treatment. For this we introduce *Multilevel Bayesian networks*, which we have applied to clinical data from family practices in the Netherlands on heart failure and diabetes mellitus. We compare the outcomes to conventional methods, which reveals a better insight of interactions between multiple diseases.

1 Introduction

Recent epidemiological research in the Netherlands indicates that more than two third of all patients older than 65 years have two or more chronic diseases at the same time; this problem, one of the most challenging of modern medicine, is referred to as the problem of comorbidity or multimorbidity. Where *comorbidity* is defined in relation to a specific index condition, the term *multimorbidity* has been introduced in chronic disease epidemiology to refer to any co-occurrence of two, but often more than two, multiple chronic or acute diseases within a person. The introduction of this term indicated a shift of interest from a given index disease (i.e. the primary disease) to individuals having multiple diseases.

There is no guarantee that, in case of a patient with multiple diseases, treating each disease individually is optimal. The need of an integrated optimal treatment for a patient with multiple diseases also implies the need for an integrated research methodology of multiple diseases. However, medical researchers often focus on an index disease rather than looking at multimorbidity in total.

Typically, regression methods are used to analyze the variance in disease variables, where researchers focus on the power of specific variables for predicting the presence or absence of specific diseases [22]. Where linear regression is used for continuous outcome variables, logistic regression is mostly used for dichotomous outcome variables. In case patients can be divided into groups, *multilevel* regression can be used to analyze the group dependent variance by adding extra variance components [7].

In contrast to using regression of fixed functional form, the patient data can also be modeled using probabilistic graphical models, such as Bayesian networks [11]. The edges of the graphical model then represent relationships between patient characteristics, pathophysiology and diagnostic tests for the disease of interest, which naturally generalizes to multiple diseases. However, multilevel modeling has not been studied in this context.

In this paper we introduce a new representation of multilevel disease models using Bayesian networks – which we call *multilevel Bayesian networks* – of which the multilevel regression model is a special case. This gives us the advantage that multiple models, e.g. of diseases, can be merged into one model, which allows examination of the interactions between them. Moreover, we apply this framework to patient data from family practices in the Netherlands. Its effectiveness is shown by comparing the model to the traditional methods based on regression analysis.

2 Multimorbidity: Context and Related Research

An Abstract Disease Model The context of multimorbidity is illustrated by Fig. 1 which provides an abstract view on the problem. The left-hand side shows the typical relationships between variables when considering a single disease. They form a hierarchical topology: genetics and environment, patient characteristics, disease, pathophysiology, and measurable variables, i.e. specific signs, symptoms and laboratory results.

If we represent multiple diseases in the same model, all kinds of interaction between variables within this model can be identified, as is illustrated at the right-hand side of Fig. 1. Mutual dependences between the two diseases may

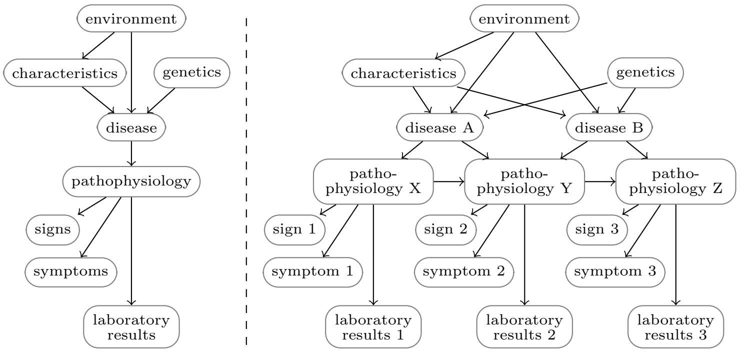

Fig. 1. Abstract model of a single disease (left) and multiple diseases (right).

concern their pathophysiology, symptoms, signs, and lab results. By modeling these interactions explicitly, better decisions can be made for patients having multiple diseases. Moreover, when considered separately, single disease models often contain a lot of overlap with each other, which may be avoided by integrating different disease models into a single model.

Normally, in scientific research, one would investigate diseases separately, resulting in different predictive values of variables shared by both diseases. Recently, multilevel regression analysis was used to investigate the influence of particular family practice variables on hypertension and diabetes mellitus, revealing an inter-practice variance in predictability [10]. However, since interactions could have an additive effect on prevalence, this yields no insight into the predictive value in case both diseases are present. Actually, we need an extra regression on the combined diagnosis to be able to conclude on such information.

In regression methods the variance of the observations is minimized with respect to the dependency between variables. Multilevel analysis also tries to explain the variance caused by grouping variables that intermediate on the lower level variables, i.e. it allows the intercept and slope, that determines the linear dependency between two variables, to alter for different groups. To analyze complex multimorbidity models one might have to deal with large datasets in which many variance is introduced. This can be due to the fact that data is collected from different kind of sources (e.g. family practices) or the data represents patients of all kind of populations (social, economic, and demographic differences). If we would ignore this, identifying interactions between disease variables such as pathophysiology and laboratory results could be difficult or erroneous.

Ultimately, we need one model that explains, both the variance, introduced in the observations, and the interactions in case of multiple diseases. If we can translate the multilevel regression models, which can deal with the variance explained by hierarchical structures themselves, to a graphical representation in such a way that we are able to connect multiple models of different diseases together, we also make them dependent on the interactions between diseases.

Related Research Much of the medical research relies on regression models which are applied to a single disease, and, thus, ignore the complexity of multimorbidity. Prevalence of multimorbidity are studied in family practices [1], sometimes with clustering of specific diseases [8]. These results illustrate the impact and complexity, but give little insight into interactions between diseases.

More advanced methods to analyze multimorbidity in particularly were not available until recently. A network analysis of pairwise comorbidity correlations for 10.000 diseases from 30 million medical records illustrated the complexity of many physiological processes in a context of patient characteristics such as ethnicity and genetic predisposition [6]. Markov blanket models and Bayesian model averaging were used in algorithms for learning patient-specific models from clinical data, to predict the outcome of sepsis or death in case of cardiovascular diseases [21]. To deal with polypharmacy, there is recent work of a Bayesian network meta-analysis to compare antihypertensive treatments in randomized controlled trials [16]. The method allows a comparison of multiple treatments where only a subset of treatments were compared in each trial. This mixed treatment comparison was facilitated with a framework of Markov models to be able to monitor disease progression [13].

Bayesian graphical modeling [18] is presented as a framework for generalized linear models, including multilevel and hierarchical models, with the aim to represent the conditional independence assumptions for parameters and observables and to make them the basis for a local computational strategy, generally based on Markov Chain Monte Carlo (MCMC) methods. It addresses solutions to deal with overdispersion, hierarchical modeling, dependency between coefficients, model criticism, making predictions, covariates and missing populations.

Although not specially designed for multimorbidity, similarity networks and Bayesian multinets [2] may offer a suitable method to represent uncertain knowledge in case of multiple diseases. An advantage of these methods is the possibility to represent asymmetric independence assertions, meaning that dependency between variables may only occur for certain values of these variables.

In the next section, basic techniques used in this paper are briefly reviewed.

3 Preliminaries

In this section we provide the basic concepts that we will use when modeling multimorbidity. Before moving on to the regression methods and Bayesian networks we first summarize basic elements of probability theory putting emphasis on multivariate probability distributions. Further on, we will discuss the issues that need to be dealt with when modeling multiple diseases.

3.1 Probability Theory

The patients's characteristics, pathophysiology, investigations, etc., can be seen as random variables each with its own distribution. Formally, random variables are denoted with uppercases, and observations with lowercases. We assume there is some joint, or multivariate, probability distribution over the set of random variables X, denoted by $P(X)$. The probability of a conjunction of two sets of variables, $X \wedge Y$, is denoted as $P(X \wedge Y)$ and also as $P(X, Y)$. The marginal distribution of $Y \subseteq X$ is then given by summing (or integrating) over all the remaining variables, i.e., $P(Y) = \sum_{Z=X \setminus Y} P(Y, Z)$. A conditional probability distribution $P(X \mid Y)$ is defined as $P(X, Y)/P(Y)$. Two variables X and Y are said to be conditionally independent given a third variable, Z, if $P(X \mid Y, Z) = P(X \mid Z)$.

In case a variable X is discrete, the variable is bounded by a finite set of possible values x, a probability is then denoted by $P(X = x)$. In case the outcome space of a variable X is the set of real numbers $\mathbb{R}$ or a subset thereof, one uses the probability $P(X \leq x)$.

3.2 Linear Regression

In general, in medical research, we have a dataset of observations of a number of patients, and we could see them as possible outcomes of the random variables. The variables can be split up in several domains. We can distinguish outcome variables, denoted as O_i, and explanatory variables, denoted as E_i. Some explanatory variables act on a group level, i.e. they have the same value for each individual within a certain group, which are denoted as L_i.

Linear regression tries to fit the observations of a random continuous variable (assuming it is normally distributed) into a linear model. This is done using an algorithm, e.g. a least square method, that minimizes the defiance of the observations with respect to the model parameters (the variance). Typically, we want to explain an observation o with respect to explanations e_i assuming that the observations o are possible outcomes of a random variable O. If the vectors $(1, e_1, \ldots, e_i, \ldots, e_n)^T$ are explanations, linear regression often estimates the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_i, \ldots, \beta_n)^T$, such that

$$P(O \mid e) \sim \mathcal{N}(\mu, \Sigma), \text{ with } \mu = \beta^T e \tag{1}$$

for every explanation e. Linear regression only makes sense in case of continuous variables. In case of disease variables this mostly only accounts for physical measurements. For example, a linear relation between two different kind of blood measurements BM_1 and BM_2, e.g. the low density lipoprotein (LDL) and high density lipoprotein (HDL) blood values, could be modeled as:

$$P(BM_1 \mid BM_2 = bm_2) \sim \mathcal{N}(\beta_0 + \beta_1 bm_2, \Sigma)$$

For more details about linear regression and other regression methods, especially in the medical area, one is referred to [22].

3.3 Multilevel Regression

In multilevel regression, part of the variance is explained due to group effects, i.e. the intercept and slope of the linear dependencies is allowed to alter amongst different groups. Now suppose we have a set of observations l_j, with $1 \leq j \leq m$, that have the same value within a certain group of patients, and based on that we can divide the patients into k groups. We could simply add these variables to the regression model as extra predictors. If we have $e = (1, e_1, \ldots, e_n, l_1, \ldots, l_m)^T$ as possible multivariate outcome, and β as $(\beta_0, \beta_1, \ldots, \beta_n, \beta_{n+1}, \ldots, \beta_{n+m})^T$ we keep a model as defined in Equation (1), having $n + m + 1$ degrees of freedom.

Multilevel regression, however, offers a different approach. For each k^{th} group we define a linear regression model, with O_k as random outcome variable, and allow dependency of the regression coefficients on the variables l_j and certain deviation from the overall mean. With $e = (1, e_1, \ldots, e_n)^T$, $l = (1, l_1, \ldots, l_m)^T$, $\beta_k = (\beta_{k0}, \ldots, \beta_{kn})^T$, $\delta_k = (\delta_{k0}, \ldots, \delta_{kn})^T$, Γ_k a matrix consisting of γ_{ij}^k, and $\delta_{ki} \sim \mathcal{N}(0, \Sigma_\delta)$, the model becomes:

$$P(O_k \mid e, g) \sim \mathcal{N}(\mu, \Sigma), \text{ with } \mu = (\delta_k + \Gamma_k g)^T e \tag{2}$$

The model is now more complex and the number of degrees of freedom is $k(n+1)(m+2)$. For example, if we extend our previous example by grouping on gender represented by a variable *gen*, and allow an influence of gender on the relation between the two blood measurements, the model then becomes:

$$P(BM_1 \mid \mathit{male}, bm_2] = \mathcal{N}(\delta_0^m + \gamma_0^m + (\delta_1^m + \gamma_1^m)bm_2, \varSigma)$$
$$P(BM_1 \mid \mathit{female}, bm_2] = \mathcal{N}(\delta_0^f + \gamma_0^f + (\delta_1^f + \gamma_1^f)bm_2, \varSigma)$$

The parameters of multilevel regression models can be estimated using an restricted iterative generalized least square (RIGLS) method, which coincides with restricted maximum likelihood (REML) in Gaussian models [3]. It estimates the parameters by alternating the optimizing process between the fixed parameters (γ_{kij}) and the stochastic parameters (δ_{ki}) until convergence is reached, and is equivalent to the maximum likelihood estimation in standard regression.

3.4 Generalized Regression Models

The former model assumes that the random outcome variable O is normally distributed. But suppose we want to consider a dichotomous outcome variable with only the possible values 'yes' and 'no'. An approach to deal with non-normally distributed variables is to include the necessary transformation and the choice of the appropriate error distribution explicitly into the model. This class of statistical models are called generalized linear models. They are defined by three components: an outcome variable O that has an expected value $\mathrm{E}[O|e]$, a linear additive regression equation that produces an unobserved (latent) predictor η of the outcome variable O, and a link function that links the expected values of the outcome variable O to the predicted values for η. In logistic regression the link function is given by $\eta = \mathrm{logit}(\mathrm{E}[O|e]) = \log \frac{\mathrm{E}[O|e]}{1-\mathrm{E}[O|e]}$. The logistic multilevel model then becomes:

$$\mathrm{logit}(\mathrm{E}[O_k] \mid e, l) = (\delta_k + \varGamma_k l)^T e$$

The conditional probability in case of logistic regression is then defined as:

$$P(O_k \mid e, l) = \frac{1}{1+e^{-x}}, \text{ with } x = (\delta_k + \varGamma_k l)^T e \qquad (3)$$

For example, we are interested in the predictive value of blood measurement BM_1 and BM_2 to an dichotomous outcome variable such as disease $D1$ with possible values 'yes' or 'no'. The multilevel logistic regression model with respect to gender then becomes:

$$\mathrm{logit}(\mathrm{E}[D_1] \mid \mathit{male}, bm_1, bm_2) = \delta_0^m + \gamma_0^m + (\delta_1^m + \gamma_1^m)bm_1 + (\delta_2^m + \gamma_2^m)bm_2$$
$$\mathrm{logit}(\mathrm{E}[D_1] \mid \mathit{female}, bm_1, bm_2) = \delta_0^f + \gamma_0^f + (\delta_1^f + \gamma_1^f)bm_1 + (\delta_2^f + \gamma_2^f)bm_2$$

Parameters for dichotomous outcomes are estimated with marginal and penalized quasi-likelihood (MQL/PQL) algorithms [4]. Alternatively MCMC methods such as Gibbs Sampling can be used [15].

3.5 Bayesian Networks

Bayesian networks offer an effective framework for knowledge representation and reasoning under uncertainty [11]. Formally, a *Bayesian network*, or BN, is a tuple $\mathcal{B} = (G, X, P)$, with $G = (V, E)$ a directed acyclic graph (DAG), $X = \{X_v \mid v \in V\}$ a set of random variables indexed by V, and P a joint probability distribution. X is a Bayesian network with respect to the graph G if P can be written as a product of the probability of each random variable, conditional on their parent variables:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{v \in V} P(X_v = x_v \mid X_j = x_j \text{ for all } j \in \pi(v)) \quad (4)$$

where $\pi(v)$ is the set of parents of v (i.e. those vertices pointing directly to v via a single arc). If there are continuous variables, the definition is similar, and can be defined by using the probability density function. While the conditional probabilities could be estimated using regression methods [14], parameter and structure learning methods for Bayesian networks are readily available [9].

For example, suppose we have two binary variables, D for disease present yes/no and G for gender, both having an direct effect on the blood measurements BM_1 and BM_2. Besides that BM_1 affects BM_2 also directly. Using marginalization, we obtain the probability for BM_2 by:

$$P(BM_2) = \sum_{D} \sum_{G} \sum_{BM_1} P(BM_2 \mid BM_1, D, G) P(BM_1 \mid D, G) P(D) P(G)$$

Conditional independence in Bayesian networks is an important concept when modeling multimorbidity. When considering three vertices u, v and w we can distinguish certain types of dependencies:

- v is a tail-tail vertex ($u \leftarrow v \rightarrow w$)
- v is a head-tail vertex ($u \rightarrow v \rightarrow w$)
- v is a head-head vertex ($u \rightarrow v \leftarrow w$)

For the first two situations we obtain independence between X_u and X_w if we condition on X_v, i.e. $P(X_u \mid X_w, X_v) = P(X_u \mid X_v)$, also denoted as $X_u \perp\!\!\!\perp X_w \mid X_v$, whereas $X_u \not\perp\!\!\!\perp X_w \mid \varnothing$. In the third situation, the situation is reversed, as X_u and X_w are unconditionally independent, whereas they become dependent when conditioning on X_v, i.e., $X_u \perp\!\!\!\perp X_w \mid \varnothing$ and $X_u \not\perp\!\!\!\perp X_w \mid X_v$. The Markov blanket (MB) of a vertex contains all the variables that shield the vertex from the rest of the network, meaning that if all variables within the MB can be observed, this is the only knowledge needed to predict the behavior of that vertex [11].

It is appealing to define disease variables as binary variables, i.e. the disease of interest is present yes or no. Socio-economical and demographic variables are often categorical (sometimes ordered) or numerical (e.g. age). Laboratory investigations are often continuous (especially blood measurements), but can be discretized, e.g. blood glucose levels could be defined as normal, subclinical, and clinical. Variables can be dependent on each other, or independent, which

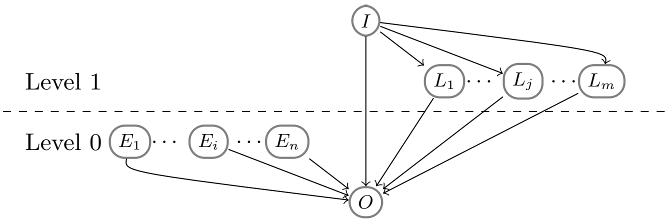

Fig. 2. Bayesian network of multilevel regression.

can be represented in a Bayesian network, defining variables as vertices and dependencies as edges. If the structure is unknown it can be learned.

Ideally, we expect to obtain some kind of hierarchical topology in the learned structure, just as described in Fig 1. In fact, we can put restrictions into the learning algorithm to force such a topology. If we consider the disease variables, an association might be present between them, but there's a chance we could make them conditional independent if we observe the environmental and patient's characteristics variables, i.e. they serve as tail-tail variables with respect to disease variables. At the other hand looking at the laboratory results, those might act as head-head variables with respect to diseases, therefore we cannot make the diseases conditional independent looking solely to laboratory results.

4 Multilevel Bayesian Networks

In this section, we introduce the multilevel Bayesian network (MBN) formalism as interpretation of multilevel regression. First, we briefly explore the relation between multilevel regression models and Bayesian networks. Then, we generalize this by allowing more structure within the model. We discuss the building and learning of such models and compare this to the regression approach.

4.1 Multilevel Regression Analysis as a Bayesian Network

In multilevel regression, the outcome variable O depends on the explanations $e = e_1, \ldots, e_n$ and $l = l_1, \ldots, l_m$. In a Bayesian network approach, we model O as a conditional probability distribution given the set of parents $E_1, \ldots, E_n$, $L_1, \ldots, L_m$, and $I_1, \ldots, I_k$, i.e., we now interpret the explanations and group explanations as instantiations of random variables. The variables I_j, with $1 \leq j \leq k$ is an indicator variable for grouping of objects at a certain level j. Fig. 2 then shows the corresponding Bayesian network, assuming independent predictors.

Clearly, this model is not realistic in case of multimorbidity domains. There is no structure present between predictors and we have only one outcome variable of interest. In general, the opposite is more likely to be true, i.e. multiple outcome variables, multiple dependencies between predictors and variables that are both predictor and outcome variable. While discriminative learning algorithms, such as regression, are good for prediction, they do not provide insight into the domain, nor can they be used to model interactions necessary in the case of multimorbidity. Bayesian networks have the ability to give such insight, by allowing dependencies between variables.

4.2 Multilevel Bayesian Networks in General

The idea of a multilevel Bayesian network is that I variables split the domain into different categories with a deterministic effect on variables that are constant within a category (L). Some variables (E) are group-independent, though structure may exist between these variables. Other variables (O) depend both on grouping and other variables in the same or higher levels. The Bayesian network is constrained in the sense that no edges exist from a lower-level variable to a higher-level variable. This ensures that we keep the hierarchical structure obtained with multilevel regression methods (see Fig. 3). A *multilevel Bayesian network* is defined as a tuple (G, N, I, E, O, L, P) such that:

- (G, V, P), where $V = I \cup E \cup O \cup L$, forms a Bayesian network;
- I, E, O, and L are pair-wise disjoint;
- $N \in \mathbb{N}$ denotes the number of levels on top of the base level 0;
- $I = \{I_1, \ldots, I_N\}$ are variables such that each value of such a variable contains a group. It holds that I_j is the only parent of I_{j-1} in G for all $1 \leq j \leq N$;
- $E = \{E^0, \ldots, E^N\}$ where each E^j is a set of variables corresponding to level j, such that if $(V \to E_i^j) \in G$, then $V \in (E \cup O)^{j+k}$, with $k \geq 0$;
- $O = \{O^0, \ldots, O^N\}$ where each O^j is a set of variables corresponding to level j, such that if $(V \to X_i^j) \in G$, then $V \in (E \cup O)^{j+k}$ or $V \in L^{j+k+1}$ $k \geq 0$;
- $L = \{L^1, \ldots, L^N\}$ where each L^j is a set of group variables corresponding to a level j. If $L_i^j \in L^j$ is a group variable, then it holds that (i) $(I_j \to L_i^j) \in G$, (ii) there are no other variables V such that $V \to L_i^j$, and (iii) $P(L_i^j \mid I_j)$ is deterministic.

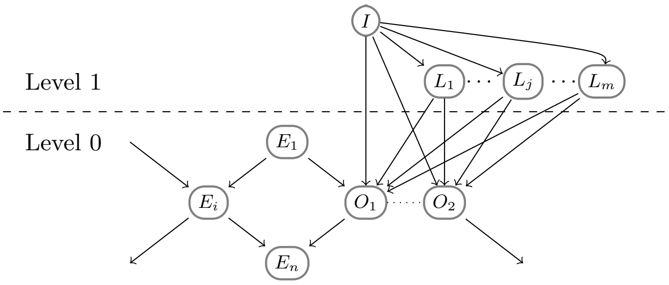

Fig. 3. Bayesian network of multilevel regression with an improved structure between predictors (E_i) and outcome variables (O_i).

4.3 Building Multilevel Bayesian Networks

In order to build the structure between intra-level variables, we can make use of two approaches. We can either model the structure manually based on existing medical knowledge or learn the structure from data. Structure learning of Bayesian networks offers a suitable method to learn these dependencies. The constraints imposed by the multilevel Bayesian network can be captured by blacklisting and whitelisting edges, which can be incorporated into a wide range of structure learning algorithms (see e.g., [17]). For example, the necessary edges between I and other variables are whitelisted, whereas edges from a lower level to a higher level are all blacklisted. The parameters can be learned using standard Bayesian network techniques. Compared to multilevel regression models, it is also possible to use a Bayesian approach for learning the parameters [19] and therefore include even more domain knowledge to the model.

Model validation can be achieved by using information criteria such as the Deviance Information Criteria and the Bayesian Information Criteria [20]. Alternatively, standard cross validation (e.g. k-fold cross validation) is a robust method to validate regression and Bayesian models [12].

5 Modeling Inter-practice Variation

5.1 Problem domain and data

In this paper, we apply the MBN approach for modeling inter-practice variations for predicting heart failure and diabetes mellitus. Data was collected by the Netherlands Information network of General Practice (LINH). In 1996, they started as a register of referrals of general practitioners to medical specialists. Information about contacts and diagnoses, prescriptions, referrals and laboratory and physiological measurements are extracted from the information systems. The LINH database contains information of routinely collected data from approximately 90 general practices. Unless patients moved from practices, and practices opted out, longitudinal data of approximately 300.000 distinct patients are stored. Patients under 25 were excluded, because of their low probability on multimorbidity. Practices who recorded during less than six month were also excluded from statistical analysis. Eventually, we used data of 218333 patients. Morbidity data were derived from diagnoses, using the international classification of primary care (ICPC) and anatomical therapeutic chemical (ATC) codes.

5.2 Results

Our main variables of interest are heart failure and diabetes mellitus. The predictors are shown in Table 1, with the urbanity of the practice's area as higher level variable. Multilevel logistic regression leaves us then with five separate models, for which the parameters are estimated using MLWin [5]. To obtain the parameters of the MBN interpretation we ran a MCMC method which is available in the WinBUGS software [19]. Dichotomous variables are modeled using

	RIGLS		MCMC-fixed		MCMC-learn	
Diabetes Mellitus	β	Odds	β	Odds	β	Odds
Intercept	-5.800 (0.3%)		-5.678 (0.3%)		-5.866 (0.3%)	
Age	0.029	1.029	0.028	1.028	0.063	1.065
Gender (ref = male)	-0.090	0.914	-0.089	0.915	-0.222	0.801
Overweight/obesity	0.545	1.725	0.513	1.671	1.189	3.282
Lipid disorder	1.862	6.437	1.855	6.392		
Hypertension	1.736	5.675	1.758	5.800		
Atherosclerosis	-0.047	0.954	-0.052	0.949		
Heart failure	0.124	1.132	0.178	1.194		
Retinopathy	2.225	9.253	2.269	9.669		
Angina pectoris	-0.387	0.679	-0.409	0.665		
Stroke / CVA	-0.262	0.770	-0.269	0.766		
Renal disease	0.162	1.176	0.183	1.200		
Cardiovasc. symptoms	-0.165	0.848	-0.163	0.850		
Urbanity (ref=urban)						
strongly urban	0.232	1.261	0.243	1.275	-0.326	0.722
modestly urban	0.390	1.477	0.399	1.490	0.389	1.476
little urban	0.362	1.436	0.342	1.408	-0.934	0.393
not urban	0.388	1.474	0.230	1.259	-0.313	0.731
Model validation average accuracy (cross validation)	89%		89%		88%	
Heart Failure	β	Odds	β	Odds	β	Odds
Intercept	-11.373 (0.0%)		-11.20 (0.0%)		-11.24 (0.0%)	
Age	0.101	1.106	0.101	1.106	0.105	1.111
Gender (ref=male)	-0.195	0.823	-0.204	0.815	-0.160	0.852
Overweight/obesity	0.524	1.689	0.470	1.600		
Diabetes mellitus	0.228	1.256	0.726	2.067	0.330	1.391
Lipid disorder	0.159	1.172	-0.832	0.435		
Hypertension	0.728	2.071	0.425	1.530	0.963	2.618
Atherosclerosis	0.482	1.619	0.231	1.260	0.655	1.925
Retinopathy	0.270	1.310	0.099	1.104		
Angina pectoris	0.795	2.214	0.781	2.184		
Stroke / CVA	0.328	1.388	0.334	1.397		
Renal disease	0.630	1.878	0.632	1.881	0.720	2.054
Cardiovasc. symptoms	0.954	2.596	0.969	2.636		
Urbanity (ref=urban)						
strongly urban	0.135	1.145	0.147	1.158		
modestly urban	0.166	1.181	0.176	1.192		
little urban	0.352	1.422	0.375	1.456		
not urban	0.289	1.335	0.276	1.318		
Model validation average accuracy (cross validation)	89%		89%		95%	

Table 1. Parameter estimations and cross validation of multilevel analysis for Heart Failure and Diabetes Mellitus. RIGLS = restrictive iterative general least square method for the multilevel logistic regression model, MCMC-fixed = Monte Carlo Markov Chain method for the multilevel Bayesian network without structure learning, MCMC-learn = same as MCMC-fixed but with structure learning.

a Bernoulli distribution. Parameter estimates and the average accuracy of predicting heart failure and diabetes mellitus are presented in in Table 1. Validated using a 10-fold validation, the table shows that the MBN model is in line with results obtained by multilevel regression.

The next step is structure learning of predictors and outcome variables while maintaining the multilevel structure as mentioned in Section 4. Diseases are obviously not a cause of practice and patient characteristics such as age and gender. The *bnlearn* package [17] in the statistical software R provides both constraint based as scoring algorithms to learn the structure of a Bayesian network. The score based methods reveal the most appealing structure for the available data. See Fig. 4a for the resulting Bayesian network structure.

Some of the directions of certain edges is opposite to what the domain experts would expect, e.g. angina pectoris and heart failure is pointing towards atherosclerosis, but in reality the latter is seen as a cause of the first and not a

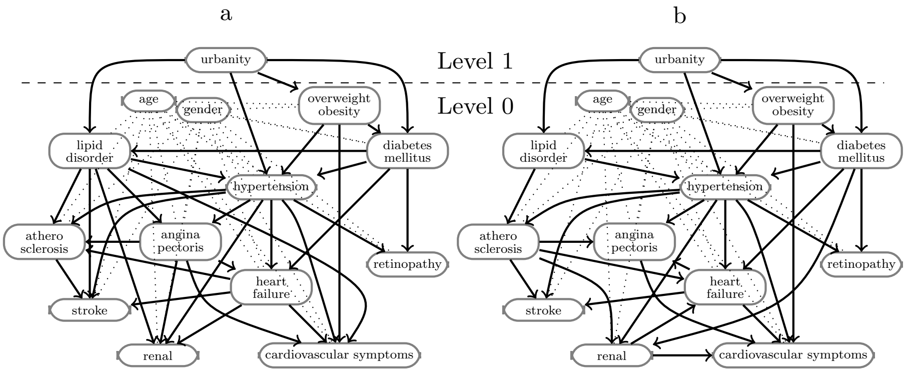

Fig. 4. Structure learning of Heart Failure and Diabetes in Family Practices, (a) with inter-level restrictions only, and (b) with intra-level restrictions (expert opinions / evidence from other research) as well.

	obesity /overw.	diabetes mellitus	lipid disorder	hyper- tension	athero- sclerosis	renal disease	heart failure	angina pectoris	stroke /CVA	retino- pathy
Learned	89%	88%	84%	86%	96%	95%	95%	94%	96%	97%
Restricted	89%	88%	85%	87%	96%	95%	95%	95%	96%	97%

Table 2. Accuracy for predicting diseases in a multilevel Bayesian network for the model containing both heart failure and diabetes mellitus. The 'Learned' model corresponds with Fig 4a, and the 'Restricted' model corresponds with Fig 4b.

consequence. Probably, this opposite direction is due to the fact that atherosclerosis is mostly diagnosed clinically by interpreting symptoms and signs of the disease. By incorporating domain knowledge into the model and re-running the structure learning algorithm, we obtain the model as shown in Fig. 4b.

We have learned the probability distributions of both the learned and restricted model. Using a 10-fold cross validation we the calculated the accuracy of predicting not only diabetes mellitus and heart failure, but also for the other diseases present in the MBN. See Table 1 and 2 for an overview of the results. The accuracy of prediciing diabetes mellitus is similar to the previous models, whereas the accuracy of predicting heart failure is 6% better. The accuracies for the other disease variables, ranging between 84% and 97%, are slightly better in the restricted model. Looking into interactions between predictors, the structured model is more accurate when looking to heart failure. Table 3 shows the estimated and true prevalences of heart failure in the presence of multiple comorbidities. The estimations of the structured model are closer to the actual data. Clearly, the problem with the regression model is that it does not recognize the fact that the prevalence of heart failure is independent of obesity when conditioned on hypertension and diabetes mellitus.

	obesity				no obesity			
	diabetes		no diabetes		diabetes		no diabetes	
Heart Failure	hypertension	no hypertension	hypertension	no hypertension	hypertension	no hypertension	hypertension	no hypertension
Multilevel Regression	10	5.1	8.1	4.1	6.5	3.2	5.2	2.6
Multilevel Network	9.1	0.0	4.1	0.7	9.3	0.3	5.2	0.6
Calculated from data	10	0.0	4.3	0.7	10.3	0.3	5.5	0.6

Table 3. Prevalences (in percentages) of heart failure in the presence of obesity, diabetes and hypertension, based on model parameters compared to actual values.

6 Discussion

In this paper we introduced Bayesian networks as an interpretation of multilevel analysis. Using patient data from family practices, its predictive value for heart failure and diabetes mellitus is just as good compared to traditional methods such as multilevel regression analysis, despite a reduced number of predictors.

The advantage of multilevel Bayesian networks is that it allows multiple outcome variables within one model, reducing redundancy of multiple multilevel regression models. Furthermore, we can add intra-level structures between variables giving extra insight into dependencies. Bayesian networks can be used to model conditional independence between variables, as we have seen with heart failure. We could perform a complete structure learning of the data, ignoring the hierarchy of variables.

But in practice the model is then very prone to assign causality from lower level variables to higher level variables, which in fact is not possible if we define the hierarchy properly. In case of patient data within a multimorbidity setting it is appealing to use the hierarchical topology: genes and environment (e.g. urbanity) – patient characteristics (e.g. age, gender, habits) – pathophysiology (diseases, syndromes) – and symptomatology (e.g. symptoms, signs, laboratory results), which can be modeled well using multilevel Bayesian networks.

The disadvantage of using patient data from family practices, is that it is driven by the actions of the physician. Since most diabetic patients are sent to an eye doctor, obviously we will find a overestimated relation between diabetes and retinopathy. Data of specific pathophysiologic tests are not available (yet), so diagnoses are not strict but act with a certain probability depending on the specificity of the tools used in family practice. Future work will also focus on patient data retrieved from randomized controlled trials, with the additional difficulty to learn multiple parameters using low patient counts.

Furthermore, since the data available will never provide a full causal model, it is important to make use of expert input. Besides putting restrictions on existing variables, one might also introduce variables that are missing from the data, but which may add crucial explanatory power. This is possible in BNs, and thus MBNs can also use the same expertise to quantify the probabilistic relationships involving these missing variables even though no data exists for them. So, multilevel Bayesian networks enforces a kind of supervised structural learning with respect to variance explained by higher level variables.

References

1. M Fortin, et al. Prevalence estimates of multimorbidity: a comparative study of two sources. BMC Health Services Research 10:111, 2010.
2. D Geiger, D Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. Artificial Intelligence 82:45-76, 1996.
3. H Goldstein. Restricted unbiased iterative generalised least squares estimation. Biometrika 76:622-623, 1989.
4. H Goldstein, J Rabash. Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society (Series A) 159:505-512, 1996.
5. H Goldstein, W Browne, J Rasbash. Multilevel modeling of medical data. Statistics in Medicine 21(21):3291-3315, 2002.
6. CA Hidalgo, N Blumm, AL Barabasi, NA Christakis. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol. 2009;5:e1000353.
7. JJ Hox. Multilevel Analysis: techniques and applications, 2^{nd} ed. Routledge 2010.
8. A Marengoni, D Rizzuto, HX Wang, B Winblad, and L Fratiglioni. Patterns of chronic multimorbidity in the elderly population. J Am Geriatr Soc 57:225-230, 2009.
9. RE Neapolitan. Learning Bayesian networks. Prentice Hall, 2004.
10. MMJ Nielen, et al. Inter-practice variation in diagnosing hypertension and diabetes mellitus: a cross-sectional study in general practice. BMC Fam Pract 10:1-6, 2009.
11. J Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.
12. R Picard, D Cook. Cross-validation of Regression Models. Journal of the American Statistical Association 79(387):575-583, 1984.
13. MJ Price, NJ Welton, AE Ades. Parameterization of treatment effects for meta-analysis in multi-state Markov models. Statistics in Medicine 30:140-151, 2011.
14. F Rijmen. Bayesian networks with a logistic regression model for the conditional probabilities. Int. J. Approx. Reasoning 48(2):659–666, 2008.
15. MH Seltzer, et al. Bayesian analysis in applications of hierarchical models: issues and methods. Journal of Educational and Behavioral Statistics 21:131-167, 1996.
16. S Sciaretta, F Palano, G Tocci, R Baldini, M Volpe. Antihypertensive treatment and development of heart failure in hypertension. Arch Intern Med.171:384-394, 2011.
17. M Scutari. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software 35(3):122, 2010.
18. DJ Spiegelhalter. Bayesian Graphical Modelling: A Case-Study in Monitoring Health Outcomes. Applied Statistics 47-1:115-133, 1998.
19. DJ Spiegelhalter, A Thomas, N Best, D Lunn. WinBUGS User Manual; Version 1.4. MRC Biostatistics Unit, Cambridge UK, 2001.
20. DJ Spiegelhalter, NG Best, BP Carlin, A van der Linde. Bayesian measures of model complexity and fit. J.R. Statist. Soc. B 64-4:583-639, 2002.
21. S Visweswaran, DC Angus, M Hsieh, L Weissfeld, D Yealy, GF Cooper. Learning patient-specific predictive models from clinical data. J Biom Inform.43:669-85, 2010.
22. E Vittinghoff, DV Glidden, SC Shiboski, CE McCulloch. Regression Methods in Biostatistics: linear, logistic, survival and repeated measures models. Springer 2005.