Artificial Intelligence in Medicine 57 (2013) 171–183

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Multilevel Bayesian networks for the analysis of hierarchical health care data

Martijn Lappenschaar[a,*], Arjen Hommersom[a], Peter J.F. Lucas[a], Joep Lagro[b], Stefan Visscher[c]

[a] Radboud University Nijmegen, Institute for Computing and Information Sciences, PO Box 9010, 6500 GL Nijmegen, The Netherlands
[b] Radboud University Nijmegen Medical Centre, Department of Geriatric Medicine, PO Box 9101, 6500 HB Nijmegen, The Netherlands
[c] Netherlands Institute for Health Services Research (NIVEL), PO Box 1568, 3500 BN Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 15 December 2011
Received in revised form
14 December 2012
Accepted 16 December 2012

Keywords:
Bayesian network
Multilevel analysis
Disease prediction
Multimorbidity
Inter-practice variation
Cardiovascular disease

ABSTRACT

Objective: Large health care datasets normally have a hierarchical structure, in terms of levels, as the data have been obtained from different practices, hospitals, or regions. Multilevel regression is the technique commonly used to deal with such multilevel data. However, for the statistical analysis of interactions between entities from a domain, multilevel regression yields little to no insight. While Bayesian networks have proved to be useful for analysis of interactions, they do not have the capability to deal with hierarchical data. In this paper, we describe a new formalism, which we call multilevel Bayesian networks; its effectiveness for the analysis of hierarchically structured health care data is studied from the perspective of multimorbidity.
Methods: Multilevel Bayesian networks are formally defined and applied to analyze clinical data from family practices in The Netherlands with the aim to predict interactions between heart failure and diabetes mellitus. We compare the results obtained with multilevel regression.
Results: The results obtained by multilevel Bayesian networks closely resembled those obtained by multilevel regression. For both diseases, the area under the curve of the prediction model improved, and the net reclassification improvements were significantly positive. In addition, the models offered considerable more insight, through its internal structure, into the interactions between the diseases.
Conclusions: Multilevel Bayesian networks offer a suitable alternative to multilevel regression when analyzing hierarchical health care data. They provide more insight into the interactions between multiple diseases. Moreover, a multilevel Bayesian network model can be used for the prediction of the occurrence of multiple diseases, even when some of the predictors are unknown, which is typically the case in medicine.

1. Introduction

Health care research is often done using clinical data that contain a hierarchical structure—they have *levels* as its said—as the data have been obtained from different practices, hospitals, or regions. Since patients within the same practice are often more alike than two randomly chosen patients, they will likely have some correlation on variables related to the practice. Statistical analyses that ignore these correlations will lead to results that are statistically invalid [1]. Commonly used statistical techniques such as logistic regression do not allow incorporating the characteristics of the different levels in the hierarchy. Therefore, multilevel regression methods are often used to analyze such data. The books [2,3] offer an overview of such methods.

In the artificial intelligence literature, probabilistic graphical models, such as Bayesian networks [4], have had a significant impact on the modeling and analysis of the patient data [5]. The edges in the graphical model represent probabilistic relationships between specific patient variables for a disease of interest. Bayesian networks allow for the integration of medical domain knowledge, and clinical expertise can be modeled explicitly. Moreover, clinical knowledge derived from clinical health care data can be used to further refine and validate the model.

In this paper, we combine *multilevel* modeling and learning with Bayesian network modeling. This can be useful in complex domains, for example, when studying the problem of *multimorbidity*, i.e., the epidemiology of patients with multiple diseases. Multimorbidity is often analyzed using multilevel regression, as it requires a large amount of data coming from different sources in order to study the interaction between diseases. Moreover, it is a typical problem where Bayesian networks can be useful, as expert knowledge is needed, and representing multiple diseases requires scaling up to models containing a large number of variables.

Since Bayesian networks have already been successfully applied to model single diseases [5–11], and also for multiple diseases [12–16], the research question is whether and how it is

* Corresponding author. Tel.: +31 (0) 638896321.
E-mail address: mlappens@cs.ru.nl (M. Lappenschaar).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.12.007

possible to adopt the multilevel approach for Bayesian networks. In that way we would be able to explore complex health care data that is hierarchically structured using Bayesian networks with the advantage that, in contrast to multilevel logistic regression, models are obtained that offer a clear representation of the interactions between multiple diseases.

The main contribution of this paper is that it introduces a new representation of multilevel disease models using Bayesian networks, which we call *multilevel Bayesian networks*. It has the advantage that it is at least as powerful as multilevel logistic regression, yet supports, in contrast to multilevel logistic regression, gaining new insights into the interactions between multiple diseases.

Using patient data from family practices in The Netherlands, we applied this framework to obtain a prediction model for multiple chronic diseases, namely diabetes and heart failure. The effectiveness of multilevel Bayesian networks has been studied by comparing the resulting model to the traditional models based on multilevel regression analysis.

2. Related research

Multimorbidity is the health care problem where we focus on in this paper, although multilevel Bayesian networks may have other applications as well. We start, therefore, by introducing the research context.

Although in the current aging society multimorbidity is the norm rather than something rare, in medicine there is still a focus on single diseases with respect to their comorbidities, rather than that multimorbidity is considered in total. This is often done by studying the prevalence and significance of specific factors for predicting the presence or the absence of specific diseases, typically by applying (multilevel) regression methods where the variance of the observations is minimized with respect to a linear or logistic model. Where multimorbidity should be studied by exploring the interactions between diseases with associated signs and symptoms in their full generality, in practice current research explores this only in a very restrictive fashion.

For example, prevalence of multimorbidity has been studied in family practices [17,18], sometimes by clustering of specific diseases [19]. Multimorbidity indices are a way to measure specific types of multimorbidity within a population. A systematic review of these indices can be found in [20]. These methods illustrate the size, impact and complexity of multimorbidity, but give little insight into interactions between diseases.

Multilevel regression has many applications in the social sciences and in medicine; however, it was not especially designed to model multimorbidity [21–23]. In [24] complex hierarchical patient data were used to analyze the predictive value of cardiovascular diseases for hypertension and diabetes mellitus. Since both diseases are analyzed separately, the results only give a preliminary view on correlations between cardiovascular diseases.

Various Bayesian network models for multiple disease have been developed since the beginning of the 1990s. Examples are Pathfinder [12,13], Hepar II [15] and MUNIN [25]. They deal with multiple diseases, although belonging to the same class. One of few existing exceptions is QMR-DT [26,27], as it covers a broad subset of internal medicine. However, it was never meant for actual use. All these Bayesian network models have been constructed based on expert opinion and engineering background knowledge. They did only incorporate *known* disease interactions; they were not meant for uncovering *new* disease interactions. This explains why dealing with multilevel data was not seen as a problem. In this paper we make an important step forwards in this respect, as Bayesian network models are learned in order to gain insight into the interactions between diseases. Without the capability to deal with hierarchical data, using multilevel methods, such learning results are statistically unsound.

Bayesian networks have also been used in algorithms for learning patient-specific models from clinical data to compare mixed treatments and to predict disease progression [28,29]. Somewhat confusingly, the adjective 'hierarchical' is also used in connection to Bayesian networks. For example, nested, hierarchical Bayesian network allow one to define genetic models that can be reused [30]. Hierarchical Bayesian networks have also been proposed as an aggregating abstraction [31] that clusters variables closely related to each other. This all closely relates to object-oriented Bayesian networks [32], but there is no relationship to multilevel analysis where the hierarchy stands for nested data from different groups.

Eventually, one would preferably obtain models for health care data that can handle multimorbidity, and have the ability to be personalized, i.e., put observations on the patient into the underlying probabilistic model and obtain updated parameters that specifically account for that patient. Such personalized models help to obtain specific advice that relates to the patient's health status. The probabilities of the underlying model could be extracted from existing clinical research or from available patient data, using a valid method that takes interactions between diseases into account.

To illustrate the type of relationships that can occur, we show in Fig. 1 at the left-hand side the typical relationships between variables for a single disease, and at the right-hand side the integration of multiple diseases into one graphical model. Representing multiple diseases in one model avoids redundancy of separate representations and has the advantage that it shows where diseases interact. Mutual dependences may concern diseases, therapies, pathophysiology, symptoms, signs, and lab results, and modeling interactions explicitly, allows us to make better decisions for patients having multiple diseases. In fact, the architecture of networks such as MUNIN [25] is similar, as it also models diseases in terms of their pathophysiology and patient findings.

3. Preliminaries

In this section, the basic concepts are introduced that we will use in the following sections. Before moving on to Bayesian networks and multilevel regression we first review basic probability theory putting emphasis on multivariate probability distributions.

3.1. Probability theory

Disease variables can be seen as random variables, either discrete or continuous, each with their own distribution. Random variables are denoted by uppercases, e.g., X, and lowercases, e.g., x, indicate their values. Binary variables have the values x and $\bar{x}$. We assume there is a multivariate probability distribution over the set of random variables X, denoted by $P(X)$. The joint probability distribution of two disjoint sets X and Y is denoted as $P(X, Y)$.

Furthermore, a probability distribution is defined by a probability density function f_X for the continuous case, or a probability mass function f_X, for the discrete case. The marginal distribution of $Y \subseteq X$ is then given by summing (or integrating) over all the remaining variables: $P(Y) = \sum_{Z=X\setminus Y} P(Y, Z)$. A conditional probability distribution $P(X|Y)$ is defined as $P(X, Y)/P(Y)$, for positive $P(Y)$. Corresponding conditional density or mass functions are denoted by $f_{X|Y}$. Two variables X and Y are said to be conditionally independent given a third variable Z, if $P(X|Y, Z) = P(X|Z)$, for any value of Y, also denoted as $X \perp\!\!\!\perp_P Y|Z$. If, in contrast, these variables are (conditionally) dependent, this is denoted by $X \not\perp\!\!\!\perp_P Y|Z$

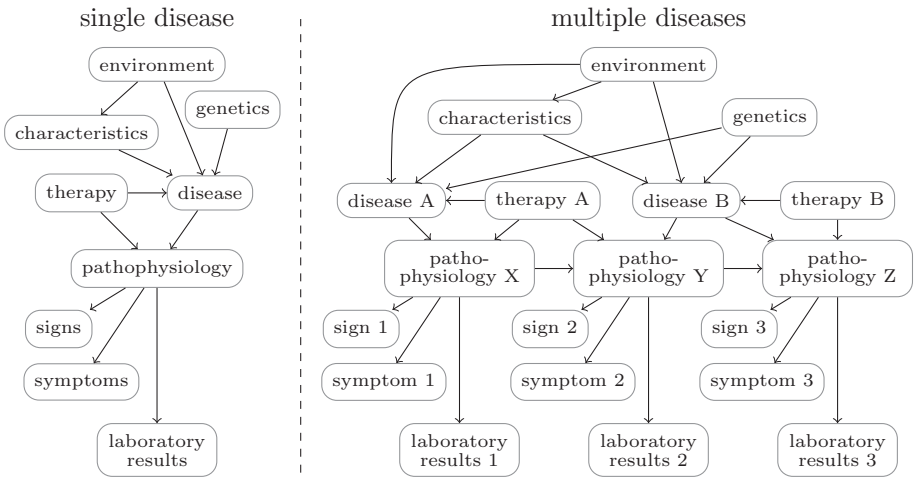


Fig. 1. Abstract model of a single disease (left) and multiple diseases (right).

3.2. Bayesian networks

Bayesian networks offer an effective framework for knowledge representation and reasoning under uncertainty [4]. A *Bayesian network*, or BN for short, is a tuple $\mathcal{B} = (G, X_V, P)$, with $G = (V, A)$ a directed acyclic graph, or DAG for short, with vertices or nodes V_G (also abbreviated to V) and arcs $A_G \subseteq V_G \times V_G$, $X_V = X_{v \in V}$ a set of random variables indexed by V, and P a joint probability distribution. This distribution P can be written as the product of the local probability of each random variable, conditional on their parent variables in the graph G:

$$P(X_V) = \prod_{v \in V} P(X_v | X_{\pi(v)}) \tag{1}$$

where $\pi(v)$ is the set of parents of v (i.e., those vertices pointing directly to v via a single arc). Learning methods for both the parameters as well as the graphical structure of a Bayesian networks are readily available [33].

Blockage of paths in the associated graph G of a Bayesian network, defined as *d-separation* of variables and denoted by $A \perp\!\!\!\perp_G B|C$ (any undirected path, if it exists, from a vertex in A to a vertex in B is blocked by vertices in C) [4], implies conditional independences of the corresponding random variables:

$$A \perp\!\!\!\perp_G B|C \rightarrow X_A \perp\!\!\!\perp_P X_B|X_C$$

i.e., P is *faithful* to G. This property can be exploited to study the problem of multimorbidity. Since models of multimorbidity typically contain many more variables than single-disease models, it is useful to select subsets of variables for predicting a particular disease. The relevant subset can be obtained by determining the *Markov blanket* (MB) of a vertex v: the set of vertices such that v is d-separated of all other vertices given the set of vertices in the Markov blanket [34]. In a BN, the Markov blanket of a vertex is the set of parents, children, and parents of children. Usually, we will not distinguish between variables and their corresponding vertices.

In multimorbidity it is of interest to study in which way diseases interact. For example, diseases D and D' might be unconditionally dependent of each other, i.e., $D \not\perp\!\!\!\perp_P D'|\varnothing$, but they could become independent if an environmental factor F is taken into account, $D \perp\!\!\!\perp_P D'|F$. This means that the factor F offers a complete explanation of the interaction between the disease D and D'. Moreover, the MB of a disease D are all factors, possibly other diseases, that are relevant to predict this disease D.

3.3. Multilevel regression

To analyze multimorbidity problems one has to deal with large datasets in which variance is introduced by the fact that the data have been collected from different sources, such as family practices and populations, either social, economic, or demographic. If we would ignore this, identifying interactions between disease variables, such as pathophysiology and laboratory results, could be difficult and even erroneous.

While Bayesian networks model a joint probability distribution, regression methods estimate conditional distributions. Linear regression tries to estimate a linear dependency between the observations of a random continuous variable (assuming it is normally distributed), denoted by O, and a set of (non-random) *explanatory variables*, denoted by e. This is done by using an optimization algorithm, such as the least square method, that minimizes the deviation of the observations with respect to the model parameters.

If, additionally, the data is hierarchically structured, then at each level, the data can be split into *groups*. Characteristics of each group are modeled by additional (non-random) *level variables*, denoted by l. For example, if the different practices are modeled by a grouping variable, a variable such as urbanity that will be shared among practices is modeled by such a level variable. Multilevel analysis tries to explain the variance caused by level variables that have an influence on the explanatory variables e. For example, if we use linear regression, the intercept and slope, that determine the linear dependency between two variables, may alter for different groups.

More precisely, in multilevel regression we wish to explain an observation o with respect to explanations e and l, assuming that the observations o are possible outcomes of a random variable O. Let us first assume that there are only two levels to cope with the grouped data. The explanations e represent the first level, i.e., they can be different for each individual. The second level then represents the groups, which are characterized by the explanations l. The explanations l can thus only differ per group, and together with the explanations e they describe each individual.

Let there be r groups with n first-level explanations and m second-level explanations. Then, for each qth group at the second level we define a linear regression model for O, and allow dependency of the regression coefficients on the variables l_j and certain deviation from the overall mean. With $e = (1, e_1, \ldots, e_i, \ldots, e_n)^T$, $l = (1, l_1, \ldots, l_j, \ldots, l_m)^T$, i.e., $n + m$ explanations, $\delta_q = (\delta_{0q}, \ldots, \delta_{nq})^T$ (the second level noise), for $q = 0, \ldots, r$, and β a matrix consisting of components β_{ij} (the effect of l_j on the explanation e_i), the model then becomes: $E[O_q|e, l] = (\delta_q + \beta l)^T e$, which, if the noise is normally

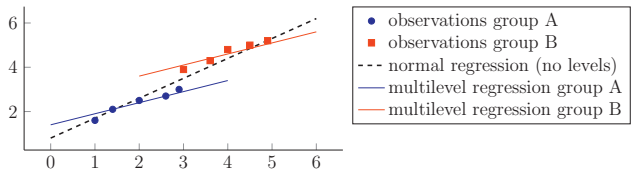


Fig. 2. Multilevel regression, showing that the effect of x on y (the slope) is in fact lower as computed from normal regression. This effect is due to the fact that multilevel regression allows different a priori estimates (β_0) for each group.

distributed, can be interpreted as a conditional probability distribution:

$$P(O_q|e, l)\sim\mathcal{N}(\mu_q, \sigma) \tag{2a}$$

$$\mu_q = (\delta_q + \beta l)^T e \tag{2b}$$

for $q = 0, \ldots, r$, where the expectation of the outcome variable $E[O_q|e, l] = \mu_q$.

In this model, the outcome for each group is dependent of explanatory variables e weighed by the coefficients β, the level variables, and random variables δ_{iq}, where for each i, the δ_{iq} are normally distributed with expectation zero, and correlated with a $\delta_{iq'}$. These correlations ensure that observations for one group have an impact on other groups through this hierarchical structure.

Generally, multilevel models assume homogeneity of variance for all observations on the first level, i.e., σ is constant, and does not depend on e, l, and q. Likewise, it also assumed that the variance on the second level is homogeneous, i.e., the variance of δ_{iq} is equal to σ_i^2, and the covariance of δ_{iq} and $\delta_{i'q}$ is equal to $\sigma_{ii'}^2$, and thus not group specific. But there is no reason why this should be true in all applications. An alternative is to allow heteroscedasticity, i.e., heterogeneity of variances among groups on at least one of the levels. Heteroscedasticity, however, requires additional modeling when estimating the different variances [35–37], and is not described in detail in this paper.

Adding the observations l_j simply to the regression model as additional explanatory variables, i.e., $e = (1, e_1, \ldots, e_n, l_1, \ldots, l_m)^T$, with corresponding regression parameters, i.e., $\beta = (\beta_0, \beta_1, \ldots, \beta_n, \beta_{n+1}, \ldots, \beta_{n+m})^T$, we obtain a one-level regression model with $n + m + 1$ degrees of freedom, which corresponds to standard linear regression. The number of degrees of freedom in the multilevel model is $q(n + 1)(m + 2)$. Fig. 2 compares standard regression and multilevel regression on a synthetic dataset with observations divided into two groups.

The concept can be extended to more levels, e.g., three levels. If the q subgroups can be grouped further into s meta-groups, we can define a three-level model, with $l_1 = (1, l_{21}, \ldots, l_j, \ldots, l_{2m_1})^T$, and $l_2 = (1, l_{21}, \ldots, l_{2k}, \ldots, l_{2m_2})^T$ as the second, and third level variables respectively (the first level is the evidence e), and allow dependency of β on the third level variables as well. The coefficient β is now a three-dimensional array consisting of components β_{ijk}. If the vector γ_{qs}, consisting of elements γ_{iqs}, represents the third level noise (with homogeneity of variances), the model becomes:

$$P(O_{qs}|e, l)\sim\mathcal{N}(\mu_{qs}, \sigma) \tag{3a}$$

$$\mu_{qs} = (\delta_q + ((\gamma_{qs} + \beta l_2)^T l_1)^T e \tag{3b}$$

where again the expectation $E[O_{qs}|e, l] = \mu_{qs}$.

This last model assumes the random outcome variable O to be normally distributed, but in case that O is dichotomous this no longer holds. In this case a specific transformation of the outcome variable, e.g., the logistic function, is assumed to be linear dependent of the explanatory variables. For logistic regression the transformation is given by:

$$\text{logit}\, E[O|e] = \log \frac{E[O|e]}{1 - E[O|e]},$$

and the logistic multilevel model therefore becomes:

$$\text{logit}\, E[O_{qs}|e, l] = (\delta_q + ((\gamma_{qs} + \beta l_2)^T l_1)^T e.$$

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

$$P(O_{qs}|e, l)\sim\text{Bernoulli}(p) \tag{4a}$$

$$\mathit{logit}\, p = (\delta_q + ((\gamma_{qs} + \beta l_2)^T l_1)^T e \tag{4b}$$

When actually doing the multilevel regression we might not want (or expect) an effect of certain higher levels variables on *all* lower level variables. In that case the corresponding component β_{ijk} is fixed to zero, i.e., it is omitted from the model.

Multilevel regression requires less parameters in comparison to standard regression, where the higher level variables are modeled as explanatory variables [3]. Parameters of multilevel regression models can be estimated using an iterative generalized least square (IGLS) method. IGLS is a least square method that estimates the parameters by alternating the optimizing process between the fixed parameters (β_{ij}) and the stochastic parameters (δ_{iq}) until convergence is reached. Goldstein [38] proved that this method is equivalent to the maximum likelihood estimation in standard regression, and improved it to restricted iterative generalized least square (RIGLS) which coincides with restricted maximum likelihood (REML) in Gaussian models [39]. Parameters for dichotomous outcomes are estimated with marginal and penalized quasi-likelihood (MQL/PQL) algorithms [40,41]. Alternatively Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling can be used [42]. Further information and comparison of Bayesian and likelihood-based methods for fitting multilevel models can be found in [43]. Note that, a regression method always tries to fit the model on observed variables only, i.e., it does not consider unobserved variables. For more details about multilevel regression models one is referred to [3].

4. Dealing with multilevel data by Bayesian networks

In this section, we introduce the *multilevel Bayesian network* (MBN) formalism as a new model-based representation of multilevel data. As mentioned in the introduction, this combines the multilevel methodology, used in multilevel regression, with Bayesian networks, in such way that we are able to analyze interactions and probabilistic dependencies between multiple diseases, using patient data obtained from multiple sources, such as family practices.

4.1. Basic ideas

The advantage of a Bayesian network over regression models is that all variables are treated as uncertain, where in regression, including multilevel regression, only the outcome variable is treated as uncertain. If one is primarily interested in the interaction between all relevant variables, and not only in prediction of outcome, in the context of multiple diseases, this is convenient way to model multiple diseases. Furthermore, as multilevel regression models can be seen as conditional probability distributions, they can be used as a factor in a Bayesian network (cf. Eq. (1)). In this section, we explore this relationship by varying the amount of structure in such models and compare this to the multilevel regression approach. However, the first challenge that must be met is

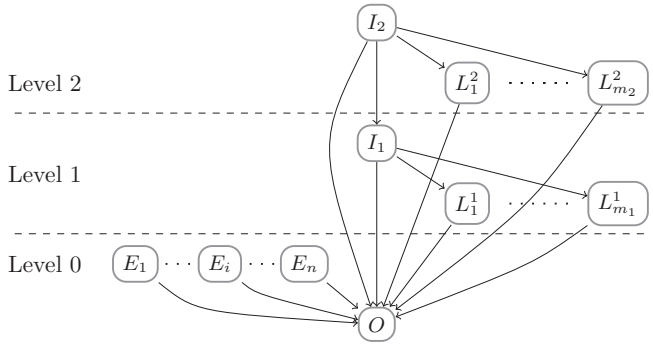


Fig. 3. Bayesian network representation of a multilevel regression model.

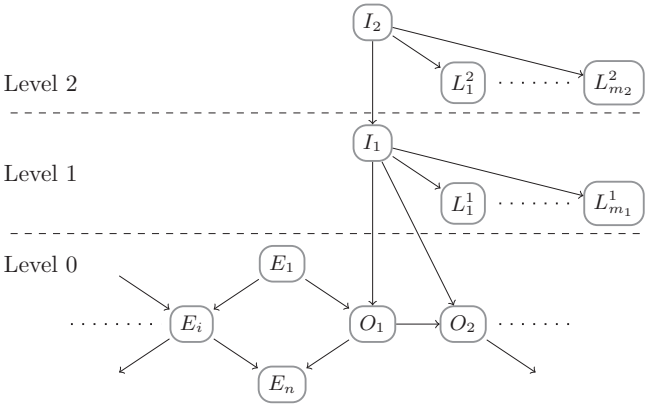


Fig. 4. Multilevel Bayesian network with 3 levels and discrete variables.

the incorporation of multilevel methods in the Bayesian-network framework.

In multilevel regression, the random outcome variable O depends on the vectors of explanations of (non-random) variables, i.e., $e = (e_1, \ldots, e_n)$ and $l^j = (l_1^1, \ldots, l_{m_j}^j)$, with $j = 1, \ldots, m$, (sub)groups q, and $m+1$ different levels. For a Bayesian network approach, we model O as a conditional probability distribution given the set of parents $\{E_1, \ldots, E_n\} \cup \bigcup_{j=1}^{m} L^j$, with $L^j = \{L_1^j, \ldots, L_{m_j}^j\}$, and an *indicator variables* I_j, where $j = 1, \ldots, m$, that selects the group of objects at a certain level j. Fig. 3 shows the corresponding Bayesian network with three levels, assuming no further dependence between variables. Clearly, this model is still too restrictive for most health-care applications, as no structure is present between the explanatory variables and we have only one outcome variable of interest.

The idea of a multilevel Bayesian network is that the indicator variables I split the domain into different categories with a deterministic effect on the group variables L that are constant for a given category chosen by I. If not present, I variables can be constructed, e.g., by the Cartesian product in case of categorical L variables. However, multilevel analysis, and thus a multilevel Bayesian network, is typically designed for hierarchically structured data, and then the indicator I variables are part of the database definition.

Some of the explanatory variables are group-independent, though structure may exist between these variables. These variables correspond with the set of variables E in an MBN. Other variables, depend both on grouping and other variables at the same or higher levels. These variables correspond with the set of variables O in an MBN. 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 present in multilevel regression methods. Because of the deterministic relations we are able to simplify the structure of the MBN using the following property.

Lemma 1. *Let X and Y be two random variables such that Y is deterministically dependent of X, i.e., there exists some function f such that $Y = f(X)$. Then, for all sets of random variables Z disjoint of X and Y it holds that $Z \perp\!\!\!\perp_P Y|X$.*

Proof. Take some arbitrary Z. If it is a discrete distribution, then it holds that:

$$P(Z|X) = \sum_Y P(Z, Y|X) = \sum_Y P(Z|X, Y)P(Y|X)$$

By the relationship between X and Y, it holds $P(Y|X) = 1$ if $Y = f(X)$, and 0 otherwise, so it follows that:

$$P(Z|X) = P(Z|X, f(X)) = P(Z|X, Y)$$

Similarly, for continuous distributions, we have:

$$p(Z|X) = \int p(Z, Y|X)\mathrm{d}Y = \int p(Z|X, Y)p(Y|X)\mathrm{d}Y$$
$$= \int p(Z|X, Y)\delta(Y - f(X))\mathrm{d}Y = p(Z|X, f(X)) = p(Z|X, Y)$$

where δ is the Dirac delta function. □

We can apply this lemma to our initial MBN for two cases. Since $P(L_i^j|I_j)$ is deterministic, we obtain $O \perp\!\!\!\perp_P L_i^j|I_j$. The implication of this, is that no arcs exist between the group vertices in L and the outcome and explanatory vertices in $O \cup E$. Since the probability distribution $P(I_{j+1}|I_j)$ is deterministic too, we obtain $O \perp\!\!\!\perp_P I_j|I_1$ for all j. The implication of this is that within the indicator vertices I there are only arcs from I_{j+1} to I_j, for all j, and between the indicator vertices I and outcomes O there are only direct arc from I_1 to any O_i. These restrictions greatly simplify the structure of an MBN. When making predictions based on the parameters of the MBN, the indicator variables are mostly unknown. However, the structure still allows us to use the higher level variables to explain the outcome variable.

We now give a precise definition of MBNs. To shorten the definition, members of the various sets S are denoted by S_i, and $S \setminus \{S_i\}$ with $S - S_i$.

Definition 1. A Bayesian network $\mathcal{B} = (G, X_V, P)$ is a *multilevel Bayesian network*, or MBN for short, if its set of vertices V is described by the tuple (m, O, E, L, I), with pairwise disjoint sets $O, E, L, I \subseteq V_G$, such that:

- $m \in \mathbb{N}$ denotes the *number of levels* of the MBN, where level 0 is called the *base level*;
- O, the set of *outcome variables*, is at base level such that if $(V \to O_i) \in A_G$, then $V \in E \cup (O - O_i) \cup I$;
- E, the set of *explanatory variables*, is at base level, such that if $(V \to E_i) \in A_G$, then $V \in (E - E_i) \cup O$;
- $L = \{L^1, \ldots, L^m\}$, where each L^j is a set of group variables at level $j \geq 1$. For group variable L_i^j it holds that
  1. $(V \to L_i^j) \in A_G$ implies that $V = I_j$;
  2. $P(L_i^j|I_j)$ is deterministic.
- $I = \{I_1, \ldots, I_m\}$ are *indicator variables*, such that I_j is the only parent of I_{j-1} in G, for all $1 \leq j \leq m$, and $P(I_{j-1}|I_j)$ is deterministic;
- $X_V = \{X_v | v \in (I \cup E \cup O \cup L)\}$.

Fig. 4 offers graphical illustration of the definition. Note that, within one MBN multiple diseases can be modeled as outcome variable. By Lemma 1, the outcome variables O are independent

of the level variables L given the value of the I variables by lemma. However, this does not imply these are variables are meaningless. Once the parameters of the MBN are learned, it can be used to estimate the variance that is introduced by such a level variable on the probability distribution of outcome variables, without knowing the value of the indicator variable.

4.2. Probability distributions for multilevel Bayesian networks

Without taking into account the level variables, the probability of the outcome variables O conditioned on the explanatory variables E can be obtained by

$$P(O = o|E = e) = f_{O|E}(o|e; \beta) = \frac{f_{O,E}(o, e; \beta)}{\sum_o f_{O,E}(o, e; \beta)},$$

if O is discrete, and

$$P(O \leq o|E = e) = \int_{-\infty}^{o} f_{O|E}(x|e; \beta)\mathrm{d}x = \frac{\int_{-\infty}^{o} f_{O,E}(x, e; \beta)\mathrm{d}x}{\int_{-\infty}^{\infty} f_{O,E}(x, e; \beta)\mathrm{d}x}$$

if O is continuous. The parameter β represents the parameters typically used for a specific distribution, e.g., $\beta = (\mu, \sigma)$ in case $f_{O,E}(o, e; \beta)$ is a Gaussian distribution with mean μ and variance σ.

In a multilevel Bayesian network the grouping variable splits the conditional probability distributions between an outcome variables and its explanatory variables into multiple (countable) distributions keeping them closely related, i.e., only the distribution type dependent parameters differ between groups. In case O is discrete we obtain $P(O = o|E = e, I = i) = f_{O|E}(o|e; \beta_i)$, and likewise, if O is continuous we obtain $P(O \leq o|E = e, I = i) = \int_{-\infty}^{o} f_{O|E}(x|e; \beta_i)\mathrm{d}x$.

For example, in case O and $E = e$ are both discrete and O is binary with a Bernoulli distribution with parameter $\beta = p_{e,i}$, we obtain:

$$P(O = o|E = e, I = i) = f_{O|E,I}(o|e, i; \beta) = \text{Bernoulli}(p_{e,i})$$

$$= \begin{cases} p_{e,i} & \text{if } O = o \\ 1 - p_{e,i} & \text{otherwise} \end{cases}$$

In case O and E are both continuous and O follows a Gaussian distribution, we obtain the probability density function:

$$f_{O|E,I}(o|e, i; \beta) = \mathcal{N}(\mu_{e,i}, \sigma)$$

Just as in multilevel linear regression, a linear dependency between E and O can be obtained if $\mu_{e,i} = \beta_i e$, also for E being a discrete variable.

In case O is discrete and E is continuous a link function is used in multilevel regression, to keep the linearity in the model, of which the logistic function is the most popular one. The probability mass function for such a discrete variable with a continuous parent is:

$$f_{O|E,I}(o|e, i) = \frac{\exp(\beta^i_{o0} + \beta^i_{o1} e)}{\sum_o \exp(\beta^i_{o0} + \beta^i_{o1} e)}$$

For binary outcome variables this reduces to:

$$f_{O|E,I}(o|e, i) = [1 + \exp(\beta^i_0 + \beta^i_1 e)]^{-1}$$

5. Experimental methodology

In the previous section, the basic ingredients of multilevel Bayesian networks were outlined. In this section, we take the step in making the technique practically useful. At the end of this section, we demonstrate that the methodology works by using synthetic data. In the next section, the same is done, but then for a dataset obtained from a public health registry containing patient data from general practices.

5.1. Parameter learning

Because we have incorporated the multilevel regression model as factors in the model, we can make use of multilevel regression to estimate the outcome variables. This has the advantage that we exploit the correlation between different groups (if it exists) and therefore requires less data per group than a standard Bayesian network learning algorithm needs for parameter learning per group. For multilevel-level logistic regression models, it is recommended to use a minimum group size of 50 with at least 50 groups to produce valid estimates [44]. An exact inference algorithm for parameter estimation in networks with discrete children of continuous parents is proposed in [45]. Compared to multilevel regression models, it is also possible to use a Bayesian approach for learning the parameters [46] and therefore include even more domain knowledge to the model.

5.2. Model validation

Possible criteria to validate the model parameters are the Akaike information criteria (AIC) [47], the Bayesian information criteria (BIC) [48], and the deviance information criteria (DIC) [49]. The AIC and BIC are widely accepted decision criteria, but computationally expensive when dealing with large amounts of data and MCMC methods. This problem is overcome using the DIC, which calculates deviance residuals, that sum up to the deviance statistic, along with the MCMC process. Unfortunately, in disease mapping, DIC is in favor of overparameterized models, especially when using large datasets [50].

Alternatively, an approximation method proposed by [51] can be used, which works very well for large data sets in an MCMC setting. It uses replication of the stochastic parameters and the outcome variables for a specified part of the data along with the MCMC simulation based on the remaining part of the data. The replicate outcome variables can then be compared to the real outcomes, allowing us to assess the predictability of the model.

Although computationally expensive as well, standard cross validation (e.g., k-fold cross validation) is a robust method to validate regression and Bayesian models [52], and receiver operating characteristic (ROC) analysis can be used to validate accuracy and precision of the model parameters. Recently, a new measure was introduced, the net reclassification improvement (NRI), offering additional incremental information compared to the area under the curve (AUC) within an ROC analysis [53], which provides more insight into risk prediction.

5.3. Structure learning

In order to build the structure between 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., [54]). For example, the necessary edges between I_1 and all variables $O_i \in O$ are whitelisted, whereas edges from a lower level to a higher level are all blacklisted.

A systematic approach to identify statistically significant edges in a network, has been developed by Friedman et al. using bootstrap resampling and model averaging [55]. The empirical probability of an edge, defined as the fraction of occurrences in the networks learned from bootstrapped samples, are known as edge intensities (or strengths), and can be interpreted as the degree of confidence that the edge is present in the true network structure describing the true dependence structure of the original data. Scutari et al. propose

a statistically motivated estimator for the confidence threshold minimizing a specific norm between the cumulative distribution function of the observed confidence levels and the cumulative distribution function of the confidence levels of the unknown true network [56]. Classical norms are the rectilinear distance, denoted as the L_1 norm, and the Euclidean distance, denoted as the L_2 norm [57].

5.4. Artificial multimorbidity example with synthetic data

Suppose we have the variables D_1, D_2, and D_3 that model whether the diseases D_1, D_2, and D_3 are present, a genetic variable G, and two demographic vertices L_1 and L_2 that model certain environmental conditions. Furthermore, let the demographics be variables obtained from higher levels in a hierarchically structured dataset, i.e., L_1 and L_2 are level-2 and level-3 variables respectively. For example, the variables D_1, D_2, and D_3 could represent diseases like *diabetes*, *retinopathy*, and *hypertension*. The variable G could represent *gender*, or a specific *gene*, and the grouping variable I_1 could represent a division in *practices* with *type* as L_1, and I_2 a division in *area* with *urbanity* as L_2.

There are fifty *practices* ($I_1 \in \{1, \ldots, 50\}$) and five *areas* ($I_2 \in \{1, 2, 3, 4, 5\}$). The variable *type* ($L_1$) has 5 possible values and the variable *urbanity* (L_2) is binary. The deterministic relations between them are:

$$I_2 = \begin{cases} 1 & \text{if } I_1 \in \{1, \ldots, 10\} \\ 2 & \text{if } I_1 \in \{11, \ldots, 20\} \\ 3 & \text{if } I_1 \in \{21, \ldots, 30\} \\ 4 & \text{if } I_1 \in \{31, \ldots, 40\} \\ 5 & \text{if } I_1 \in \{41, \ldots, 50\} \end{cases}$$

and

$$U_1 = \begin{cases} 1 & \text{if } I_1 \bmod 10 \in \{0, 1\} \\ 2 & \text{if } I_1 \bmod 10 \in \{2, 3\} \\ 3 & \text{if } I_1 \bmod 10 \in \{4, 5\} \\ 4 & \text{if } I_1 \bmod 10 \in \{6, 7\} \\ 5 & \text{if } I_1 \bmod 10 \in \{8, 9\} \end{cases}$$

and

$$U_2 = \begin{cases} 0 & \text{if } I_2 \in \{1, 2\} \\ 1 & \text{if } I_2 \in \{3, 4, 5\} \end{cases}$$

We sampled 10,000 patients uniformly over the fifty practices and determined its respective values for the other higher level variables. The binary variable G is binomially sampled with a probability of 0.50. The diseases D_1, D_2 and D_3 are sampled as follows.

$$D_1 = \text{Binomial}(0.50 + 0.01G + \mathcal{N}(\mu_q, \sigma_q))$$

$$D_2 = \text{Binomial}(0.20 + \mathcal{N}(\mu_s, \sigma_s))$$

$$D_3 = \text{Binomial}(0.20 + 0.1D1 + 0.2D2 + 0.2D1D2 + \mathcal{N}(\mu_{qs}, \sigma_{qs}))$$

with $q = 1, \ldots, 50$, and $s = 1, \ldots, 5$, corresponding to the number of *practices* and *areas*. The distributions $\mathcal{N}(\mu_q, \sigma_q)$ and $\mathcal{N}(\mu_{qs}, \sigma_{qs})$ are randomly sampled from a $\mathcal{N}(0, 0.1)$ distribution, μ_s is 0.25, 0.30, 0.35, 0.40, and 0.45, for $s = 1, \ldots, 5$ respectively, and $\sigma_s = 0.01$.

Applying multilevel regression, if we, for example, only allow an influence of the level-2 and level-3 variables on the intercept

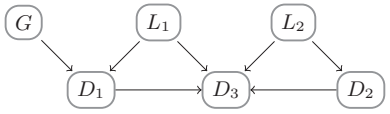


Fig. 5. Bayesian network representing probabilistic dependencies between certain diseases (D_1, D_2, D_3), a genetic variable G, and some demographics (L_1, L_2).

and the regression coefficient of the explanatory variable D_1, the multilevel regression model becomes:

$$P(D_{3qs}|d_1, d_2, g, l_1, l_2) \sim \text{Bernoulli}(p) \tag{5.1a}$$

$$\text{logit}\, p = \beta_{0qs} + \beta_{1qs}d_1 + \beta_2 d_2 + \beta_3 g \tag{5.1b}$$

$$\beta_{0qs} = \beta_{00s} + \beta_{01s}l_1 + \delta_{0q} \tag{5.2a}$$

$$\beta_{1qs} = \beta_{10s} + \beta_{11s}l_1 + \delta_{1q} \tag{5.2b}$$

$$\beta_{00s} = \beta_{000} + \beta_{001}l_2 + \gamma_{00s} \tag{5.3a}$$

$$\beta_{01s} = \beta_{010} + \beta_{011}l_2 + \gamma_{01s} \tag{5.3b}$$

$$\beta_{10s} = \beta_{100} + \beta_{101}l_2 + \gamma_{10s} \tag{5.3c}$$

$$\beta_{11s} = \beta_{110} + \beta_{111}l_2 + \gamma_{11s} \tag{5.3d}$$

With $\delta_{iq} \sim N(0, \sigma_{iq})$ and $\gamma_{iqs} \sim N(0, \sigma_{iqs})$. Substituting Eq. (5.3) into (5.1), and Eq. (5.2) into (5.1), Eq. (5.1) becomes:

$$\begin{aligned} \text{logit}\, p = {} & (\beta_{000} + \beta_{001}l_2 + \gamma_{00s} + (\beta_{011}l_2 + \gamma_{01s} + \beta_{010})l_1 + \delta_{0q}) \\ & + (\beta_{100} + \beta_{101}l_2 + \gamma_{10s} + (\beta_{111}l_2 + \gamma_{11s} + \beta_{110})l_1 + \delta_{1q})d_1 \\ & + \beta_2 d_2 + \beta_3 g \end{aligned}$$

Since L_1 and L_2 are discrete variables, and have the same value within a group, we can rewrite this into:

$$\begin{aligned} \text{logit}\, p = {} & (\beta'_0 + \beta'_{0s} + \gamma'_{0s} + \beta'_{0qs} + \gamma'_{0qs} + \beta'_{0q} + \delta_{0q}) + (\beta'_1 + \beta'_{1s} + \gamma'_{1s} \\ & + \beta'_{1qs} + \gamma'_{1qs} + \beta'_{1q} + \delta_{1q})d_1 + \beta_2 d_2 + \beta_3 g \end{aligned}$$

Now, assume that using structure learning (without using the indicator variables) it is observed that $\pi(G) = \pi(L_1) = \pi(L_2) = \varnothing$, $\pi(D_1) = \{G, L_1\}$, $\pi(D_2) = \{L_2\}$, and $\pi(D_3) = \{D_1, D_2, L_1, L_2\}$. Fig. 5 then shows the corresponding Bayesian network and the joint distribution $P(V)$ is given by

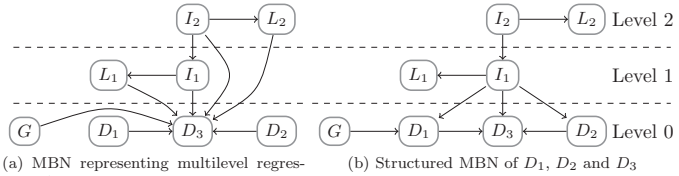
$$P(D_3|D_1, D_2, L_1, L_2)P(D_1|L_1, G)P(D_2|L_2)P(L_1)P(L_2)P(G)$$

To predict whether a disease D_3 is present given that L_1, L_2 and G are known, we have by Eq. (1) and standard probability theory:

$$P(D_3|L_1, L_2, G) = \sum_{D_1, D_2} P(D_3|D_1, D_2, L_1, L_2)P(D_1|L_1, G)P(D_2|L_2)$$

Since the Markov blanket of D_3 is $\{D_1, D_2, L_1, L_2\}$, any information about the genetic variation of a person is irrelevant, i.e., since $D_3 \perp\!\!\!\perp_P G|D_1$ we obtain: $P(D_3|D_1, D_2, L_1, L_2, G) = P(D_3|D_1, D_2, L_1, L_2)$.

Applying the MBN techniques, Fig. 6 shows the corresponding MBN representations. One can see that in the multilevel regression



(a) MBN representing multilevel regression of D_3 (b) Structured MBN of D_1, D_2 and D_3

Fig. 6. MBN representations of the example in Fig. 5. (a) MBN representing multilevel regression of D_3. (b) Structured MBN of D_1, D_2 and D_3.

Table 1
Probability estimations of D_3 conditioned on D_1, D_2 and L_1.

	$L_1=1$	$L_1=2$	$L_1=3$	$L_1=4$	$L_1=5$
(a) True probability distributions in the test set					
$D_1=0, D_2=0$	0.150	0.175	0.200	0.225	0.250
$D_1=0, D_2=1$	0.250	0.275	0.300	0.325	0.350
$D_1=1, D_2=0$	0.350	0.375	0.400	0.425	0.450
$D_1=1, D_2=1$	0.650	0.675	0.700	0.725	0.750
(b) Multilevel logistic regression (Eq. (5))					
$D_1=0, D_2=0$	0.135	0.137	0.171	0.187	0.198
$D_1=0, D_2=1$	0.279	0.281	0.319	0.335	0.346
$D_1=1, D_2=0$	0.410	0.414	0.479	0.505	0.523
$D_1=1, D_2=1$	0.632	0.635	0.676	0.692	0.702
(c) Structured multilevel Bayesian network (Fig. 6(b))					
$D_1=0, D_2=0$	0.164	0.148	0.189	0.228	0.218
$D_1=0, D_2=1$	0.271	0.242	0.286	0.304	0.331
$D_1=1, D_2=0$	0.330	0.398	0.442	0.453	0.466
$D_1=1, D_2=1$	0.653	0.680	0.747	0.735	0.749

network (Fig. 6(a)) only D_3 is modeled as an outcome variable of interest, as where in the structured model (Fig. 6(b)) D_1 and D_2 are modeled as outcome variables as well (still being an explanatory variables of D_3). As a consequence of Theorem 1, the disease variables in Fig. 6(b) do not have edges from I_2 and L_i directed toward themselves.

The comparison between the multilevel regression technique and the structured multilevel Bayesian network is outlined in Table 1, showing the probability of disease D_3 in the presence of L_1, D_1 and D_2. Parameters of the multilevel regression model are obtained with the MLWin software, in which the algorithms described at the end of Section 3.3 are implemented [58]. Parameters of the MBN are learned using the *bnlearn* package [54] in the statistical software R.

Using AIC and BIC, the most accurate multilevel logistic regression model allows random intercepts and random slopes on D_1 for each entry of $L1$. Although the probabilities derived from the MBN are closer to the true probabilities, the area under the curves (AUCs) within an ROC analysis are close together, i.e., 0.725 and 0.712 for the MBN and multilevel regression respectively. In the multilevel regression all variables are used for prediction, whereas for the MBN only the variables of the Markov blanket are used for prediction.

The net reclassification improvement is in favor of the MBN, i.e., the NRI is 0.2144 ($p<0.001$). Thus, on average the MBN is significantly better then the multilevel regression approach in this synthetic example. This due to the fact that an MBN is able to give an exact solution with respect to a dependency structure between variables and its observations. Multilevel regression does not have these dependency constraints, which possibly favors overfitting the model.

6. Modeling inter-practice variation in multimorbidity

Normally, in scientific research, one would investigate diseases separately, resulting in different predictive values of variables shared by both diseases. For example, multilevel regression analysis was recently used by Nielen et al. to investigate the influence of particular family practice variables on hypertension and diabetes mellitus, revealing an inter-practice variance in predictability [24]. However, since interactions could have an additive effect on prevalence, this yields no insight into the predictive value in case both diseases are present. In fact, we need an extra regression model on the combined diagnosis of hypertension and diabetes together to be able to make such conclusions.

In this paper, we will use the research of Nielen et al. as starting point. Firstly, we compare the parameter estimations of an unstructured MBN with multilevel regression. Secondly, we compare the predictive power of a structured MBN with multilevel regression.

6.1. Description of the models

To evaluate if the parameter estimations of an MBN are comparable with a multilevel regression we analyzed models for both diabetes mellitus and heart failure. Nielen et al. analyzed hypertension instead of heart failure. However, besides the validation of parameter estimations, we also want to investigate the predictive power for diseases that have a different onset during life. Heart failure is known to be associated with diabetes mellitus and hypertension [59], and its risk management involves almost the same variables [60]. Since the onset of hypertension and diabetes mellitus is typically earlier in the patient's life than the onset of heart failure it is in our interest if the finally structured MBN follows these associations.

We used five models for the analysis. The first two models are the multilevel regression models for predicting either diabetes mellitus (model MLR-DM) or heart failure (model MLR-HF) using data which is grouped by practice, where the urbanity of the practice is modeled as higher level variable. The next two models (MBN-DM and MBN-HF) are the corresponding unstructured MBNs for the first two models, assuming no further dependencies between variables exist (cf. Fig. 3), and that the urbanity is independent of the disease, given the practice (cf. Lemma 1). Finally, we consider a structured model (MBN-STR) which contains both diseases as well as structure between the outcome and explanatory variables, which we call intra-level structure.

All five models use practice and urbanity as higher level variables. Since the practices use different types of information systems, one might argue this is of influence on the predictions. To model this, a second level grouping variable (the used information system) can be incorporated on top of the first level grouping variable (practice). However, it turns out that there is no significant benefit when doing so. Therefore this idea is omitted for further analysis.

6.2. Research problem and data

The patient data was routinely collected by the Netherlands information network of general practice (LINH). In 1996, they started as a registry 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. Currently, the LINH database contains information of routinely collected data from approximately 90 family practices out of several different information systems. 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 218,333 patients from 82 Dutch general practices, meaning an average number of patients around 2650 per practice. Morbidity data were derived from diagnoses, using the international classification of primary care (ICPC) and anatomical therapeutic chemical (ATC) codes.

6.3. Unstructured MBNs compared to multilevel regression

For both the multilevel regression models MLR-DM and MLR-HF we estimated the parameters using MLWin [58]. For the models MBN-DM and MBN-HF we used MCMC simulation, available in the WinBUGS software [46]. All variables were discretized and modeled using a Bernoulli distribution. Parameter estimates using a 10-fold cross validation are presented in Table 2. As expected, the results of the unstructured MBN models are similar to the results obtained by

Table 2
Parameter estimations of explanatory (parent) variables, represented as odds ratios, using cross validation in a multilevel analysis for diabetes mellitus and heart failure (MLR = multilevel regression, MBN = multilevel Bayesian network, DM = diabetes mellitus, HF = heart failure).

	Diabetes mellitus		Heart failure	
Model	MLR-DM	MBN-DM	MLR-HF	MBN-HF
Age	1.029	1.028	1.106	1.106
Gender (ref = male)	0.914	0.915	0.823	0.815
Overweight/obesity	1.725	1.671	1.689	1.600
Diabetes mellitus	–	–	1.256	1.260
Lipid disorder	6.437	6.392	1.172	1.183
Hypertension	5.675	5.800	2.071	2.067
Peripheral artery disease	0.954	0.949	1.619	1.530
Heart failure	1.132	1.194	–	–
Retinopathy	9.253	9.669	1.310	1.104
Angina pectoris	0.679	0.665	2.214	2.184
Stroke/CVA	0.770	0.766	1.388	1.397
Renal disease	1.176	1.200	1.878	1.881
Cardiovascular symptoms	0.848	0.850	2.596	2.636
Urbanity (ref = urban)				
Urban	1.000	1.000	1.000	1.000
Strongly urban	1.261	1.275	1.145	1.158
Modestly urban	1.477	1.490	1.181	1.192
Little urban	1.436	1.408	1.422	1.456
Not urban	1.474	1.259	1.335	1.318

multilevel regression, showing that multilevel Bayesian networks are a valid alternative method for multilevel analysis.

6.4. Composition of the structured MBN

The structure of the MBN-STR model is learned using the *bnlearn* package [54] in the statistical software R, which provides various methods for structure learning. We have restricted the search of Bayesian networks to those that satisfy the multilevel structure by using white- and blacklists. See Fig. 7 for the resulting Bayesian network structure. Note that indeed there is only a dependency between consecutive levels, and that this is solely through the grouping variables. Furthermore, it turned out that only a subset of the disease variables depends on the practice variable, of which diabetes mellitus is amongst them whereas heart failure is not. So technically diabetes mellitus is an outcome variable and heart failure is an explanatory variable within the definition of an MBN.

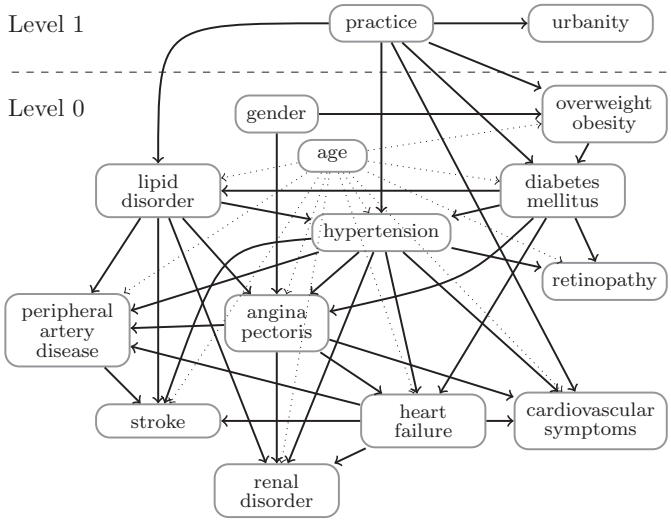


Fig. 7. Structure learning without any domain knowledge of cardiovascular diseases and diabetes mellitus in family practices. The dotted arcs are arcs from 'age' in order to make the model more readable.

However, since all variable can be treated as uncertain we can still use the model to make predictions for heart failure.

Some of the directions of certain edges is opposite to what the domain experts would expect, e.g., angina pectoris is pointing toward peripheral artery disease (PAD), but in reality this is seen as a comorbidity due to atherosclerosis, which itself is not present in the model. Therefore, we also incorporated some domain knowledge [59,60] into the model and allowed a geriatric specialist and two physicians to validate the model. Removed edges are: *angina pectoris* → *PAD*, *angina pectoris* → *renal disease*, *heart failure* → *PAD*, and *practice* → *cardiovascular symptoms*. The edge *heart failure* → *renal disease* is reversed. The final model is showed in Fig. 8, along with the prior probability distributions for patients aged over 65 years. However, these results are of a preliminary nature, and we did not study the validity of the structured model further.

Using bootstrapped samples to validate the strengths of the edges, most edges shown in the network of Fig. 2 appear in more than 95% of the networks learned from the samples. The only edges with a percentage lower than 95% is *renal disease* → *heart failure* (0.73%). Most of the edges not present in the originally learned structure have an appearance close to 0%.

In this model the prevalence rate of diagnosed diabetes mellitus in practices varies between 0.008 and 0.135, with mean 0.077 and standard deviation 0.025. The prevalence of heart failure varies between 0.001 and 0.059, with mean 0.019 and standard deviation 0.011. Fig. 9 shows the same model as in Fig. 8, but now conditioned on hypertension and diabetes, i.e., both diseases are present. In this case probabilities are more or less doubled (or tripled in case of lipid disorder), indicating the population of elderly patients with both hypertension and diabetes have twice the chance of getting an additional cardiovascular disease when compared to the general elderly population. For this population, i.e., diabetics with hypertension, the prevalence of heart failure varies between 0.001 and 0.230, with mean 0.086 and standard deviation 0.049.

Finally, the conditional probability distribution of a disease node can be used to uncover interactions between diseases. If we calculate the probability of angina pectoris (ap) in the presence of both hypertension (ht) and dyslipidemia (dl), we obtain: $P(ap|ht, dl) \approx 16\,\%$. It turns out that this is much higher than one can expect from the other probabilities: $P(ap|ht, \overline{dl}) \approx 7\%$, $P(ap|\overline{ht}, dl) \approx 5\%$ and $P(ap|\overline{ht}, \overline{dl}) \approx 1\%$. We can do this exercise for an arbitrary disease and (a subset of) its parents in the MBN structure. For example, when looking at heart failure (hf), there is an interaction between hypertension and diabetes mellitus (dm): $P(hf|ht, dm) \approx 9\,\%$, $P(hf|ht, \overline{dm}) \approx 5\%$, $P(hf|\overline{ht}, dm) < 1\%$ and $P(hf|\overline{ht}, \overline{dm}) < 1\%$, which suggest that the effect of diabetes on heart failure is only of clinical significance in the presence of hypertension.

6.5. Comparison of the structured MBN with multilevel regression

Besides the estimation of odds, a more practical question is how well the model can be used for prediction. For this, we compared the predictive performance of the MBN-STR model to multilevel regression analysis for single diseases, i.e., the models MLR-DM and MLR-HF.

For the multilevel regression method, we used all the predictors, while for the MBN-STR model, we can restrict ourselves to the Markov blankets (cf. Section 3.1) of the diseases and higher level variables where necessary. For diabetes mellitus, the MB consists of practice, age, gender, obesity, lipid disorder, hypertension, heart failure, retinopathy, and renal disorder. However, making predictions in a multilevel model we treat the indicators, i.e., the practice, as uncertain, and instead we have to use the urbanity for prediction as well. The MB of heart failure on the other hand consists of age,

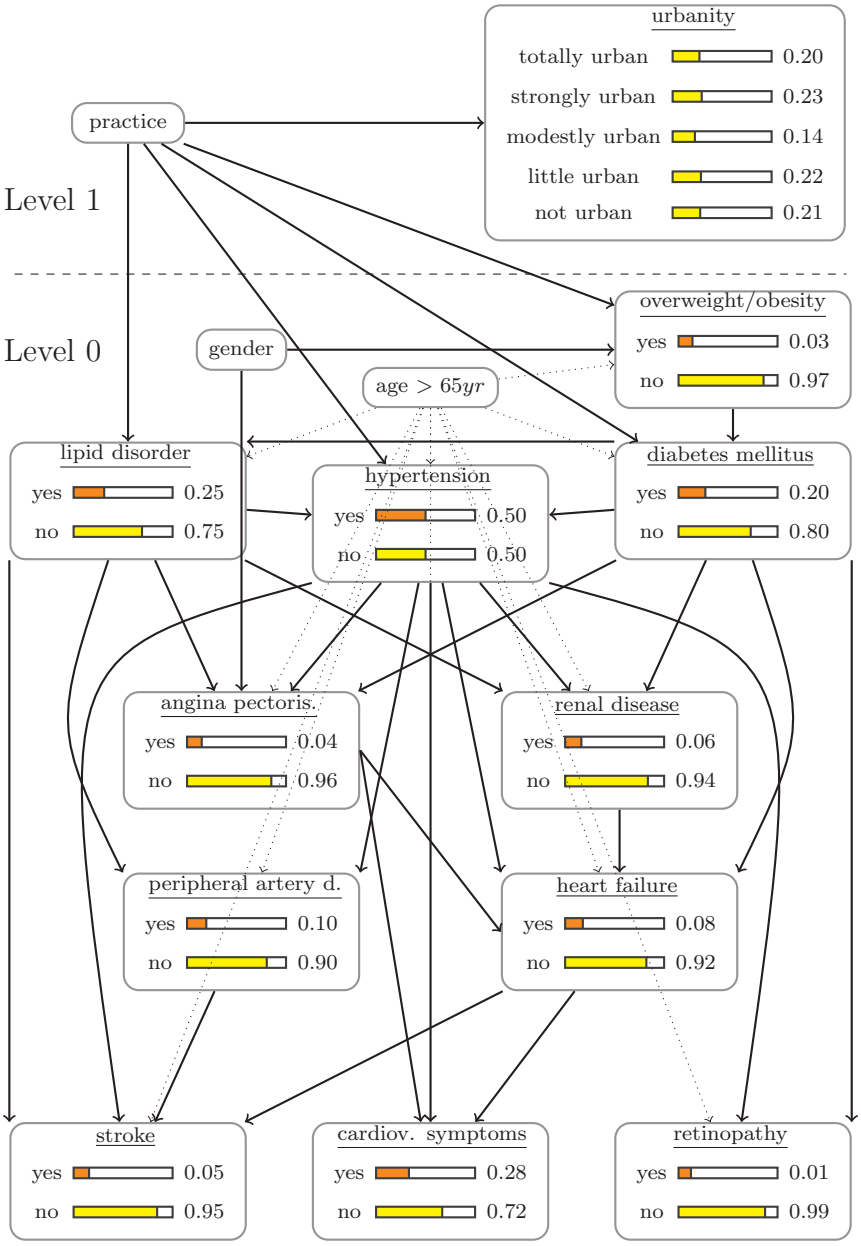


Fig. 8. Structured MBN with prior probability distributions for patients aged >65 years, using domain knowledge (expert opinions/evidence from other research) of cardiovascular diseases and diabetes mellitus in family practices. The dotted arcs are arcs from 'age' and 'gender' in order to make the model more readable.

gender, lipid disorder, diabetes mellitus, hypertension, peripheral artery disease, angina pectoris, stroke, renal disorder, and cardiovascular symptoms. For heart failure no higher level variables are needed for prediction when the diseases that vary along such variables are known, e.g., obesity, hypertension, and diabetes.

To measure the accuracy of the predictions we performed an ROC analysis (see Fig. 10). When comparing the AUC between multilevel regression and the MBN-STR model, the ones for the MBN-STR model are slightly better with a difference of approximately 1%. For the MBN-STR they are approximately 0.90 and 0.84 for diabetes mellitus and heart failure respectively. For the MLR-DM it is 0.89 and for the MLR-HF it is 0.83. When performing a net reclassification improvement analysis for the MBN-STR model compared to the multilevel regression models MLR-DM and MLR-HF, the NRI is significantly positive in both cases, i.e., the NRI is 0.723 ($p < 0.001$) for diabetes and 0.075 ($p < 0.01$) for heart failure.

7. Discussion

In this paper, we have presented a new approach to model multilevel data, and applied this to health care data of general practices. As we have discussed, such data often contain a hierarchical structure, which can be modeled by using different levels of data, e.g., patient data collected from multiple general practices. Since traditional multilevel regression methods only allow one outcome variable each time, which is unpractical in the context of multiple diseases, we combined Bayesian networks with multilevel analysis yielding multilevel Bayesian networks, which allows uncertainty of all disease variables into one model.

Furthermore, we can add intra-level structures between variables giving extra insight into probabilistic dependencies and interactions. Moreover, certain domain knowledge can be incorporated, e.g., edges between pathophysiology and its corresponding lab results are always pointing to the latter, making the model more

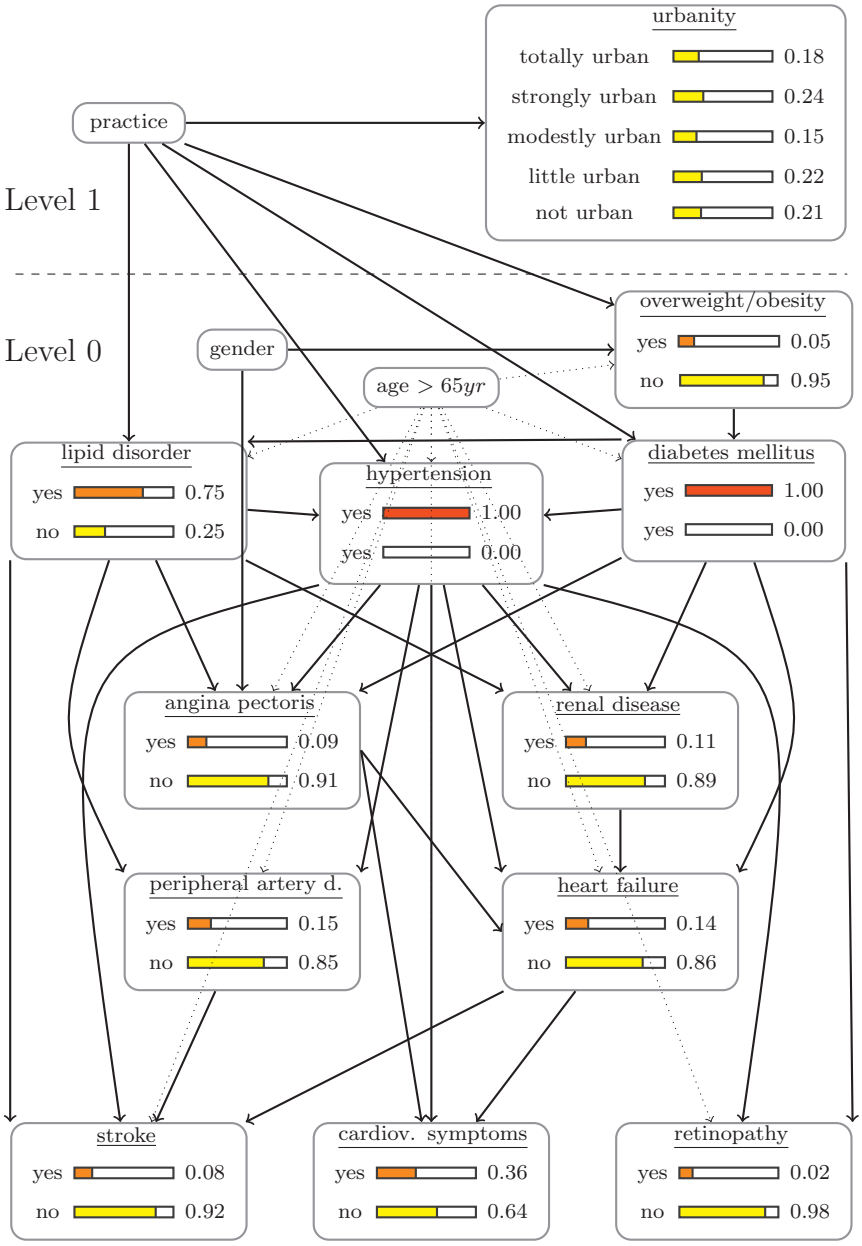


Fig. 9. Structured MBN (cf. Fig. 8) with posterior probability distributions for patients with both hypertension and diabetes (aged >65 years).

easy to interpret. Such domain knowledge can be used during the learning of the structure of a Bayesian network by restricting the search space.

In this paper, we have shown that a multilevel Bayesian network can do the same as traditional multilevel regression methods. We do not claim exact equivalence, but using synthetic data and a real-world application of MBNs with clinical patient data from family practices, we showed the empirical equivalence of a traditional multilevel regression model to an unstructured MBN. Furthermore, structured MBNs provide insight into the relationship between multiple diseases and allows for studying multiple diseases at the same time, avoiding the redundancy of regression methods (when used to analyze multiple disease in the same variable set).

Although it is not our main purpose to provide a better classifier, the predictive value of a structured MBN is just as good as multilevel regression analysis, despite a reduced number of predictors, i.e., the Markov blanket. Both in the synthetic example and the real life applications of diabetes mellitus and heart failure, there is a small improvement in the AUC and a significantly positive NRI. Bootstrapped samples showed that the strength of the edges between disease variables in the network representation of diabetes mellitus and heart failure is mostly close to 100%, meaning we can be confident about the found structure.

Using the learned MBN we are able to condition on certain disease variables, e.g., when conditioning on hypertension and diabetes, the MBN reveals that chances on obtaining another cardiovascular disease, such as heart failure, is more or less doubled. This 'personalization' of the network could be seen as a step forward to personalized clinical guidelines, as mentioned in the introduction, making the MBN a promising tool in the new domain of multimorbidity. Further research will focus on the application of the MBN framework to relevant clinical questions within Public Health and the related multimorbidity issues.

Finally, 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

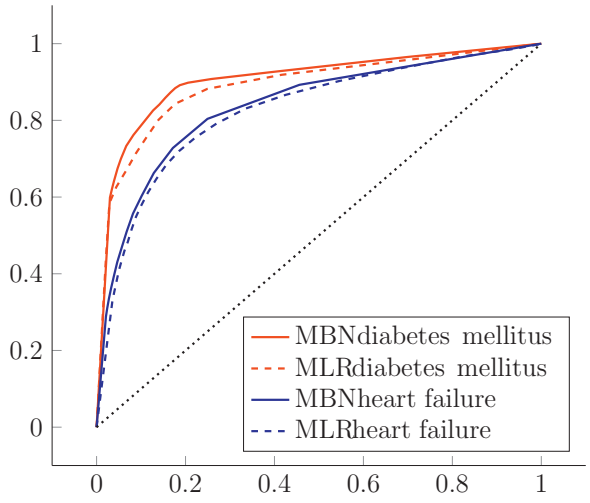


Fig. 10. ROC analysis of a structured multilevel Bayesian network (MBN) and multilevel regression (MLR) for diabetes mellitus and heart failure.

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. As an example, atherosclerosis may be added to the model, and, using the method proposed in [61], this variable may capture important combinations of observations, e.g., peripheral artery disease along with a cardiac disease such as angina pectoris. This may improve the prediction performance of these models further.

References

[1] Rice N, Leyland A. Multilevel models: applications to health data. Journal of Health Services Research & Policy 1996;1(3):154–64.
[2] Austin P, Goel V, Walraven C. An introduction to multilevel regression models. Canadian Journal of Public Health 2001;92(2):150–4.
[3] Hox J. Multilevel analysis: techniques and applications. New York, USA: Routledge; 2010.
[4] Pearl J. Probabilistic reasoning in intelligent systems. San Francisco, CA, USA: Morgan Kaufmann; 1988.
[5] Lucas P, van der Gaag L, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine 2004;30(3):201–14.
[6] Aussem A, de Morias S, Corbex M. Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks. Artificial intelligence in Medicine 2012;54(1):53–62.
[7] Flores M, Nicholson A, Burnskill A, Korb K, Mascaro S. Incorporating expert knowledge when learning Bayesian network structure: A medical case study. Artificial Intelligence in Medicine 2011;53(3):181–204.
[8] Korver M, Lucas P. Converting a rule-based expert system into a belief network. Medical Informatics 1993;18(3):219–41.
[9] Lucas P, Boot H, Taal B. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. Methods of Information in Medicine 1998;37:206–19.
[10] Lucas P, de Bruijn N, Schurink K, Hoepelman I. A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. Artificial Intelligence in Medicine 2000;19(3):251–79.
[11] Velikova M, Samulski M, Lucas P, Karssemeijer N. Improved mammographic cad performance using multi-view information: a Bayesian network framework. Physics in Medicine and Biology 2009;54:1131–47.
[12] Heckerman D, Nathwani B. Toward normative expert systems. Part I—The pathfinder project. Methods of Information in Medicine 1992;31:90–105.
[13] Heckerman D, Horvitz E, Nathwani B. Toward normative expert systems. Part II—Probability-based representations for efficient knowledge acquisition and inference. Methods of Information in Medicine 1992;31:106–16.
[14] Olesen K, Andreassen S. Specification of models in large expert systems based on causal probabilistic networks. Artificial Intelligence in Medicine 1993;5(3):269–81.
[15] Oniśko A, Druzdzel M, Wasyluk H. Extension of the hepar II model to multiple-order diagnosis. In: Intelligent information systems—advances in soft computing series. Heidelberg, Germany: Springer-Verlag; 2000. p. 303–13.
[16] Paul M, Andreassen S, Nielsen A, Tacconelli E, Almanasreh N, Fraser A, et al. Prediction of bacteremia using treat, a computerized decision-support system. Clinical Infectious Diseases 2006;42:1274–82.
[17] van den Akker M, Buntinx F, Metsemakers J, Roos S, Knottnerus J. Multimorbidity in general practice: prevalence, incidence and determinants of co-occurring chronic and recurrent diseases. Journal of Clinical Epidemiology 1998;51:367–75.
[18] Fortin M, Hudon C, Haggerty J, van den Akker M, Almirall J. Prevalence estimates of multimorbidity: a comparative study of two sources. BMC Health Services Research 2010;10:111.
[19] Marengoni A, Rizzuto D, Wang H, Winblad B, Fratiglioni L. Patterns of chronic multimorbidity in the elderly population. Journal of the American Geriatrics Society 2009;57:225–30.
[20] Diederichs C, Berger K, Bartels D. The measurement of multiple chronic diseases—a systematic review on existing multimorbidity indices. Journals of Gerontology Series A: Biological Sciences and Medical Sciences 2011;66(3):301–11.
[21] Bocquier A, Cortaredona S, Nauleau S, Jardin M, Verger P. Prevalence of treated diabetes: geographical variations at the small-area level and their association with area-level characteristics. A multilevel analysis in southeastern France. Diabetes and Metabolism 2011;37(1):39–46.
[22] Diehl K, Schneider S. How relevant are district characteristics in explaining subjective health in Germany? A multilevel analysis. Social Science and Medicine 2011;72(7):1205–10.
[23] Henriksson G, Weitoft G, Allebeck P. Associations between income inequality at municipality level and health depend on context—a multilevel analysis on myocardial infarction in Sweden. Social Sciences and Medicine 2010;71(6):1141–9.
[24] Nielen M, Schellevis F, Verheij R. Inter-practice variation in diagnosing hypertension and diabetes mellitus: a cross-sectional study in general practice. BMC Family Practice 2009;10:1–6.
[25] Suojanen M, Andreassen S, Olesen K. A method for diagnosing multiple diseases in MUNIN. IEEE Transactions on Biomedical Engineering 2001;48(5):522–32.
[26] Shwe M, Middleton B, Heckerman D, Henrion M, Horvitz E, Lehmann H, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. Methods of Information in Medicine 1998;30(4):241–55.
[27] Middleton B, Shwe M, Heckerman D, Henrion M, Horvitz E, Lehmann H, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. II—Evaluation of diagnostic performance. Methods of Information in Medicine 1991;30:256–67.
[28] Price M, Welton N, Ades A. Parameterization of treatment effects for meta-analysis in multi-state Markov models. Statistics in Medicine 2011;30:140–51.
[29] Visweswaran S, Angus D, Hsieh M, Weissfeld L, Yealy D, Cooper G. Learning patient-specific predictive models from clinical data. Journal of Biomedical Informatics 2010;43:669–85.
[30] Spiegelhalter D. Bayesian graphical modelling: a case-study in monitoring health outcomes. Applied Statistics 1998;47(1):115–33.
[31] Gyftodimos E, Flach P. Hierarchical Bayesian networks: an approach to classification and learning for structured data. In: Vouros G, Panayiotopoulos T, editors. Methods and Applications of Artificial Intelligence. Vol. 3025 of Lecture Notes in Computer Science. Samos, Greece: Springer; 2004. p. 291–300.
[32] Koller D, Pfeffer A. Object-oriented Bayesian networks. In: Geiger D, Prakash P, Shenoy P, editors. Proceedings of the thirteenth conference on uncertainty in artificial intelligence. Providence, RI, USA: Morgan Kaufmann; 1997. p. 302–13.
[33] Neapolitan R. Learning Bayesian networks. Upper Saddle River, New Jersey, USA: Prentice Hall; 2004.
[34] Cowell R, Dawid A, Lauritzen S, Spiegelhalter D. Probabilistic networks and expert systems. New York, USA: Springer; 1999.
[35] Brown W, Draper D, Goldstein H, Rasbash J. Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. Computational Statistics and Data Analysis 2002;39:203–25.
[36] Goldstein H. Heteroscedasticity and complex variation. Encyclopedia of Statistics in Behavioral Science 2005;2:790–5.
[37] Korendijk E, Maas C, Moerbeek M, van der Heijden P. The influence of misspecification of the heteroscedasticity on multilevel regression parameter and standard error estimates. Methodology 2008;2(4):67–72.
[38] Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 1986;73(1):43–56.
[39] Goldstein H. Restricted unbiased iterative generalised least squares estimation. Biometrika 1989;76:622–3.
[40] Breslow N, Clayton D. Approximate inference in generalized linear mixed models. Journal of Statistical Computation and Simulation 1993;88:9–25.
[41] Goldstein H, Rabash J. Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society (Series A) 1996;159:505–12.
[42] Seltzer M, Wong W, Bryk A. Bayesian analysis in applications of hierarchical models: issues and methods. Journal of Educational and Behavioral Statistics 1996;21:131–67.
[43] Browne W, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis 2006;1(3):473–514.
[44] Moineddin R, Matheson F, Glazier R. A simulation study of sample size for multilevel logistic regression models. BMC Medical Research Methodology 2007;7(34).
[45] Lerner U, Segal E, Koller D. Exact inference in networks with discrete children of continuous parents. In: Breese J, Koller D, editors. Proceedings of the 17th conference in uncertainty in artificial intelligence. San Francisco, CA, USA: Morgan Kaufmann; 2001. p. 319–28.

[46] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS user manual, version 1.4. Cambridge, UK: MRC Biostatistics Unit; 2001.
[47] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19(6):716–23.
[48] Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6(2):461–4.
[49] Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society (Series B) 2002;64(4):583–639.
[50] Plummer M. Penalized loss functions for Bayesian model comparison. Biostatistics 2008:1–17.
[51] Marshall E, Spiegelhalter D. Approximate cross-validatory predictive checks in disease mapping models. Statistics in Medicine 2003;22:1649–60.
[52] Picard R, Cook D. Cross-validation of regression models. Journal of the American Statistical Association 1984;79(387):575–83.
[53] Pencina M, D'Agostino Sr R, D'Agostino Jr R, Vasan R. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. Statistics in Medicine 2008;27:157–72.
[54] Scutari M. Learning Bayesian networks with the bnlearn R package. Journal of Statistical Software 2010;35(3):122.
[55] Friedman N, Goldszmidt M, Wyner A. Data analysis with Bayesian networks: a bootstrap approach. In: Laskey K, Prade H, editors. Proceedings of the 15th annual conference on uncertainty in artificial intelligence (UAI-99). Stockholm, Sweden: Morgan Kaufmann; 1999. p. 206–15.
[56] Scutari M, Nagarajan R. On identifying significant edges in graphical models. In: Hommersom A, Lucas P, editors. Proceedings of workshop on probabilistic problem solving in biomedicine. Bled, Slovenia: Springer-Verlag; 2011. p. 15–27.
[57] Kolmogorov A, Fomin S. Elements of the theory of functions and functional analysis. New York, USA: Graylock Press; 1957.
[58] Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. Statistics in Medicine 2002;21(21):3291–315.
[59] Ho K, Pinsky J, Kannel W, Levy D. The epidemiology of heart failure: the Framingham study. Journal of the American College of Cardiology 1993;22(4):6–13.
[60] Wiersma T, Smulders Y, Stehouwer C, Konings K, Lanphen J, Banga J, et al. Multidisciplinary guideline on cardiovascular risk management. Houten, The Netherlands: Bohn Stafleu van Loghum; 2011.
[61] van der Gaag L, Bolt J, Loeffen W, Elbers A. Modelling patterns of evidence in Bayesian networks: a case-study in classical swine fever. In: Computational intelligence for knowledge-based systems design. Vol. 6178 of Lecture Notes in Artificial Intelligence. Springer; 2010. p. 675–84.