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 $X \not\perp\!\!\!\perp_P Y|Z$