QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL
Qual. Reliab. Engng. Int. 2008; **24**:291–302
Published online 6 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/qre.893

Research

Modeling and Analysis of Disease and Risk Factors through Learning Bayesian Networks from Observational Data

Jing Li[1,*,†], Jianjun Shi[2] and Devin Satz[3]

[1] *Department of Industrial Engineering, Arizona State University, Tempe, AZ 85287-5906, U.S.A.*
[2] *Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, U.S.A.*
[3] *Synchronous Knowledge Inc., Falls Church, VA 22041, U.S.A.*

This paper focuses on identification of the relationships between a disease and its potential risk factors using Bayesian networks in an epidemiologic study, with the emphasis on integrating medical domain knowledge and statistical data analysis. An integrated approach is developed to identify the risk factors associated with patients' occupational histories and is demonstrated using real-world data. This approach includes several steps. First, raw data are preprocessed into a format that is acceptable to the learning algorithms of Bayesian networks. Some important considerations are discussed to address the uniqueness of the data and the challenges of the learning. Second, a Bayesian network is learned from the preprocessed data set by integrating medical domain knowledge and generic learning algorithms. Third, the relationships revealed by the Bayesian network are used for risk factor analysis, including identification of a group of people who share certain common characteristics and have a relatively high probability of developing the disease, and prediction of a person's risk of developing the disease given information on his/her occupational history. Copyright © 2007 John Wiley & Sons, Ltd.

Received 23 September 2007; Accepted 26 September 2007

KEY WORDS: case–control study; causal inference; Bayesian network; epidemiology

1. INTRODUCTION

Modeling and analysis of the relationships between a disease and its potential risk factors is important for effective disease diagnosis, control, and prevention in epidemiology. General studies under this objective can be classified into two categories: those based on experiments and those based on observational data. Compared with experimental studies in which the factors of interest are controlled and

*Correspondence to: Jing Li, Department of Industrial Engineering, Arizona State University, Tempe, AZ 85287-5906, U.S.A.
†E-mail: jinglz@asu.edu

their impacts on disease occurrences are evaluated, observational studies are performed based on population sampling, where the investigator has no control over the data collection but utilizes the existing record information. Observational studies are especially advantageous in situations such as when (i) the target disease may result in severe health damage; (ii) risk factor control is inconvenient or even infeasible; and (iii) a large number of potential risk factors need to be studied simultaneously. In all of these cases, it is almost impossible to conduct experiments.

Meanwhile, due to the advancement of computer technologies, a large amount of real-world data can now be collected and warehoused, including patients' clinical records regarding disease diagnoses and treatments; demographic attributes such as gender, age, and race; and behavioral information such as occupational histories (i.e. job type, job location, and time period for performing the job). This data-rich environment creates new opportunities for observational epidemiology. However, there is a lack of effective tools for analyzing the data, discovering the relationships, and making intelligent decisions.

Traditionally, the assessment of the relationship between a disease and a potential risk factor in an observational study was achieved using a variety of association measures[1]. Also, there were statistical methods, such as logistic regression[2], to quantify both the association and the uncertainty surrounding the estimation of the association. However, these measures and methods paid little attention to the nature of the risk factor and the disease; that is, they did not distinguish between the cause and the effect. Furthermore, the apparent association could even be non-causal due to the confounding effects of other variables. Although the confounding effects can be eliminated by certain data-separation techniques such as stratification[1], the success of the elimination has to be based on pre-identification of the confounding variables. A complicated scenario arises when the problem is data mining in nature. In this case, most of these traditional approaches may not be effective. Data mining is not targeted at studying the association between the disease and one risk factor, but searches for interesting relationships between the disease and all potential risk factors simultaneously. Without any *a priori* knowledge about what variables confound which associations, these traditional approaches may not be able to distinguish among different associations (i.e. causal or non-causal) in an effective and efficient manner.

The problem formulation for discovering the relationships between a disease and its potential risk factors in epidemiology is similar to that for data-driven modeling and analysis of the relationships between process variables and product quality variables in quality engineering. In the latter case, Bayesian networks based on massive observational data have been identified as effective tools for studying dependent, independent, and causal relationships[3,4]. Recent years have witnessed the application of quality engineering methods in non-traditional domains such as health care[5,6] and computer networks[7,8]. As a similar effort, this paper adopts the concepts and methods of Bayesian networks, and combines them with the uniqueness of epidemiologic studies to model and analyze the relationships between a disease and its potential risk factors.

A Bayesian network is an explicit representation of the dependent and independent relationships among variables. Depending on the interpretation, it can also represent causalities[9]. Previous research in epidemiology has focused on using a developed Bayesian network, usually constructed by *a priori* knowledge, to infer the presence of confounding factors and to compute the strength of the associations[1]. Little has been found on how to develop a Bayesian network from observational data. A Bayesian network developed from observational data provides a way to simultaneously study the causal relationships between the disease and all the potential risk factors of interest. Another important advantage of a Bayesian network is that *a priori* knowledge from the specific application domain can be incorporated into the learning algorithms of the Bayesian network to improve the learning accuracy and efficiency.

This paper presents an application of Bayesian networks in an epidemiologic study with the objective of identifying the causal relationships between disease occurrences and patients' occupational histories. The proposed approach emphasizes the integration of medical domain knowledge with data preparation and analysis. The focus is on studying the following issues of the disease through the learning and inference

DOI: 10.1002/qre

of a Bayesian network:

- *Occupational distribution*, i.e. whether and how disease occurrences are associated with certain occupations.
- *Geographical distribution*, i.e. whether and how disease occurrences are associated with certain locations (i.e. geographical location of a job conducted).
- *Time-sensitive occupational distribution*, i.e. whether and how the association with an occupation varies over time.
- *Time-sensitive geographical distribution*, i.e. whether and how the association with a location (i.e. geographical location of a job conducted) varies over time.
- *Role of demographic attributes*, i.e. whether the association with an occupation or a location is due to the confounding effects of other demographic attributes such as gender and age, and what the confounding factors are.

A thorough understanding of the aforementioned issues of the disease helps identify 'who,' 'where,' and 'when,' i.e. what demographic attributes increase the likelihood of developing the disease, in which location was the disease developed, and when was the disease developed. All these provide supporting information to discover the 'what,' i.e. the cause of the disease.

The remainder of this paper is organized as follows. Section 2 describes the data. Section 3 introduces important considerations in data preprocessing. Section 4 demonstrates the use of a Bayesian network to learn the relationships between disease occurrences and potential risk factors. Section 5 discusses how to make inferences based on the Bayesian network. Section 6 provides case studies based on real-world data. Finally, conclusions are given in Section 7.

2. DESCRIPTION OF THE DATA

This research utilizes two data sets. One contains the clinical records of patients (clinical data set), and the other contains information on the patients' occupational histories (occupational data set). In the clinical data set, each row corresponds to one medical appointment and each column corresponds to one clinical or demographic attribute of the patient, including the patient's ID, appointment date, the diagnostic result, age, gender, race, marital status, etc. In the occupational data set, each row corresponds to one patient identified by the patient's ID; and contains the jobs he/she had, the locations of the jobs in chronological order up to the time when the data set was created, and the time period (i.e. the starting and ending dates) of each job. To link the diagnostic result (i.e. the disease) with a patient's occupational history, the clinical and occupational data sets are merged using patient ID as the common variable, and a combined data set is created. This research is based on the combined data set. The data sets, their contents, and their relationships are described in Figure 1.

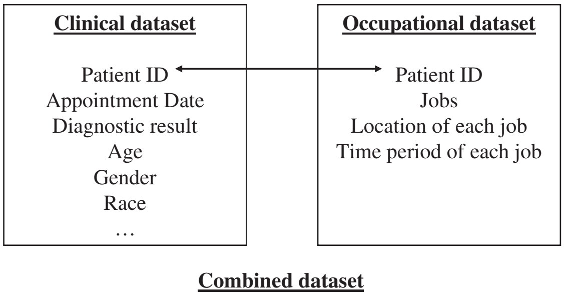


Figure 1. Description of the data

3. DATA PREPROCESSING

To obtain a meaningful Bayesian network, data preprocessing is critical. During preprocessing, medical domain knowledge can be utilized to transform the variable, de-noise the data, and remove redundant information, in order to facilitate the developing of a Bayesian network. Several important issues are discussed in this section.

3.1. Hierarchical grouping of the diagnostic results

The variable 'diagnostic result' contains the disease that a patient was diagnosed with. Different diagnostic results can be grouped together if they are subtypes of a more general type of disease. For example, both yellow fever and mosquito-borne viral encephala belong to arthropod-borne viral diseases. The grouping can be further refined, resulting in a three-level disease hierarchy. The diseases at different levels may have different relationships with the occupational history. For example, the geographical distribution for the occurrences of yellow fever may be different from that of arthropod-borne viral diseases. To obtain a complete understanding of how the occupational history may impact disease occurrences, diseases on all three levels have been studied in this research.

3.2. Handling multiple appointment records

In the combined data set, a patient may have more than one appointment record (i.e. multiple rows) with identical diagnostic results. These appointments may all correspond to one occurrence of the disease, where the first appointment in chronological order was the one when the diagnosis was made, and subsequent appointments were made for continuous treatments of the disease. Because the induction of the disease must occur prior to the time of the first appointment, the patient's occupational history after the first appointment is uninformative for identifying the cause of the disease. Thus, there is no need to keep all the appointment records after the first one.

In some cases, multiple appointment records of one patient could also be generated due to disease recurrences or re-infections. Therefore, efforts need to be made to distinguish between the multiple appointments associated with a single disease occurrence and those associated with multiple disease occurrences. By assuming that the successive appointments in the situation of a single disease occurrence have smaller time intervals than those in the situation of multiple disease occurrences, the following criterion is implemented: t_{A_i} and $t_{A_{i+1}}$ denote the time of two successive appointments A_i and A_{i+1}, respectively, and Δt_M can be obtained from medical awareness on the treatment cycle of the disease.

If $t_{A_{i+1}} - t_{A_i} < \Delta t_M$, then both A_i and A_{i+1} correspond to one single disease occurrence. Otherwise, each of them corresponds to a different disease occurrence.

3.3. Estimating the induction time of disease

The combined data set contains a patient's occupational history spanning from the first day his/her information was recorded to the time when the data analysis was initiated. Not all parts of this occupational history are relevant to the disease occurrence. For example, the occupation he/she held after the medical appointment is non-informative. Even for the occupations prior to the appointment, only those during a certain time period could possibly impact the disease occurrence. This period, denoted by $[t_I^L, t_I^U]$, is an estimate of the induction time of the disease based on medical domain knowledge, where t_I^L and t_I^U are the estimated lower and upper bounds of the induction time, respectively. The estimation procedure is shown in Figure 2. For a given appointment time, t_A, an estimate of the time when the symptom of the disease appears t_S, is first obtained. This can be done either through the medical awareness of how long the symptom can be tolerated by the patient after it appears, or by retrieving more detailed diagnostic information from additional data sets containing the patient's description of his/her disease development. Furthermore, for most diseases in epidemiology, there is a defined incubation period, which is the time between the contraction of a disease and the first appearance of

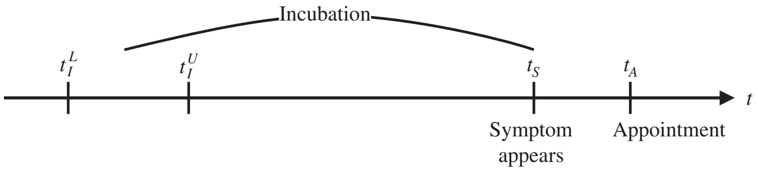


Figure 2. Estimation of the induction time

a symptom. Based on the incubation period, an interval estimate of the induction time $[t_I^L, t_I^U]$ can be achieved.

Because the length of the time interval $[t_I^L, t_I^U]$ is generally much smaller than the time period of a patient staying in one occupation at one location, it is reasonable to assume that his/her occupation and the location of the occupation remain unchanged within $[t_I^L, t_I^U]$. Therefore, although the patient's entire occupational history may contain a series of changes in occupations and locations, only one occupation and one location are kept for each record.

3.4. *Selecting a control group*

The observational research presented in this paper is a case–control study. Selecting an appropriate control group is important to avoid reaching biased conclusions regarding the relationships between the disease and risk factors. One distinct feature of this study is that there are records only for the people who were sick and hospitalized, but not for the healthy population. Therefore, the subjects in the control group have to be selected from the same source of subjects in the case group, i.e. the patients with diseases different from the disease of the case group in the combined data set. The following principles are used to find such a 'diseased' control group.

(i) If a patient is diagnosed with a disease that is associated with any of the risk factors of interest, he/she cannot be selected as a subject in the control group.

Wacholder *et al.*[10] pointed out that the best strategy regarding the selection of the diseases in forming a control group is to exclude all conditions likely to be related to the risk factors of interest. This is based on the consideration that if the disease of the subjects in the control group is related to the risk factors, the association between the disease of the case group and the risk factors may be masked or twisted. Although theoretically attractive, in practice this principle is difficult to implement, as it requires *a prior* knowledge about the relationship between the disease of a control group candidate and the risk factors. Thus, a second principle is proposed as a supplement.

(ii) If a patient is diagnosed with the disease of the case group or is diagnosed with a disease sharing a common ancestor with the disease of the case group in the disease hierarchy, he/she is not selected as a subject in the control group.

It is reasonable to believe that the diseases belonging to the same disease family are more likely related to the risk factors in the same way as those in different families. Although applying this principle cannot completely exclude the unqualified subjects from the control group defined by principle (i), it narrows the scope of the investigation and helps find the qualified subjects.

(iii) Multiple control groups are preferable to a single control group, with each control group corresponding to a different disease.

The data analysis can be conducted based on each individual control group and then the results are compared. If a certain relationship is found to be concordant across the control groups, a high confidence in the existence of this relationship can be achieved. However, when the results are discordant, additional efforts have to be made to determine which result is correct. Multiple control groups may also be helpful when each group serves a different purpose. For example, if the disease (other than

DOI: 10.1002/qre

the disease of the case group) of a control group is known to have little association with location, this control group can be used to study the relationship between the disease of the case group and location. Another control group may be used to study the relationship between the disease of the case group and occupation if the disease of this control group is known to be independent of occupation. The patients from different control groups can also be pooled together to form one control group. The data analysis based on this pooled control group can effectively avoid bias when an individual control group is used.

3.5. *Data visualization*

Data visualization is important for achieving an initial understanding of data distributions and variable relationships. In this section, several graphical techniques are presented which target the visualization of the time-sensitive geographical and occupational distribution of the disease.

3.5.1. *Multi-location temporal distribution (MLTD) graph and multi-occupation temporal distribution (MOTD) graph*

For each disease occurrence i, let $(t_I^L)_i$ and $(t_I^U)_i$ denote the starting and ending dates of the estimated induction period of the disease, and z_i denote the location, i.e. the location of the patient during this period. The MLTD graph is constructed with the x- and y-axes representing the locations and the calendar time, respectively. If the calendar time is in units of days, there will be $(t_I^U)_i-(t_I^L)_i+1$ points plotted in the graph for the disease occurrence i, with the x-coordinates all being z_i and y-coordinates being $(t_I^L)_i, (t_I^L)_i+1, \ldots, (t_I^U)_i$. To avoid the overlap of the points from different disease instances, two random numbers, U_1 and U_2, are generated and added to the x- and y-coordinates of each point, respectively. Thus, the resulting points plotted in the graph have the coordinates z_i+U_1 and $(t_I^L)_i+k+U_2$, $k=0, \ldots, (t_I^U)_i-(t_I^L)_i$, where U_j, $j=1,2$, follows a uniform distribution with the range $[-a_j, a_j]$, $0<a_j<0.5$, and a_j can be adjusted so as to achieve the clearest graph possible.

Under certain circumstances, it may be more favorable to use larger units for the calendar time such as weeks or months, especially when the data are collected continuously over a long time span. Thus, using days as the units will result in a graph that is extremely lengthy in the vertical direction. As the units become coarser, fewer points will be produced from each disease instance. The y-coordinates of the points can be inferred from the induction period.

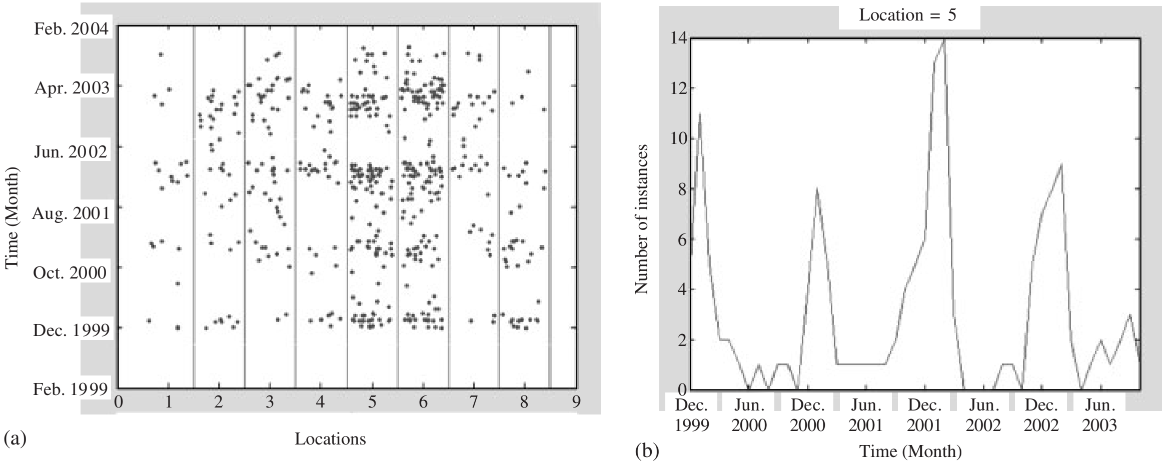


Figure 3. (a) MLTD graph and (b) ULTD graph

DOI: 10.1002/qre

The MOTD graph can be developed in a similar manner. The MLTD/MOTD graph displays the temporal distribution of the disease at each location/occupation and also tells how the temporal distribution varies across different locations/occupations. For example, in Figure 3(a), the distribution of the particular disease has a cyclic fluctuation, that is, most instances cluster around certain months of a year. This pattern is consistent across different locations.

It should be pointed out that inspections on the temporal distribution of the disease in the MLTD/MOTD graph should be focused on the pattern or trend of each location/occupation and the contrast between locations/occupations, and any interpretations based on the absolute number of points should be avoided. For example, the number of points on a MLTD/MOTD graph does not reflect disease prevalence because more points could result from a long induction period.

3.5.2. Uni-location temporal distribution (ULTD) graph and uni-occupation temporal distribution graph

After certain patterns have been identified from the MLTD/MOTD graph, it may be desirable to further examine the parameters of the patterns. Examples of these parameters include the slope of an increasing trend and the period of a cyclic fluctuation. To achieve this objective, a graph that plots the number of disease instances vs time for each individual location (or occupation) can be constructed. For example, Figure 3(b) displays such a graph for location 5 in Figure 3(a). It shows that the disease is prevalent in three months of a year, i.e. December, January, and February.

4. LEARNING THE RELATIONSHIPS BETWEEN A DISEASE AND POTENTIAL RISK FACTORS USING A BAYESIAN NETWORK

A Bayesian network provides an explicit representation of the dependent and independent relationships among the variables in a problem. Depending on the interpretation, it can also represent causalities. Bayesian networks have been extensively used in genetics[11], ecology[12], social science[13], and physical science[14].

A Bayesian network is a directed acyclic graph (DAG), as shown in Figure 4. Each variable defines a node. An arc with an arrow points from X_j to X_i if X_j is a parent (i.e. direct cause) of X_i. For each nodeX_i, there is a conditional probability distribution $P(X_i|Parents(X_i))$ quantifying the dependence between X_i and its parents.

The process of developing a Bayesian network from observational data includes learning the structure (i.e. the DAG) and learning the parameters (i.e. the conditional probability distributions) given a structure. The existing algorithms for learning the structure can be divided into two categories: constraint-based approaches and score-based approaches. This paper focuses on the first category. Within this category, the algorithms can be further classified depending on whether the variables are discrete or continuous. Since the disease and most of the potential risk factors are measured at discrete scales, the learning algorithm for discrete variables is adopted in this research, which follows two basic principles[15]. Let $\mathbf{V}$ denote the set of variables, $\mathbf{V}=\{X_1,\ldots,X_n\}$. Any subset of $\mathbf{V}$ is denoted by $\mathbf{T}$. $X_i \coprod X_j|\mathbf{T}$ means X_i is independent of X_j given $\mathbf{T}$.

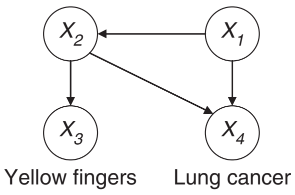


Figure 4. A Bayesian network

DOI: 10.1002/qre

$\neg A$ means that statement A does not hold. The principles are as follows:

Principle I: Let $X_i, X_j \in \mathbf{V}$. Then X_i and X_j are adjacent if and only if for every $\mathbf{T} \subset \mathbf{V}$ such that $X_i, X_j \notin \mathbf{T}$, $\neg X_i \coprod X_j | \mathbf{T}$.

Principle II: If X_i and X_k are adjacent, X_j and X_k are adjacent, but X_i and X_j are not, then orient the arcs among such triplets as $X_i \rightarrow X_k \leftarrow X_j$ if and only if for every $\mathbf{T} \subset \mathbf{V}$ such that $X_i, X_j \notin \mathbf{T}$ and $X_k \in \mathbf{T}$, $\neg X_i \coprod X_j | \mathbf{T}$.

The first principle discovers all arc connections among variables and the second identifies arc directions. These two principles have been implemented in a computational algorithm called PC (Peter and Clark)[16]. Taking the BN structure in Figure 4 as an example, the PC algorithm is briefly introduced here.

- Start with a fully connected undirected graph in which each variable is linked with all other variables by undirected arcs.
- The undirected arc between one pair of variables is removed when conditional independence is found. For example, in Figure 4, the arc between X_3 and X_4 is removed because X_3 is independent of X_4 given their common cause X_2, and the arc between X_3 and X_1 is removed because X_3 is independent of X_1 given X_3's parent X_2. Other undirected arcs can be removed similarly.
- For the undirected arcs that cannot be removed in the previous step, the following rule is used to orient the arcs: each triple of variables $X_i - X_j - X_k$ is oriented as $X_i \rightarrow X_j \leftarrow X_k$, if X_i and X_k are found to be independent given a set of variables which do not contain X_j. Aside from this major rule in orientation, Meek[17] introduced some supplementary rules such as orienting the remaining undirected arcs in a way that no cycles are created in the BN. In addition, some orientations may be achieved through knowledge or first principles.

PC can be improved by incorporating a partial or complete temporal order of the variables, which is readily available in most epidemiologic studies. For example, gender and age are naturally determined and thus they are always the earliest variables; moreover, the disease is always later than the risk factors. The basic idea for incorporating the temporal order into the learning is that if X_i occurs after X_j, it cannot be a cause of X_j.

After the structure is obtained, the parameters can be learned based on the structure and the data. Learning parameters can be achieved by the EM algorithm in the cases of complete data sets and the data sets with missing data[18].

5. INFERENCE BASED ON A BAYESIAN NETWORK

5.1. Qualitative inference

After a Bayesian network is obtained, the DAG can be used to qualitatively interpret the association between the disease and a risk factor in the following ways.

(i) The association is causal, i.e. the risk factor is a cause of the disease if and only if there is *directed path* from the risk factor to the disease. A directed path from X_i to X_j is a sequence of variables beginning with X_i and ending with X_j such that for every pair of variables X_k and X_l, adjacent in this sequence and occurring in that order, there is an arc pointing from X_k to X_l. For example, age (X_1) and smoking (X_2) affect the development of lung cancer (X_4) in Figure 4, while yellow fingers (X_3) is not a cause of lung cancer (X_4).

(ii) A variable is independent of its non-descendants given its parents. This statement is called the causal Markov condition[16]. X_i is a descendant of X_j if there is a directed path from X_j to X_i. The causal Markov condition can be used to infer (conditional) independence relationships between variables. For example, in Figure 4, X_3 is independent of X_1 given X_2. An interpretation of this independence is that knowing whether a person smokes (X_2) is sufficient to assess his/her risk of having yellow fingers (X_3), regardless of his/her age (X_1).

(iii) To determine whether confounding factors exist, a two-step procedure is followed[1]: (a) Delete all arcs from the risk factor that points to any other nodes. (b) In the reduced graph, determine whether there is any *unblocked backdoor path* from the risk factor to the disease. If such a path exists, the association between the disease and the risk factor is due to the confounding effect of other variables. Otherwise, there is no confounding effect.

An unblocked backdoor path is defined as follows. A backdoor path from X_i to X_j is a path that begins by leaving X_i along an arc whose arrow points into X_i, and then continues to X_j regardless of the arc directions. A path is blocked if it contains at least one collider. A node X_i on a path is called a *collider* if the arcs of the path entering and leaving X_i both have arrows pointing into X_i.

For example, the association between yellow fingers X_3 and lung cancer X_4 is due to the confounding effect of smoking X_2 because $X_3 \leftarrow X_2 \rightarrow X_4$ is an unblocked backdoor path from X_3 to X_4. $X_3 \leftarrow X_2 \rightarrow X_4$ is an unblocked backdoor path because it does not contain a collider.

5.2. *Quantitative inference*

Quantitative inference can be achieved based on the parameters of the Bayesian network. A joint distribution over all the variables can be computed as a function of the parameters, i.e. $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$[19]. Based on the joint distribution, any conditional probability $P(\mathbf{U}|\mathbf{W})$, $(\mathbf{U}, \mathbf{W} \subset \mathbf{V})$ can be computed. One conditional probability, $P(D=1|\mathbf{R}=\mathbf{r})$, is of particular interest, where $D=1$ represents ‘having the disease’ and $\mathbf{R}$ is a set of risk factors, as it provides an assessment of a person’s risk of developing the disease. In addition, computing $\hat{\mathbf{R}} = \arg\max_{\mathbf{r}} P(D=1|\mathbf{R}=\mathbf{r})$ provides a way to identify the group of people who share some common characteristics and have the highest probability of developing the disease.

6. *CASE STUDY*

Two case studies are presented in this section, each corresponding to one particular disease of interest. The data used in the case studies are collected from a real-world data warehouse, which has synchronized occupational data and clinical data for more than 100 000 data records. There are five potential risk factors. The variables in the case studies include disease X_4 ($X_4=1$: case; $X_4=2$: control), gender X_1 ($X_1=1$: male; $X_1=2$: female), age group X_2 ($X_2=1$: $18 \leq \text{age} < 25$; $X_2=2$: $25 \leq \text{age} < 35$; $X_2=3$: $35 \leq \text{age} < 45$), occupation X_3 ($X_3=1$–8), location X_5 ($X_5=1$–6), and estimated induction time X_6. Note that the ages of patients are aggregated into three age groups (i.e. X_2) according to medical definitions. The estimated induction time (X_6) is transformed depending on the objective of the study. For example, if the seasonal fluctuation of the disease prevalence is of interest, X_6 should be aggregated into 12 months or four quarters of a year; if the trend of the disease prevalence is of interest, the entire time axis can be divided into several intervals using discretization approaches such as equal bin size or equal frequency[20]. Visualization of the data, i.e. the MLTD/MOTD graph, gives a hint on how to transform X_6. For example, Figure 3(a) plots partial instances of the disease in the first case study, and it is obvious that the disease prevalence has seasonal fluctuation. Thus, X_6 is grouped into four quarters of a year. In the second case study, X_6 is divided into four intervals by the equal-frequency approach, that is, $X_6=1$: [6/2001–1/2002], $X_6=2$: [2/2002–7/2002], $X_6=3$: [8/2002–3/2003], $X_6=4$: [4/2003–8/2003].

A partial temporal order of the variables is specified as follows and further utilized by the PC algorithm to improve the learning. The variables belong to three temporal layers, with gender and age on the earliest layer, time, location, and occupation on the second earliest layer, and disease on the latest layer.

6.1. *Case study 1*

Figure 5 shows the structure and partial parameters ($P(X_4|X_2, X_6)$) of the Bayesian network for this particular disease. The following conclusions can be reached.

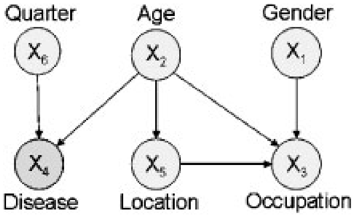
DOI: 10.1002/qre

(i) Age and quarter affect disease occurrences. Location and occupation are associated with the disease but do not affect disease occurrences, because the associations are due to the confounding effect of age. Gender is not associated with the disease.

(ii) $P(X_4=1|X_2=x_2, X_6=x_6)$ is computed and shown in the table of Figure 5, which can be used to predict a person's risk of developing the disease given his/her age and the time (i.e. quarter) for performing his/her job. In addition, $\arg\max_{x_2,x_6} P(X_4=1|X_2=x_2, X_6=x_6)=(1, 1)$, which means people in the youngest age group (i.e. $X_2=1$: $18\leq \text{age}<25$) in the first quarter of a year ($X_6=1$) have the highest probability to develop this disease.

6.2. Case study 2

The structure of the Bayesian network for this particular disease is shown in Figure 6. (The parameters are not shown here.) The following conclusions can be inferred:



X2	X6	P(X4=1| X2, X6)
1	1	0.3116
1	2	0.1147
1	3	0.0839
1	4	0.1913
2	1	0.1667
2	2	0.0057
2	3	0.022
2	4	0.03
3	1	0.1849
3	2	0.031
3	3	0.0248
3	4	0.0886

Figure 5. The structure and partial parameters of the Bayesian network in case study 1

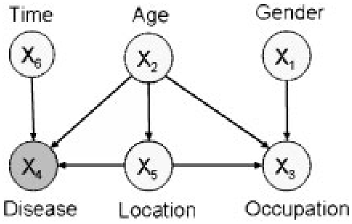


Figure 6. The structure of the Bayesian network in case study 2

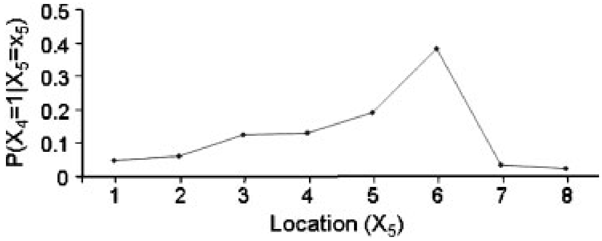


Figure 7. Prediction of a person's risk of developing the disease given locations in case study 2

DOI: 10.1002/qre

(i) Age, time, and location affect disease occurrences. Occupation is associated with the disease but does not affect disease occurrences, because the association is due to the confounding effect of age and location. Gender is not associated with the disease.

(ii) $P(X_4=1|X_5=x_5)$ is computed and plotted in Figure 7, which can be used to predict a person's risk of developing the disease given the location of his/her job. In addition, $\arg\max_{x_2,x_5,x_6}(X_4=1|X_2=x_2, X_5=x_5, X_6=x_6|)=(1,6,4)$, which means people in the youngest age group (i.e. $X_2=1$: $18\leq$ age<25) at location 6 (i.e. $X_5=6$) during the time interval 4/2003–8/2003 (i.e. $X_6=4$) have the highest probability of developing this disease.

7. CONCLUSION

This paper presents a systematic approach to discovering the relationships between disease occurrences and patients' occupational histories based on real-world data. A Bayesian network is employed to learn the relationships. Data pre-processing is discussed as an important step in supporting the Bayesian network learning and ensuring meaningful results. The proposed approach addresses the integration of medical domain knowledge with the data preparation and analysis.

Based on the Bayesian networks, qualitative inferences can be made to distinguish among different types of associations (causal or non-causal) between the disease and risk factors. Quantitatively, given a person's demographic and occupational attributes, such as gender, age, and occupation, his/her risk of developing the disease can be predicted; in addition, it is possible to identify a group of people who share some common descriptive characteristics (e.g. a certain gender or/and occupation) and have the highest probability of developing the disease, which is important for disease control and prevention in epidemiology.

The proposed approach requires that sufficient data are collected for the variables, which is a challenging issue in some cases, such as when (i) the data set contains many missing values and (ii) the disease to be studied is not prevalent so that only a few cases are gathered. Further work will focus on studying how to derive meaningful dependent, independent, and causal relationships with missing values and small sample size.

Acknowledgements

The authors would like to thank the researchers and analysts at Synchronous Knowledge Inc., specifically Wesley Watkins, Rick Reichard, Corey Ritchie, Stavros Kalamatianos, Olga Brazhnik, and Rick Reichard, for their invaluable support and contribution to this research.

REFERENCES

1. Jewell NP. *Statistics for Epidemiology*. Chapman & Hall/CRC: Boca Raton, FL, 2004.
2. Shoukri MM, Edge VL. *Statistical Methods for Health Sciences*. CRC Press: Boca Raton, FL, 1996.
3. Li J, Shi J. Knowledge discovery from observational data for process control using causal Bayesian networks. *IIE Transactions* 2007; **39**(6):681–690.
4. Lerner U. Hybrid Bayesian networks for reasoning about complex systems. *PhD Thesis*, Stanford University, 2002.
5. Montgomery DC. SPC research—Current trends. *Quality and Reliability Engineering International* 2007; **23**(5): 515–516.
6. Woodall WH. The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology* 2006; **38**(2):89–104.
7. Ye N, Chen Q. Attack-norm separation for detecting attack-induced quality problems on computers and networks. *Quality and Reliability Engineering International* 2006; **23**(5):545–553.
8. Kwon DW, Ko K, Vannucci M, Reddy ALN, Kim S. Wavelet methods for the detection of anomalies and their application to network traffic analysis. *Quality and Reliability Engineering International* 2006; **22**(8):953–956.

DOI: 10.1002/qre

9. Buntine W. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering* 1996; **8**(2):195–210.
10. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case control studies I–III. *American Journal of Epidemiology* 1992; **135**:1019–1050.
11. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 2002; **7**:601–620.
12. Borsuk ME, Stow CA, Reckhow KH. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modeling* 2004; **173**:219–239.
13. Dai H, Korb KB, Wallace CS, Wu X. A study of causal discovery with weak links and small samples. *Proceeding of the Fifteenth International Joint Conference on Artificial Intelligence*, San Francisco, CA, 1997; 1304–1309.
14. Gutiérrez JM, Cano R, Sordo CM. Applications of Bayesian networks in meteorology, *Advances in Bayesian Networks*, Gámez JA *et al*. (eds.). Springer: New York, 2004; 309–327.
15. Korb KB, Nicholson AE. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC: London, U.K., 2003.
16. Spirtes P, Glymour C, Scheines R. *Causation*, *Prediction and Search*. MIT Press: Cambridge, MA, 2000.
17. Meek C. Causal inference and causal explanation with background knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, 1995; 403–410.
18. Lauritzen SL. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* 1995; **19**(2):191–201.
19. Krause P. Learning probabilistic networks. *The Knowledge Engineering Review* 1998; **3**(4):321–351.
20. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. *Proceedings of the Twelfth International Conference*, San Francisco, CA, 1995.

Authors' biographies

Jing Li is an Assistant Professor in the Department of Industrial Engineering at Arizona State University. She received her PhD in industrial and operations engineering from the University of Michigan. Her primary research interests are applied statistics, data mining, and causal modeling and inference for process control, especially modeling and analyzing massive high-dimensional data sets in complex systems for improving the quality of products and processes. Her work has been applied to manufacturing and health-care problems.

Jianjun Shi is a Professor in the Department of Industrial and Operations Engineering at the University of Michigan. He received his PhD in mechanical engineering from the University of Michigan. His research interests focus on the fusion of advance statistics and domain knowledge to develop methodologies for modeling, monitoring, diagnosis, and control of complex manufacturing systems. His research has been funded by the National Science Foundation, National Institute of Standards and Technology, Advanced Technology Program, General Motors, Daimler-Chrysler, Ford, Lockheed-Martin, Honeywell, and various other industrial companies and funding agencies.

Devin Satz was the former President of Synchronous Knowledge Inc. (SKI), currently known as IMS Government Solutions, a wholly owned subsidiary of IMS Health Incorporated since May 2005. He was formerly a Major in the U.S. Air Force and the Administrator and Chief Information Officer for the Air Force Medical Operations Agency.

DOI: 10.1002/qre