BUILDING BAYESIAN NETWORKS BASED ON PATIENT REPORTED OUTCOME QUESTIONNAIRES FOR MUSCULO-SKELETAL CONDITIONS

KAS-İSKELET RAHATSIZLIKLARI İÇİN HASTA TARAFINDAN BİLDİRİLEN SONUÇ ÖLÇEKLERİNİ TEMEL ALAN BAYES AĞLARI GELİŞTİRİLMESİ

HAKAN YÜCETÜRK

ASST. PROF. DR. BARBAROS YET
Supervisor

Submitted to Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfilment of the Requirements
for the Award of the Degree of Master of Science
in Industrial Engineering

2020

ABSTRACT

BUILDING BAYESIAN NETWORKS BASED ON PATIENT REPORTED OUTCOME QUESTIONNAIRES FOR MUSCULO-SKELETAL CONDITIONS

HAKAN YÜCETÜRK

MASTER OF SCIENCE, DEPARTMENT OF INDUSTRIAL ENGINEERING
SUPERVISOR: ASST. PROF. DR. BARBAROS YET
June 2020, 74 Pages

Machine learning (ML) which is a branch of artificial intelligence (AI), has been an important approach used in the medical domain. ML approaches learn from historical data to evaluate and predict patient status. These approaches have been successful in medical domains, such as radiology and dermatology, where a large amount of data exists with clearly labelled patient outcomes. However, such clearly labelled outcome data do not exist in large amounts in most medical domains. Patient reported outcome measures (PROMS) are the primary way to assess patient outcomes in many medical areas. Filling in PROMs regularly and repetitively can be difficult due to time and cognitive-load requirements. Considering that some PROMs contain over 30 questions, collecting large amounts of patient outcome data can be difficult in these domains. This study proposes an approach for collecting patient outcome data with less time and cognitive-load requirements. In this context, an ML approach called Bayesian networks (BNs) is used to predict patient outcomes with missing PROM inputs, and to identify the most informative PROM questions for specific patients. Also, random questions were selected from the PROMs and these questions were used to determine the patient status. The obtained estimation results were compared with the estimation results obtained by using the most informative questions. The proposed approach has been applied to PROMS used in the musculo-skeletal domain. Results were evaluated by cross validation method. Cross-validation results show that the proposed approach can accurately predict patient outcomes with fewer PROM questions.

Key words: Bayesian Networks, Musculo-Skeletal Disorders, Structure Learning, Information Theory, Artificial Intelligence

ÖZET

KAS-İSKELET RAHATSIZLIKLARI İÇİN HASTA TARAFINDAN BİLDİRİLEN SONUÇ ÖLÇEKLERİNİ TEMEL ALAN BAYES AĞLARI GELİŞTİRİLMESİ

HAKAN YÜCETÜRK

YÜKSEK LİSANS, ENDÜSTRİ MÜHENDİSLİĞİ BÖLÜMÜ
TEZ DANIŞMANI: DR. ÖĞR. ÜYESİ BARBAROS YET
Haziran 2020, 74 Sayfa

Yapay zekânın bir kolu olan makine öğrenimi, tıp alanında kullanılan önemli bir yaklaşım olmuştur. Makine öğrenimi yaklaşımları, hastanın durumunu değerlendirmek ve tahmin etmek için geçmiş verilerden öğrenir. Makine öğrenimi yöntemleri, radyoloji ve dermatoloji gibi, hasta sonuçlarının veride net bir şekilde belirtildiği ve yüksek miktarda veri kümelerinin bulunduğu alanlarda başarılı olmuştur. Fakat birçok tıp alanında bu şekilde yüksek miktarda temiz bir hasta sonucu verisinin bulunması mümkün değildir. Çoğu alanda hasta sonucuna ilişkin veriler, hasta tarafından bildirilen sonuç ölçütleri (PROMs) olarak adlandırılan tıbbi anketler aracılığıyla toplanır. Hastaların PROM düzenli ve yinelemeli olarak PROM doldurmaları gerektirdiği zaman ve bilişsel yük yüzünden zor olabilir. Bazı PROM'ların 30 veya daha fazla soru içerdiği düşünüldüğünde, bu alanlarda yüksek miktarda hasta sonucu verisi toplanması güçtür. Bu çalışmada, hastalardan daha az ve bilişsel yük gerektirerek, yüksek doğrulukta hasta çıktısı toplanması için bir yaklaşım önerilmektedir. Bir makine öğrenmesi yöntemi olan Bayes ağları kullanılarak, hastalara en çok bilgi veren PROM sorularının sorulmasına olanak verilmekte ve eksik PROM sorularıyla da hasta sonuçları tahmin edilmektedir.

Ayrıca, PROM'lardan rasgele sorular seçilmiş ve hasta durumunu belirlemek için bu sorular kullanılmıştır. Elde edilen tahmin sonuçları, en bilgilendirici sorular kullanılarak elde edilen tahmin sonuçlarıyla karşılaştırılmıştır. Geliştirilen yaklaşım kas-iskelet rahatsızlıkları alanında kullanılan PROM'lara uygulanmış, sonuçları çapraz validasyon yöntemiyle değerlendirilmiştir. Sonuçlar incelendiğinde, hasta durumunun az sayıda PROM sorusuyla yüksek doğrulukla tahmin edilebildiği görülmüştür.

Anahtar Kelimeler: Bayes Ağları, Kronik Kas ve İskelet Rahatsızlıkları, Yapı Öğrenme, Bilgi Teorisi, Yapay Zekâ

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Barbaros YET for his guidance, endless support, advice. During my research, I am grateful to him for finding solutions to all my problems and encouraging me. I am so thankful to have had the opportunity to work with him.

Also I would like to thank Dr. Ceren TUNCER ŞAKAR for her help and support.

I would like to thank Hacettepe Industrial Engineering lecturers who have given me all kinds of support during my Master's degree study.

Finally I would like to thank my family who were with me at every moment of my life and who support me all the time.

This work was supported by an Institutional Links grant, ID 352394702, under the Turkey – Newton - Katip Çelebi Fund partnership. The grant is funded by the UK Department of Business, Energy and Industrial Strategy (BEIS) and The Scientific and Technological Research Council of Turkey (TUBITAK - ID 217M518) and delivered by the British Council. For further information, please visit www.newtonfund.ac.uk

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND ABBREVIATIONS

Symbols

R^2	Pearson correlation

Abbreviations

AbsDiff	Absolute Difference
AI	Artificial Intelligence
ALARM	A Logical Alarm Reduction Mechanism
ASP	Asymptomatic Patient
ATC	Acute Traumatic Coagulopathy
AUC	Area Under Curve
BN	Bayesian Network
CDSS	Clinical Decision Support System
CI	Conditional Independence
COPD	Chronic Obstructive Pulmonary Disease
CPT	Conditional Probability Tables
CSI	Central Sensitization Inventory
CWM	Construction Waste Management
DAG	Directed Acyclic Graph
DBM	Distribution Based Method
DBN	Dynamic Bayesian Network
DEMATEL	Decision Making Trial and Evaluation Laboratory
EM	Expectation Maximization
EQ-5D	European Quality of Life-5 Dimensions
FABQ	Fear Avoidance Beliefs Questionnaire
FHSQ	Foot Health Status Questionnaire
FPR	False Positive Rates
GFH	General Foot Health
GS	Grow Shrink
HC	Hill Climbing

MAE	Mean Absolute Error
MBC	Multidimensional Bayesian Network Classifier
MCDM	Multi Criteria Decision Making
MCID	Minimum Clinically Important Difference
MID	Minimum Important Difference
MII	Most Informative Inputs
MIPM	Mixed Integer Programming Model
ML	Machine Learning
MMHC	Min-Max Hill Climbing
MSE	Mean Squared Error
NPT	Node Probability Tables
NSCLC	Non-Small Cell Lung Cancer
PCS	Pain Catastrophizing Scale
PDQ-39	Parkinson’s Disease Questionnaire
PROM	Patient Reported Outcome Measure
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SF-12v2	Short Form Health Survey
SP	Symptomatic Patient
SVM	Support Vector Machine
TPR	True Positive Rates
TSW	Trauma SCAN Web
VP	Very Poor

1. INTRODUCTION

Artificial intelligence (AI) is transforming the way decisions made in the medical domain. Early medical applications of AI methods focused on using rule-based systems [1]. In such systems, a set of if-then based rules is used to define domain knowledge. However, in this approach, there are two problems. Firstly, a large number of rules have to be defined to model moderately complex systems. If the system does not perform as desired, even more rules need to be added. The addition of new rules to the system may cause the system to become inconsistent as they may change the logical relations between the existing rules. Secondly, and more importantly, rule systems do not incorporate uncertainty whereas medical decisions are inherently uncertain [2] [3].

More recently, machine learning (ML) approaches have become the primary approach in medical AI. ML uses computer algorithms to learn from past observations and data, and provides predictions and decision support accordingly [4]. ML algorithms are based on data mining [5], optimization [6], and statistical [7] techniques, and they are successfully used in many domains [8] [9] [10]. Probabilistic ML approaches, such as those based on Bayesian Networks (BNs), can handle uncertainty as well.

Medicine is one of the primary application domains of ML approaches. Successful medical applications of ML include oncology [11], diagnosing dermatoglyphic [12], thyroid [13], liver-related diseases [14] and cardiology [15]. Radiography and dermatology have been the main success stories of medical ML. These domains have a large amount of data where the patient outcomes (i.e. benign or malign) are clearly labelled. However, such clearly labelled patient outcomes data are not present in many other clinical domains which could potentially benefit from AI. ML could not provide successful results in most of these domains as the current ML techniques struggle to provide accurate results when there is a lack of clearly labelled data [16].

In many medical domains, patient outcomes are measured by using medical questionnaires. These are called patient reported outcome measures (PROMS). PROMs are used to measure a wide variety of patient outcomes including biomedical status, psychosocial status, and quality [17]. PROMS are typically composed of Likert or visual scale questions. Some PROMS may have over 30 questions [18] [19] [20], and the patients may be asked to fill them in regularly. Therefore, cognitive-load and time requirements for filling in these questionnaires can be difficult for patients to give accurate responses regularly [21]. This makes it difficult to collect a large amount of high-quality patient outcome data which can be used to train ML and AI tools.

This study focuses on developing intelligent PROMs that can learn from the patient's previous responses to PROM questionnaires, and that asks only the most informative questions to the patient. ML approaches can potentially be used for learning the relations between questions in the PROMs. Therefore, it can provide a way to understand the patients' health status with fewer inputs. This can be to assess whether the questions asked to a patient are sufficient to understand the patient's current situation or whether further questions should be asked in a dynamic and automated way. Bayesian networks (BNs) offer a suitable ML tool for this approach as they have algorithms for learning the causal and associational relationships among variables [22]. This can be used to learn the relations between different PROM questions. Moreover, BNs have also inference algorithms for computing probabilistic inference. This can be used to identify the most informative PROM questions to identify the state of a specific patient.

This study proposes an approach to build BN models for PROMs to predict the patient outcome when some PROM inputs are missing and to identify the most informative PROM questions for predicting patient-specific outcomes. We apply this approach to multiple PROMs from the musculo-skeletal disorders domain. Since the condition of each patient and their responses to PROM are different from each other, the most informative questions suggested by our approach can differ between different patients. The proposed approach asks PROM questions in iterations until the stopping rule is satisfied. It uses the posteriors calculated by the BN model and conditional entropy to calculate the most informative questions given the patient's previous answers in each iteration. The stopping rule assesses whether the uncertainty about the predicted outcomes is below a certain threshold. Therefore, the proposed approach can potentially reduce the time and effort required for collecting patient outcome data. To create relevant BN models, we use the patient data obtained from PROMs conducted by experts. We use data-driven algorithms to create BN models.

The PROM data in the study were collected by the physiotherapy unit located at Queen Mary University of London. Data are completely anonymous and provided by the physiotherapy unit of Queen Mary University of London. Therefore, further ethical permissions are not required for this data set.

In the remainder of this thesis, the second chapter presents an overview of BNs. Bayes' theorem and elements of it are described, then BNs introduced. BN parameters and parameter estimation methods are described as well as connection types in BNs. Necessary algorithms to build BN models are presented. Finally, Bayesian inference methods are examined in detail.

In the third chapter, previous medical applications of BNs are reviewed, and the previous studies that use the questionnaire data to create the BN model are examined. Early and modern applications of medical BNs that were established by using expert knowledge, data, and a combination of both have been analyzed.

In the fourth chapter, an approach to build BN models based on PROMs is shown. This chapter presents an overview of learning from PROMs, and shows the steps to create BN models and methods for evaluation. Also, an algorithm that is based on information theory elements to calculate estimates is shown.

In the fifth chapter, the proposed approach is applied to case studies. Methods that were shown in the fourth chapter have been used to learn different BN models. Prediction accuracies of different BN models were calculated. In addition, the patient status was estimated by the random questions in the PROMs. The estimation results obtained through random questions were compared with the estimation results of the method we proposed. At the end of this chapter, different models established by using different approaches have been discussed along with the validation results. In the final chapter, the conclusions of this thesis are presented.

2. BAYESIAN NETWORKS

This section provides an overview of BNs. Sections 2.1 and 2.2 introduces the Bayes theorem and BNs. In Sections 2.3 and 2.4 independence assertions in BNs and term d-separation are examined. In Section 2.5, algorithms and techniques to build BN models are reviewed. Finally, in Section 2.6 BN parameters and parameter estimation methods are discussed.

2.1. Bayes' Theorem

Bayes' theorem calculates the conditional probability $P(H|E)$ from its inverse conditional probability $P(E|H)$ is shown in Equation.(1.1).

$$P(H|E) = \frac{P(E|H)\ x\ P(H)}{P(E)} \tag{2.1}$$

where E and H are two separate events and $P(E)\neq 0$ and $P(H)\neq 0$. The Bayes theorem provides a way of revising existing 'prior' beliefs given new data and evidence $P(H|E)$ when H is interpreted as a hypothesis and E is interpreted as the data or evidence about a concept [23]. In other words, Bayes' Theorem enables individuals to update their subjective beliefs based on objective data. Owing to this useful feature, this theorem is used in various fields including AI and ML [24], time series models [25], medical field [24] [26], as well as data mining [27]. However, representing the problem and making Bayesian computations for a large and interrelated set of variables can be challenging. BNs offers a suitable environment for representing and making calculations of complex probability distributions in such cases.

2.2. Introduction to Bayesian Networks

BNs are graphical models and they represent the joint probability distributions. BNs can be used to define networks of causal and associational relations, and use a graphical framework to identify these causal relations [28]. The graphical structure of BNs is suitable to represent expert knowledge about causal relations. Using the graphical structure, we can evaluate the independence assertions imposed on the joint probability distribution. Data-driven algorithms are available to learn the BN partially or completely from data. Therefore, BNs offer a suitable modelling approach for combining data and expert knowledge.

A BN includes two essential components: one of them is a directed acyclic graph (DAG), the other one is node probability tables (NPT) [28]. Nodes represents variables in the graph, and arcs denotes the dependencies among these variables. Nodes and arcs create structure of DAG [29]. An example of a DAG is shown in Figure 1.

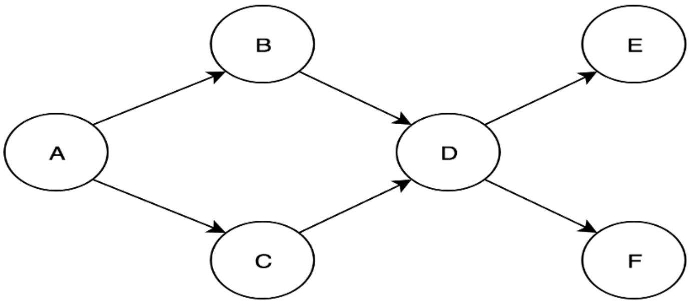


Figure 1 An Example Directed Acyclic Graph

Nodes in the Figure 1 denotes variables and arcs between them displays relationships of the variables. Direct causal relation which is represented by the arc from *A* to *B* means that *A* has influence on *B*. In that case, it is said that *A* is a parent node of node *B*. In the same context, node *B* is child node of node *A*. Since node *A* doesn't have any parent node, it is said that node *A* is a root node. The essential part of the DAG is that there is no loop between nodes. There is connection from *A* to *C* and *C* to *D*, in this case there cannot be arc from *D* to *A*, so that circular reasoning is avoided [28].

Given that parent nodes, the conditional probability of the child node is represented by NPTs which are associated probability tables of nodes in directed graphs. NPTs consist of probability values of nodes which have different states. For any root node, NPT is defines the marginal probability distribution of that node. For variables with parents, NPT defines the conditional probability distribution of the node [30]. An example for NPT shown in Table 1. Table 1 shows an example of the parameters of a node with two parents. BN variables can also be continuous, however, we only focus on discrete BNs in this study.

In a BN, any subset of variables can be observed, and the posterior probability of variables could be calculated using inference algorithms. These variables are the other unobserved variables given the observed ones. Observing a variable is also called ‘entering an evidence on a variable’. This enables the BN to be used as an inference or a predictive tool. Efficient inference algorithms [31] are widely implemented in commercial and open-source BN software, therefore they are not the focus of this study.

Table 1 An Example Node Probability Table

	B	***b1***	***b1***	***b2***	***b2***
	C	*c1*	*c2*	*c1*	*c2*
D	*d1*	0.4	0.3	0.2	0.5
	d2	0.6	0.7	0.8	0.5

2.3. Connection Types and d-Separation in Bayesian Networks

Connections between variables in a BN can be classified into three types: i.e. serial, diverging and converging connections [32]. These three types of connections can be used to describe the conditional independency assertions encoded in a BN.

2.3.1. Serial Connections

An example of a serial connection between 3 variables is shown in Figure 2. In this case, if *K* has no evidence (observation), then *H* and *L* are dependent on each other. Given that the state of *K* is unknown, there is an evidential influence between *H* and *L*. On the other hand, if the state of *K* is known, then states of *H* and *L* become independent each other; this means *H* and *L* are conditionally independent of each other given *K*. In this context, if the evidence is entered in node *K*, information flow between *H* and *L* is interrupted. If node *K* has no observation, when any evidence is entered node *H* or *L*, this would update the states of the other nodes in the causal network [33] [34].

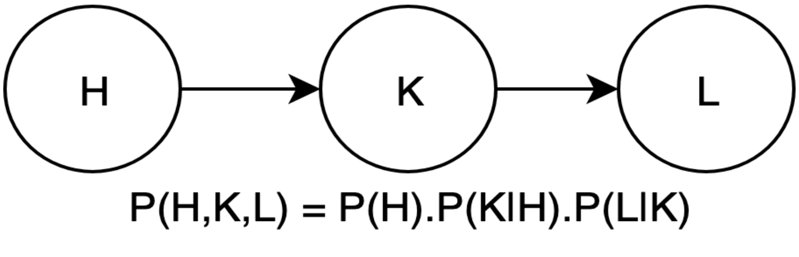


Figure 2 Serial Connection in Bayesian Network

2.3.2. Converging Connections

In a converging connection which is denoted in Figure 3, if the state of *K* is not known, then *H* and *L* are independent of each other. *H* and *L* have an evidential influence on node *K*, however, information cannot be transferred from *H* to *L*. In this case, parent nodes *H* and *L* are independent. However, if the evidence is entered in node *K*, parents of *K* become conditionally dependent given *K*. In this case, evidence entered on the parents of node *K* will update the probability distribution of the other parents of node *K* which are not instantiated. This situation is called "explaining away" [35] [34].

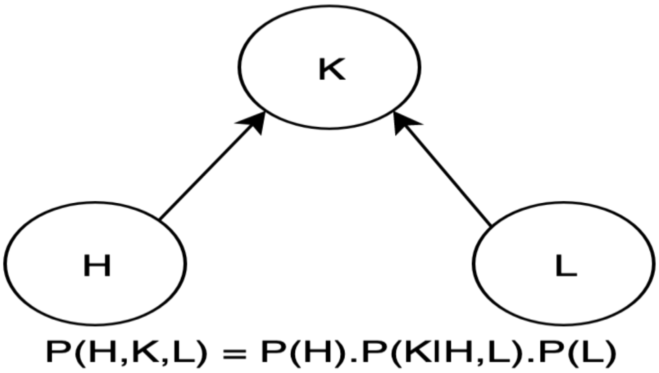


Figure 3 Converging Connection in Bayesian Network

2.3.3. Diverging Connections

In diverging connections shown in Figure 4, information can be transmitted between all the children of *K* if the state of *K* is unknown. If evidence entered into a child of *K*, the information will be transmitted to the other children through *K*. However, the children of *K* become independent once *K* is known [35] [34].

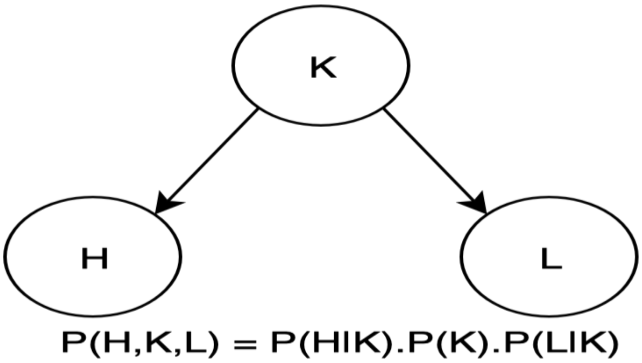


Figure 4 Diverging Connection in Bayesian Network

2.4. D-Separation

D-separation formally defines the conditional independence (CI) assertions for serial, converging, and diverging relations. The letter “D” stands for dependence. Given that two nodes *H* and *L* are d-separated relative to a node set of *C*, it can be said that these two nodes are independent conditional on *C*. In other words, node *H* and node *L* are conditionally independent on *C* if there is information about *C* and information about *H* gives no additional information about *L* or information about *L* gives no additional information about *H*.

Let *H*, *L* and *V* be three distinct node sets in BN structure *G*. *H* and *L* are d-separated given *V*, denoted $dsep_G(H; L|V)$, if and only if there is no active path between any node $H \in H$ and $L \in L$ given *V*. If there is an active path between two nodes, any information can be transmitted between these two nodes. Given that evidence nodes *E*, an active path requires two conditions:

- For every converging connection ($X \rightarrow Y \leftarrow Z$) on the path, node *Y* or one of its child nodes is in node set *E*.
- Other nodes on the path can’t be in node set E.

If *H* and *L* are not d-separated, they are d-connected [34].

In Figure 5 a larger BN example is shown. The evidence is entered in node *F*. Given this case if the evidence is entered in node *A*, it affects the state of node *B*. On the other hand, information cannot be transmitted to node *H* since variable *F* is blocked. However, information can be transferred to node *C* due to the serial connection.

In the context of d-separation, a minimal set of nodes that makes a node d-separated from the rest of the nodes in the BN is called the Markov blanket. Children of any node in the network, parents of the node, and parents of its children create the Markov blanket for that node [34].

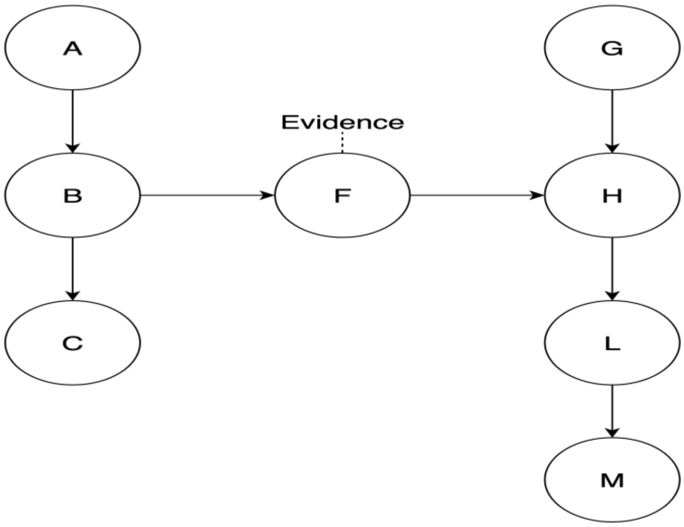


Figure 5 A Bayesian Network with *F* Instantiated.

2.5. Building Bayesian Network Structure

Creating BNs consists of two phases [36];

- Determining the structure of the BNs
- Obtaining NPTs of each node in the structure

The BN structure and NPTs in BNs can be determined by expert knowledge, data, or a combination of both. Defining a network structure by experts could be complex especially if number of variables that need to be defined are large. In such cases, the BN structure and parameters of each node are learned from data through data-driven approaches.

There are several data-driven algorithms used to learn BN structure from data. These algorithms are divided into three categories: constraint-based algorithms, score-based algorithms, and hybrid algorithms that combine constraint and score-based techniques.

2.5.1. Score-Based Algorithms

Score-based algorithms are used to find the BN structure which optimizes a performance score. Often heuristic search algorithms are used [37]. Bayesian information criterion (BIC) is a commonly used score for this task. In statistics, it is sometimes referred as Schwarz information criterion. BIC aims to find the structure that maximizes the likelihood but it also has a penalty term for additional relations to avoid overly complex structures and overfitting [38]. The Akaike information criterion (AIC) is another similar score used for this task. AIC also rewards likelihood and penalizes complexity but with different coefficients [39].

Hill climbing (HC) is a commonly used score-based algorithm. HC uses a greedy approach to find optimize the performance score. It adds, removes and changes the directs of arcs in the BN model to optimize the score. At every step, it chooses the arc operation that provides the highest increase in the score [40]. HC gets stuck at local optima or plateau situations where any single arc operations does not increase the score. For avoiding from this case random starting points can be chosen and algorithm can keep searching until the global maximum point is found [41]. The pseudocode of the HC is shown in Table 2.

Table 2 Pseudocode of the HC Algorithm

```
procedure Hill Climbing($D$)
  Input: Data $D$
  Output: DAG
for every variable $X$ in set $V$ do
Start with empty graph or random solution
$S$ = adding, subtracting, and changing the direction of the arc
$H_x$ = HC($D$,$X$,$S$)
If score difference f($S$) $\leq 0$
   Return the highest scoring DAG
End If
End For
end procedure
```

Tabu-Search algorithm also aims to maximize the score but it avoids doing operations that reverse the arc operations in a certain number of most recent previous operations. This aids the algorithm to avoid getting stuck at a local optimum [42]. The pseudocode of the Tabu is denoted in Table 3.

Table 3 Pseudocode of the Tabu Search Algorithm

```
procedure Tabu Search(D)
  Input: Data D, m : number of operations, C: stopping condition
  Output: DAG
for every variable X in set V Loop do
    n_iter = 0
    Start with empty graph or random solution
    S = adding, subtracting, and changing the direction of the arc
    Tx = Tabu(D,X,S,num_iter)
    n_iter = n_iter +1
    If n_iter = m
        Return the DAG
End For
If C is met
    Return the DAG
Else return Loop
End If
end procedure
```

2.5.2. Constraint-Based Algorithms

Constraint-based algorithms aim to identify conditional independence assertions between the variables in the data and builds a structure that is consistent with them. They apply statistical tests to the data to identify conditional independence. Constraint-based algorithms can be used for learning some causal relations in the form of converging relations from observational data [43]. Table 4 shows the pseudocode for the Grow-Shrink (GS) algorithm which is a widely used constraint-based algorithm [44]. There are two phases in this algorithm. The first one is the growing phase, the second is the shrinking phase. The algorithm starts with an empty set *S*, in the growing phase, each variable that is dependent with *X* is added to the set *S*. In the shrinking phase conditional independence between each *X* and each variable within *S* are tested to identify the Markov Blanket *B*(*X*) of variable *X*. Note that, *B*(*X*) is the minimal set of variables that makes *X* independent of the rest of the variables in the BN when they are observed [44].

Table 4 Pseudocode of the GS Algorithm

```
procedure Grow-Shrink (GS) Markov Blanket
  Input: Empty Set: $S$, Set of Variables: $U$
  Output: Markov Blanket of variable X
While every $Z \in U - \{X\}$ such that Z is dependent of $X \mid S$ do
$S \leftarrow S \cup \{Z\}$
While every $Z \in S$ such that $Z$ is independent of $X \mid (S - \{Z\})$ do
$S \leftarrow S - \{Z\}$
B($X$) $\leftarrow S$.
return B($X$)
end procedure
```

Another constraint-based algorithm is the PC algorithm. There are two different steps in the PC; the first step is the process of learning the skeleton structure from data. The second step is the process of aligning undirected edges to obtain the DAG structure [45]. Let Z be a subset and include all the neighbors of X and Y and let S be an empty set. In the skeleton structure learning step, to decide on about removing an edge or not, CI tests are applied to a fully connected network structure. CI testing for each edge is done. The tests check whether node X and node Y are independent conditionally on Z. By considering levels that are based on conditioning set sizes which are denoted by "d", CI tests are organized. These sizes are also called depth. When depth equals to zero, CI tests are applied to all vertices pairs conditioning on S. Edges between pairs are deleted if there is independency between two variables, and the algorithm continues with remaining edges. The level of depth increases progressively by one each time. When the level of depth becomes greater than the size of the vertices that are tested, the algorithm stops. The PC algorithm is efficient because when any edge is removed, adjacent sets of a certain node are updated [45]. The pseudocode of the PC is shown in Table 5.

Table 5 Pseudocode of the PC Algorithm

```
procedure PC
  Input: Empty Set: $S$, Depth:$d = 0$
  Output: DAG
For each edge in DAG do
  Start with fully connected network
  Independence test I($X$,$Y$|$S$)
  Remove edge $X - Y$ if $X$ does not depend on $Y$ given $S$
  Number of element: $s(S) = S + 1$, $S$ = {each neighbor of test nodes}
  $d = d + 1$
  If $d > s(S)$
    Return DAG
  End If
End For
end procedure
```

2.5.3. Hybrid Algorithms

Min-Max Hill Climbing (MMHC) algorithm is a hybrid algorithm. In hybrid algorithms, both score-based and constraint-based techniques are used. MMHC defines the frame BN structure by using a constraint-based algorithm, and the direction of the arcs are determined by maximizing the scoring function [46]. Unlike the standard HC algorithm, arc addition operation is only performed if the arc was discovered by Min-Max Parents and Children (MMPC) algorithm. The pseudocode of the MMHC is shown in Table 6.

In the first step, a heuristic search algorithm called MMPC is used which is shown in the third step of the pseudocode. MMPC is an algorithm that is used to obtain the skeleton structure of BN. Children and parents set of nodes of every variable X are found by MMPC and they are assigned to the PC_x variable. PC in the third step of pseudocode refers to children and parents set of variable X. After obtaining children and parents set of variable X, the standard HC algorithm is applied to find DAG with the highest score.

Table 6 Pseudocode of the MMHC Algorithm

```
procedure Max-Min Hill Climbing($D$)
  Input: Data $D$
  Output: DAG
for every variable $X$ in set $V$ do
    $PC_x$ = MMPC ($D$,$X$)
End For
Start with empty graph and apply Hill-Climbing algorithm by adding, subtracting, and changing the direction of the arcs. If $S \in PC_x$ adding operation is performed ($S \rightarrow X$).
Return the highest scoring DAG
end procedure
```

2.6. Defining Parameters of Bayesian Networks

NPTs of a BN can be defined by using data or knowledge of expert. The maximum likelihood estimation approach can be used to learn the parameters from data [47]. Alternatively, Bayesian parameter estimation approaches can be used to learn from data based on prior information about the parameters [23]. Methods which are used to learn parameters enable to combine expert knowledge and data have been proposed [48].

If the data is a missing data set or there is no data available for some variables; the expectation maximization (EM) algorithm can be applied. EM is an iterative algorithm and has two steps. In step “E”, started by assigning random values to the parameters of BN and the calculation is made according to parameters. At each iteration, the expected value is calculated by taking the missing data into consideration and the missing data is updated according to the expected value, and data were completed by this expected value. In step “M”, the maximum likelihood of the parameters was estimated. EM does not necessarily converge to a global maximum of likelihood, it can get stuck at local maximum [49].

Computation complexity in BNs increases with the size of parameters. To decrease parameter size and complexity in BN, different methods could be used. These methods could be used to decrease the size of parameters that are required to inferred from experts or learned from data. One way to simplify NPTs is to add an intermediate node between parent and child node. This process is called "parent divorcing" [50]. Noisy-OR and Noisy-Max gates can also be used to simplify NPTs [51] [52]. These models are useful to decrease the size of parameters in the BN model by assuming independence between and no interaction of effects between the parent nodes.

3. CLINICAL AND QUESTIONNAIRE BASED BAYESIAN NETWORKS

This section reviews the previous medical applications of BN models (Section 3.1), and previous studies that used questionnaire data for BN learning (Section 3.2).

3.1. Medical Bayesian Networks

Medical decision making involves diagnosing patients in uncertain conditions and selecting the treatment with the best possible outcomes. Bayesian networks are suitable for this kind of reasoning as they can make diagnostic and causal inferences under uncertainty. Therefore, since the late 1980s, BNs have been widely used for medical applications [53]. In this section, some articles containing developed clinical Bayes networks will be examined. The readers are referred to [54] for a detailed review.

Medical BN models can be created based on expert knowledge using knowledge engineering approaches, or purely through data using ML methods. Also, some medical BNs can be constructed by combining expert opinions and learning from data.

Early applications of medical BNs were purely based on expert knowledge. Heckerman et al. [55] developed Pathfinder project in 1989, which is a system based on purely expert knowledge for diagnosing hematopathology and diseases that appear in different lymph nodes. In their study, they developed a methodology for finding relations between findings and disease. They constructed a large probabilistic network after identifying features and diseases. Relationships between features and diseases were identified by experts beforehand. The accuracy of the model was compared earlier version of the program. The diagnostic accuracy of probabilistic model was measured by using two score-based metrics (expert-rating, formal decision-theoretic).

Beinlich et al. [56] developed an application, called ALARM (A Logical Alarm Reduction Mechanism), which is used to research techniques of reasoning in BNs. It was used to calculate differential diagnosis probabilities by using observations. They used medical knowledge in order to create a graphical structure that consists of diagnoses, findings, and intermediate variables. Purely expert-based knowledge was used to create a relevant network. Two distinct algorithms were applied to BN; one of them is a message-passing algorithm by Pearl [56] and the other one is the Lauritzen-Spiegelhalter [31] algorithm. These algorithms were used for probability computations and probability updating.

Another purely expert-driven BN for the medical domain was developed by Oniśko & Druzdzel [57]. They developed a BN model called in order to diagnose liver disorders. Also, six other BNs were developed by using knowledge of experts. Variables including liver diseases and findings were used in the structure of the model. Expert elicitation was used to determine relationships among variables. In the model validation phase, leave one out cross validation was used and diagnostic accuracy of all models was tested.

A knowledge-driven BN model was constructed by Sanders & Aronsky [58] to detect asthma attacks. They also used data from the pediatric emergency department. They learned BN parameters by using this data. They used Netica 2.0 software to develop three BN models which were based on expert knowledge and parameter learning was done via the same software. They assessed the prediction performance of BN through receiver operating characteristic (ROC) curves by using three-fold cross validation. Calculated area under curve (AUC) values were compared for evaluation. A comparison of the networks built by expert knowledge was made and the network with the highest prediction performance was chosen.

Advances in BN learning algorithms increased the number of purely data-driven applications of medical BNs. In 2009, Himes et al. [59] used data-driven approaches to construct BN in this context. In their study, the prediction of chronic obstructive pulmonary disease (COPD) is performed through BN. In order to find BN, they discovered different network models, and each of them was scored by their probability when the data was entered as evidence. They chose a model which had a maximum posterior probability. The network found by using K2 score-based algorithm. The focus of the study was about finding the nodes that have a direct effect on COPD and BN was constructed according to these nodes. Model validation was done by performing five-fold cross validation.

Another data-based BN was developed by Zheng et al. [60] in 1999 to diagnose breast cancer. Results of mastography findings, physical examinations, and related properties as well as clinical histories of patients were used to construct three Bayesian belief networks. The BNs in this study were constructed by applying ML algorithms. A training set was used to build BN considering dependence and independency of the selected variables. A BN with the highest prediction performance was chosen. The network in the study was built via Hugin software. Related probabilities were calculated from data that was selected for BN learning. Five-fold cross validation method was used to obtain prediction performance of BN which was performed by using the program ROCFIT. Zhao & Weng [61] developed a BN model to predict pancreatic cancer. Experts in the study determined variables in the model through a literature review. Risk factors that can affect pancreatic cancer were categorized under some variables which will be used later in BN development. The BN model was developed via expert knowledge. The prediction accuracy of the BN model developed in the study was compared with two different models using validation data. According to ROC curve results, the current BN model outperformed the other models and shows high prediction accuracy.

Petousis et al. [62] focused on the prediction of lung diseases. Dataset taken from lung cancer patients was used to create different dynamic Bayesian networks (DBNs) by using different processes such as forward and reversed arrow approaches in order to examine patient status. Variables in DBN were obtained from previous studies. Variables also included personal questions as well as questions such as family and personal cancer history. In order to reduce the dimension size of NPTs, some variables aggregated into one variable. States of variables were discretized by experts in advance. Since the dataset includes some missing data EM algorithm was applied in order to compute NPTs. DBNs were also compared to expert-driven model and other data-driven models such as logistic regression models. Model validations were done by applying the ten-fold cross validation method to overcome overfitting. Model comparisons were made by taking AUC results into account. According to results, DBNs outperformed expert-driven and logistic regression models.

In 2014, Jiang et al. [63] created a BN model to predict patient survivorship in the case of breast cancer and they used purely data-driven approaches. A developed model was used to predict survivorship for each year individually. The developed model could also handle large data. The model in the study was compared to other models and according to significance testing, it showed better performance than others. Five-fold cross validation method was used to test model robustness and related results are obtained for evaluation.

Hybrid approaches that benefit from expert knowledge and data have been used in the medical domain. These approaches use the graphical structure of the BN to encode expert knowledge. Then, BN structure and probability distributions are learned by using data and parameter algorithms. In 2005, Hoot & Aronsky [64] developed a BN to predict survival rates. In the study, the UNOS database which is a nationwide organization providing the suitable conditions to organ transplantation was used. After determining necessary predictors of survival for BN by using knowledge of clinical experts and by doing literature search relevant BN was constructed. Predictors that are used in BN were chosen from the pre-transplant variables that were available for the transplantation process. Analysis of performance and data operations was done through Matlab, whereas the creation of BN and relevant simulation was done via Netica software. A final BN model was created using data from 2000-2001, and three-fold cross validation was performed for prediction performance tests using an independent set which consists of data from 2002. Evaluation of prediction performance was done by using ROC curves and relevant AUC values.

A knowledge-based BN model was developed by Ahmed et al. [65] in 2009, was used for assessing patients suffer from abdominal and chest injury. Trauma SCAN-Web (TSW) which is a computer-based decision support system that is used for evaluating patients that have trauma resulting from the chest and abdominal injury was used in the study. All relevant probabilities were learned from data and with the help of expert knowledge, BN was constructed. In the study, there were also two networks that were created by ML approaches. One of BNs was including external wound variables, the other one didn't. And prediction performances of original BN for TSW were compared to performances of two BNs. Data in the study that were obtained from three hospitals and prediction performances were measured using ten-fold cross validation. Accuracy results of BNs were compared according to AUC values. BNs that were used by a decision support system, was determined sufficient to evaluate penetrating injuries.

Survival prediction model for non-small cell lung cancer (NSCLC) patients developed by Jochems et al. [66] is another example of an expert-based BN model. In the study, they developed the BN model to predict the survival of the patient. Experts defined the network structure in advance and variables of the BN model and they also determined relationships among variables. The conditional probability tables (CPTs) of each node (variable) were calculated by the use of a maximum likelihood technique. Maximum likelihood is a technique that is based on the EM algorithm. The BN model in the study was constructed through the JSMILE framework. Prediction performance analysis of data was done in R by using ROC curves. They performed five-fold cross validation on an independent set of data. According to prediction performances, the current BN model in the study outperformed previously developed models for this study. Jayasurya et al. [67] also predicted the survival rate of lung cancer patients. BN and support vector machine (SVM) models were constructed by using patient data. All variables were predefined by experts and the Markov chain Monte Carlo algorithm was used to create the BN model by using missing data. Parameter estimation was done through the EM algorithm. Two different models were compared using validation sets and comparisons were made according to AUC results. According to performance metrics, the BN model showed better prediction accuracy.

Yet et al. [68] combined both clinical data-based approach and expert knowledge to establish a BN model that consists of multiple variables. One of the main aims of the study was that the developed technique provides a way to obtain parameters from complex models. But especially it provides a way to learn parameters from the reduced size of the dataset. They used the developed technique in a case study to estimate the survival rate of patients with severely injured lower extremities. Experts in the domain determined relevant variables that will be used in the BN model. They also determined relationships between variables. They also proposed a method that combines data and univariate meta-analysis for learning parameters from the BN model that has multiple parents. The BN model developed in the study was compared with other models created through data-driven algorithms as well as different scoring models. They compared the performance of models using ten-fold cross validation method. According to AUC results, the developed model in the study outperformed other models.

Yet et al. [69] proposed a method for building clinical BNs that can make reasoning and predict in the presence of latent variables. They used both expert knowledge and data in order to create relevant models. Then, they applied the developed methodology to a case study. Developed BN model was used to predict acute traumatic coagulopathy (ATC) which is a blood disorder that increases the risk of death. Domain experts developed a model with variables including latent variables. Experts in the domain also elicited causal relationships between variables. They used the EM algorithm to learn the parameters of the model. By using a ten-fold cross validation method, model accuracy was tested. They measured model accuracy, discrimination, calibration. By examining AUC results and Brier scores, they evaluated the performance of the model.

Yet et al. [70] also proposed a clinical decision support system (CDSS) by using BNs in 2013. They tried to assist clinicians in Warfarin therapy management, which is therapy about preventing disorders including pulmonary embolism and atrial fibrillation. They developed a CDSS with the collaboration of Swedish hospital groups that are dealing with this therapy. The model they developed could help clinicians to make decisions about adjusting dose and follow-up intervals. It could also help clinicians to investigate variation causes and help them to evaluate risks related to therapy. They built the model structure with the help of medical literature in the domain, and with the assistance of physicians and nurses that were working in the therapy. They used Genie-SMILE software to build a BN structure. Model parameters were learned from a dataset of patients that received therapy for more than 14 days. Some parameters between some factors and variables were elicited by experts because data in the study were not enough to learn parameters from data. The variables and states of the variables were identified by nurses. Since the number of probabilities was high Noisy-OR/MAX gate models were used to deal with this problem. After they constructed a model and learned parameters of it, model validation and verification were tested. They applied the ten-fold cross validation model to measure prediction accuracy. They compared decisions that were made in the past cases and evaluate the model performance by looking at whether the model predicted the same decisions about dose adjustment as in the past.

Although BNs have been widely popular in the clinical domain, the use of BNs for PROM data has been limited. The following section reviews the BN models developed from questionnaire data in medical and other domains.

3.2. Questionnaire Based Bayesian Networks

Variables and the data required in the BN models can also be obtained from questionnaires. Fortini et al. [71] used a household survey and analyzed its variables to create a BN model. Data was obtained through conducting the survey in UK households and the BN model was built by the Bayes ware Discoverer program. After the data were obtained, by using the K2 greedy search algorithm four different BN models were obtained. Four different global precision values (i.e. α =1, 5, 10, 20) were used in order to build BN models. Accuracy measures of four different models were compared and according to evaluation, the most accurate model (α = 5) was chosen. In the same context, Constantinou et al. [72] created a BN model from questionnaires to provide medical decision support. Expert knowledge was used in the study to develop the BN model structure and data were collected from complex and incomplete questionnaires. One of the main aims of the study was to show how to build BN when the data is limited. They used different methods and compared predictive performances by considering AUC results.

Kaya & Yet [73] proposed a new method that uses a Multi-Criteria Decision Making (MCDM) approach called Decision Making Trial and Evaluation Laboratory (DEMATEL) to create BN models. DEMATEL uses survey questions to elicit causal relations, and their proposed method transformed the DEMATEL results into a BN model. The applied a proposed method to a case study. The robustness of the BN model was assessed through applying sensitivity analysis.

In 2017, a survey-based BN was developed by Bakshan et al. [74] for providing decision support for construction waste management (CWM). They created and applied a questionnaire to field workers in construction sites. They combined survey questions under certain factors and BN was obtained by using these factors. An associated BN model was created by determining factors that have a direct effect on behaviors towards CWM with the help of experts. Data collected from the survey was used to learn the CPTs in the BN model. Sensitivity analysis for single and multiple factors was done to see which factor or factors affect the behavior of workers toward CWM.

Borchani et al. [75] used Multi-dimensional Bayesian network classifiers (MBCs). MBCs are the models that have probabilistic graphs. These models were developed to deal with multi-dimensional classification problems. Features of the one vector or dataset are assigned to other features of the different datasets. The difference from normal classificationproblems is that features of the dataset are not assigned to single class value, instead features of the dataset are assigned multiple values. Since patients can have more than one disease, using the MBCs model would be a suitable option in this case. In the study, they used Markov blanket based algorithm for the model mentioned. The algorithm they used called HITON focuses on building the Markov blanket around each class variable. They tried to predict European Quality of Life-5 Dimensions (EQ-5D) by using items in the Parkinson’s Disease Questionnaire (PDQ-39). They firstly used algorithm and performed evaluations on generated and yeast data. Yeast is the multivariate dataset that was obtained from the ML repository. Two datasets were used to evaluate prediction performances. They also compared Markov blanket based algorithm with other BN based algorithms such as class-bridge decomposable, independent Markov blankets, PC, and k nearest neighbor. Then, they finally applied algorithms to real-world Parkinson’s disease data set. The five-fold cross validation method was used for evaluating the performance of alternative BNs. They used mean and global accuracy as well as mean squared error (MSE), mean absolute error (MAE), Pearson correlation (R^2), the absolute difference (AbsDiff), to see whether the model accurately predicts or not. In general, it was seen that the Markov blanket based model showed better results for almost all performance metrics.

Le & Doctor [76] developed a survey-based BN model for predicting the health status of patients. They used EQ-5D and Short Form Health Survey(SF-12v2) questionnaires in the study to obtain the structure of the relevant BN model. They tried to predict each of the items in the EQ-5D survey by using items in the SF-12v2 survey. Data obtained from questionnaires were used to learn the structure of the BN model by using the constraint-based method. They also learned parameters from available data. BN model was compared with other methods such as OLS, CLAD, MNL. Methods mentioned are the mapping methods in the literature. They used validation sets to evaluate the prediction performance of models. They also used statistical metrics such as MAE, MSE, and R^2 to make comparison between models. According to results, the BN model showed more accurate prediction performance.

Marvin et al. [77] created a BN to predict the effects of Nanomaterials and putting them in order according to their impacts. Variables in the BN, states of the variables and causal relationships among variables were determined by experts and by conducting a comprehensive literature review. Domain experts completed a questionnaire, and their opinions are asked to determine the final set of variables. They defined the initial structure of the BN model by the information that was obtained through the collaboration of experts. Then, they built the final structure of BN by using the EM algorithm. They also learned the model parameters using the same algorithm. They validated the created model by using an independent set of data collected from the relevant publications.

Susana et al. [78] benefited from data that was obtained through conducting a national survey based on working conditions in order to establish BN. In their study, the main aim of creating BN was to predict work-related accidents. Before creating the BN model, variables of them were determined by experts. The questions in the survey are classified under different factors including occupational accidents factor. By using these factors, occupational accidents which is an outcome variable was predicted. They have determined the factors that affect the occupational accident factor the most with this study. One of the constraint-based algorithms which is necessary path condition was used in order to obtain relevant BN. They used two-fold cross validation method to evaluate the prediction accuracy of the BN model. They have evaluated the performance of the BN model by considering the AUC result.

Blodgett & Anderson [79] developed a BN model to predict customer complaint behaviors. The BN model was built through Bayesian Knowledge Discoverer software. The dataset for building the BN structure was collected by conducting a questionnaire about customer attitude towards the store. They applied the k-means clustering algorithm and classified questions in the survey under multiple factors. The BN model structure associated with these factors were learned from data by using the TETRAD software. They used CI tests to determine causal relationships between factors and create the BN model structure. They used the what-if analysis to test the performance of the model. Chakraborty et al. [80] used both expert knowledge and survey data to create a BN model to predict customer satisfaction. A survey was conducted to people who use public transport. The variables, states of variables and causal relationships for the BN model were determined from the survey together with domain experts. The parameters of the BN model were calculated by a combination of expert elicitation, and data-driven approaches. Validation of the model was done based on expert opinion. The experts evaluated whether the results of the BN model matched with their opinions.

Another BN model for the customer satisfaction model was developed by Salini & Kenett [81]. They used customer satisfaction surveys to create the BN model. They aimed to create a BN model to predict the overall customer satisfaction and to determine whether customers would buy a product again. The variables in the BN were determined first. Then states of variables are defined based on the survey. The BN model structure was created by integrating expert knowledge and the Greedy Thick Thinning algorithm. Anderson et al. [82] also focused on building BNs about customer satisfaction and loyalty. Their dataset was obtained from a questionnaire, which was created with the collaboration of firm managers, academicians, and consultants. They created different BN model structures, firstly by identifying causal relationships between variables with PC algorithm. Then, they evaluate Bayesian scores of the models they found and chose the BN model that has the highest score. They applied the ten-fold cross validation method to assess the prediction accuracy of the final BN model.

Mohammadfam et al. [83] constructed a BN model for improving the safety behaviors of workers in workplaces. It was another illustration of questionnaire-based models. They tried to predict the safety behavior of employees. The data in the study were collected from a questionnaire filled by employees which was conducted at power plant constructions in Iran. They conduct factor analysis to the questionnaire data to define the variables in the BN model. They have used both expert elicitation and structure learning algorithms to create BNs. They have found out that there was no significant difference between the BN models that were obtained through these approaches. Therefore, they chose to use the BN model structure that was created by using expert knowledge. They learned BN parameters from data by using the EM algorithm after they determined the BN structure. They evaluated the accuracy of the BN model by using confusion tables.

In summary, questionnaire data has been a promising resource for developing BN models and several techniques has been developed. However, reviewed studies did not offer a general approach for predicting outcomes with PROM data, and for finding the most informative variables in a PROM. In the next section, we propose a novel approach to deal with these issues.

4. PROM FRAMEWORK FOR BAYESIAN NETWORKS

In this section, a way to prepare medical data for BN and a novel algorithm to calculate estimates of patient status are presented together. We will also talk about the steps of the method that we developed. The flowchart of the proposed method is shown in Figure 5 in a sequential order.

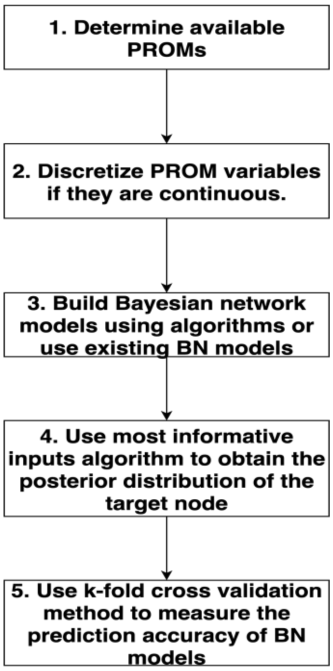

1. Determine available PROMs
2. Discretize PROM variables if they are continuous.
3. Build Bayesian network models using algorithms or use existing BN models
4. Use most informative inputs algorithm to obtain the posterior distribution of the target node
5. Use k-fold cross validation method to measure the prediction accuracy of BN models

Figure 6 Flowchart of Proposed Method

In the first step, available PROMs that will be used in the study are selected. Data in the study are obtained by conducting these PROMs. Then variables in the PROMs are examined. In the second step, the variables to be estimated in the PROMs need to be discretized if they are continuous. An overview of PROM data and the need for discretizing their factor scores is described in Section 4.1. The techniques to discretize factors scores into binary and multinomial states are shown in Sections 4.2 and 4.3. The third step shown in the flowchart is to build BN models using learning algorithms and using a predefined BN structure template. In Section 4.4, methods to learn BN models are briefly discussed. An algorithm to identify the most informative variables for prediction in the BN is described in Section 4.5. The last step in the flowchart is conducting a k-fold cross validation method to measure the prediction accuracy of models by comparing real values and estimates and it is presented in Section 4.6.

4.1. Overview of PROMs

This section will give an overview of PROMs and the requirements for learning BN models based on PROM data. For this purpose, the foot health status questionnaire (FHSQ), i.e. one of the PROMs we use in the case study in Section 5, will be used as an example.

PROMs are questionnaires filled in by the patients to measure and assess different aspects of their health. By using PROMs, symptoms of patients, their medical status, their life quality are evaluated for further decision making [17]. Earlier PROMs focused on evaluating the effectiveness of treatments. However, they are now generally used to evaluate their health status and the outcomes of the selected care. PROMs can be repetitively at different time stages of care to dynamically measure the patient status.

PROMS are usually composed of Likert or visual scale questions. Scores can be calculated for different patient factors, such as severity or functionality, based on these questions. These scores represent the overall measurement for the associated factor of the particular patient. Some PROMs consist of several questions; it may be difficult for patients to fill them in regularly. This can be one of the limitations of PROMs [21].

FHSQ is a PROM that was developed to examine the foot health status of patients and how this affects their lives. It consists of 13 questions with a 5-point Likert scale, where a score of 1 represents the worst status, while a score of 5 represents the best status for the associated questions [84]. FHSQ questions are shown in Table 7.

Table 7 Foot Health Status Questionnaire

	None	**Very Mild**	**Mild**	**Moderate**	**Severe**
1. What level of foot pain have you had during the past week?	5	4	3	2	1
	Never	**Occasionally**	**Fairly Many Times**	**Very Often**	**Always**
2. How often have you had foot pain?	5	4	3	2	1
3. How often did your feet ache?	5	4	3	2	1
4. How often did you get sharp pains in your feet?	5	4	3	2	1

	Not at all	**Slightly**	**Moderately**	**Quite a bit**	**Extremely**
5. Have your feet caused you to have difficulties in your work or activities?	5	4	3	2	1
6. Were you limited in the kind of work you could do because of your feet?	5	4	3	2	1
7. How much does your foot health limit you in walking?	5	4	3	2	1
8. How much does your foot health limit you from climbing stairs?	5	4	3	2	1
	Strongly Agree	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**
9. It is hard to find shoes that do not hurt my feet.	5	4	3	2	1
10. I have difficulty finding shoes that fit my feet.	5	4	3	2	1
11. I am limited in the number of shoes I can wear.	5	4	3	2	1
	Excellent	**Very Good**	**Good**	**Fair**	**Poor**
12. How would you rate your overall foot health?	5	4	3	2	1
13. In general, what condition would you say your feet are in?	5	4	3	2	1

FHSQ questions are associated with certain factors indicating the general health status of the patient: 4 of these questions are related to pain, 4 questions are related to function, 3 questions are related to footwear and 2 questions are related to general foot health (GFH) factor. A score is calculated for each factor based on the answers of these questions. These scores are in a continuous scale ranging from (0 – 100), where a total score of 100 indicates the best foot status, while the total score of 0 indicates the worst foot status.

The purpose of the BNs that we will develop in the study is to predict these factors by using a subset of the PROM questions. Our BNs aim to classify the patients into groups according to the severity of their condition. Although higher factor scores indicate good outcomes and lower scores indicate bad outcomes; it is not clear what should be the thresholds for classifying the patients based on these scores. In other words, it is not certain scores below which value indicates poor health. For example, people with poor biomedical status may have high scores due to their personality or psychosocial state.

Therefore, “measurement uncertainty" associated with these scores and the underlying health needs to be taken into account [85].

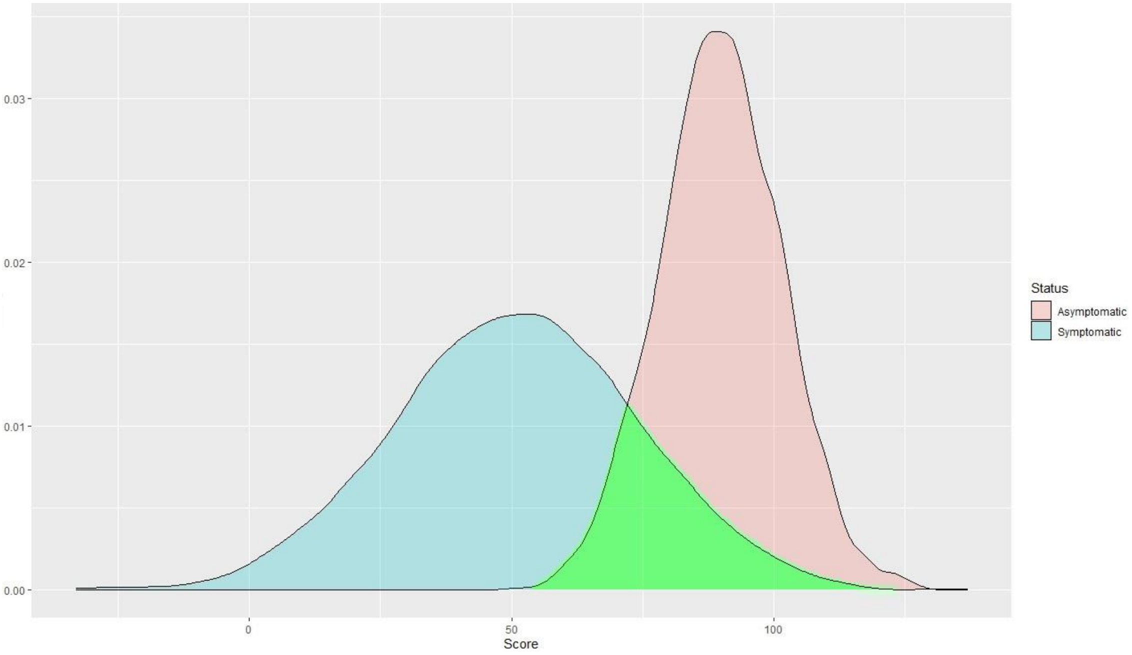


Figure 7 FHSQ Pain Factor Symptomatic and Asymptomatic Patient Scores

Figure 7 shows a graphical illustration of this uncertainty associated with the FHSQ factor scores and the underlying condition. The blue density plot shows the factor score distribution of the people with symptomatic foot problems. The red plot shows the score distribution of the people with asymptomatic foot problems. Note that, these two distributions overlap as shown in the green area. Some symptomatic patients can report high scores, and similarly, some asymptomatic patients can report low scores.

When learning BNs for classification, we need to discretize these factors into either binary normal/unhealthy states, or multinomial states like worse/bad/good. However, the measurement uncertainty mentioned before makes this discretization challenging. To perform binary discretization, a threshold (cut-off) must be determined between normal and symptomatic patients' scores. If discretization is done to obtain more than two states, the intervals of these states must be determined. To discretize these scores, we use a define a ‘cut-off’ point for distinguishing healthy and sick patients. We then use the minimum clinically important difference (MCID) parameter in a mixed-integer programming model for the discretization of continuous score variables in PROMs. In the following sections, the outline of the method used to create the Bayes network, including the approaches to determine the binary and multinomial discretization, is described.

4.2. Determining Minimal Clinically Important Difference

One of the main issues of using questionnaires for decision support is interpreting the PROMs. Most PROM results are numerical scores that reflect patient status. However, clinicians may not be sure how to interpret these scores for evaluating the patient and comparing the outcomes before and after treatment. Natural variation and randomness in patient responses make this more challenging.

MCID studies aim to find the smallest difference that indicates a clinically important change in the patient's state [86]. In some studies, also defined MCID as the minimal important difference (MID). Therefore, both terms could be used. MCID is helpful for clinical decision support since it provides information about whether there is a significant improvement in patient condition. In addition, it enables clinicians to assess patients' conditions.

In literature, there are several ways to determine the MCID of a PROM. One of them is an anchor-based method. It is based on comparing patient status before and after treatment. Conducting medical questionnaires and asking general questions to patients about how they feel before and after treatment is an essential part of this method [87]. The patient's answers are recorded as scores. Experts compare differences in scores of patients after treatment. As a result of treatment, the average score of patients who feel better and who feel the same are examined. The difference between these two average scores is used to determine MCID. This method includes different features that vary from patient to patient, therefore one patient may make progress in the healing process as a result of treatment, however, another patient may need more time to make progress as a result of the same treatment. Progress in the outcome scores could be significant in a statistical way, whereas this does not mean that it has to be significant in a clinical way. In large sample size, even changes that can be counted as small, make a statistically significant difference. Therefore, in the process of making decisions, clinicians determine if relevant treatments achieve clinically important outcomes or not [88].

Anchor-based method's main advantage is that there is a regular exchange of information between patients and doctors, patients inform clinicians for specifying clinically meaningful changes. However, if some patients show progress in the healing process and conditions of some patients the remain same, using the anchor-based method would not be a suitable option. Calculations in this method are not made according to patients that show progress in healing process vs patients whose status remains the same, on the contrary, they are made according to patients which show some progress and which show significant improvements [89].

Another method in determining MID is the Delphi method. It is based on consensus resulting from experts who share their ideas and brainstorm in order to determine MID. The council that consists of experts makes discussion about the result of a study. Experts examine this result separately, as a result of this examination they express their decisions about MID for the relevant study. Average of their feedbacks about estimates of MID is taken and summarized and this summary is sent back to each of them. This process continues until they achieve consensus. [90].

MCID values found in the literature with this method can be used to work with clinical questionnaires. However, we can obtain MCID value by using our data in the studies. For this purpose, we can use distribution-based methods (DBMs). DBMs determine MCID by using data based on standard deviation (SD) [91], whereas the anchor-based method depends on the judgment of experts. However, both methods can have similar results in the same study. Since DBM is based purely on calculations of statistical metrics and it is related to data in the study, using this method in every study may not be the most suitable option [92].

Norman et al. [93] suggest MCID to be determined as approximately one half of the SD of outcome measure. In their study, a total of 33 different studies which were conducted to determine MCID value were examined. The studies they examined include different medical questionnaires that consist of different scoring scales. In each study, the anchor-based method was used to determine MCID value. Averages and SDs of patient scores before and after treatment were obtained. And they have seen that for each study MCID values obtained through an anchor-based method, consistently equal to approximately half of the SD value of pre-treatment patient scores. In this thesis, we follow Norman et al.'s approach of determining MCID as half of the SD value.

4.3. Using Minimal Clinically Important Difference to Discretize Continuous Variables

After obtaining MCID values by applying methods in Section 4.1, we use these values to discretize the continuous variables. Variables in the study are obtained from the medical questionnaires. MCID and cut-off points could be used to determine discrete states of each factor in the PROMs [94] [95].

In the literature, there may not always be a cut-off point for the relevant scales. For such cases, the cut-off point for variables can be determined using the distributions of variables in the data. In order to determine cut-off point for medical scales, using distributions of patient data would be convenient. To do that, we can use the means of symptomatic patients (SPs) as a cut-off points. But we cannot take mean scores directly and determine it as a cut-off point.

If we want to use the means and SDs that were obtained from distributions, we first determine MCID for relevant variables in data by using DBM. By using means and SDs of symptomatic and asymptomatic patients (ASPs), we determine the cut-off point where two distributions have equal density. We can see the procedure of obtaining cut-off points using means and SDs of SPs and ASPs that were obtained from data in Figure 8.

In addition to this information, if the scale in hand is desired to be discretized to more than two states, MCID scores and SDs can be used for cut-off points obtained from literature and data respectively. We defined a mixed-integer programming model (MIPM) for determining additional states. Model parameters, objective function, and decision variables are defined below with constraints.

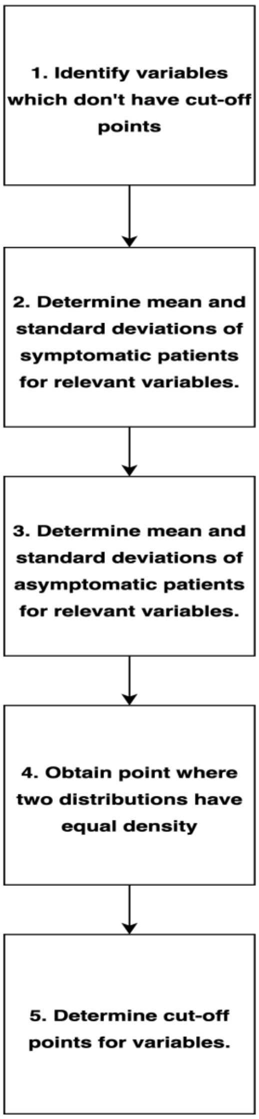

1. Identify variables which don't have cut-off points
2. Determine mean and standard deviations of symptomatic patients for relevant variables.
3. Determine mean and standard deviations of asymptomatic patients for relevant variables.
4. Obtain point where two distributions have equal density
5. Determine cut-off points for variables.

Figure 8 Procedure of Obtaining Cut-off Points for Variables

Decision Variables:

d_i: Length of interval *I*, continuous variable between *LB* and *UB*

y_i: Binary variable that ensures one *di* coincides with *T*

f_{ij}^{+}: Variable used for determining the absolute value of the difference between d_i and d_j

f_{ij}^{-}: Variable used for determining the absolute value of the difference between d_i and d_j

Parameters:

T: threshold for symptomatic and asymptomatic patients

MCID: Minimal clinically important difference

n: Number of intervals

M: A big number

UB: Upper bound of continuous scale

LB: Lower bound of continuous scale

Objective function

$$\text{Min} \sum_{i}^{n}\sum_{j}^{n} f_{ij}^{+} + f_{ij}^{-}$$

Constraints:

$$\sum_{i=1}^{n} d_i = UB \qquad (4.1)$$

$$di \geq MCID \quad \forall i \in \{1, \dots, n\} \qquad (4.2)$$

$$\sum_{l=1}^{i} d_l \geq T - (1 - y_i)M \quad \forall i \in \{1, \dots, n\} \qquad (4.3)$$

$$\sum_{l=1}^{i} d_l \leq T + (1 - y_i)M \quad \forall i \in \{1, \dots, n\} \qquad (4.4)$$

$$\sum_{i=1}^{n} y_i = 1 \qquad (4.5)$$

$$d_i - d_j = f_{ij}^{+} - f_{ij}^{-} \qquad \forall\, i, j \ \in \{1, \dots, n\} \qquad (4.6)$$

$$y_i \in \{0, 1\} \qquad \forall\, i \ \in \{1, \dots, n\} \qquad (4.7)$$

$$f_{ij}^{+}, f_{ij}^{-} \geq 0 \qquad \forall\, i, j \in \{1, \dots, n\} \qquad (4.8)$$

$$LB \leq di \leq UB \qquad \forall\, i \in \{1, \dots, n\} \qquad (4.9)$$

Objective function aims to minimize the absolute difference between interval lengths. In other words, we want to get as equal intervals as possible in terms of lengths. With Constraint (4.1), the sum of the intervals is ensured to be equal to the upper limit. Constraint (4.2) provides the condition that each interval variable is greater than or equal to MCID. With Constraint (4.3), (4.4) and (4.5), it was ensured that one of the interval borders must coincide with the threshold value. With Constraint (4.6), absolute value of the differences between interval lengths is defined. Constraint (4.7) is binary constraint defined for variable y_i. With Constraint (4.8), bounds of f_{ij}^{+}, f_{ij}^{-} variables are defined. With constraint (4.9), bounds of interval lengths are defined. As a result of this model, we get interval lengths and we define new interval and states for relevant PROM variables that are based on continuous scale.

4.4. Learning Bayesian Network Models

After discretizing continuous variables, we learn the BN models by using the score-based, constraint-based, and hybrid algorithms described in Sections 2.6. In Section 5, we use the HC, Tabu Search, and MMHC algorithms to learn the BN models as some of the most widely used score-based and hybrid learning algorithms. However, other structure learning algorithms can also be applied. Alternatively, BN structures for PROMs can be built based on expert knowledge.

4.5. Identifying Most Informative Variables for Predicting Outcomes

In our algorithm, we will use entropy and mutual information to identify the most informative variables for predicting the patient outcomes in the BN. In information theory, the amount of uncertainty of any variable denotes entropy of that variable [96]. The entropy of a discrete random variable L is shown in Equation. (4.10).

$$H(L) = -\sum_{\forall l \in L} P(l)\, log_2 P(l) \tag{4.10}$$

where l represents the states of L. The entropy of a continuous variable is calculated as shown in Equation. (4.11).

$$h(f) = -\int f(x)\, ln\big(f(x)\big)\, dx \tag{4.11}$$

Entropy provides a measure of uncertainty for a random variable. The entropy value of an event is maximum if the uncertainty of a random event is maximum, in other words, when outcomes of an event have equal probability. However, the entropy value is minimum if the event is known, in other words, if an outcome has the probability of 1 [97].

Conditional entropy is the uncertainty of one variable given knowledge of another. Conditional entropy of L given M, *i.e.* $H(L|M)$ is calculated as follows for discrete variables:

$$H(L|M) = -\sum_{L,M} P(L,M)\, log_2 P(L|M) \tag{4.12}$$

where $P(L|M)$ defined as the conditional probability of L given M. Conditional entropy represents the entropy of variable L when the outcome of M is known [97].

Mutual information $I(L,M)$ of two random variables L and M represents the amount of information provided when the outcome of one of these variables is known. Note that, when further information about one of these random variables is obtained, the entropy of the other random variable either decreases or does not change [96]. Therefore, mutual information is greater than or equal to 0. Mutual information is defined as follows:

$$I(L,M) = H(L) - H(L|M) = H(M) - H(M|L) = H(L) + H(M) - H(L,M) \tag{4.13}$$

The mutual information that is shown in Equation. (4.13) is symmetric: $I(L,M) = I(M,L)$. If the mutual information is small or near zero, that means random variable L and random variable M are independent or close to being independent. If the mutual information value is maximal, one of the random variables is almost determined by the other, that is, they are not independent [98].

4.5.1. Algorithm for Finding Most Informative Inputs

In this section, we show an algorithm that provides us the most informative questions for predicting the outcome based on mutual information.

Firstly, target nodes that are aimed to be predicted, and input nodes that will be answered by the patient are defined. Next, the BN is computed, and the entropy of the target nodes and the mutual information between target and input nodes are calculated. Input node that has the maximum mutual information provides the highest amount of information for predicting the outcome. Therefore, the node with the maximum mutual information is asked to the patient, the patients' answer is instantiated in the BN, and the BN is calculated. Since this node is instantiated, it is removed from the list of input nodes. Afterward, the stopping conditions are checked, and if they are not satisfied, these steps are repeated with the remaining nodes in the list of input nodes.

The first stopping condition in our case is when the difference between mutual information becomes less than or equal to a threshold that was chosen in advance. This threshold value represents the case that the remaining inputs do not provide sufficient information to predict the outcomes. The second condition is reaching the maximum number of questions that are allowed to be asked. Once the stopping conditions are met, the posterior distribution of the target nodes is calculated and shown to the user. The flowchart of the algorithm is shown in Figure 9.

Once the posterior distributions of the target nodes are calculated, these can be compared to the actual data of the target node to assess the predictive accuracy of our approach. Process of comparison actual values and estimates is described in detail in Section 4.6.

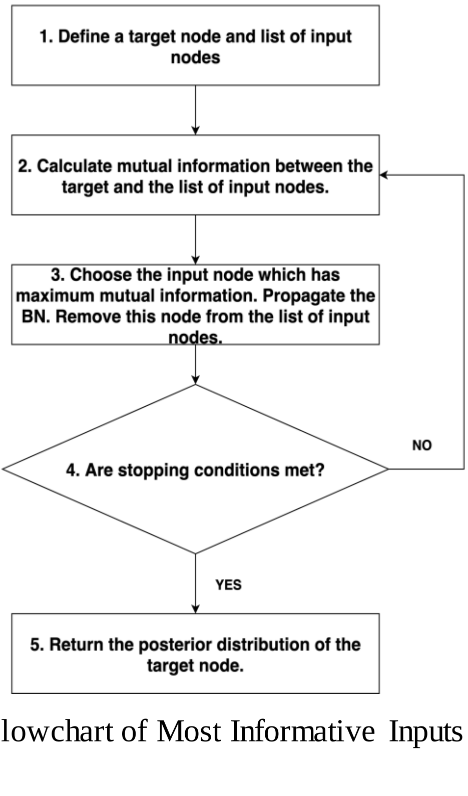

1. Define a target node and list of input nodes
2. Calculate mutual information between the target and the list of input nodes.
3. Choose the input node which has maximum mutual information. Propagate the BN. Remove this node from the list of input nodes.
4. Are stopping conditions met?
NO
YES
5. Return the posterior distribution of the target node.

Figure 9 Flowchart of Most Informative Inputs Algorithm

4.6. Prediction Accuracy of Bayesian Network Model

Cross validation is a technique to validate models by using independent data sets [99]. Cross validation method is based on the division of samples into construction and validation sub-samples. Construction data set is used to learn models, whereas validation sub-sample in other words unknown test data is used to test model validity. The main aim to use this method is obtaining the prediction accuracy of models when using unknown data to get estimates. Also dividing data into validation and train sets prevents overfitting problem. Overfitting occurs when the model is learned and validated with the same data set. In such situations, the model provides biased results as it overfits to its training data [100]. To use the whole data for training and validation without overfitting k-fold cross validation method can be used. This method divides data into equal k parts. Firstly, the k-th part is chosen for validation, then the rest of the parts are chosen for learning models. For each part in the data set, operations are repeated. To obtain the final result, the validation results of each part are combined [101].

We used Receiver Operating Characteristic (ROC) curves to evaluate our model accuracy in k-fold cross validation. Since the predictions of BNs are probability values, a certain threshold value is needed. After determining the threshold value, it is then used to evaluate the discriminating power of a BN model for predicting a value. For example, suppose a BN is designed for predicting the presence of a disease, and its predicted posterior probability for this disease 0.65. Whether this probability represents the presence of disease or absence of disease depends on the threshold value defined for predictions of this BN.

ROC curves are used to examine the overall performance of the BN, rather than assessing it for a specific threshold. ROC curves are used to test the discriminating power of different models at various thresholds. At various thresholds, true positive rates (TPR) and false positive rates (FPR) of predictions are compared in the graph. True positive rate is called sensitivity, and it is rate or probability that the classifier in the model correctly predicts positive values. Specificity also called true negative rate. It is the rate that the classifier in the model predicts negative values correctly [102]. In this context, $\text{FPR} = 1 - \text{Specificity}$.

The area under the ROC curve (AUC) can be used to provide an overall performance measure based on the ROC curve. The diagonal line starting at the top right corner and extending to the bottom left corner of the ROC curve is called the ‘no discrimination line’. The AUC value of this line is 0.5. AUC values are ranging from 0.5 to 1, where 0.5

indicates the worst prediction performance while 1 indicates the best prediction performance. AUC of the top left corner (0,1), represents the best prediction performance. Therefore, AUC values represent the accuracy of the model [103].

5. CASE STUDY & RESULTS

We applied the proposed approach to four different PROMs. Before learning the BN models, we discretized the factor scores for each PROM based on cut-off and MCID values as described in Section 4. MCID values were calculated from the data as described in Section 4.2. If the cut-off values for a PROM were available in previous publications we used those values, otherwise, we calculated cut-off values as described in Section 4.3.

We predict the pain, function, and GFH factors from the FHSQ questionnaire as these are the main factors related to the patient's health. Landorf and Radford [95] published the MCID scores for each domain of FHSQ. In this study, for the pain domain of the questionnaire, the score of 13 was determined as MCID. Scores of 7, -2, and 9 were determined as MCID for the function, footwear, and GFH domains respectively. MCID score of -2 for footwear shows that this factor was not as sensitive as other factors. Hence, it was seen that footwear factors are independent from pain, GFH, and function domains. It couldn't detect the improvement in the patient status which was achieved from interventions. An anchor-based approach with a 15 point Likert scale was used to determine MCID for this questionnaire. Cut-off points for the symptomatic and asymptomatic patients for the domains of FHSQ were not found in previous studies.

We applied the approach shown in Figure 8 for FHSQ domains. After applying this procedure, we determined the cut-off scores of 72.2 for the pain domain, of 73.3 for the function domain, of 55.2 as for the GFH domain respectively. Therefore, for these factors, scores that are below cut-off points can be interpreted as worse foot state, while higher scores can be interpreted as better foot state. According to results obtained from DBM, MCID scores determined as 11.5, 15, and 12.5 for pain, function, and GFH respectively. For all medical questionnaires in this study, the mean scores of SPs and ASPs obtained from literature and data are shown in Table 8.

Table 8 Mean Scores of Symptomatic and Asymptomatic Patients

	Literature		Data	
	Symptomatic	*Asymptomatic*	*Symptomatic*	*Asymptomatic*
FHSQ-Pain	52.6[104]	89[105]	50.1 ± 23	89.7±11.5
FHSQ-Function	60.7[104]	99[105]	56.1 ± 30	92.3±13.8
FHSQ-Footwear	44.4[106]	68.8[106]	52.1± 25	37.6±26.6
FHSQ-GFH	26.5[104]	79[105]	37 ± 25	75±18.9
FABQ-Work	26.7[107]	4.4[108]	11.2 ± 9.3	6.7 ± 7.7
FABQ-Physical Activity	14.0[107]	3.0[108]	14.2 ± 5.5	9.1 ± 6.5
PCS	12.5[109]	8.24[110]	14.6 ± 12.0	10.8 ± 10.3
CSI	27[111]	21.55[112]	30.4 ± 15.0	27.1 ± 13.6

In order to discretize FHSQ scores for building the BN model, MIPM described in Section 4.3 is used. We discretized FHSQ to five states. The discretized states for the pain factor were as follows: the interval [0 – 18.06) is the *bad* state, [18.06 – 36.1) is *very poor (VP)*, [36.1 – 54.1) is *poor*, [54.1 – 72.2) is *fair*, and [72.2 – 100] is the *good* state. For function factor, the interval from [0 - 18.33) stated as *bad*, from [18.33 - 36.67) stated as *VP*, from [36.67 - 55.01) stated as *poor*, from [55.01 - 73.3) stated as the *fair* state, and from [73.3 – 100] stated as the *good* state. For the GFH factor, the interval from [0 - 18.41) stated as *bad*, from [18.41 - 36.83) stated as *poor*, from [36.83 - 55.2) stated as *fair*, from [55.2 - 77.6) stated as *good* state, and from [77.6 – 100] stated as the *excellent* state. We did not build a model for predicting footwear due to insensitivity of MCID of this variable [95].

In order to obtain accuracy measures of BN models, we used the 5-fold cross validation method. We used the most informative inputs (MII) algorithm shown in Figure 9 to predict the factors in each fold. We used AUC values to evaluate the performance of different factors and learning techniques.

FHSQ data in the study consist a total of 409 observations. By using the MII algorithm, for each patient or observation, the most informative 2,4 and 6 questions were determined. Note that, at the calculation stage of the estimates, the most informative 2,4 and 6 questions could be different for each patient in the data. The validation results were obtained separately for the number of inputs determined. Three different pre-defined algorithms including HC, Tabu as well as MMHC were used to obtain BN models.

The second medical questionnaire used in the study is the fear avoidance beliefs questionnaire (FABQ). It is a questionnaire that is used to determine patients who have chronic low back pain. Patients feel fear as a result of their pain and avoid doing physical activity. FABQ was created to measure pain and physical activity variables [113]. There are 16 questions and 2 factors in this medical questionnaire. A 7 point Likert scale was used ranging from a score of 0 to 6, where a score of 0 indicates the expression of strongly disagree, while a score of 6 indicates the expression of strongly agree. One of the factors is work and the other one is physical activity. Work factor is based on a continuous scale and total score of it ranging from (0 - 42*)*, where the total score of 0 shows the best health status, while the score of 42 indicates strong fear avoidance belief. The physical activity factor is also based on a continuous scale. The total score of this factor ranging from 0-24, where a total score of 0 indicates the best health status, while a score of 24 indicates strong fear avoidance belief.

By using DBM, for our data, MCID score for work factor determined as 4.65, and MCID score for physical activity factor determined as 2.75. The cut-off point for work factor was found as 29 and for physical activity factor, it was found as 13 [114]. Therefore, scores below the cut-off point show good health status, whereas scores above the cut-off point show bad health status. If the scale in hand is desired to be discretized to more than two states, we can use MIPM in this situation as we did for FHSQ. Using additional cut-off points obtained from MIPM, for physical activity factor, the interval from [0 - 4.33) stated as *excellent*, from [4.33 - 8.66) stated as *good*, from [8.66 - 13) stated as *fair*, from [13 - 18.5) stated as *poor*, and from [18.5 – 24] stated as *bad* state. For work factor, the interval from [0 - 9.66) stated as *excellent*, from [9.66 - 19.33) stated as *good*, from [19.33 – 29) stated as *fair*, from [29 - 35.5) stated as *poor*, and from [35.5-42] stated as *bad* state.

FABQ data in the study consist a total of 397 observations. By using the MII algorithm, for each patient or observation, the most informative 2,4 and 6 questions were determined. The validation results were obtained separately for the number of inputs determined. Three different pre-defined algorithms including HC, Tabu, and MMHC were used to obtain BN models.

The third medical questionnaire in the study is PCS. PCS is a questionnaire for investigating the catastrophizing effects that cause pain in the musculo-skeletal system.

Their study focused on patients that have subacute pain. Patients are asked how they feel and what they think while they are experiencing pain. PCS composed of 13 items and it is based on continuous scale. A 5 point Likert scale was used ranging from score of 0 to 4, where a score of 0 indicates the expression of not at all, while a score of 4 indicates the expression of all the time. The total score is ranging from (0 - 52*)*, whereas a score of 0 indicates the best health status, while a score of 52 indicates worst health status.

In order to obtain MCID from data DBM was used. After applying DBM, MCID score for questionnaire determined as 6. Cut-off point was determined as 24 [115]. Therefore, scores below the cut-off point stated as good health status, while scores above the cut-off point stated as bad health status. In order to obtain different states and intervals for a continuous scale, MIPM was applied. Using additional cut-off points obtained from MIPM, for the interval from [0 – 12) stated as *excellent*, from [12 – 24) stated as *good*, from [24 - 33.3) stated as *fair*, from [33.3 - 42.6) stated as *poor*, and from [42.6 – 52] stated as *bad* state.

PCS data in the study consist a total of 398 observations. By using the MII algorithm, for each patient or observation, the most informative 2,4 and 6 questions were determined. The validation results were obtained separately for the number of inputs determined. HC, Tabu, and MMHC algorithms were used to obtain relevant BN models.

The last medical questionnaire in the study is the central sensitization inventory (CSI). It is a self-administered questionnaire and the main aim is to measure the status of patients who have diseases such as neck injury, migraine, and diseases resulting from the damaged nervous system. It consists of 25 questions. A 5 point Likert scale was used ranging from a score of 0 to 4, where a score of 0 indicates the expression of never, while a score of 4 indicates the expression of always. It is based on a continuous scale ranging from (0 – 100), where a total score of 0 indicates the best health status, while a score of 100 indicates the worst health status [116]. Neblett et al. [117] conducted a study and they found a cut-off point for this questionnaire. They specified that the cut-off point for this questionnaire was 40.9.

As a result of applying DBM, the MCID score for the questionnaire determined as 7.5 for this scale. In order to obtain different states and cut-off points for scale, MIPM was applied. Using additional cut-off points obtained from MIPM, for the interval from [0 - 20.45) stated as *good*, from [20.45 - 40.9) stated as *fair*, from [40.9 - 70.45) stated as *poor*, from [70.45 - 100] stated as *bad* state. It is not convenient to add more than four states as the total CSI score in the study does not have continuous values of more than 80.

CSI data in the study consist a total of 396 observations. For CSI data, it has been considered appropriate to use the five-fold cross-validation method in the validation phase. By using MII algorithm, for each patient or observation, the most informative 2,4 and 6 questions were determined. The validation results were obtained separately for the number of inputs determined. HC, Tabu, and MMHC algorithms were used to obtain relevant BN models.

With the questions determined by the MII algorithm, PROM factors were estimated. In addition, factor estimates were made with randomly determined questions. Cross validation results of the two approaches were obtained for 2,4 and 6 input sample sizes, then the results were compared. The PROM factor estimation with random inputs was repeated 50 times for 2,4 and 6 input sample sizes, then the average of estimation results was taken.

FHSQ, FABQ, CSI and PCS patient data to be used in case analysis are completely anonymous and taken from the Queen Mary University of London dataset.

Cut-off points, MCID scores obtained from DBM and states for each variable can be seen in Table 9.

Table 9 MCID, Cut-off Points and States of Each Variable

	***0.5* SD / MCID**	**Cut-off**	***Bad***	***Very Poor***	***Poor***	***Fair***	***Good***
FHSQ-Pain	11.50	72.2	[0-18.06)	[18.06-36.1)	[36.1-54.1)	[54.1-72.2)	[72.2-100]
FHSQ-Function	15	73.3	[0-18.33)	[18.33-36.67)	[36.67-55.01)	[55.01-73.3)	[73.3-100]
			Bad	***Poor***	***Fair***	***Good***	***Excellent***
FHSQ-GFH	12.5	55.2	[0-18.41)	[18.41-36.83)	[36.83-55.2)	[55.2-77.6)	[77.6-100]
			Excellent	***Good***	***Fair***	***Poor***	***Bad***
FABQ-Physical Activity	2.75	13	[0-4.33)	[4.33-8.66)	[8.66-13)	[13-18.5)	[18.5-24]
FABQ-Work	4.65	29	[0-9.66)	[9.66-19.33)	[19.33-29)	[29-35.5)	[35.5-42]
PCS	6	24	[0-12)	[12-24)	[24-33.3)	[33.3-42.6)	[42.6-52]
			Good	***Fair***	***Poor***	***Bad***	
CSI	7.5	40.9	[0-20.45)	[20.45-40.9)	[40.9-70.45)	[70.45-100]	

5.1. FHSQ Results

MII algorithm cross validation results of the pain and function domains for 2,4 and 6 input sample sizes are shown in Table 10, Table 11, and Table 12 respectively. Random inputs cross validation results of the pain and function domains for 2,4 and 6 input sample sizes are shown in Table 13, Table 14, and Table 15 respectively. MII algorithm results of the GFH factor are shown in Table 16. Since results of 2,4 and 6 input sizes were the same for the GFH factor they are shown in the same table. Random inputs cross validation results of the GFH factor for 2,4 and 6 input sample sizes are shown in Table 17, Table 18, and Table 19 respectively.

Table 10 MII Algorithm Cross Validation Results of Pain and Function Domains for 2 Input

	Poor	***Fair***	***Good***	***VP***	***Bad***
Pain HC	0.903	0.908	0.976	0.975	0,989
Function HC	0,950	0.907	0.978	0.950	0.969
Pain Tabu	0.903	0.908	0.976	0.975	0.989
Function Tabu	0.950	0.907	0.978	0.950	0.969
Pain MMHC	0.753	0.845	0.962	0.885	0.935
Function MMHC	0.937	0.908	0.975	0.925	0.961

Table 11 MII Algorithm Cross Validation Results of Pain and Function Domains for 4 Input

	Poor	***Fair***	***Good***	***VP***	***Bad***
Pain HC	0.894	0.911	0.984	0.979	0.985
Function HC	0.952	0.930	0.992	0.952	0.969
Pain Tabu	0.894	0.911	0.984	0.979	0.985
Function Tabu	0.952	0.930	0.992	0.952	0.969
Pain MMHC	0.758	0.847	0.962	0.882	0.928
Function MMHC	0.936	0.908	0.975	0.928	0.961

Table 12 MII Algorithm Cross Validation Results of Pain and Function Domains for 6 Input

	Poor	***Fair***	***Good***	***VP***	***Bad***
Pain HC	0.895	0.909	0.984	0.979	0.985
Function HC	0.952	0.929	0.992	0.951	0.968
Pain Tabu	0.895	0.909	0.984	0.979	0.985
Function Tabu	0.952	0.929	0.992	0.951	0.968
Pain MMHC	0.758	0.848	0.962	0.879	0.928
Function MMHC	0.937	0.909	0.975	0.925	0.961

Table 13 Random Inputs Cross Validation Results of Pain and Function Domains for 2 Input

	Poor	***Fair***	***Good***	***VP***	***Bad***
Pain HC	0.751	0.738	0.857	0.824	0.869
Function HC	0.834	0.757	0.871	0.846	0.856
Pain Tabu	0.750	0.745	0.860	0.841	0.862
Function Tabu	0.831	0.760	0.871	0.860	0.865
Pain MMHC	0.669	0.624	0.708	0.715	0.721
Function MMHC	0.721	0.662	0.714	0.741	0.705

Table 14 Random Inputs Cross Validation Results of Pain and Function Domains for 4 Input

	Poor	***Fair***	***Good***	***VP***	***Bad***
Pain HC	0.819	0.807	0.921	0.890	0.941
Function HC	0.892	0.812	0.926	0.919	0.907
Pain Tabu	0.821	0.802	0.922	0.894	0.931
Function Tabu	0.891	0.816	0.927	0.916	0.916
Pain MMHC	0.731	0.689	0.832	0.813	0.815
Funtion MMHC	0.819	0.757	0.821	0.838	0.817

Table 15 Random Inputs Cross Validation Results of Pain and Function Domains for 6 Input

	Poor	***Fair***	***Good***	***VP***	***Bad***
Pain HC	0.835	0.823	0.949	0.909	0.954
Function HC	0.919	0.862	0.954	0.938	0.942
Pain Tabu	0.832	0.815	0.929	0.906	0.944
Function Tabu	0.914	0.864	0.943	0.926	0.927
Pain MMHC	0.738	0.741	0.876	0.833	0.844
Funtion MMHC	0.855	0.805	0.882	0.885	0.877

Table 16 MII Algorithm Cross Validation Results of GFH Domain for 2,4 and 6 Inputs

	Poor	***Fair***	***Excellent***	***Good***	***Bad***
GFH HC	0.992	0.950	1.000	1.000	0.999
GFH Tabu	0.992	0.950	1.000	1.000	0.999
GFH MMHC	0.992	0.950	1.000	1.000	0.999

Table 17 Random Inputs Cross Validation Results of GFH Domain for 2 Input

	Poor	***Fair***	***Excellent***	***Good***	***Bad***
GFH HC	0.829	0.662	0.847	0.752	0.852
GFH Tabu	0.827	0.677	0.851	0.758	0.852
GFH MMHC	0.700	0.613	0.733	0.712	0.713

Table 18 Random Inputs Cross Validation Results of GFH Domain for 4 Inputs

	Poor	***Fair***	***Excellent***	***Good***	***Bad***
GFH HC	0.886	0.733	0.929	0.867	0.925
GFH Tabu	0.885	0.735	0.923	0.865	0.921
GFH MMHC	0.822	0.678	0.867	0.841	0.856

Table 19 Random Inputs Cross Validation Results of GFH Domain for 6 Inputs

	Poor	***Fair***	***Excellent***	***Good***	***Bad***
GFH HC	0.908	0.757	0.902	0.907	0.908
GFH Tabu	0.918	0.753	0.936	0.902	0.937
GFH MMHC	0.902	0.741	0.902	0.904	0.916

Figure 10 shows the BN models created with Tabu and HC algorithms, which includes both the pain and function factors. Figure 11 shows the BN model learned with the MMHC algorithm. For the function factor, HC and Tabu algorithms shown in Table 12 had the highest performance for predicting the state "*Good*" with an average AUC of 0.992 (see Figure 12).

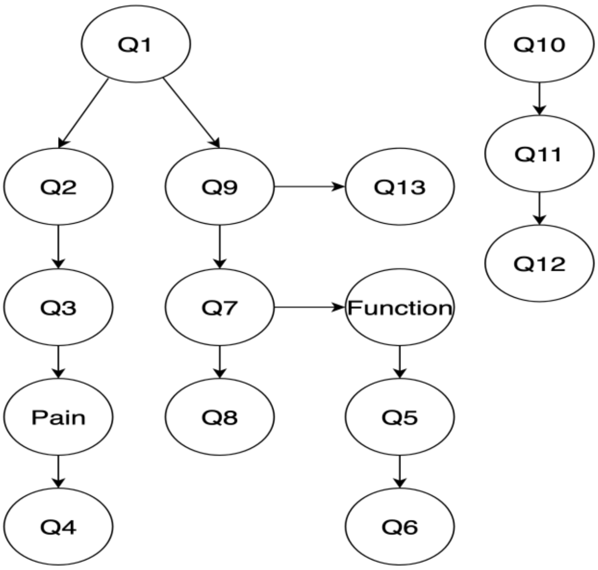


Figure 10 BN Model of Pain and Function Factors Created Through HC and Tabu Algorithms

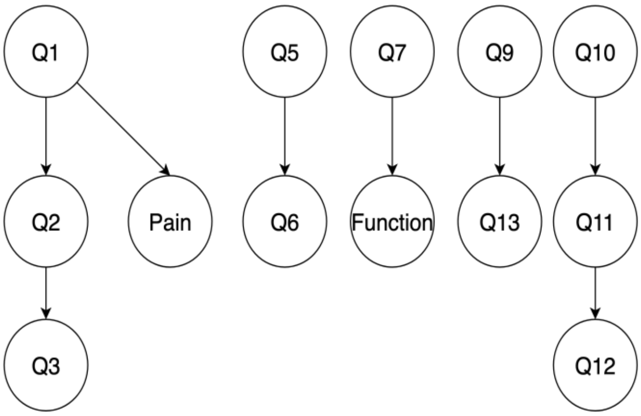


Figure 11 BN Model of Pain and Function Factors Created Through MMHC Algorithm

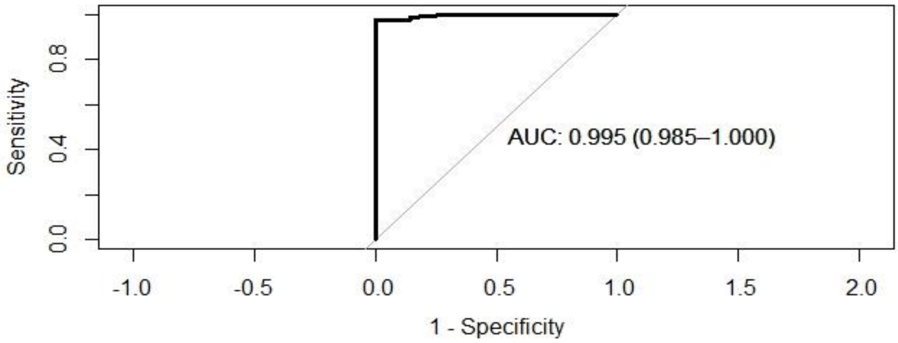


Figure 12 ROC Curve of HC and Tabu Models for Function Factor

For GFH factor, BN models created with Tabu and HC algorithms were the same. The learned model is shown in Figure 13. The learned model for GFH factor with MMHC algorithm is shown in Figure 14. The ROC curves with the highest performance for inputs were also shown as examples. For the function, pain and GFH factors, state "*Fair*" with the 0.95 mean AUC result shown in Table 16 showed one of the highest performances, and the ROC curve of models which were created by HC, Tabu and MMHC algorithms is shown in Appendix-A.

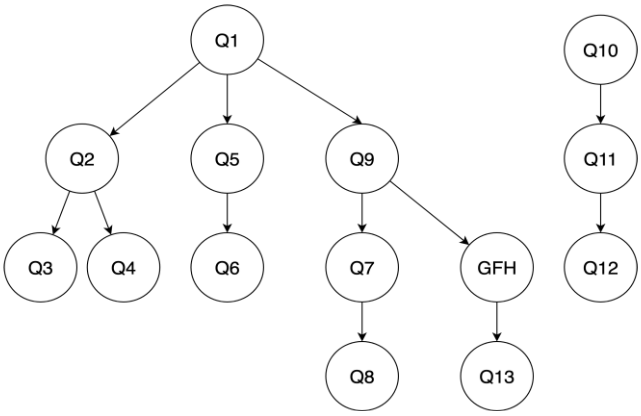


Figure 13 BN Model of GFH Factor Created Through HC and Tabu Algorithms

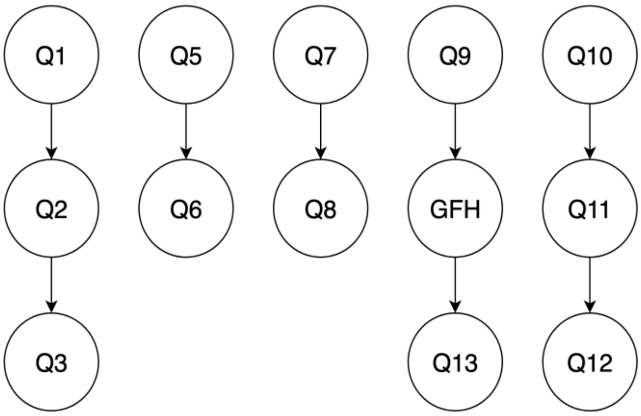


Figure 14 BN Model of GFH Factor Created Through MMHC Algorithm

5.2. FABQ Results

MII algorithm cross validation results of the physical activity and work domains for 2, 4 and 6 input sample sizes, are shown in Table 20, Table 21 and Table 22 respectively. Since physical activity and work factors have same state names, results of them were given in same table. Random inputs cross validation results of the physical activity and work domains for 2,4 and 6 input sample sizes are shown in Table 23, Table 24, and Table 25 respectively.

The BN models created by using HC, Tabu, and MMHC algorithms are shown in Figure 15, Figure 16 and Figure 17 respectively. The model learned by the MMHC algorithm was sparse; each factor was only connected to 2 questions. Therefore, adding more than 2 inputs did not affect the prediction accuracy of the BN model created through this algorithm. ROC curves with the highest performance for 6 inputs were also shown as examples. For the physical activity factor estimated by the Tabu model, the state

"*Excellent*" with the 0.983 mean AUC result shown in Table 22 showed one of the highest performances. The ROC curve of the model created by the Tabu algorithm is shown in Appendix-B. For the work factor, Tabu and HC algorithms showed similar performance. State “*Bad*” with the 0.992 mean AUC result of the HC algorithm shown in Table 22 was one of the highest performances in the validation results. The ROC curve of the HC algorithm for this state is shown in Appendix-C.

Table 20 MII Algorithm Cross Validation Results of Physical Activity and Work Factors for 2 Input

	Poor	***Excellent***	***Fair***	***Good***	***Bad***
Physical Activity HC	0.799	0.964	0.778	0.784	0.946
Work HC	0.982	0.925	0.900	0.836	0.992
Physical Activity Tabu	0.799	0.964	0.778	0.784	0.946
Work Tabu	0.982	0.925	0.900	0.836	0.992
Physical Activity MMHC	0.763	0.953	0.768	0.751	0.900
Work MMHC	0.500	0.582	0.574	0.567	0.665

Table 21 MII Algorithm Cross Validation Results of Physical Activity and Work Factors for 4 Input

	Poor	***Excellent***	***Fair***	***Good***	***Bad***
Physical Activity HC	0.870	0.977	0.836	0.816	0.974
Work HC	0.982	0.932	0.892	0.831	0.992
Physical Activity Tabu	0.899	0.983	0.887	0.840	0.973
Work Tabu	0.982	0.931	0.892	0.831	0.992
Physical Activity MMHC	0.763	0.953	0.768	0.751	0.900
Work MMHC	0.500	0.582	0.574	0.567	0.665

Table 22 MII Algorithm Cross Validation Results of Physical Activity and Work Factors for 6 Input

	Poor	***Excellent***	***Fair***	***Good***	***Bad***
Physical Activity HC	0.870	0.977	0.836	0.816	0.974
Work HC	0.986	0.930	0.892	0.831	0.992
Physical Activity Tabu	0.899	0.983	0.887	0.840	0.973
Work Tabu	0.986	0.932	0.891	0.831	0.992
Physical Activity MMHC	0.763	0.953	0.768	0.753	0.900
Work MMHC	0.500	0.582	0.574	0.567	0.665

Table 23 Random Inputs Cross Validation Results of Physical Activity and Work Factors for 2 Input

	Poor	*Excellent*	*Fair*	*Good*	*Bad*
Physical Activity HC	0.631	0.715	0.582	0.605	0.701
Work HC	0.888	0.845	0.843	0.740	0.966
Physical Activity Tabu	0.657	0.759	0.592	0.623	0.740
Work Tabu	0.884	0.849	0.848	0.745	0.957
Physical Activity MMHC	0.568	0.633	0.542	0.557	0.613
Work MMHC	0.500	0.515	0.519	0.511	0.532

Table 24 Random Inputs Cross Validation Results of Physical Activity and Work Factors for 4 Input

	Poor	*Excellent*	*Fair*	*Good*	*Bad*
Physical Activity HC	0.702	0.838	0.648	0.671	0.803
Work HC	0.910	0.897	0.880	0.787	0.990
Physical Activity Tabu	0.728	0.876	0.672	0.700	0.848
Work Tabu	0.909	0.896	0.883	0.785	0.960
Physical Activity MMHC	0.623	0.725	0.576	0.597	0.697
Work MMHC	0.500	0.529	0.534	0.525	0.500

Table 25 Random Inputs Cross Validation Results of Physical Activity and Work Factors for 6 Input

	Poor	*Excellent*	*Fair*	*Good*	*Bad*
Physical Activity HC	0.742	0.853	0.693	0.726	0.872
Work HC	0.903	0.917	0.896	0.811	0.995
Physical Activity Tabu	0.777	0.919	0.728	0.742	0.896
Work Tabu	0.921	0.900	0.900	0.814	0.963
Physical Activity MMHC	0.657	0.801	0.618	0.631	0.757
Work MMHC	0.500	0.532	0.549	0.535	0.599

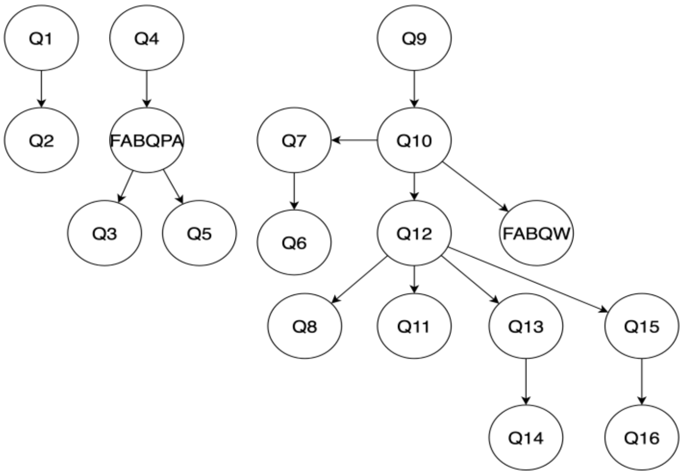


Figure 15 BN Model of FABQ Created Through HC Algorithm

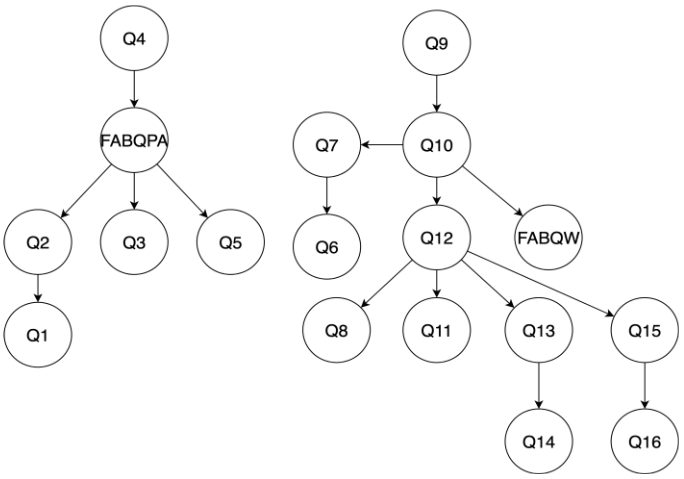


Figure 16 BN Model of FABQ Created Through Tabu Algorithm

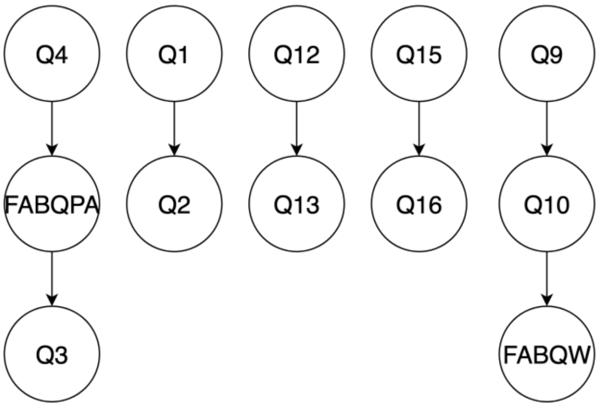


Figure 17 BN Model of FABQ Created Through MMHC Algorithm

5.3. PCS Results

MII algorithm cross validation results of PCS for 2,4 and 6 input sample sizes, are shown in Table 26, Table 27 and Table 28 respectively. Random inputs cross validation results of the PCS factor for 2,4 and 6 input sample sizes are shown in Table 29, Table 30, and Table 31 respectively.

Table 26 MII Algorithm Cross Validation Results of PCS for 2 Input

	Bad	***Excellent***	***Good***	***Poor***	***Fair***
PCS HC	0.934	0.949	0.888	0.987	0.942
PCS Tabu	0.934	0.949	0.888	0.987	0.942
PCS MMHC	0.500	0.500	0.500	0.500	0.500

Table 27 MII Algorithm Cross Validation Results of PCS for 4 Input

	Bad	***Excellent***	***Good***	***Poor***	***Fair***
PCS HC	0.970	0.974	0.929	0.989	0.990
PCS Tabu	0.970	0.974	0.929	0.989	0.990
PCS MMHC	0.500	0.500	0.500	0.500	0.500

Table 28 MII Algorithm Cross Validation Results of PCS for 6 Input

	Bad	***Excellent***	***Good***	***Poor***	***Fair***
PCS HC	0.994	0.984	0.952	0.992	0.979
PCS Tabu	0.994	0.984	0.952	0.992	0.979
PCS MMHC	0.500	0.500	0.500	0.500	0.500

Table 29 Random Inputs Cross Validation Results of PCS for 2 Input

	Bad	***Excellent***	***Good***	***Poor***	***Fair***
PCS HC	0.877	0.904	0.841	0.922	0.913
PCS Tabu	0.879	0.906	0.844	0.914	0.909
PCS MMHC	0.500	0.500	0.500	0.500	0.500

Table 30 Random Inputs Cross Validation Results of PCS for 4 Input

	Bad	***Excellent***	***Good***	***Poor***	***Fair***
PCS HC	0.922	0.923	0.899	0.919	0.927
PCS Tabu	0.942	0.935	0.898	0.938	0.929
PCS MMHC	0.500	0.500	0.500	0.500	0.500

Table 31 Random Inputs Cross Validation Results of PCS for 6 Input

	Bad	***Excellent***	***Good***	***Poor***	***Fair***
PCS HC	0.933	0.956	0.923	0.955	0.941
PCS Tabu	0.953	0.966	0.923	0.944	0.951
PCS MMHC	0.500	0.500	0.500	0.500	0.500

BN models created by using HC and Tabu algorithms are same and they are shown in Figure 18. A valid BN model could not be created using the MMHC algorithm. ROC curves with the highest performance for 6 inputs were also shown as examples. For PCS, since HC and Tabu algorithms have same prediction accuracy, ROC curves of any algorithms can be chosen for illustration. State "*Bad*" with the 0.994 mean AUC result of HC and Tabu algorithms shown in Table 28 was one of the highest performances in the validation results. ROC curve of HC algorithm for this state is shown in Appendix-D.

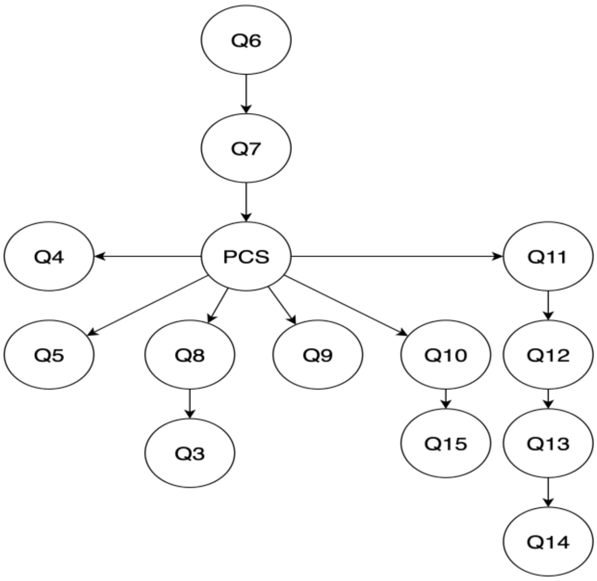


Figure 18 BN Model of PCS Created Through HC and Tabu Algorithms

5.4. CSI Results

For 2,4 and 6 input sample sizes, MII algorithm cross validation results of PCS are shown in Table 32, Table 33, and Table 34 respectively. Random inputs cross validation results of the CSI factor for 2,4 and 6 input sample sizes are shown in Table 35, Table 36, and Table 37 respectively.

Table 32 MII Algorithm Cross Validation Results of CSI for 2 Input

	Poor	***Good***	***Fair***	***Bad***
CSI HC	0.904	0.894	0.780	0.955
CSI Tabu	0.904	0.894	0.780	0.955
CSI MMHC	0.902	0.898	0.781	0.859

Table 33 MII Algorithm Cross Validation Results of CSI for 4 Input

	Poor	***Good***	***Fair***	***Bad***
CSI HC	0.945	0.943	0.859	0.983
CSI Tabu	0.945	0.943	0.859	0.983
CSI MMHC	0.902	0.898	0.781	0.859

Table 34 MII Algorithm Cross Validation Results of CSI for 6 Input

	Poor	***Good***	***Fair***	***Bad***
CSI HC	0.948	0.969	0.899	0.983
CSI Tabu	0.948	0.969	0.899	0.983
CSI MMHC	0.902	0.898	0.781	0.859

Table 35 Random Inputs Cross Validation Results of CSI for 2 Input

	Poor	***Good***	***Fair***	***Bad***
CSI HC	0.836	0.834	0.695	0.886
CSI Tabu	0.829	0.826	0.690	0.870
CSI MMHC	0.614	0.617	0.564	0.648

Table 36 Random Inputs Cross Validation Results of CSI for 4 Input

	Poor	***Good***	***Fair***	***Bad***
CSI HC	0.892	0.892	0.776	0.904
CSI Tabu	0.888	0.896	0.779	0.922
CSI MMHC	0.694	0.689	0.614	0.720

Table 37 Random Inputs Cross Validation Results of CSI for 6 Input

	Poor	***Good***	***Fair***	***Bad***
CSI HC	0.916	0.925	0.805	0.948
CSI Tabu	0.918	0.924	0.819	0.937
CSI MMHC	0.753	0.748	0.645	0.749

BN models created by using HC and Tabu algorithms are the same as shown in Figure 19. BN models created with the MMHC algorithm is shown in Figure 20. When we examine the BN model created with the MMHC algorithm, we see that the CSI factor was only connected to questions *Q9* and *Q15*. Therefore, while calculating estimations of the CSI factor these two questions were used. For the MMHC model, increasing the input size hasn’t had any effect on estimations. ROC curves with the highest performance for 6 inputs were also shown as examples. For CSI, since the HC and Tabu algorithms have the same prediction accuracy, the ROC curves of any algorithms could be chosen for illustration. State “*Bad*” with the 0.983 mean AUC result of HC and Tabu algorithms shown in Table 34, was one of the highest performances in the validation results. Other ROC curves related to the Tabu algorithm is shown in Appendix-E.

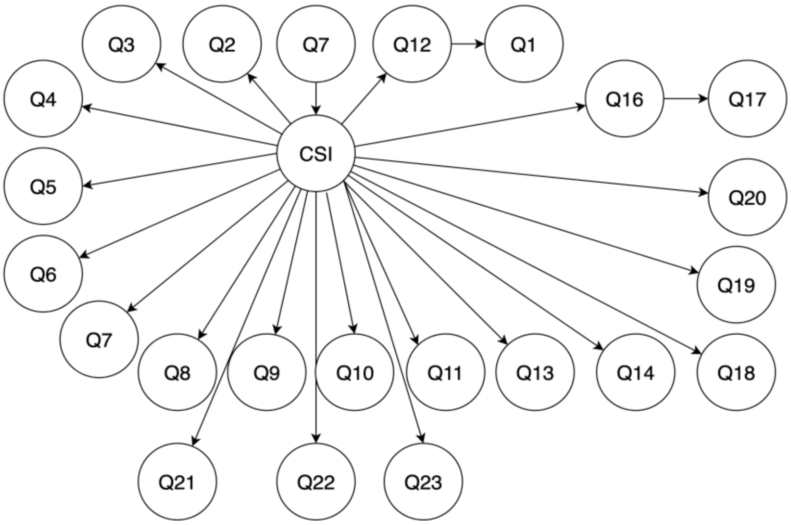


Figure 19 BN Model of CSI Created Through HC and Tabu Algorithms

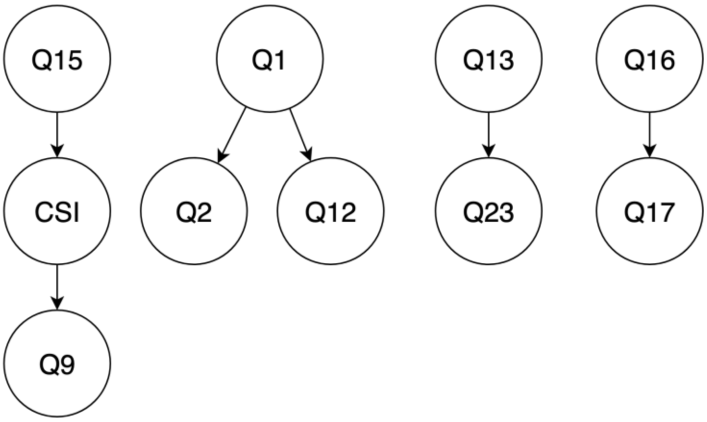


Figure 20 BN Model of CSI Created Through MMHC Algorithm

5.5. Discussion

FHSQ, FABQ, PCS, and CSI are widely used PROMs for assessing the medical status of patients. According to the validation results of FHSQ, even with 2 the most informative questions determined for each patient, overall AUC scores were more than 0.9 for pain and function factors. Since BN models learned by the HC and Tabu search algorithms were similar, similar validation results were obtained for these algorithms. If we compare the performance of data-driven models, we see that BN models, which were generally created by HC and Tabu search algorithms, outperformed the model created through the MMHC algorithm.

According to validation results of the GFH factor that is in the FHSQ, all data-driven algorithms showed about the same prediction performance. Although BN models created through HC and Tabu algorithms were different from the model created by using the MMHC algorithm, validation results were the same. It is due to the parameters of all models were the same. For all input sizes, validation results were the same for the GFH factor. Even with an input size of 2, for all models, prediction accuracies were more than %90. In fact, for some states of this factor, %100 prediction accuracy was obtained. Adding more inputs provided no additional mutual information. Due to this situation, all validation outcomes were the same.

As the number of input size increase, the entropy in other words uncertainty decreases, which eventually result in increased mutual information. However, it has been seen that the performance of some models decreased when the number of inputs increased. Decreases and increases in AUC results were, however, small often occurring at the 3rd decimal point of the numbers.

FABQ is another PROM which consists of two distinct factors. The validation results of this medical questionnaire were obtained for both factors. According to the validation results of FABQ, all algorithms except the MMHC algorithm had an overall AUC score of more than 0.85 for work and physical activity factors. Tabu search algorithm had a superior prediction performance compared to the others. This may be because this algorithm tried to avoid getting stuck at the local optimum point, unlike the HC algorithm [42]. The MMHC algorithm had a poorer prediction performance than other algorithms. This is not surprising as MMHC learned an overly simple BN structure for this case.

One of the limitations of the MMHC algorithm is that it can avoid connecting some sets of variables with the target variable in the d-separation scope. This algorithm may not consider these variables after it discovers d-separation. In contrast, the HC and Tabu algorithms can continue adding variables to the parents set even if this process doesn't increase the score significantly [46]. Therefore, BN models obtained with MMHC are simpler than models obtained by HC and Tabu algorithms. In addition, the parameters obtained by the MMHC algorithm may not be as accurate as the parameters obtained by the HC and Tabu algorithms. This may result in worse prediction performance.

Unlike previous PROMs, PCS had a single factor that is equal to the summation of scores of questions in the medical questionnaire. According to the validation results of PCS, all algorithms except the MMHC algorithm had an overall AUC score of more than 0.9. BN models learned by using HC and Tabu algorithms were the same. Similar to the previous case, MMHC had the worst performance for PCS.

CSI also has a single factor that equals the summation of the scores of its individual questions. According to the validation results of CSI PROM, the mean AUC results of the HC and Tabu algorithms were more than 0.85 and the mean AUC results of the MMHC algorithm were more than 0.78. BN models established by using HC and Tabu algorithms were the same. Therefore, obtained validation results for both algorithms were the same. Overall, HC and Tabu had slightly better results than the MMHC algorithm. As the number of inputs increased, the performance of the HC and Tabu algorithms increased as well.

For all PROMs and factors, when we examined random inputs cross validation results we have seen that we obtained lower estimation results compare to results obtained through the MII algorithm for each input sample size. In this case, we can say that the inputs determined with MII algorithm had a better performance in predicting PROM factors and patient status.

In summary, the proposed approach was able to predict different PROMs accurately using only a few inputs determined by the mutual information-based algorithms. Between data-driven algorithms, HC and Tabu search algorithms appeared to have similar results. The MMHC algorithm has generally lower estimation performance than the remaining two algorithms.

6. CONCLUSION

This thesis proposed a novel approach for finding the most informative PROM questions for predicting patient-specific outcomes, and for predicting patient outcomes with missing PROM questions. PROMs are composed of questions for measuring patient outcomes. Filling in PROMs may require a large amount of time and cognitive-load from patients since PROMs may need to be filled in repetitively and some PROMs can contain a large number of questions. The proposed approach enables the PROMs to be filled in more efficiently with fewer inputs. The proposed approach is based on BNs, and it has been applied to four PROMs in the musculo-skeletal conditions domain, i.e. FABQ, FHSQ, PCS, and CSI.

The proposed approach is composed of two main elements. The first element is the BN model representing the relations between PROM questions. We used structure learning algorithms including HC, Tabu search and MMHC to learn the BN from data. Structure learning algorithms were based on discrete variables, and we had to discretize the continuous PROM measures. We used an integer programming-based approach for this. The second element is the algorithm for determining the most informative questions. We used a conditional entropy based algorithm for this task.

According to the results of FHSQ, when data-based models are evaluated among themselves, it was seen that models created with HC and Tabu algorithm generally had the same prediction performance. These models had a superior prediction performance than the model created with MMHC algorithm. It was seen that as input sizes increased, prediction accuracies of all models either increased and remained same. As mentioned earlier in previous section, the reason for the prediction performance to remain the same is because adding more input did not have an impact on mutual information. Results of FABQ showed that model created by using Tabu algorithm has slightly better prediction performance than HC model. However, considering the results of all PROMs, it has been observed that the models created with the HC and Tabu algorithms have better prediction performances than the models created with the MMHC algorithm. In addition, PROM factors and patient status were estimated with randomly determined inputs. When all results were evaluated, it was seen that the estimation of PROM factors and patient status was more consistent with the questions determined by the MII algorithm.

This study only focused on PROMs from the foot musculo-skeletal conditions domain. This can be considered as a limitation of this study. PROMs with different domains could potentially be used to make a more comprehensive evaluation of the study. Another limitation is that methods and models in the study were used to predict only a single snap-shot status of the patients. In other words, the BN models incorporated observations from a single PROM response. In future studies, dynamic BN models could be developed so that changes in the patient's status can be followed easily. Another limitation is that only one target variable (factor) was determined in the study and the posterior distribution of that target variable was calculated. Computing the posterior probability of multiple target variables at the same time would increase the computational complexity. In future studies, after calculating the joint distribution of target nodes, the entropy of these nodes could be calculated. Then, the mutual information between target nodes and input nodes could be calculated to determine the most informative questions. In addition, a study can be conducted to show which questions are asked to which specific patients in future studies. By doing this, it is shown which question is the most informative question for that patient. This could increase the understandability of the proposed method.

Moreover, the BN models in the proposed approach only incorporated a single PROM. In future studies, BN models containing a question pool from multiple PROMs can be created. Factors of specific PROMs can be estimated using questions from different PROMs in this question pool. Alternatively, using the factors of one PROM, the factors of the other PROM can be estimated. Finally, interfaces and applications can be designed for an easier use of the developed methods and models in this study. In this way, a wider use of the proposed approach can be enabled.

REFERENCES

[1] R. Duda and E. Shortliffe, “Expert Systems Research,” *Science (80-.).*, vol. 220, no. 4594, pp. 261–268, Apr. **1983**.

[2] W. J. Clancey and R. Letsinger, “Neomycin: Reconfiguring a Rule-Based Expert System for Application To Teaching.,” vol. 2, no. May, pp. 829–836, **1981**.

[3] R. Davis, “Expert Systems: Where are we? And where do we go from here?,” *AI Mag.*, vol. 3, no. 2, p. 3, **1982**.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[5] S. Chakrabarti, M. Ester, U. Fayyad, and J. Gehrke, “Data mining curriculum: a proposal,” *Acm Sigkdd*, pp. 1–10, **2006**.

[6] B. P. De, R. Kar, D. Mandal, and S. P. Ghoshal, “Optimal selection of components value for analog active filter design using simplex particle swarm optimization,” *Int. J. Mach. Learn. Cybern.*, vol. 6, no. 4, pp. 621–636, **2015**.

[7] D. Bzdok, N. Altman, and M. Krzywinski, “Points of Significance: Statistics versus machine learning,” *Nat. Methods*, vol. 15, no. 4, pp. 233–234, **2018**.

[8] A. Graves, G. Wayne, and I. Danihelka, “Neural Turing Machines,” pp. 1–26, **2014**.

[9] J. Powles and H. Hodson, “Google DeepMind and healthcare in an age of algorithms,” *Health Technol. (Berl).*, vol. 7, no. 4, pp. 351–367, **2017**.

[10] E. Gibney, “DeepMind algorithm beats people at classic video games,” *Nature*, vol. 518, no. 7540, pp. 465–466, **2015**.

[11] T. Elomaa and N. Holsti, “An experimental comparison of inducing decision trees and decision lists in noisy domains,” in *Proceedings of the Fourth European Working Session on Learning, Montpelier, France*, **1989**, pp. 59–69.

[12] K. C. C. Chan and A. K. C. Wong, “Automatic construction of expert systems from data: A statistical approach,” in *Proc. IJCAI’89 Workshop on Knowledge Discovery in Databases*, **1989**.

[13] K. Horn, P. J. Compton, L. Lazarus, and J. R. Quinlan, “An expert system for the interpretation of thyroid assays in a clinical laboratory,” *Aust Comput J*, vol. 17, pp. 7–11, **1985**.

[14] L. Lesmo, L. Saitta, and P. Torasso, “LEARNING OF FUZZY PRODUCTION RULES FOR MEDICAL DIAGNOSIS,” D. Dubois, H. Prade, and R. R. B. T.-R. in F. S. for I. S. Yager, Eds. Morgan Kaufmann, **1993**, pp. 901–912.

[15] R. Cuocolo, T. Perillo, E. De Rosa, L. Ugga, and M. Petretta, “Current applications of big data and machine learning in cardiology,” *J. Geriatr. Cardiol.*, vol. 16, no. 8, pp. 601–607, Aug. **2019**.

[16] R. C. Deo, “Machine Learning in Medicine HHS Public Access,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, **2015**.

[17] J. Dawson, H. Doll, R. Fitzpatrick, C. Jenkinson, and A. J. Carr, “Routine use of

patient reported outcome measures in healthcare settings,” *BMJ*, vol. 340, no. 7744, pp. 464–467, **2010**.

[18] J. E. J. Ware and C. D. Sherbourne, “The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.,” *Med. Care*, vol. 30, no. 6, pp. 473–483, Jun. **1992**.

[19] E. M. Roos and L. S. Lohmander, “The Knee injury and Osteoarthritis Outcome Score (KOOS): From joint injury to osteoarthritis,” *Health Qual. Life Outcomes*, vol. 1, pp. 1–8, **2003**.

[20] A. K. Nilsdotter, L. Stefan Lohmander, M. Klässbo, E. M. Roos, and L. Stefan Lohmander -Stefan, “BMC Musculoskeletal Disorders Hip disability and osteoarthritis outcome score (HOOS) – validity and responsiveness in total hip replacement,” vol. 8, pp. 1–8, **2003**.

[21] A. Hutchings, K. Grosse Frie, J. Neuburger, J. Van Der Meulen, and N. Black, “Late response to patient-reported outcome questionnaires after surgery was associated with worse outcome,” *J. Clin. Epidemiol.*, vol. 66, no. 2, pp. 218–225, **2013**.

[22] E. H. Shortliffe, B. G. Buchanan, and E. A. Feigenbaum, “Knowledge engineering for medical decision making: A review of computer-based clinical decision aids,” *Proc. IEEE*, vol. 67, no. 9, pp. 1207–1224, **1979**.

[23] T. Koski and J. M. Noble, *Bayesian Networks*. Wiley, **2009**.

[24] E. H. Shortliffe and B. G. Buchanan, “A model of inexact reasoning in medicine,” *Math. Biosci.*, vol. 23, no. 3–4, pp. 351–379, Apr. **1975**.

[25] L. M. Berliner, “Hierarchical Bayesian Time Series Models,” *Maximum Entropy and Bayesian Methods*, no. 2, pp. 15–22, **1996**.

[26] K. G. M. Moons, G. A. Van Es, J. W. Deckers, J. D. F. Habbema, and D. E. Grobbee, “Limitations of sensitivity, specificity, likelihood ratio, and Bayes’ theorem in assessing diagnostic probabilities: A clinical example,” *Epidemiology*, vol. 8, no. 1, pp. 12–17, **1997**.

[27] Z. Pawlak, “Rough sets and intelligent data analysis,” *Inf. Sci. (Ny).*, vol. 147, no. 1–4, pp. 1–12, **2002**.

[28] N. Fenton and M. Neil, *Assessment and Decision Bayesian*. CRC Press, **2013**.

[29] S. Acid, L. M. de Campos, and J. G. Castellano, “Learning Bayesian Network Classifiers: Searching in a Space of Partially Directed Acyclic Graphs,” *Mach. Learn.*, vol. 59, no. 3, pp. 213–235, Jun. **2005**.

[30] R. E. Neapolitan, “Learning Bayesian Networks Multidimensional Computer Adaptive Testing for Patient-Reported Outcomes View project,” **2003**.

[31] S. . L. Lauritzen and D. . Spiegelhalter, “Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems,” *Society*, vol. 50, no. 2, pp. 157–224, **1988**.

[32] M. Neil, N. Fenton, and L. Nielsen, “Building large-scale Bayesian networks,” *Knowl. Eng. Rev.*, vol. 15, no. 3, pp. 257–284, **2000**.

[33] M. Kwan, K.-P. Chow, F. Law, and P. Lai, “Reasoning About Evidence Using Bayesian Networks,” in *Advances in Digital Forensics IV*, vol. 383 AICT, Boston,

MA: Springer US, **2008**, pp. 275–289.

[34] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*, vol. 50, no. 1. New York, NY: Springer New York, **2007**.

[35] H. Tse, K. P. Chow, and M. Kwan, “Reasoning about evidence using Bayesian networks,” *IFIP Adv. Inf. Commun. Technol.*, vol. 383 AICT, pp. 99–113, **2012**.

[36] H. Xiao-xuan, W. Hui, and W. Shuo, “Using Expert’s Knowledge to Build Bayesian Networks,” in *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, **2007**, pp. 220–223.

[37] M. Scutari, “Learning Bayesian networks with the bnlearn R Package,” *J. Stat. Softw.*, vol. 35, no. 3, pp. 1–22, **2010**.

[38] G. Schwarz, “Estimating the Dimension of a Model,” *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. **1978**.

[39] H. Akaike, “Information Theory and an Extension of the Maximum Likelihood Principle,” in *Nursing times*, vol. 90, no. 29, **1998**, pp. 199–213.

[40] J. A. Gámez, J. L. Mateo, and J. M. Puerta, “Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood,” *Data Min. Knowl. Discov.*, vol. 22, no. 1–2, pp. 106–148, Jan. **2011**.

[41] S. J. Russell and P. Norvig, *Artificial Intelligence A Modern Approach; PearsonEducation*. **2003**.

[42] F. Glover, “Tabu Search - Part II,” *ORSA J. Comput.*, vol. 2, no. 1, pp. 4–32, **1990**.

[43] J. Pearl, “BAYESIAN INFERENCE,” in *Probabilistic Reasoning in Intelligent Systems*, vol. 156, Elsevier, **1988**, pp. 29–75.

[44] Dimitris Margaritis, “Learning Bayesian Network Model Structure from Data,” **2003**.

[45] T. D. Le, T. Hoang, J. Li, L. Liu, H. Liu, and S. Hu, “A Fast PC Algorithm for High Dimensional Causal Discovery with Multi-Core PCs,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 9, pp. 1483–1495, Sep. **2014**.

[46] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, “The max-min hill-climbing Bayesian network structure learning algorithm,” *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, **2006**.

[47] G. A. Young, “Mathematical Statistics: An Introduction to Likelihood Based Inference,” *Int. Stat. Rev.*, vol. 87, no. 1, pp. 178–179, **2019**.

[48] Y. Zhou, N. Fenton, and M. Neil, “Bayesian network approach to multinomial parameter learning using data and expert judgments,” *International Journal of Approximate Reasoning*, vol. 55, no. 5. pp. 1252–1268, **2014**.

[49] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum Likelihood from Incomplete Data Via the EM Algorithm,” *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–22, **1977**.

[50] K. G. Olesen *et al.*, “A munin network for the median nerve—A case study on loops,” *Appl. Artif. Intell.*, vol. 3, no. 2–3, pp. 385–403, **1989**.

[51] M. Henrion, “Practical Issues in Constructing a Bayes’ Belief Network,” *Int. J. Approx. Reason.*, vol. 2, no. 3, p. 337, Mar. **2013**.

[52] O. Pourret, P. Naim, and B. Marcot, *Bayesian Networks: A Practical Guide to Applications*. Wiley, **2008**.

[53] P. Lucas, "Bayesian networks in medicine: a model-based approach to medical decision making," *Proc. EUNITE Work. Intell. Syst. patient Care*, pp. 73–97, **2001**.

[54] E. Kyrimi, S. McLachlan, K. Dube, M. R. Neves, A. Fahmi, and N. Fenton, "A Comprehensive Scoping Review of Bayesian Networks in Healthcare: Past, Present and Future," pp. 1–32, **2020**.

[55] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Update on the Pathfinder Project * Section on Medical Informatics Section on Medical Informatics Stanford University School of Medicine Stanford University School of Medicine Development of the Knowledge Base," pp. 203–207, **1989**.

[56] I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper, "The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks," no. 0, **1989**, pp. 247–256.

[57] A. Oniśko and M. J. Druzdzel, "Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems," *Artif. Intell. Med.*, vol. 57, no. 3, pp. 197–206, **2013**.

[58] D. L. Sanders and D. Aronsky, "Detecting Asthma Exacerbations in a Pediatric Emergency Department Using a Bayesian Network," pp. 684–688, **2006**.

[59] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 3, pp. 371–379, **2009**.

[60] X. H. Wang, B. Zheng, W. F. Good, J. L. King, and Y. H. Chang, "Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network," *Int. J. Med. Inform.*, vol. 54, no. 2, pp. 115–126, **1999**.

[61] D. Zhao and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction," *J. Biomed. Inform.*, vol. 44, no. 5, pp. 859–868, **2011**.

[62] P. Petousis, S. X. Han, D. Aberle, and A. A. T. Bui, "Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network," *Artif. Intell. Med.*, vol. 72, pp. 42–55, **2016**.

[63] X. Jiang, D. Xue, A. Brufsky, S. Khan, and R. Neapolitan, "A new method for predicting patient survivorship using efficient Bayesian network learning," *Cancer Inform.*, vol. 13, pp. 47–57, **2014**.

[64] N. Hoot and D. Aronsky, "Using Bayesian networks to predict survival of liver transplant patients.," *AMIA Annu. Symp. Proc.*, pp. 345–349, **2005**.

[65] B. A. Ahmed, M. E. Matheny, P. L. Rice, J. R. Clarke, and O. I. Ogunyemi, "A comparison of methods for assessing penetrating trauma on retrospective multi-center data," *J. Biomed. Inform.*, vol. 42, no. 2, pp. 308–316, **2009**.

[66] A. Jochems *et al.*, "Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 99, no. 2, pp. 344–352, **2017**.

[67] K. Jayasurya *et al.*, “Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy,” *Med. Phys.*, vol. 37, no. 4, pp. 1400–1407, **2010**.

[68] B. Yet, Z. B. Perkins, T. E. Rasmussen, N. R. M. Tai, and D. W. R. Marsh, “Combining data and meta-analysis to build Bayesian networks for clinical decision support,” *J. Biomed. Inform.*, vol. 52, pp. 373–385, **2014**.

[69] B. Yet, Z. Perkins, N. Fenton, N. Tai, and W. Marsh, “Not just data: A method for improving prediction with knowledge,” *J. Biomed. Inform.*, vol. 48, pp. 28–37, **2014**.

[70] B. Yet, K. Bastani, H. Raharjo, S. Lifvergren, W. Marsh, and B. Bergman, “Decision support system for Warfarin therapy management using Bayesian networks,” *Decis. Support Syst.*, vol. 55, no. 2, pp. 488–498, **2013**.

[71] M. Fortini, B. Liseo, A. Nuccitelli, and M. Scanu, “On Bayesian record linkage,” *Res. Off. Stat.*, vol. 4, no. 1, pp. 195–198, **2001**.

[72] A. C. Constantinou, N. Fenton, W. Marsh, and L. Radlinski, “From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support,” *Artif. Intell. Med.*, vol. 67, pp. 75–93, **2016**.

[73] R. Kaya and B. Yet, “Building Bayesian networks based on DEMATEL for multiple criteria decision problems: A supplier selection case study,” *Expert Syst. Appl.*, vol. 134, pp. 234–248, **2019**.

[74] A. Bakshan, I. Srour, G. Chehab, M. El-Fadel, and J. Karaziwan, “Behavioral determinants towards enhancing construction waste management: A Bayesian Network analysis,” *Resour. Conserv. Recycl.*, vol. 117, pp. 274–284, **2017**.

[75] H. Borchani, C. Bielza, P. Martí nez-Martín, and P. Larrañaga, “Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: An application to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson’s Disease Questionnaire (PDQ-39),” *J. Biomed. Inform.*, vol. 45, no. 6, pp. 1175–1184, **2012**.

[76] Q. A. Le and J. N. Doctor, “Probabilistic mapping of descriptive health status responses onto health state utilities using bayesian networks: An empirical analysis converting SF-12 into EQ-5D utility index in a national US sample,” *Med. Care*, vol. 49, no. 5, pp. 451–460, **2011**.

[77] H. J. P. Marvin *et al.*, “Application of Bayesian networks for hazard ranking of nanomaterials to support human health risk assessment,” *Nanotoxicology*, vol. 11, no. 1, pp. 123–133, **2017**.

[78] S. García-Herrero, M. A. Mariscal, J. García-Rodríguez, and D. O. Ritzel, “Working conditions, psychological/physical symptoms and occupational accidents. Bayesian network models,” *Saf. Sci.*, vol. 50, no. 9, pp. 1760–1774, Nov. **2012**.

[79] J. G. Blodgett and R. D. Anderson, “A Bayesian Network Model of the Consumer Complaint Process,” *J. Serv. Res.*, vol. 2, no. 4, pp. 321–338, **2000**.

[80] S. Chakraborty, K. Mengersen, C. Fidge, L. Ma, and D. Lassen, “A Bayesian Network-based customer satisfaction model: a tool for management decisions in railway transport,” *Decis. Anal.*, vol. 3, no. 1, **2016**.

[81] S. Salini and R. S. Kenett, “Bayesian networks of customer satisfaction survey data,” *J. Appl. Stat.*, vol. 36, no. 11, pp. 1177–1189, **2009**.

[82] R. D. Anderson, R. D. Mackoy, V. B. Thompson, and G. Harrell, “A Bayesian network estimation of the service-profit chain for transport service satisfaction,” *Decis. Sci.*, vol. 35, no. 4, pp. 665–688, **2004**.

[83] I. Mohammadfam, F. Ghasemi, O. Kalatpour, and A. Moghimbeigi, “Constructing a Bayesian network model for improving safety behavior of employees at workplaces,” *Appl. Ergon.*, vol. 58, pp. 35–47, Jan. **2017**.

[84] J. L. Riskowski, T. J. Hagedorn, and M. T. Hannan, “Measures of foot function, foot health, and foot pain,” *Arthritis Care Res. (Hoboken).*, vol. 63, no. S11, pp. S229–S239, Nov. **2011**.

[85] I. Lira, “Evaluating the Measurement Uncertainty: Fundamentals and practical guidance,” *Meas. Sci. Technol.*, vol. 13, no. 9, p. 1502, **2002**.

[86] H. J. Schunemann and G. H. Guyatt, “Commentary - Goodbye M(C)ID! hello MID, where do you come from?,” *Health Serv. Res.*, vol. 40, no. 2, pp. 593–597, **2005**.

[87] J. K. Kim and E. S. Park, “Comparative responsiveness and minimal clinically important differences for idiopathic ulnar impaction syndrome hand,” *Clin. Orthop. Relat. Res.*, vol. 471, no. 5, pp. 1406–1411, **2013**.

[88] O. Hägg, P. Fritzell, and A. Nordwall, “The clinical importance of changes in outcome scores after treatment for chronic low back pain,” *Eur. Spine J.*, vol. 12, no. 1, pp. 12–20, **2003**.

[89] S. D. Glassman, A. G. Copay, S. H. Berven, D. W. Polly, B. R. Subach, and L. Y. Carreon, “Defining substantial clinical benefit following lumbar spine arthrodesis,” *J. Bone Jt. Surg. - Ser. A*, vol. 90, no. 9, pp. 1839–1847, **2008**.

[90] C. Okoli and S. D. Pawlowski, “The Delphi method as a research tool: an example, design considerations and applications,” *Inf. Manag.*, vol. 42, no. 1, pp. 15–29, Dec. **2004**.

[91] E. F. Juniper, “Measuring health-related quality of life in rhinitis,” *J. Allergy Clin. Immunol.*, vol. 99, no. 2, **1997**.

[92] J. Bagó, F. J. S. Pérez-Grueso, E. Les, P. Hernández, and F. Pellisé, “Minimal important differences of the SRS-22 Patient Questionnaire following surgical treatment of idiopathic scoliosis,” *Eur. Spine J.*, vol. 18, no. 12, pp. 1898–1904, **2009**.

[93] G. R. Norman, J. A. Sloan, and K. W. Wyrwich, “Interpretation of changes in health-related quality of life the remarkable universality of half a standard deviation,” *Med. Care*, vol. 41, no. 5, pp. 582–592, **2003**.

[94] N. Friedman and M. Goldszmidt, “Discretizing Continuous Attributes While Learning Bayesian Networks,” *Icml*, pp. 157–165, **1996**.

[95] K. B. Landorf and J. A. Radford, “Minimal important difference: Values for the Foot Health Status Questionnaire, Foot Function Index and Visual Analogue Scale,” *Foot*, vol. 18, no. 1, pp. 15–19, **2008**.

[96] B. McMillan and D. Slepian, “Information Theory,” *Proc. IRE*, vol. 50, no. 5, pp. 1151–1157, **1962**.

[97] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. **1948**.

[98] P. Baudot, M. Tapia, D. Bennequin, and J. M. Goaillard, “Topological information data analysis,” *Entropy*, vol. 21, no. 9, pp. 1–31, **2019**.

[99] M. Stone, “Cross-Validatory Choice and Assessment of Statistical Predictions,” *J. R. Stat. Soc. Ser. B*, vol. 36, no. 2, pp. 111–133, **1974**.

[100] G. C. Cawley and N. L. C. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, **2010**.

[101] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, **2016**.

[102] S. Jain, M. White, and P. Radivojac, “Recovering True Classifier Performance in Positive-Unlabeled Learning,” *31st AAAI Conf. Artif. Intell. AAAI 2017*, pp. 2066–2072, Feb. **2017**.

[103] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, **2006**.

[104] V. H. Chuter, A. Searle, and M. J. Spink, “Flip-flop footwear with a moulded foot-bed for the treatment of foot pain: a randomised controlled trial,” *BMC Musculoskelet. Disord.*, vol. 17, no. 1, pp. 1–8, **2016**.

[105] U. T. Taddei, A. B. Matias, F. I. A. Ribeiro, S. A. Bus, and I. C. N. Sacco, “Effects of a foot strengthening program on foot muscle morphology and running mechanics: A proof-of-concept, single-blind randomized controlled trial,” *Phys. Ther. Sport*, vol. 42, pp. 107–115, **2020**.

[106] J. Sullivan, E. Pappas, R. Adams, J. Crosbie, and J. Burns, “Determinants of footwear difficulties in people with plantar heel pain,” *J. Foot Ankle Res.*, vol. 8, no. 1, pp. 1–7, **2015**.

[107] M. Grotle, N. K. Vøllestad, M. B. Veierød, and J. I. Brox, “Fear-avoidance beliefs and distress in relation to disability in acute and chronic low back pain,” *Pain*, vol. 112, no. 3, pp. 343–352, **2004**.

[108] K. Claeys, S. Brumagne, W. Dankaerts, H. Kiers, and L. Janssens, “Decreased variability in postural control strategies in young people with non-specific low back pain is associated with altered proprioceptive reweighting,” *Eur. J. Appl. Physiol.*, vol. 111, no. 1, pp. 115–123, **2011**.

[109] M. Cotchett, A. Lennecke, V. G. Medica, G. A. Whittaker, and D. R. Bonanno, “The association between pain catastrophising and kinesiophobia with pain and function in people with plantar heel pain,” *Foot*, vol. 32, pp. 8–14, Aug. **2017**.

[110] R. Severeijns, M. A. Van Den Hout, J. W. S. Vlaeyen, and H. S. J. Picavet, “Pain catastrophizing and general health status in a large Dutch community sample,” *Pain*, vol. 99, no. 1–2, pp. 367–376, **2002**.

[111] P. C. Wheeler, “Up to a quarter of patients with certain chronic recalcitrant tendinopathies may have central sensitisation: a prospective cohort of more than 300 patients,” *Br. J. Pain*, vol. 13, no. 3, pp. 137–144, **2019**.

[112] J. Kregel *et al.*, “The Dutch Central Sensitization Inventory (CSI),” *Clin. J. Pain*, vol. 32, no. 7, pp. 624–630, **2016**.

[113] L. Janssens *et al.*, “The effect of acute back muscle fatigue on postural control strategy in people with and without recurrent low back pain,” *Eur. Spine J.*, vol. 52, no. 2, pp. 2152–2159, **2013**.

[114] T. Inrig, B. Amey, C. Borthwick, and D. Beaton, “Validity and reliability of the Fear-Avoidance Beliefs questionnaire (FABQ) in workers with Upper Extremity injuries,” *J. Occup. Rehabil.*, vol. 22, no. 1, pp. 59–70, **2012**.

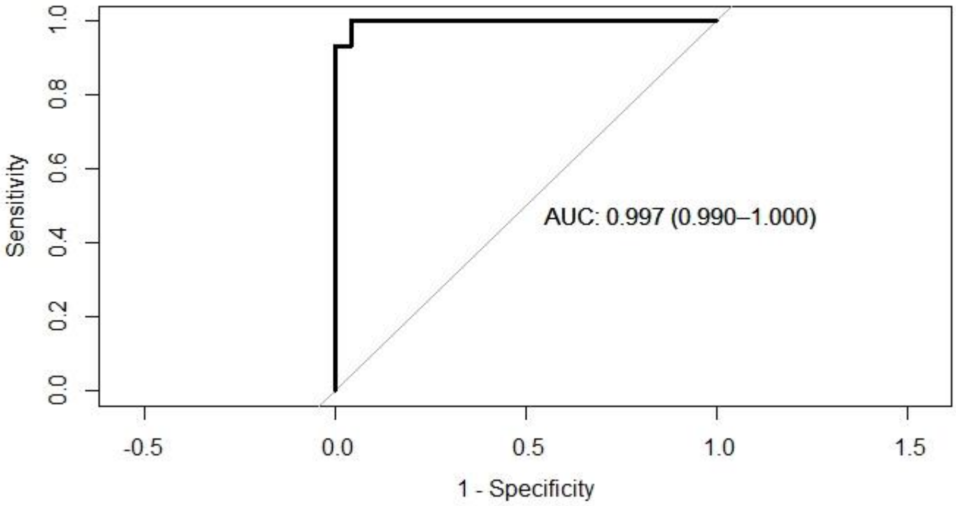
[115] W. Scott, T. H. Wideman, and M. J. L. Sullivan, “Clinically meaningful scores on pain catastrophizing before and after multidisciplinary rehabilitation: A prospective study of individuals with subacute pain after whiplash injury,” *Clin. J. Pain*, vol. 30, no. 3, pp. 183–190, **2014**.

[116] R. Neblett, “The central sensitization inventory: A user’s manual,” *J. Appl. Biobehav. Res.*, vol. 23, no. 2, pp. 1–13, **2018**.

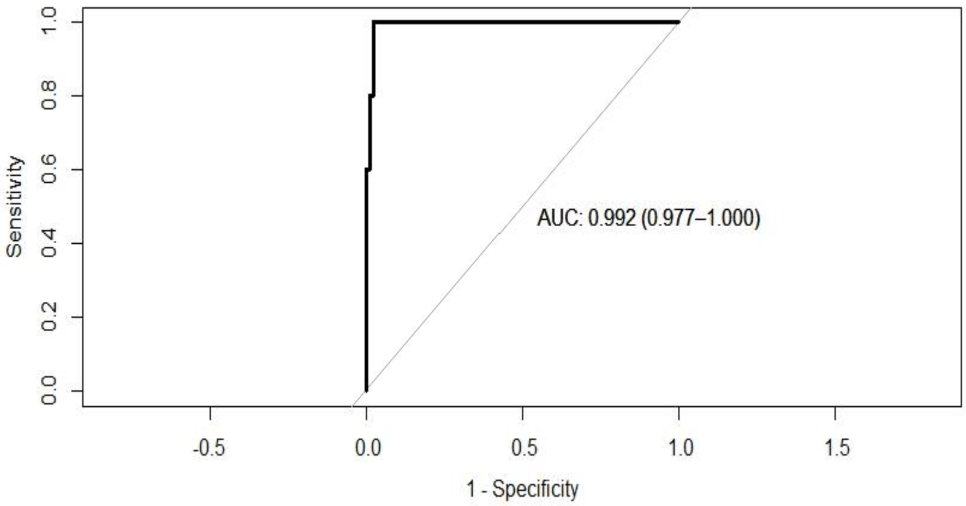
[117] R. Neblett *et al.*, “The central sensitization inventory (CSI): Establishing clinically significant values for identifying central sensitivity syndromes in an outpatient chronic pain sample,” *J. Pain*, vol. 14, no. 5, pp. 438–445, **2013**.

APPENDICES

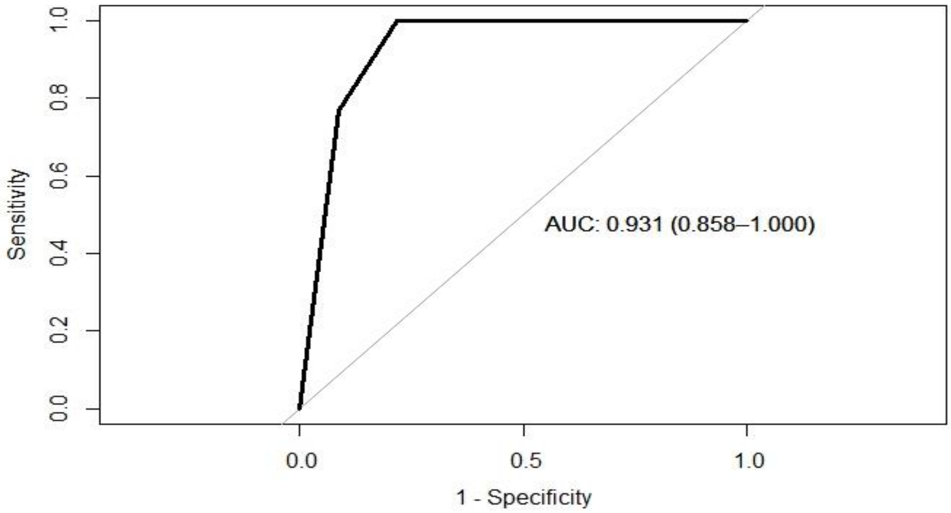
Appendix-A: ROC Curve of HC, Tabu and MMHC Models for GFH Factor



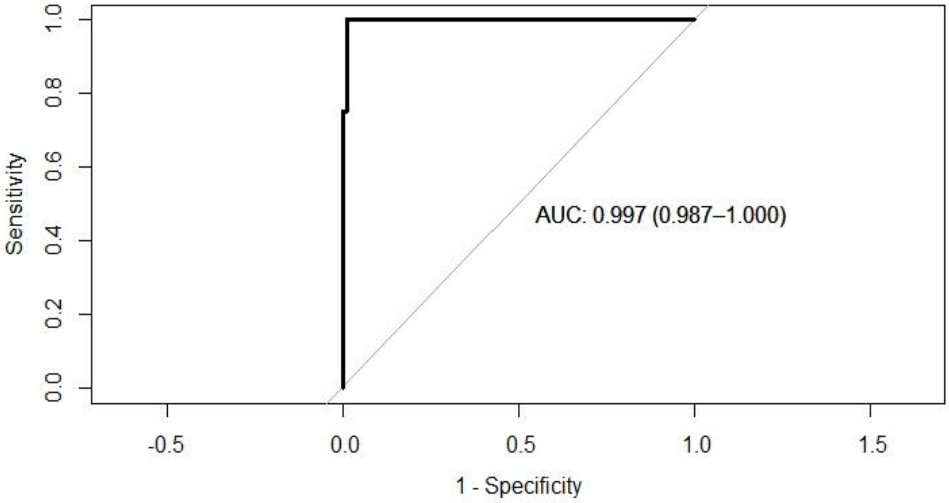
Appendix-B: ROC Curve of Tabu Model for Physical Activity Factor



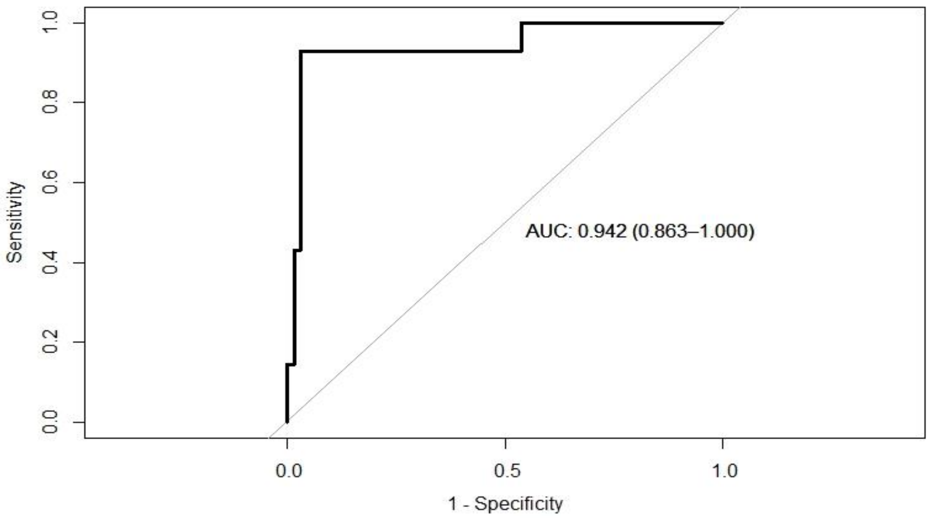
Appendix-C: ROC Curve of HC Model for Work Factor



Appendix-D: ROC Curve of HC Model for PCS



Appendix-E: ROC Curve of Tabu Model for CSI


Sensitivity
AUC: 0.942 (0.863–1.000)
1 - Specificity