Proposal

A Bayesian Network Decision Support Tool for Low Back Pain Using a RAND Appropriateness Procedure: Proposal and Internal Pilot Study

Adele Hill[1], MSc; Christopher H Joyner[2], PhD; Chloe Keith-Jopp[1,3], MRES; Barbaros Yet[4], PhD; Ceren Tuncer Sakar[5], PhD; William Marsh[2], PhD; Dylan Morrissey[1,3], PhD

[1]Sport and Exercise Medicine, Queen Mary University of London, London, United Kingdom

[2]Electronics, Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

[3]Barts Health NHS Trust, London, United Kingdom

[4]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[5]Department of Industrial Engineering, Hacettepe University, Ankara, Turkey

Corresponding Author:
Adele Hill, MSc
Sport and Exercise Medicine
Queen Mary University of London
Mile End Road
London, E1 4NS
United Kingdom
Phone: 44 7971495806
Email: a.hill@smd18.qmul.ac.uk

Abstract

Background: Low back pain (LBP) is an increasingly burdensome condition for patients and health professionals alike, with consistent demonstration of increasing persistent pain and disability. Previous decision support tools for LBP management have focused on a subset of factors owing to time constraints and ease of use for the clinician. With the explosion of interest in machine learning tools and the commitment from Western governments to introduce this technology, there are opportunities to develop intelligent decision support tools. We will do this for LBP using a Bayesian network, which will entail constructing a clinical reasoning model elicited from experts.

Objective: This paper proposes a method for conducting a modified RAND appropriateness procedure to elicit the knowledge required to construct a Bayesian network from a group of domain experts in LBP, and reports the lessons learned from the internal pilot of the procedure.

Methods: We propose to recruit expert clinicians with a special interest in LBP from across a range of medical specialties, such as orthopedics, rheumatology, and sports medicine. The procedure will consist of four stages. Stage 1 is an online elicitation of variables to be considered by the model, followed by a face-to-face workshop. Stage 2 is an online elicitation of the structure of the model, followed by a face-to-face workshop. Stage 3 consists of an online phase to elicit probabilities to populate the Bayesian network. Stage 4 is a rudimentary validation of the Bayesian network.

Results: Ethical approval has been obtained from the Research Ethics Committee at Queen Mary University of London. An internal pilot of the procedure has been run with clinical colleagues from the research team. This showed that an alternating process of three remote activities and two in-person meetings was required to complete the elicitation without overburdening participants. Lessons learned have included the need for a bespoke online elicitation tool to run between face-to-face meetings and for careful operational definition of descriptive terms, even if widely clinically used. Further, tools are required to remotely deliver training about self-identification of various forms of cognitive bias and explain the underlying principles of a Bayesian network. The use of the internal pilot was recognized as being a methodological necessity.

Conclusions: We have proposed a method to construct Bayesian networks that are representative of expert clinical reasoning for a musculoskeletal condition in this case. We have tested the method with an internal pilot to refine the process prior to deployment, which indicates the process can be successful. The internal pilot has also revealed the software support requirements for the elicitation process to model clinical reasoning for a range of conditions.

International Registered Report Identifier (IRRID): DERR1-10.2196/21804

(JMIR Res Protoc 2021;10(1):e21804) doi: 10.2196/21804

KEYWORDS

back pain; decision making; Bayesian methods; consensus

Introduction

Back pain is a prime example of the general increase in long-term musculoskeletal conditions. It has been deemed a leading cause of years lived with disability worldwide, and health care costs for treating back pain are escalating [1]. Some low back pain (LBP) cases are associated with injuries that will self-resolve, but there are a considerable number of people who live with disabling LBP. It is difficult to predict who will have a favorable outcome and who will not. Meta-analyses have consistently shown that treatment outcomes for musculoskeletal conditions are partial [2]. Studies have shown that clinicians may not recognize complexity when assessing these patients [3], and efforts made to improve diagnosis via pattern recognition or analytical approaches [4] have limited success. An artificial intelligence (AI) decision support tool that overcomes one or more of these issues would greatly improve the referral of LBP patients to the correct treatment care pathway, and the approach would be generalizable to other musculoskeletal conditions.

The potential of AI systems to address some of these issues has been recognized by The Digital Framework for Allied Health Professionals in the United Kingdom [5], which highlights the priority to develop digitally mature systems, improve outcomes, and limit variation in service delivery. Similarly, in 2019, Her Majesty's Government announced a £250 million investment in AI development in the UK National Health Service [6] for such purposes. However, many of these techniques are data driven and thus require large error-free data sets that must, crucially, also contain outcome results. Unfortunately, while there is an abundance of data within the musculoskeletal health care system, access is severely restricted and most of the data are unrefined, often being free text or hand written without the necessary outcomes. Furthermore, studies have found that clinicians are skeptical about the ability of AI to perform at the human level [7], and patient and public involvement forums have highlighted a mistrust of black box systems. Successful adoption of AI therefore requires combining clinical evidence, patient data, and expert opinion to mirror the clinical reasoning process and provide transparency regarding how predictions have been reached.

An LBP decision support tool to guide the treatment of patients has been developed with an expert-opinion Delphi-consensus approach in the Netherlands [8], with some success, although there was a suggestion that the results could be further improved with AI and machine learning techniques [9]. The previous authors particularly highlight the limitations of traditional multivariate regression models and suggest that a Bayesian network (BN) would take into account more clinical aspects of assessment. A BN is a probabilistic model that offers an ideal approach for modeling clinical reasoning as it is capable of combining expert opinion with available evidence to provide probabilistic outcomes for a given scenario [10]. These outcomes can be continuously updated to ensure the most recent clinical knowledge is being used [11] and provide explanations of the predictions arrived at [12].

In order to support clinicians in their decision making, we envisage a BN that can predict a patient's response to a given treatment or course of action when they first present with LBP (ie, in a primary care setting). In order to do this, we must first characterize the types of LBP possible, and the associated risk factors, signs, and symptoms. We propose eliciting a BN that predicts the probability of a patient having a certain LBP-related presentation and, subsequently, how likely they are to improve in response to a course of action. To be clear, this focus on characterization rather than diagnosis was deliberately taken owing to the lack of accurate diagnostic labels for the majority of people with LBP [13], but the BN was designed to identify patients with serious diagnoses, such as nerve root pain and cauda equina syndrome, where these can be identified. For reasons mentioned above, the clinical knowledge base of our BN must be elicited from experts [14,15]. We also believe that it is important not to constrain the type of LBP considered but use expert opinion to guide our focus. In this paper, we outline a protocol for a BN elicitation process that aims to balance the construction of a complex expert-driven model for the treatment of LBP while minimizing the elicitation burden placed on participants. We also describe a pilot study of this process, which highlights the subsequent results that informed the main protocol, discuss the overall methodology, and consider future implementation of the AI tool.

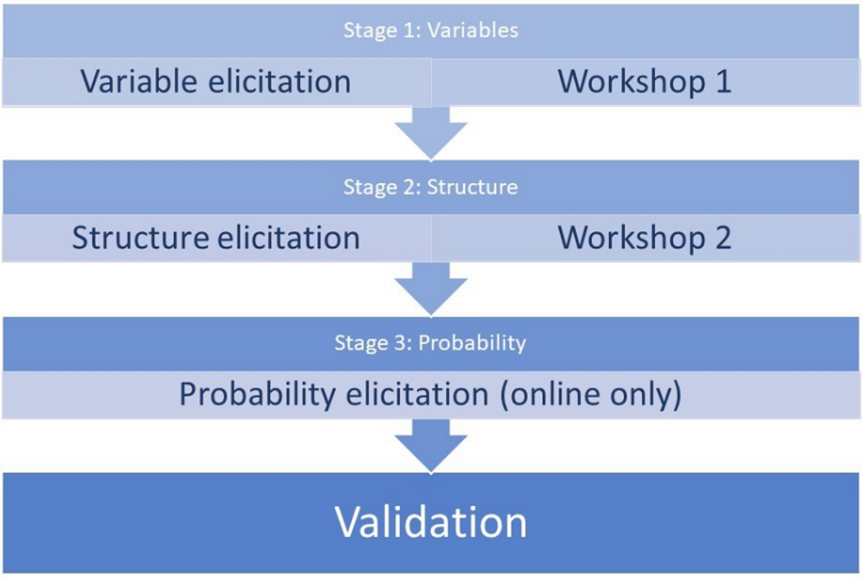
Methods

Design

A BN is comprised of the following three components: (1) variables, quantities of interest, such as age, BMI, and the presence/absence of a condition; (2) structure, the dependence of variables on each other, for instance, the condition may be more prevalent in the elderly but the BMI makes no difference; (3) probabilities, quantification of the structure, for example, the probability of having a condition given the person is of a certain age.

We therefore divide the elicitation process into three distinct stages related to those components plus a final stage intended as a rudimentary validation of the output from the process (Figure 1).

Figure 1. Study procedure.



The overall elicitation method will be a combined Delphi and RAND appropriateness procedure [10]. Participants will first answer questions individually online before contentious results are discussed together in a face-to-face workshop. We have chosen this strategy because related methods have been used with success in previous musculoskeletal studies [8,16-18], and it combines individual tasks and group discussions to explore biases [19]. We acknowledge that further rounds of discussion may help with consensus, but time constraints render this infeasible.

Ethical approval was sought from the Queen Mary University of London, and approval was granted by the Queen Mary Ethics of Research Committee (reference: QMREC2018/48027).

Participants and Recruitment

Clinicians will be recruited via clinical networks, word of mouth, and social media. There is no definitive number of participants recommended for this style of study. However, for face-to-face workshops, an optimum number of attendees has been suggested to lie between seven and 15 [19]. This number of participants is thought to represent a range of views and promote discussion, without becoming too onerous to manage and facilitate in a workshop setting. We will include extended scope/advanced practitioner physiotherapists (covering orthopedics, rheumatology, neurosurgery, and general musculoskeletal triage services), general practitioners with a special interest in musculoskeletal conditions, and sports medicine doctors. We will exclude clinicians who do not regularly manage patients with LBP or do not hold current professional registration. Participants will be provided with information about the research and be asked to provide informed written consent prior to taking part.

Data Management

Participants will not be identified in any reports or publications of the results. Any information held by the project team in relation to participants will be kept confidential and managed in accordance with the Data Protection Act, the General Data Protection Regulation (GDPR), the UK Policy Framework for Health and Social Care Research, and the conditions of the Research Ethics Committee.

Elicitation Process

Stage 1: Variable Elicitation

Ad hoc building of the BN causal structure is prone to biases and modeling error, which can lead to overly complex models that are not repeatable in later studies [20]. To mitigate this potential issue and to form concepts familiar to the clinical participants, we have constructed a conceptual model where potentially relevant factors in the assessment and treatment of LBP will be grouped into either “risk factors,” “signs and symptoms,” or “judgement factors.” A list of potential treatments will also be elicited. The categories have been connected in the fashion as shown in Figure 2 to form a “skeleton causal diagram,” which approximates the clinical reasoning process. This is necessary for the model functionality and makes the elicitation process recognizable for participants.

Conventional appropriateness scoring usually relies on a Likert scale [19]; however, it is anticipated that participants will have considerable time constraints and the number of potential variables will render the task unfeasible. We will therefore use a placement and ranking procedure. At the start, participants will see many example variables chosen by the research team (Figure 3) (reasons behind this decision are presented in the Results section). They can also create new variables. They will then be asked to place them into one of the three categories of their choosing and order the three associated lists so the variables they deem most important for the assessment of patients with LBP appear at the top. Unfortunately, this setup does create the risk of anchoring biases, but we hope this will be mitigated when the variables are voted on in the workshop.

Figure 2. Skeleton causal diagram involving the categories of variables to be considered.

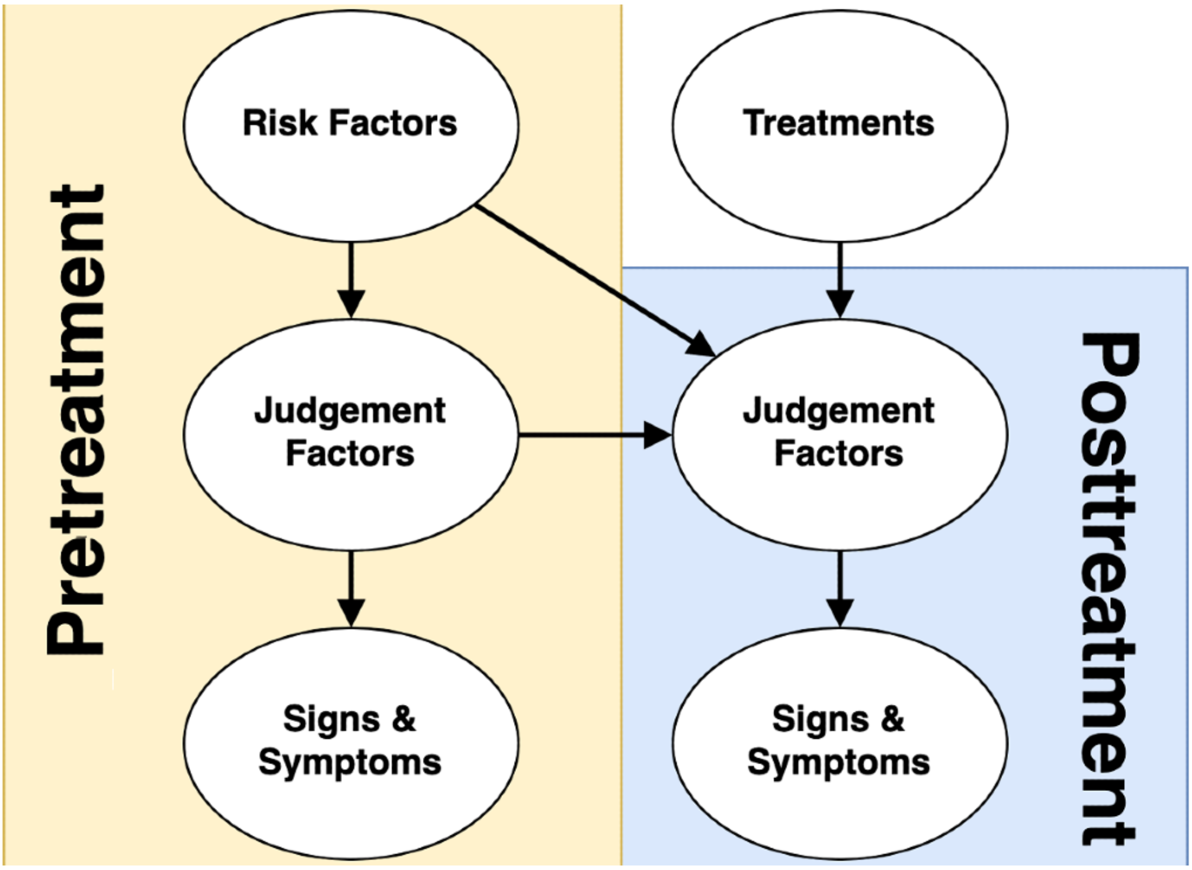
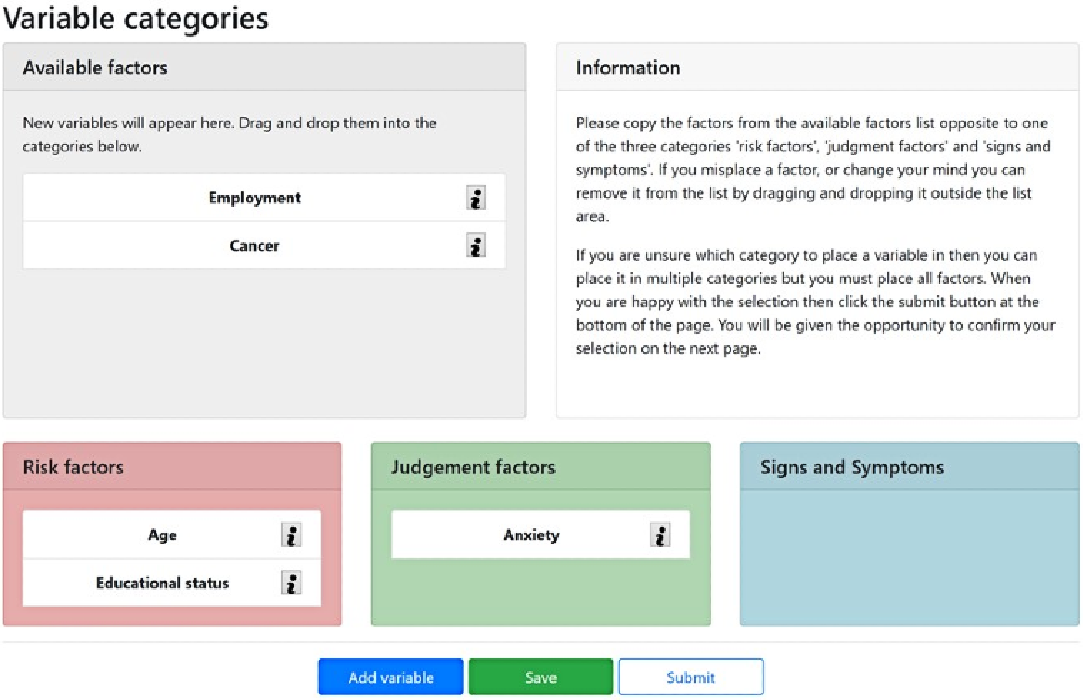


Figure 3. Mockup of the online interface for stage 1 of the process.


Variable categories

Available factors

New variables will appear here. Drag and drop them into the categories below.

Employment

Cancer

Information

Please copy the factors from the available factors list opposite to one of the three categories 'risk factors', 'judgment factors' and 'signs and symptoms'. If you misplace a factor, or change your mind you can remove it from the list by dragging and dropping it outside the list area.

If you are unsure which category to place a variable in then you can place it in multiple categories but you must place all factors. When you are happy with the selection then click the submit button at the bottom of the page. You will be given the opportunity to confirm your selection on the next page.

Risk factors: Age, Educational status

Judgement factors: Anxiety

Signs and Symptoms

To aid their understanding, participants will be given explanatory information about what constitutes risk factors, signs and symptoms, and judgement factors (Table 1), as well as many example variables from the prepopulated lists. Additionally, the meaning of each variable will be clarified by giving a question that could be answered to find out the value of the variable (such as “What is the patient’s anxiety level?”) and indication of the mode and units of measurement of the variable to ensure consistent responses.

Table 1. Categories of variables in the Bayesian network.

Category name[a]	Definition
Risk factor	These are pre-existing factors that might raise (or lower) the chances of the patient having a certain judgement factor or characterization of low back pain. They may also have an effect on the efficacy of the treatment but, crucially, are factors (like age and demographics) that cannot be changed by the treatment.
Signs and symptoms	These are factors that arise as a consequence of the patient having certain conditions or judgement factors. In contrast to risk factors, signs and symptoms are factors that may change or respond to treatment.
Judgement factors	These are clinical reasoning factors or the characterization of the patient's low back pain. They are the factors that are likely to be unmeasurable with a Patient Reported Outcome Measure or test, but which help a clinician reason about prognosis or likelihood of recovery.

[a] Explanatory information about each type of variable will be made available to participants in order to inform the categorization of factors from the prepopulated lists and those created in the elicitation.

Each variable will be included in a particular category if more than 80% of participants allocate it. From those participants, we will take the median normalized rank (0 if placed at the bottom of the list and 1 at the top) to give an indication of preference for inclusion, which will be further adjusted by an interpercentile range (IPR) measure to measure consensus (more consensus will result in less adjustment of the median value). The top-placed variables will be selected for the face-to-face stage. We anticipate around 50 variables can be taken forward without overburdening participants, but we reserve some clinical judgement to decide on the exact number. Suggested variables from the online elicitation will be consolidated by the working team, and those judged by the clinical team to be duplicates will be removed.

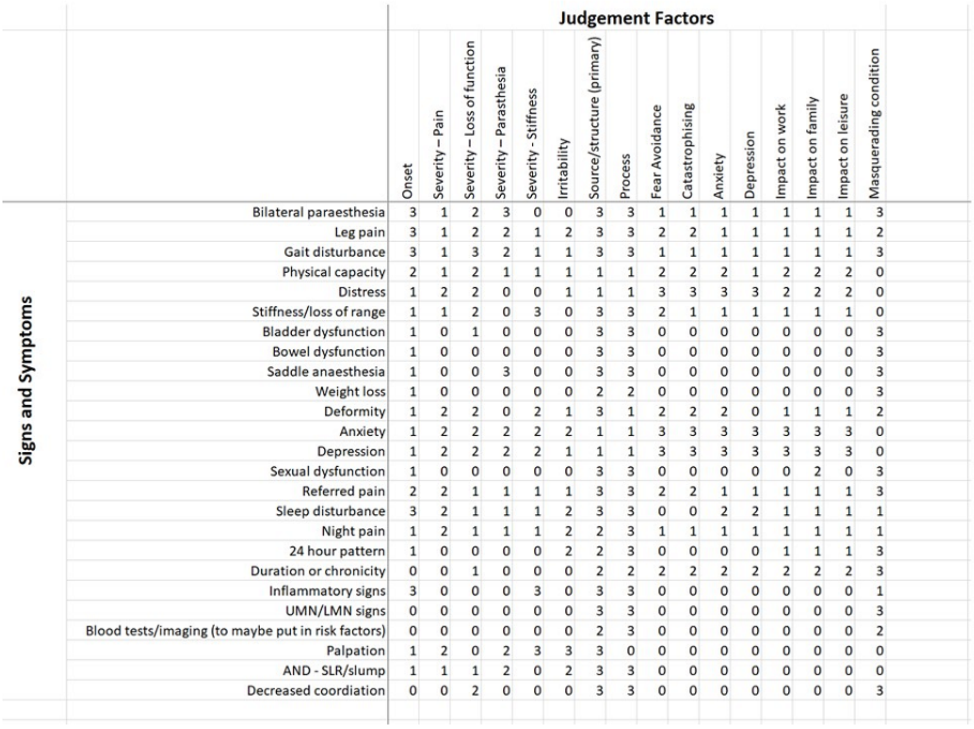
After the online session, a face-to-face workshop will be held, in which participants will use the same tool. Misconceptions or differences in understanding will be clarified by the research team. A member of the research team will facilitate discussion of the variables that do not have consensus via use of open-ended questions. The face-to-face workshops will be video recorded to capture the discussions and enable postworkshop checking and clarification. The participants will then be given the opportunity to categorize and rank the variables for a second time, on the basis of the discussion. We anticipate that the mixture of not allowing new variables, the refined interface, and having numerous variables fixed in categories will reduce the elicitation burden while still maintaining an opportunity for group discussions, according to the conventional RAND process.

Stage 2: Structure Elicitation

The second stage seeks to connect individual variables linking risk factors to judgement factors, and judgement factors to signs and symptoms. This will be achieved by grids (Figure 4). In the online elicitation, participants will be presented with a blank grid and asked to give each relationship a strength score between 0 and 3 in the appropriate cell. The definitions of the scores are as follows: 0, no relationship; 1, X *sometimes* has a *small* effect on Y; 2, X *sometimes* has a *large* effect on Y or X *always* has a *small* effect on Y; and 3, X *always* has a *large* effect on Y, and they reflect the nuances of the BN probabilities.

We will again use the median and an 80% IPR as descriptive statistics. To make the workshop manageable (but still allow for discussion), we will fix the median score of those variables with an IPR of 1 or less (ie, the cells in the grids cannot be edited), and all others will be edited again by the participants following discussions. A subsequent overall score will be taken from the workshop for each relationship in a similar manner to the overall score in stage 1. We will then only keep the strongest relationships to form the structure of the BN, and any variables not connected to others will be discarded.

Figure 4. An example grid for stage 2. Relationship strengths are placed in the cells. The definitions of the scores are as follows: 0, no relationship; 1, X sometimes has a small effect on Y; 2, X sometimes has a large effect on Y or X always has a small effect on Y; and 3, X always has a large effect on Y.

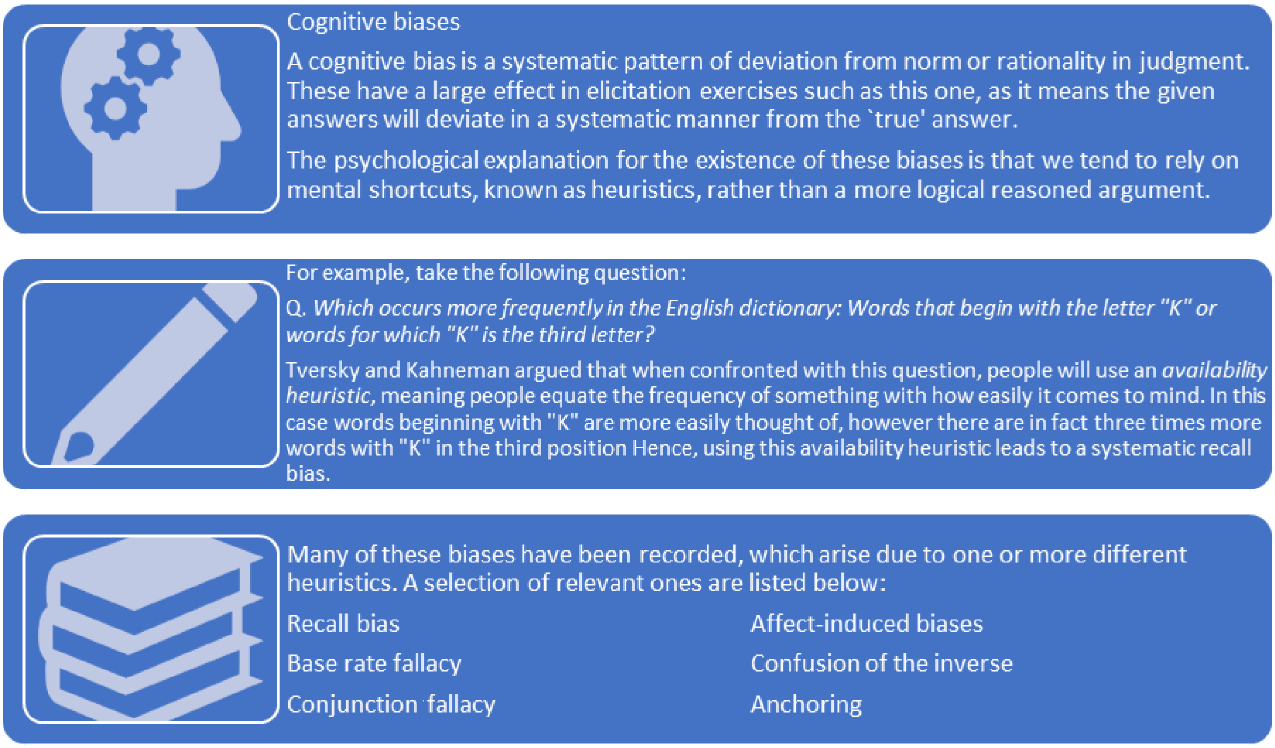


		Judgement Factors															
		Onset	Severity – Pain	Severity – Loss of function	Severity – Parasthesia	Severity - Stiffness	Irritability	Source/structure (primary)	Process	Fear Avoidance	Catastrophising	Anxiety	Depression	Impact on work	Impact on family	Impact on leisure	Masquerading condition
Signs and Symptoms	Bilateral paraesthesia	3	1	2	3	0	0	3	3	1	1	1	1	1	1	1	3
	Leg pain	3	1	2	2	1	2	3	3	2	2	1	1	1	1	1	2
	Gait disturbance	3	1	3	2	1	1	3	3	1	1	1	1	1	1	1	3
	Physical capacity	2	1	2	1	1	1	1	1	2	2	2	1	2	2	2	0
	Distress	1	2	2	0	0	1	1	1	3	3	3	3	2	2	2	0
	Stiffness/loss of range	1	1	2	0	3	0	3	3	2	1	1	1	1	1	1	0
	Bladder dysfunction	1	0	1	0	0	0	3	3	0	0	0	0	0	0	0	3
	Bowel dysfunction	1	0	0	0	0	0	3	3	0	0	0	0	0	0	0	3
	Saddle anaesthesia	1	0	0	3	0	0	3	3	0	0	0	0	0	0	0	3
	Weight loss	1	0	0	0	0	0	2	2	0	0	0	0	0	0	0	3
	Deformity	1	2	2	0	2	1	3	1	2	2	2	0	1	1	1	2
	Anxiety	1	2	2	2	2	2	1	1	3	3	3	3	3	3	3	0
	Depression	1	2	2	2	2	1	1	1	3	3	3	3	3	3	3	0
	Sexual dysfunction	1	0	0	0	0	0	3	3	0	0	0	0	0	2	0	3
	Referred pain	2	2	1	1	1	1	3	3	2	2	1	1	1	1	1	3
	Sleep disturbance	3	2	1	1	1	2	3	3	0	0	2	2	1	1	1	1
	Night pain	1	2	1	1	1	2	2	3	1	1	1	1	1	1	1	1
	24 hour pattern	1	0	0	0	0	2	2	3	0	0	0	0	1	1	1	3
	Duration or chronicity	0	0	1	0	0	0	2	2	2	2	2	2	2	2	2	3
	Inflammatory signs	3	0	0	0	3	0	3	3	0	0	0	0	0	0	0	1
	UMN/LMN signs	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	3
	Blood tests/imaging (to maybe put in risk factors)	0	0	0	0	0	0	2	3	0	0	0	0	0	0	0	2
	Palpation	1	2	0	2	3	3	3	0	0	0	0	0	0	0	0	0
	AND - SLR/slump	1	1	1	2	0	2	3	3	0	0	0	0	0	0	0	0
	Decreased coordiation	0	0	2	0	0	0	3	3	0	0	0	0	0	0	0	3

Stage 3: Probability Elicitation

The third stage seeks to endow the BN structure with quantitative probabilities for making predictions. Crucially, although cognitive biases appear in all stages, the probability elicitation is particularly susceptible because of the quantitative nature and known issues with probability estimation [21,22]. Cognitive bias training will therefore be given prior to starting the third stage. This will include questions from unrelated fields that are susceptible to biases likely to arise within the elicitation, such as recall bias and confusion of inverse probabilities [21,22]. A brief explanation of how the bias may lead to systematic deviations in the estimates will be presented for each question (Figure 5). Moreover, we will word questions so that participants are aware of the target population and are thinking about frequency estimates, rather than individual patients (Figure 6).

Figure 5. Examples of questions to be given in the cognitive bias training.



Cognitive biases

A cognitive bias is a systematic pattern of deviation from norm or rationality in judgment. These have a large effect in elicitation exercises such as this one, as it means the given answers will deviate in a systematic manner from the `true' answer.

The psychological explanation for the existence of these biases is that we tend to rely on mental shortcuts, known as heuristics, rather than a more logical reasoned argument.

For example, take the following question:

Q. *Which occurs more frequently in the English dictionary: Words that begin with the letter "K" or words for which "K" is the third letter?*

Tversky and Kahneman argued that when confronted with this question, people will use an *availability heuristic*, meaning people equate the frequency of something with how easily it comes to mind. In this case words beginning with "K" are more easily thought of, however there are in fact three times more words with "K" in the third position Hence, using this availability heuristic leads to a systematic recall bias.

Many of these biases have been recorded, which arise due to one or more different heuristics. A selection of relevant ones are listed below:

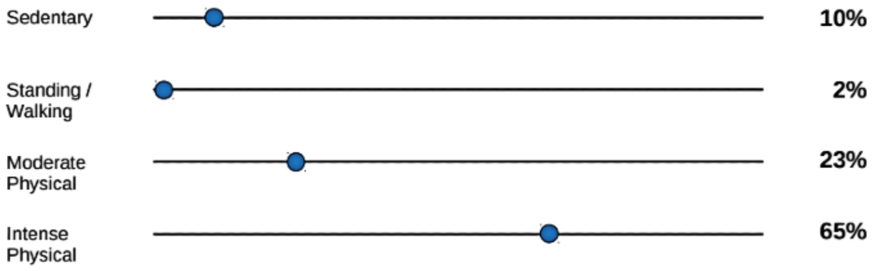
Recall bias
Base rate fallacy
Conjunction fallacy

Affect-induced biases
Confusion of the inverse
Anchoring

The variables in the BN can be binary (eg, true and false), labeled, that is, nominal with states without any order (eg, muscle, tissue, and tendon), or ranked, that is, ordinal with states with increasing or decreasing order (eg, high, medium, and low). Additionally, some variables (eg, age) will be independent of others (known as prior variables), whereas others will be dependent (eg, the prevalence of a condition may be influenced by the age of a patient).

For binary and labeled variables, sliders will be used to allow participants to assign probabilities individually to each state. For example (Figure 6), the question “What proportion of people work in the described type of job?” has a series of nominal answers, which can be represented with individual percentages adding up to 100. In contrast, ranked variables have a smooth distribution, so participants will be asked to estimate the most likely value and the extent of the variation around that value [23].

Figure 6. Example of direct probability elicitation for the question “What proportion of people work in the described type of job?”.



For independent (prior) variables, a simple question similar to that in Figure 6 will suffice, whereas for conditional variables, the states of related variables will be needed, for example, “What is the probability of a patient having a fracture *given* that they have an *age* of *80+* and experienced *trauma*?”

For binary/labeled variables, the answers from the probability elicitation will be combined using logistic regression and generalized noisy OR [23] techniques to generate the probability tables for each variables. For ranked variables, we will use the answers to estimate a truncated normal distribution [24]. This reduces the number of parameters compared with other methods [25,26] and so has the advantage of avoiding overfitting, but certain nuances of the relationships are missed. Outlying answers will be identified using an IPR-based metric and removed so as not to bias the regression approach.

Stage 4: Validation

A subset of clinical experts will be asked to give their judgement regarding the decision-making process for a number of given scenarios, which will then be compared against the model output, otherwise known as “face validity” [27]. Further validation will likely be required to establish the efficacy of the BN in practice; however, this is outside the scope of the study.

Internal Pilot

A much reduced version of the above protocol has already been conducted in an internal pilot study with three clinical colleagues (AH, CKJ, and DM) from Queen Mary University of London, who regularly treat patients with LBP. This was to test whether a suitable model could feasibly be constructed within a framework where the elicitation burden was not too high.

They completed all three stages of the elicitation process. During and after each stage of the process, a discussion took place regarding overall impressions of the stage in order to learn lessons and refine the process. These discussions captured clinical perspectives on the design of the questions and elicitation, feedback on what may or may not be acceptable to clinical participants, and appropriate redesign to ensure that the requirements for the modeling process were still being met. The differences in methods are discussed below, with lessons learned reserved for the subsequent Results section.

Stage 1: Variable Elicitation

The creation, placement, and ranking of variables were performed individually in Word documents (Microsoft Corp), with no suggested variables available from the start. Responses were collated by the computer science team and then discussed in a workshop meeting to form consensus regarding the categorization and definition of variables. Because of limited participant numbers, a final ranking was obtained via the discussion.

Stage 2: Structure Elicitation

This stage was performed as described in the main protocol, except for definitions of the relationship strength. Owing to limited participants, the mean and SD were used as descriptive statistics instead. A simpler definition of relationship strength (Table 1) was also used, with 0, 1, 2, and 3 simply described as “none,” “weak,” “medium,” and “strong,” respectively.

Stage 3: Probability Elicitation

Clinical colleagues were first asked to complete example questions based on the methods reported previously [26,28] on paper and provide their feedback. This guided the development of a primitive online tool to automatically generate the questions and record answers. Despite basic methods for easing the elicitation, the number of variables, states, and connections led to upwards of 700 questions. Completion was only feasible by splitting the questions among the three clinical members of the team and avoiding subsequent workshop discussions. Unfortunately, elicitation of probabilities associated with the treatments was abandoned owing to the volume of questions and the scope of the BN restricted to predicting judgment factors prior to treatment.

Results

Overview

The below text documents the outcomes of the pilot study that influenced the subsequent makeup of the described protocol. We discuss each stage separately.

Stage 1: Variable Elicitation

Many variables were suggested, with most having similar definitions that could be combined easily, and there was general agreement about inclusion. However, the following issues became apparent:

- Using Word documents to collate the results was cumbersome. It was also felt that the crude method could adversely affect the engagement of participants.
- There was some misunderstanding regarding the categories and how they fitted into the BN.
- There was misunderstanding and poor explanation of computer science terminology, particularly what was meant by “definition” and “states” of a variable.
- Discrepancies arose in choosing categories for certain variables. For example, “depression” could be considered a risk factor for LBP, a sign and symptom in cases where depression was a result of LBP, or a judgement factor where depression was thought to be the main driver of the patient’s condition.
- There was no formal quantitative procedure for ranking the variables.
- The time taken to discuss all the suggested variables was too long.

The use of an online interface has been influenced by the issues described. We have decided to present the participants with a number of prepopulated variables based on the pilot study together with a review of relevant literature [29-31]. This is to help them understand the terminology, allow us to gather better statistics, and reduce elicitation burden. In addition, this would highlight the difficultly in placing certain variables (eg, the “depression” example above). One of the prepopulated examples will highlight this issue to allow the participants to decide on the best handling of these variables. As described in the methods, participants will still be given the ability to add as many variables as they deem appropriate to help avoid biasing results toward the pilot.

Stage 2: Structure Elicitation

The process of filling in the grids to define relationships was understandable and streamlined; however, the main issue concerned the time burden, arising as a result of too many variables passed from stage 1. The following issues were also identified:

- The mean and SD measure frame consensus in terms of averaged values and so can be affected by outliers, which is not suitable for the full study.
- Similar to stage 1, the Excel files were cumbersome to manage and made preserving anonymity difficult to avoid potential biases.
- The wording of the relationships (none to strong) did not convey the nuances of the relationships between variables that could occur in the BN.

Again, the decision to use an online interface came from these issues of data collection, engagement of the participants, and anonymity of responses. For the main protocol, we have updated the definition of the relationships and introduced more appropriate statistical measures.

Stage 3: Probability Elicitation

The use of the online interface was, in general, a success, making the acquisition of data more streamlined and user friendly. This success, along with issues identified in previous stages of the pilot, was the reason for our decision to transfer all stages of the process online. However, beyond the sheer number of questions, we identified the following issues:

- Cognitive biases were mentioned but not adequately explained or motivated, and relevant examples were not included.
- Numerous issues with the questions and online interface were raised, including (1) eliciting the distributions required different questions depending on the variable type; (2) incorrect/strange wording of the questions; (3) questions about individual probabilities rather than frequency estimates (the former is more cognitively challenging); (4) lack of information about the related variables was not included, meaning participants had to refer elsewhere, thus slowing the process down; and (5) technical terminology displayed for ranked variables, such as “mean” and “variance,” was not very intuitive.

As mentioned in the protocol, we have introduced targeted cognitive bias training to help overcome those specific issues. Interface issues will also be addressed during development. The main issue is the elicitation burden, and the pilot has shown that any follow-up workshop would be too time consuming for participants. Therefore, instead, possible outlying answers will be identified using an IPR-based metric and removed so as not to bias the regression approach.

Stage 4: Informal Validation

Following the probability elicitation, a tentative BN was constructed using the AgenaRisk software [32] and compared against fictitious case study scenarios. This was in order to check whether the process was capable of returning a suitable model that would be meaningful to clinicians. The scenarios represented patients presenting with signs of serious underlying pathology, inflammatory pathology, and nonspecific LBP. The BN appeared to reason in a similar qualitative manner to that which the clinicians would expect.

Discussion

LBP Clinical Reasoning

We believe we have created a process that compliments the clinical reasoning process familiar to domain experts by reflecting our mathematical variables into distinct recognizable categories and asking them to describe how variables within those categories are related. Enumerating probability relationships is an area in which we anticipate experts will be least comfortable; those values are often implicitly known but almost never explicitly expressed in clinical practice. We have attempted to mitigate this by identifying common cognitive biases [21,22] and designing the elicitation to reduce the burden.

Giving this structure to the elicitation and subsequently the BN will provide confidence to future users about the "reasoning" performed by the tool. We hope the role of clinicians within the process will help alleviate concerns among users about its technical nature. Second, the structure is relatable and can be interrogated by the end user to understand the outcomes from the tool. Additionally, although the overall BN outline (Figure 2) has been designed specifically for this project, we believe it to represent clinical reasoning in a broad array of musculoskeletal conditions.

Methodology for Eliciting Expert Opinion

There exists a wealth of literature concerning the elicitation of BNs and associated probabilities [14,15,24-28]. However, we are not aware of previous attempts to elicit comparably sized BNs with such limited time constraints on experts. To increase the chances of success, many of the individual components will use tried and tested methods, such as Noisy OR [23] and ranked node approximation [24]. Nevertheless, combining them all together in this fashion poses a major challenge.

Using the Delphi/RAND methodology [10] will help mitigate biases via group discussion and give participants an opportunity to reach a consensus. Considering an approach from the report by Ritchie et al [33], the discussion will need to remain flexible enough to explore in depth the contentious issues raised by the participants, with open questioning to avoid bias from the team that has already piloted the method. Additionally, there will likely need to be exploration of divergent views, drawing attention to elicited themes where participants have disagreed.

The very basic validation process already conducted gives encouragement that this method is viable for the purpose of eliciting a clinical reasoning BN for LBP; however, the study team recognizes the inherent bias in conducting an internal validation. As mentioned, our Delphi/RAND process has been designed to help remove such biases, but it still remains to be seen whether the process conducted with external experts will yield a clinically credible BN.

Other Uses for the Methodology

This methodology, should it be successful in the main study, could provide a new framework for developing decision support for other musculoskeletal problems and for other complex interventions in medicine. Additionally, it could provide a learning tool for aspiring expert clinicians to confront their cognitive biases, as it did for the pilot study colleagues.

Conclusion

We propose a protocol for developing a complex expert-driven BN decision support tool while minimizing the elicitation burden on experts, who, as professional clinicians, have limited time to offer. A basic version of the elicitation method has been internally tested with a small group of clinician researchers in a pilot study, which has yielded credible results. The proposed protocol aims to establish consensus among users by using appropriate scoring metrics and subsequent workshops to draw consensus in a Delphi-like process. This will establish the content of the model as well as the inclusion of probability values to enable scenario testing. The initial skeleton BN, derived from the internal pilot, performed well enough with simple validation to suggest that a robust BN may be achievable from implementation of the protocol.

Acknowledgments

This report represents independent research arising from a Predoctoral Clinical Academic Fellowship, Adele Hill (ICA-PCAF-2018-01-137) supported by the National Institute for Health Research and Health Education England. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the Department of Health. This work was supported by the CAP-AI program funded through the European Regional Development Fund and Barts Charity. 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.

Authors' Contributions

AH was the main author of this protocol. CHJ was the author of the data analysis and provided overall technical advice and review. CKJ provided methodology advice. WM supervised the technological aspects of this protocol. BY and CTS have advised on numerous aspects of the probability elicitation and provided peer review of the project. DM is the supervisor of this project and provided quality review.

Conflicts of Interest

None declared.

References

1. O'Sullivan P, Caneiro JP, O'Keeffe M, O'Sullivan K. Unraveling the Complexity of Low Back Pain. J Orthop Sports Phys Ther 2016 Nov;46(11):932-937. [doi: 10.2519/jospt.2016.0609] [Medline: 27802794]

2. Barton CJ, Lack S, Hemmings S, Tufail S, Morrissey D. The 'Best Practice Guide to Conservative Management of Patellofemoral Pain': incorporating level 1 evidence with expert clinical reasoning. Br J Sports Med 2015 Jul 25;49(14):923-934. [doi: 10.1136/bjsports-2014-093637] [Medline: 25716151]
3. Hill JC, Vohora K, Dunn KM, Main CJ, Hay EM. Comparing the STarT Back Screening Tool s Subgroup Allocation of Individual Patients With That of Independent Clinical Experts. The Clinical Journal of Pain 2010;26(9):783-787. [doi: 10.1097/ajp.0b013e3181f18aac]
4. Norman G, Eva K. Diagnostic error and clinical reasoning. Med Educ 2010 Jan;44(1):94-100. [doi: 10.1111/j.1365-2923.2009.03507.x] [Medline: 20078760]
5. Kelsey T, Cavendish W. Personalised Health and Care 2020: Using Data and Technology to Transform Outcomes for Patients and Citizens-A Framework for Action. GOV.UK. 2014. URL: https://www.gov.uk/government/publications/personalised-health-and-care-2020 [accessed 2020-04-30]
6. Boseley S. Boris Johnson pledges £250m for NHS artifical intelligence. The Guardian. 2019. URL: https://www.theguardian.com/society/2019/aug/08/boris-johnson-pledges-250m-for-nhs-artificial-intelligence [accessed 2020-04-30]
7. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial Intelligence and the Future of Primary Care: Exploratory Qualitative Study of UK General Practitioners' Views. J Med Internet Res 2019 Mar 20;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
8. van Hooff ML, van Loon J, van Limbeek J, de Kleuver M. The Nijmegen decision tool for chronic low back pain. Development of a clinical decision tool for secondary or tertiary spine care specialists. PLoS One 2014 Aug 18;9(8):e104226 [FREE Full text] [doi: 10.1371/journal.pone.0104226] [Medline: 25133645]
9. van Hooff ML, van Dongen JM, Coupé VM, Spruit M, Ostelo RWJ, de Kleuver M. Can patient-reported profiles avoid unnecessary referral to a spine surgeon? An observational study to further develop the Nijmegen Decision Tool for Chronic Low Back Pain. PLoS One 2018 Sep 19;13(9):e0203518 [FREE Full text] [doi: 10.1371/journal.pone.0203518] [Medline: 30231051]
10. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Amsterdam, Netherlands: Elsevier; 1988.
11. Yet B, Perkins ZB, Tai NRM, Marsh DWR. Clinical evidence framework for Bayesian networks. Knowl Inf Syst 2016 Mar 22;50(1):117-143. [doi: 10.1007/s10115-016-0932-1]
12. Kyrimi E, Mossadegh S, Tai N, Marsh W. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making. Artif Intell Med 2020 Mar;103:101812. [doi: 10.1016/j.artmed.2020.101812] [Medline: 32143808]
13. Krismer M, van Tulder M, Low Back Pain Group of the BoneJoint Health Strategies for Europe Project. Strategies for prevention and management of musculoskeletal conditions. Low back pain (non-specific). Best Pract Res Clin Rheumatol 2007 Feb;21(1):77-91 [FREE Full text] [doi: 10.1016/j.berh.2006.08.004] [Medline: 17350545]
14. Nunes J, Barbosa M, Silva L, Gorgônio K, Almeida H, Perkusich A. Issues in the Probability Elicitation Process of Expert-Based Bayesian Networks. In: Enhanced Expert Systems. London, UK: IntechOpen; 2018.
15. Werner C, Bedford T, Cooke RM, Hanea AM, Morales-Nápoles O. Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. European Journal of Operational Research 2017 May;258(3):801-819. [doi: 10.1016/j.ejor.2016.10.018]
16. Ferguson FC, Brownlee M, Webster V. A Delphi study investigating consensus among expert physiotherapists in relation to the management of low back pain. Musculoskeletal Care 2008 Dec;6(4):197-210. [doi: 10.1002/msc.126] [Medline: 18311847]
17. Salari H, Ostovar R, Esfandiari A, Keshtkaran A, Akbari Sari A, Yousefi Manesh H, et al. Evidence for Policy Making: Clinical Appropriateness Study of Lumbar Spine MRI Prescriptions Using RAND Appropriateness Method. Int J Health Policy Manag 2013 Jun;1(1):17-21 [FREE Full text] [doi: 10.15171/ijhpm.2013.04] [Medline: 24596832]
18. Jackson A, Hettinga D, Mead J, Mercer C. Using consensus methods in developing clinical guidelines for exercise in managing persistent low back pain. Physiotherapy 2009 Dec;95(4):302-311 [FREE Full text] [doi: 10.1016/j.physio.2009.08.001] [Medline: 19892095]
19. Fitch K, Bernstein S, Aguilar M, Burnand B, LaCalle J, Lázaro P. The RAND/UCLA Appropriateness Method User’s Manual. 2001. URL: https://www.rand.org/content/dam/rand/pubs/monograph_reports/2011/MR1269.pdf [accessed 2020-01-14]
20. Kyrimi E, Neves M, McLachlan S, Neil M, Marsh W, Fenton N. Medical idioms for clinical Bayesian network development. J Biomed Inform 2020 Aug;108:103495 [FREE Full text] [doi: 10.1016/j.jbi.2020.103495] [Medline: 32619692]
21. O'Hagan A, Buck C, Daneshkhah A, Eiser JR, Garthwaite P, Jenkinson D, et al. Uncertain Judgements: Eliciting Experts' Probabilities. Hoboken, NJ: John Wiley & Sons; 2006.
22. Montibeller G, von Winterfeldt D. Individual and Group Biases in Value and Uncertainty Judgments. In: Dias L, Morton A, Quigley J, editors. Elicitation. International Series in Operations Research & Management Science, vol 261. Cham: Springer; 2018:377-392.
23. Srinivas S. A Generalization of the Noisy-Or Model. In: Uncertainty in Artificial Intelligence: Proceedings of the Ninth Conference (1993). Burlington, MA: Morgan Kaufmann; 1993:208-215.

24. Fenton NE, Neil M, Caballero JG. Using Ranked Nodes to Model Qualitative Judgments in Bayesian Networks. IEEE Trans. Knowl. Data Eng 2007 Oct;19(10):1420-1432. [doi: 10.1109/tkde.2007.1073]
25. Røed W, Mosleh A, Vinnem JE, Aven T. On the use of the hybrid causal logic method in offshore risk analysis. Reliability Engineering & System Safety 2009 Feb;94(2):445-455. [doi: 10.1016/j.ress.2008.04.003]
26. Podofillini L, Dang V. A Bayesian approach to treat expert-elicited probabilities in human reliability analysis model construction. Reliability Engineering & System Safety 2013 Sep;117:52-64. [doi: 10.1016/j.ress.2013.03.015]
27. Pitchforth J, Mengersen K. A proposed validation framework for expert elicited Bayesian Networks. Expert Systems with Applications 2013 Jan;40(1):162-167. [doi: 10.1016/j.eswa.2012.07.026]
28. Fenton N, Neil M. Comparing risks of alternative medical diagnosis using Bayesian arguments. J Biomed Inform 2010 Aug;43(4):485-495 [FREE Full text] [doi: 10.1016/j.jbi.2010.02.004] [Medline: 20152931]
29. Premkumar A, Godfrey W, Gottschalk MB, Boden SD. Red Flags for Low Back Pain Are Not Always Really Red. The Journal of Bone and Joint Surgery 2018;100(5):368-374. [doi: 10.2106/jbjs.17.00134]
30. Downie A, Williams CM, Henschke N, Hancock MJ, Ostelo RWJG, de Vet HCW, et al. Red flags to screen for malignancy and fracture in patients with low back pain. Br J Sports Med 2014 Oct;48(20):1518. [doi: 10.1136/bjsports-2014-f7095rep] [Medline: 25257627]
31. Jones M, Rivett D. Clinical Reasoning in Musculoskeletal Practice - E-Book. Amsterdam, Netherlands: Elsevier; 2019.
32. AGENARISK. URL: https://www.agenarisk.com [accessed 2020-04-30]
33. Ritchie J, Lewis J, McNaughton Nicholls C, Ormston R. Qualitative Research Practice: A Guide for Social Science Students and Researchers. Thousand Oaks, CA: SAGE Publications Ltd; 2013.

Abbreviations

AI: artificial intelligence
BN: Bayesian network
IPR: interpercentile range
LBP: low back pain

Edited by G Eysenbach; submitted 07.07.20; peer-reviewed by N Fenton, J Lurie; comments to author 15.08.20; revised version received 01.09.20; accepted 10.11.20; published 15.01.21

Please cite as:
Hill A, Joyner CH, Keith-Jopp C, Yet B, Tuncer Sakar C, Marsh W, Morrissey D
A Bayesian Network Decision Support Tool for Low Back Pain Using a RAND Appropriateness Procedure: Proposal and Internal Pilot Study
JMIR Res Protoc 2021;10(1):e21804
URL: http://www.researchprotocols.org/2021/1/e21804/
doi: 10.2196/21804
PMID: 33448937