Delphi Elicitation for Strategic Risk Assessment

Bayesian Intelligence Technical Report 2016/1

Ann Nicholson, Steven Mascaro, Shreshth Thakur, Kevin Korb
Clayton School of Information Technology, Monash University

Ross Ashman
DSTO

Jan 2016

Abstract

The ability to perform timely and systematic risk assessment is critical to the success of planning for the future, whether for an individual or an organisation. While a variety of techniques currently exist for performing risk assessments, many are of a qualitative and subjective nature, while other more robust techniques are quantitative and consultative, particularly those based on Bayesian networks, but time consuming.

This report describes a semi-automated and online knowledge engineering technique for building Bayesian networks with the cooperation of experts for strategic risk assessment — that is, the assessment of new, emerging or shifting external risks that impact on long term planning. The approach is demonstrated on the case study of tuberculosis. Tuberculosis is not prevalent in Australia but is common in surrounding Asian countries and is therefore a potential threat. By combining the Delphi method of expert elicitation with standard Bayesian network knowledge engineering techniques, we build a simple Bayesian network model of tuberculosis which is robust and which minimises the demand on both modellers and experts. We find that this approach, while still at an early stage of development and lacking refinement, shows a surprising amount of promise.r

Contents

1 Introduction

Strategic risk assessment has been an area of growing interest in many areas, particularly in business and defence (Deloitte, 2013; Frühling, 2007). While no clear definition of strategic risk exists, the term is commonly used to describe any significant, typically external, risk that impacts on long term planning. For defence, this may include the emergence of new threats or new technologies, agents (state or non-state) that develop significantly strengthened capabilities, changes in the allegiances and strategies of others as well as strategic planning failures. For business and industry, this may include changes in the competitive or regulatory environment, changes in the market (e.g., systematic changes in consumer preferences or behaviour), or faulty assumptions underlying strategic decisions.

Several tools to assess risks are available (Kelly and Smith, 2009), however most are tailored to specific problems and extending them to other problem domains is not straightforward. In this study, we propose Bayesian networks (BNs) built through a knowledge engineering process to achieve this aim. Bayesian networks are not problem specific and, hence, this process can be extended to any problem associated with risk. Here, we focus on building a BN for the latent tuberculosis (TB) problem with the assistance of domain experts.

Although Australia is a low TB prevalence area compared to global standards, TB remains an issue to be addressed given increased travel and immigration. Countries close to Australia in the Asia Pacific region have high rates of TB. Therefore, TB management remains paramount with a focus on limiting risk by detecting and treating TB early in latent form, therefore, minimising TB transmission (VicHealth, 2014).

Our simple BN for the TB problem was created using the Delphi elicitation method (Hsu and Sandford, 2007; Linstone and Turoff, 1975). The elicitation was carried out with two pieces of software: one for the structure, developed especially for this project; and another for the parameters using a software package called Bayesian Delphi, developed by one of the authors for an earlier project (Wintle et al., 2013). The elicitation process involved the use of a facilitator who managed the online consultation with tuberculosis domain experts about their preferences regarding the structure and parameters of the BN.[1] The resulting BNs from all the experts were analysed collectively and a combined BN was provided back to the experts to reconsider their original decisions. The experts then provided revised BNs, which were once again built into a final combined BN. The analysis and evaluation below makes use of this final BN, as well as data recorded for the experts' actions during the elicitation.

We begin with a review of the relevant literature, including elicitation and model-building techniques used in strategic risk assessment, qualitative modelling, and knowledge engineering in BNs. This is followed by an outline of the methodology, which explains the elicitation process. Following this, details of the online tools used to survey the experts are provided. The section on trial elicitation discusses how the elicitation on latent TB was carried out and each phase of the elicitation process is reported in this section. The following section contains the analysis of the results of the elicitation. The paper concludes with an evaluation of the network produced and the elicitation process.

2 Literature Review

2.1 Strategic Risk Assessment

Qualitative risk assessment has been the dominant trope in strategic risk assessment. However, risk is classically the probability-weighted potential costs put into play by some action –

[1]The facilitator role was shared across some of the authors. All communication with the experts was conducted via email or the online software described above.

i.e., it concerns the expected values of available actions. It seems counter-intuitive to expect qualitative methods, which by definition ignore probabilities, to do justice to the concept of risk assessment. In this paper we review a variety of elements widely used in strategic risk assessment, including expert elicitation methods and the techniques of general morphological analysis, and consider how they may be applied in the context of building and using Bayesian networks for strategic risk assessment. In the end, we will have a methodology for incorporating the insights of qualitative strategic risk assessment in a process that includes quantitative assessment, thereby supporting a richer framework for their application.

In futures studies the "scenario" has been described as the archetypal product (Bishop et al., 2007). There are many scenario development techniques used in future studies such as Delphi, General Morphology Analysis, Trend Analysis and Back-casting, among others. Bishop et al. (2007) list twenty three scenario techniques and their attributes. Of these only three (option development & evaluation, general morphological analysis and sensitivity analysis) use computers to carry them out. Surprisingly, the quantitative technique probability trees was described as only optionally using computers; Bishop et al. make the point that there is an opportunity to make greater use of software in crafting scenarios. We will explore the integral use of computers in building Bayesian risk assessment models, including the use of dynamic Bayesian Networks to explicitly represent dynamic processes.

2.2 Qualitative Modeling

2.2.1 General Morphological Analysis

General Morphological Analysis is a general method for qualitative modeling that attempts to investigate relationships in multi-dimensional problems that seem to defy quantification. The aim of GMA is to identify and investigate the total set of possible relationships contained in a given complex problem (Ritchey, 2002).

The method attempts to identify parameters of a problem by assigning each parameter a range of appropriate values or conditions. A morphological box, also known as a Zwicky box, is then formed by setting each parameter against the others in an n-dimensional matrix where each cell contains a particular value or condition and thus identifies a particular state of the problem.

Fritz Zwicky pioneered development of GMA in the 60s (Zwicky and Zwicky, 1969) and in the past two decades GMA has been computerised and extended, and is now applied to areas such as developing scenario and strategy laboratories, complex policy and planning issues, and analyzing organizational and stakeholder structures (Ritchey, 2002).

Computer support for GMA was added by the Swedish Defence Research Agency in 1995, in the form of Windows software known as MA/Carma. The authors claim the software significantly extends GMA's functionality and areas of application and adds interactive, non-quantitative inference models (Ritchey, 2002).

2.2.2 Qualitative Bayesian Models

At least as far back as the paper by Helmer and Rescher (1959), suggested to have influenced the rise of Delphi (Linstone and Turoff, 1975), the relationship between the exact and inexact sciences has been explored in the context of integrating qualitative and quantitative modelling.

Currently, techniques used for futures or scenario analysis such as General Morphological Analysis or Field Anomaly Relaxation produce futures that are static representations of a possible state or set of states; all possibilities are treated as equal and invariant. Bayesian networks provide a mechanism through Bayes theorem to investigate change that may be incorporated into causal models of scenarios or future states. These can also potentially be dynamic, in

either the sense of temporal evolution or user interaction or both. We aim to provide some guidance on how Bayesian methods could be utilised for dynamic scenario modelling of futures for estimating strategic risk.

Qualitative probabilistic networks (QPNs) have been around since at least Wellman (1990). Wellman describes Bayesian networks without probability parameters and instead with arcs labeled as positive or negative, denoting a monotonic relation between nodes that is increasing or decreasing respectively. The QPN representation is vastly simpler to elicit than a full Bayesian network, having only one binary parameter per arc and not allowing for any interactive effects. QPNs are equally abstractions from BNs and from the linear path models of Wright (1934) and provide a straightforward vehicle for qualitative Bayesian modeling. Lucas (2005) describes a version of qualitative probabilistic networks supporting specific patterns of causal interactions, leading to networks of complexity intermediate between the original QPNs and BNs.

Bashari et al. (2009) describe state transition models (STMs), providing a simple and versatile means for developing dynamic models. Bashari et al. point out that because they are purely descriptive diagrams, they have limited decision support and learning capability. They can, however, be used to visualize or investigate the impact of drivers of change on the likelihood of a transition to occur. Bashari et al. then demonstrate an approach that combines a state and transition model with a Bayesian network to provide a relatively simple and updatable dynamics model that can accommodate uncertainty and be used for scenario, diagnostic, and sensitivity analysis. They develop an STM and extend it into a BN with state transitions and factors influencing each transition, and in the process provide a general framework for facilitating the transition from STMs to BNs.

Nicholson and Flores (2011) propose a combination of STMs and DBNs that overcome some of the limitations of the approach by Bashari et al., including provision of an explicit representation of the next state, while retaining its advantages, such as the explicit representation of transitions.

2.3 Quantitative Modeling

In "Should Probabilities be used with Scenarios", Stephen Millet states:

> Very few scenario analysts have employed either cross-impact analysis or probabilities to generate scenarios as forecasts of alternative futures in comparison with the dominating intuitive scenario writing approach. The practitioners of intuitive scenarios have been largely successful in convincing subsequent generations of scenario planners that probabilities cannot be used with scenarios yet the argument has been hardly closed. (Millett, 2009)

He summarizes arguments commonly raised as objections against the use of probabilities, which include:

1. Scenarios should be used for identifying possible and preferred futures, not likely futures, an inherent danger of using probabilities

2. The use of probabilities implies too much precision and distracts from the storytelling qualities of scenarios

3. Forecasts may capture trends, but they cannot capture the discontinuities of change that come from intuition, imagination, and the story qualities of scenarios

4. Teams can achieve consensus on plausible scenarios but rarely can reach agreement on probabilities of occurrence

He rebuts these objections in turn, and concludes that probabilities should be used when there is sufficient time and resources, the scenario team is comfortable with Bayesian statistics, the corporate culture values quantitative methods and is sceptical of purely qualitative ones, and the managers embrace the use of probabilities and appreciates their strengths and weaknesses.

2.3.1 Bayesian Networks

Bayesian Networks (BNs) are directed acyclic graphs (DAGs), consisting of nodes (variables) and arcs (arrows) that represent system variables and their causes and effects. They are an amalgamation of qualitative and quantitative components, the DAGs representing qualitative relations within the system and conditional probability tables quantifying the dependencies between variables in the graph (Pearl, 1988; Korb and Nicholson, 2010).

BNs are a good candidate for modelling dynamic futures. Following Wooldridge (2003), some of their advantages for this are:

1. BNs are particularly useful for making probabilistic inference about model domains that are characterized by inherent complexity and uncertainty.

2. Due to their Bayesian formalism, BNs provide a rational technique to combine both subjective (e.g., expert opinion) and objective (e.g., measurement data) information. The flexible nature of BNs also means that new information can easily be incorporated as it becomes available.

3. BNs are helpful for challenging experts to articulate what they know about the model domain and to knit those influences into dependency networks. The graphical (visual) nature of BNs facilitates the easy transfer of understanding about key linkages.

4. Given their network structure, BNs successfully capture the notion of modularity, i.e., a complex system is built by combining simpler parts. You can start them off small, with limited knowledge about a domain and grow them (add additional variables) as you acquire new knowledge.

5. BNs facilitate informed decision-making in the face of incomplete and imperfect understanding.

Baran and Jantunen (2004) reiterate Wooldridge's view while adding that a BN approach provides a framework for effective dialogue between stakeholders as well as a learning tool for understanding the consequences of decisions. Kragt (2009) also points out that BNs can be useful decision support tools, as they allow an assessment of the relative changes in outcome probabilities that are associated with changes in management actions or system parameters – i.e., they are good for sensitivity analysis.

Regular objections to BNs include:

1. Because BNs are acyclic they cannot contain cycles or feedback loops, unlike (for example) State Transition Models (STMs) (Bashari et al., 2009).

2. When eliciting probabilities, probability tables that result from the combination of several driving variables should be as manageable as possible. Therefore, the [number of parent] variables must be as few as possible, usually no more than four (Baran and Jantunen, 2004).

3. BNs assume a simple attribute-value representation, that is, each problem instance involves reasoning about the same fixed number of attributes, with only the evidence values changing from problem instance to problem instance (Costa and Kathryn, 2006).

4. While Bayesian models are a useful way to model expert knowledge, it may be difficult to get experts to agree on the structure of the model and on which nodes are important to include (Kragt, 2009).

5. A Bayesian network is only as useful as its prior knowledge is reliable. Selecting an inappropriate set of distributions to describe the data has a notable effect on the quality of the resulting network.

There are good rebuttals to these complaints, however:

1. Dynamic Bayesian networks (DBNs) can directly reflect cycles or feedback loops, by treating them as generating sequences of time slices. To be sure, the corresponding DBNs are more complicated than STMs, but they can also represent far richer causal processes, including arbitrary interactions and non-stationary processes.

2. While probability tables can require many parameters, and removing less important variables is usually helpful, many practical, elicited and deployed BNs regularly contain variables with more than 4 parents.

3. While settling on a fixed BN for a problem is common, there is no necessity in it. Structures and parameters can be generated by more general rules (as is the case with DBNs).

4. Getting agreement on a model is difficult with or without a BN — BNs just make the disagreements explicit. Methods for producing a combined understanding (without necessarily resolving disagreements), such as the Delphi process, can help.

5. Bayesian net modeling does not stop after birthing a single model. Data, additional elicitation and expert validation in any combination should always be applied to nurture, strengthen and grow the model, iteratively.

The growing sophistication of Bayesian computational methods has led to a dramatic increase in the breadth and complexity of Bayesian applications (Garthwaite et al., 2005). Given the ability to incorporate expert opinion with historical and other quantitative data, it may be possible to use BNs for areas which largely rely on expert opinion, such as futures and scenario planning.

2.4 Strategic Risk Assessment with BNs

For exploring scenarios, in particular dynamic ones, Bayesian networks have a number of attributes that make them highly desirable. They are able to deal with uncertainty, incomplete knowledge and subjective beliefs. They bring together quantitative and qualitative information into a unified model. They can be modified or updated as data are acquired or expert opinions change. BNs that are explicitly causal can be used to explore possible interventions or management practices, while dynamic BNs can be used to explore feedback scenarios.

Often, a causal map is created to better understand a domain or process. It is possible to derive BNs from such causal maps, but four major modelling issues have to be considered: conditional independencies, reasoning underlying the link between concepts, distinction between direct and indirect relations, and eliminating circular relations (Table 1).

Cinar and Kayakutlu (2010) created scenarios for energy policies using causal BNs, in order to take advantage of their ability to cope with missing values and their combination of probabilistic and causal semantics. They used structured methods described by Nadkarni and Shenoy (2004) for transforming causal maps to BNs. These include structured interviews to

Property	Causal Map	Bayesian Network
Conditional Independence	No	Yes
Inductive/Deductive Logic	Both	Deductive
Direct/Indirect Relations	Both	Direct
Loops	Yes	No

Table 1: Properties of causal maps and Bayesian Networks

elicit adjacency matrices. Once constructed, the BN was parameterized with data and three scenarios were investigated by varying certain input parameters, such as investing in different energy types, and observing the effects on greenhouse emissions and energy imports. They concluded that BNs are effective for complex strategic planning.

In 2012 a comprehensive review of methods used for modeling ecosystems was written by Andrea White (White, 2012). White covered such diverse methods as causal maps, fuzzy cognitive maps, STMs and Bayesian networks, documenting the strengths and weaknesses of each modeling method for use in the management of Victorian parks. White concluded that BNs are the best tool for capturing interactions, while being simple enough for operational use and communicating to stakeholders. On the other hand, BNs were deemed less efficient in terms of time and other resources needed to construct the model. Much of this time was taken up by parameterising the model. It was suggested that BN's be reserved for specific management issues, where the management problem is complex, there are diverse understandings of causality and the impacts of intervention, and there is a need to develop a common understanding amongst stakeholders (White, 2012). White also states that "BNs allow the clear articulation of threats, hence decisions to be made by a management team can be focused, the analytical rationale for management options defendable, and the protocol for monitoring success and failures explicitly established."

2.5 Building Models with Expert Elicitation

Expert elicitation is a common technique for building models, whether qualitative or quantitative. Expert elicitation involves several challenges. There is evidence that overconfident political experts perform only marginally better than random chance when predicting the future (Tetlock, 2005). There are several other well studied cognitive biases that can adversely influence the results (Tversky, 1974).

Many methods for eliciting and aggregating expert opinion attempt to overcome the previously noted difficulties. Groups can, under certain circumstances, provide better predictions than the average of their individual judgements (Surowiecki, 2004) and this effect appears to be even greater in quantitative judgement (Schultze et al., 2012). Clemen and Winkler (1999) classify the elicitation and aggregation processes of expert assessments into mathematical and behavioral approaches, and Ouchi and Bank (2004) discuss the advantages and disadvantages of various methods within each class.

Wintle et al. (2013) treat group interaction, diversity and improving the judgments of groups by implementing a Delphi process, concluding that group interaction can improve forecasts. As Delphi methods are popular and well established, and the basis of our own approach here, we will now review them.

2.5.1 Delphi Methods for Elicitation

An Air Force-sponsored Rand Corporation study titled 'Project Delphi', started in the early 1950s, explored the use of expert opinion and obtaining consensus (Linstone and Turoff, 1975). Expert opinion had previously been used frequently in forecasting, but when experts were

consulted in groups there were significant weaknesses, largely due to psychological factors such as the presence of dominant personalities, the desire for peer approval and an unwillingness to change opinions that had been publicly expressed (Brown, 1968).

The Delphi method mitigates these factors. Instead of direct confrontation and debate between experts, communications are routed through a moderator, typically called the facilitator. The facilitator conducts rounds of individual interrogations sequentially, interspersed with feedback derived from other group members. All members of the group are invited to give reasons for their expressed opinions and these reasons are available for critique by the rest of the group while maintaining anonymity (Brown, 1968). Each expert is given the chance to revise his/her opinion based on the others' reviews and the process is repeated, typically for several rounds until a reduced spread of opinions is achieved (Ouchi and Bank, 2004).

The justification for Delphi is the same now as when it was originally created: it is a useful technique for maximizing the knowledge sharing and understanding of experts, while minimizing the influence of personalities (Linstone and Turoff, 1975; Sahal and Yee, 1975).

The following features are seen as distinguishing Delphi methods from other techniques (Rowe and Wright, 2001):

1. Anonymity, eliminating much of the bias due to social psychology.
2. Feedback, allowing experts to defend their judgments and to respond to their peers' assessments and arguments.
3. Iteration, allowing for the operation of group judgment. There are rarely more than one or two iterations, i.e., two or three rounds in total.
4. Statistical aggregation of results.

Ideal numbers for Delphi Clearly, for any given task there must be an optimal number of participants, but it is unclear what that number might be in general. MacMillan and Marshall (2006) state that a group of ten individuals is considered appropriate for a Delphi expert panel, quoting Crance (1987). Hodgetts (1977) suggests an ideal number of eight participants, but neither Crance nor Hodgetts provide justification. With modern web based implementations, Delphi has been successfully used with hundreds of participants (Hejblum et al., 2001).

2.6 Knowledge Engineering

Knowledge engineering with BNs refers to building reliable Bayesian networks (Korb and Nicholson, 2010). The process is inspired by the software development life cycle from software engineering and is outlined as follows:

- *Build the BN*: includes determining the structure, eliciting the parameters and possibly determining utilities (in case of decision networks). This is usually done with the help of experts or via data mining techniques.
- *Validate*: includes accuracy testing and sensitivity analysis. The network needs to be evaluated for whether or not the network is appropriate for the problem at hand. This ties in closely with the accuracy of the network. For example, testing a query node to determine whether or not the predictions satisfy the experts' intuitions. Sensitivity analysis is about analysing how sensitive the network is to changes in parameters and inputs. This information can be used to improve the network through further research or more robust modeling.

- *Test*: BNs, like any complex software, require extensive testing. There are three levels of testing commonly used. Firstly, *Alpha* testing is usually carried out by experts and developers who were not directly involved in the development of the BN. This is followed by external users (potentially experts not involved in building the BN) in the *Beta* testing phase to iron out defects. Finally, *Acceptance* testing is carried out by the end-users.

- *Apply*: once thoroughly tested, the BNs can be put to use in practice. This potentially involves refining the network with experience gained from a practical point of view.

- *Refine*: while a model can be sufficient, no model is ever 'final'. Further improvements to the BN should be incorporated and tested as new information comes to hand, uses change, and new opportunities for testing and further development arise.

A large part of the effort, so far, has focused on building the BN and in particular the set of techniques which combine expert knowledge and automated methods. The subsequent steps in this process are rarely followed in a systematic manner but Korb and Nicholson (2010) emphasise and demonstrate the need to move in this direction.

2.7 Summary

Potential advantages of developing a BN model capable of dynamic futures or scenario prediction are that a wide range of scenarios can be explored in a variety of ways based just on current understanding. BNs can be modified or updated as time goes by and as certain events change, these can be fed back into the model to reassess the likelihood of particular scenarios materialising.

For exploring scenarios, particularly dynamic ones, BNs have a number of attributes that make them a preferred approach. They are able to deal with uncertainty, incomplete knowledge and subjective data. They bring together quantitative and qualitative information, as well as data and expert and stakeholder knowledge into a single entity.

BNs also have potential weaknesses, such as not dealing easily with feedback cycles and requiring significant effort in populating the graphs with probabilities. However, these drawbacks can be easily managed with improved software tools, and there seems to be a growing interest in applying Bayesian networks to scenario modelling for decision making. This interest will no doubt lead to a greater understanding of how to circumvent real and perceived limitations in practice.

3 The Delphi BN Elicitation Method

We propose an online process for constructing a BN that involves the structured elicitation of knowledge from a group of experts. The process can be used to build a BN for strategic risk assessment or for any other domain. It assumes that a set of variables relevant to the domain has already been identified, and proceeds with an online Delphi-based procedure for aggregating expert knowledge to create the structure and parameters that qualitatively and quantitatively tie these variables together. There are four main groups of people involved in the construction of the BN:

1. Experts

2. Facilitators

3. Modellers

4. Consumers

The experts are those with primary knowledge of the domain. They need not be experts in any strict sense, but must be capable of contributing domain knowledge to the process in some way (perhaps through further research). The facilitators are responsible for communicating with the experts about what needs to be done, encouraging discussion and presenting results and analysis back to the experts. The modellers are responsible for constructing the BN based on the expert responses received. Finally, there are the model consumers, who commission the model, have final say on scope and purpose and ultimately are the ones that will use the model. We won't discuss the role of consumers below, however, in general, modellers should work closely with consumers to ensure the model fulfils their needs.

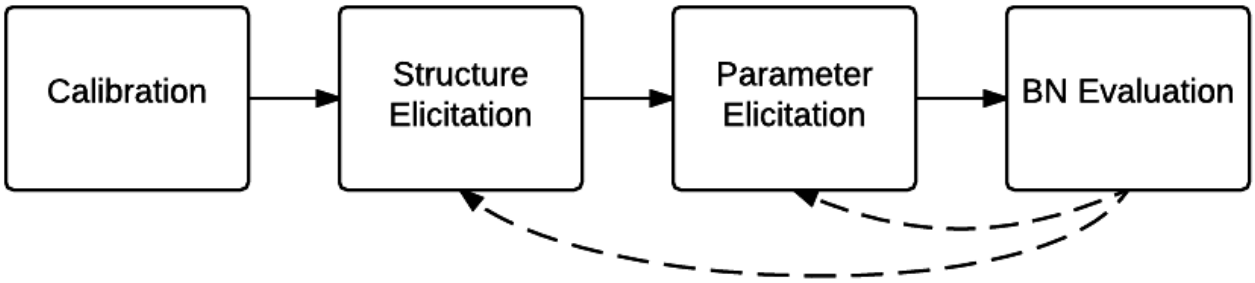


Figure 1: Overview of the elicitation process

The construction process itself consists of four major phases, shown in Figure 1. The purpose of the calibration phase is to ensure that experts share a common language and understanding of the elicitation process with both facilitators and modellers. After calibration, experts are presented with the real problem and, in the structure elicitation phase, are asked to answer questions around the causal structure of the problem. The third phase sees experts provide the values for the structural parameters defined in the previous phase. In the final phase, the modellers validate the model against known outputs and employ experts to assess the model as a whole. The evaluation may lead to the need for iteration, with the structure or parameter elicitation being revisited. Each of these phases is described in more detail below.

3.1 Phase 1: Calibration

The calibration phase of the elicitation process is used to develop a common language between the experts, facilitators and modellers. It is key to ensuring that the experts understand how the elicitation process works, and to ensure that they are familiar with the terminology and concepts that will be needed in the later stages of the elicitation. The calibration phase can deal with the terms and concepts involved in the structural elicitation phase, the parameter elicitation phase or both dependent on the backgrounds and needs of the experts. It is important to perform at least some form of calibration with experts, since it also performs the function of familiarising experts with the software.

The modellers should choose or construct a model that is simple, intuitive and for which an answer is fully (or mostly) known by the modellers but not (immediately) known by the experts. The model should be simple and intuitive enough that almost anyone (regardless of expertise) should be capable of correctly answering questions about the model. (A generic example is described in Section 4.2). A description of the elements (variables) of the model should be given together with a series of questions around what causal relationships exist between the described elements. Calibration then follows a simple procedure:

1. Modellers and facilitators choose calibration questions from a known simple network.
2. Facilitators email instructions to experts.
3. Experts answer questions, preferably with immediate (automated) feedback provided on right or wrong answers.
4. Facilitators discuss any misunderstandings with experts via email if necessary.

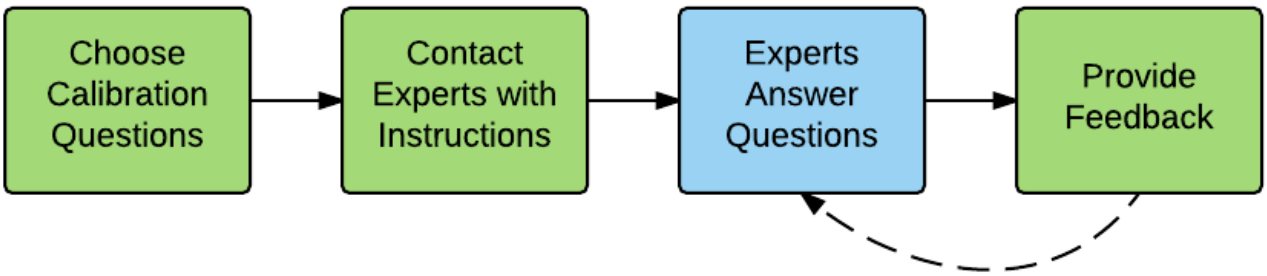


Figure 2: Process for calibration phase

3.2 Phase 2: Structure Elicitation

Once calibration is complete, the process can move on to the real domain, beginning with the building of the structure. The very first step is to choose the variables that will be a part of the process.

3.2.1 Variable Selection

The elicitation process described here does not specify how variables are identified and selected. Nonetheless, there are some general principles to keep in mind when selecting variables for the elicitation.

It is good practice to start with a target variable (or a set of target variables) that one wants to learn about. In a strategic risk context, an example of a target variable may be "Technology X is developed in 30 years time". It is important that the target variables be clearly defined. Of course, the states of the variables should at least be mutually exclusive and exhaustive. But in addition each state should be defined well enough that the corresponding state in the real world can easily be identified.

Further variables should be chosen on the basis of their relevance and importance to the target variables. They need not all be causes, nor directly linked to the target variables. The key point is that modellers ought to have good reason to believe that the variables provide relevant information to the problem.

Since simplicity in BN models is a virtue, variables should be chosen carefully, with each additional variable being well justified for inclusion. This is especially true given the resource-intensive nature of structure and parameter elicitation. It is possible that some variables are connected in straightforward and intuitive ways that don't require consulting experts — these can be included in the model directly by the modellers, but should be omitted from the elicitation process.

After variable selection is complete, we can move to a structured process to elicit the remainder of the network.

3.2.2 Causal Structure

After variable selection, we need to determine the causal structure of the network — that is, to identify which variables *directly influence* which others. To do so, we create a survey for our experts and then aggregate the results. To automate the construction of this survey, we use the fact that any graph can be encoded as a two dimensional matrix, where each entry in the matrix corresponds to an arc in the graph. Since a BN is also a (directed acyclic) graph, it too can be encoded as a matrix. This is shown visually in Figure 3, whereby the directed relationships between variables are encoded as entries in a two dimensional $n \times n$ matrix, where n is the number of variables in the network. The structure of a causal BN, as understood by a single expert, can be derived by iterating over every cell in the matrix, and asking: Does X influence Y?

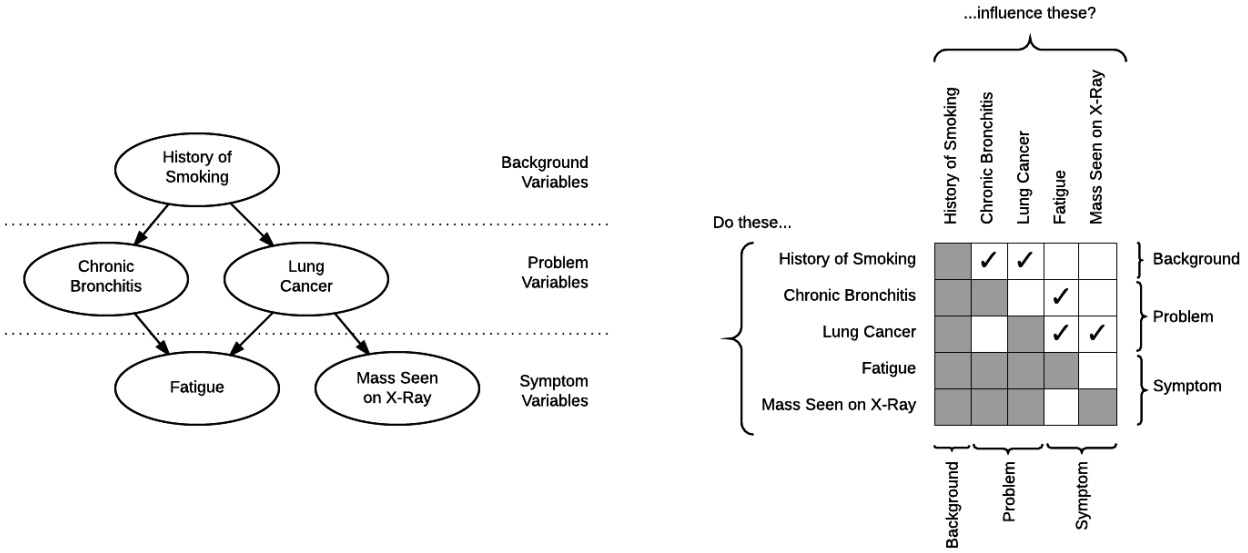


Figure 3: An example of placing variables into tiers to reduce the number of structure questions. Without tiers, this network would require 20 possible questions, while with tiers, it requires 12

One of the most immediate problems that arises with this method is the large number of questions that need to be answered. To reduce the number of questions, we can segment variables into temporal tiers. For example, in a health context, disease variables can influence symptom variables, but symptom variables rarely influence disease variables. The separation of variables in this way reduces the number of possible relationships in the matrix, thus reducing the number of questions required by the survey, sometimes dramatically.

Depending on the specific questions that we ask the experts, there are several ways in which the expert responses about causal structure can be combined. For instance, one possibility may be to ask the question, Does X influence Y?, for each variable, and allow experts to answer Yes, No or Don't Know. We can then sum just the positive Yes responses, or much better, we can counterbalance the Yes and No responses, counting +1 for each positive response and -1 for each negative response. This latter method is described by Serwylo (2016). In addition, we could ask experts for the perceived *strength* of the influence (if present) and their own confidence in their response. Properly calibrated, this would allow for a useful weighted mixture of the expert responses.

3.2.3 Adding the Delphi Protocol

We embed the process above within a Delphi protocol. In particular, we iterate the process for two rounds. In the first round, experts complete their surveys in *isolation* from other experts. The expert responses for this round are anonymised and combined by the facilitator in some way and presented back to the experts for discussion in the second round. If the experts are working directly with the BN structure, the aggregated BN itself can also be presented to the experts, though this is not essential, and in some cases may be undesirable. The facilitator should, at this point, highlight anything interesting or controversial in the first round results, and direct experts to give closer attention to those less certain or more controversial aspects of the structure. In the second round, experts should discuss each question in the survey with each other and revise their responses in the light of information provided by other experts.

3.2.4 Process for Structure Elicitation

1. Modellers prepare by selecting variables and tiers and creating N questions.
2. A new round is opened, and M experts are invited to answer questions (and provide comments), but in isolation from other experts.
3. Experts answer the questions. They can change their answers and comments as many times as they wish before the round is over.
4. Facilitator collects MxN answers and provides experts with feedback. This may or may not include a summary BN structure.
5. Facilitator repeats the process from Step 2 to 4, but this time allowing the experts to see each other's answers and comments.
6. The modellers create a single summary BN structure, using some automated method of aggregation (such as Serwylo's counting method). This model may then be subject to further manual refinement.

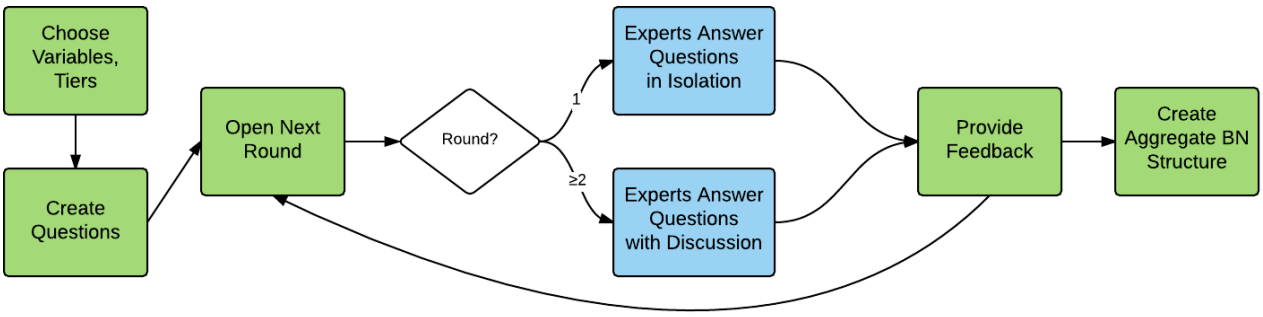


Figure 4: Process for structure elicitation phase

3.3 Phase 3: Parameter Elicitation

Once the structure elicitation is complete, the process moves on to parameter elicitation.

3.3.1 Parameter Selection

For almost any realistic BN, there will be far too many parameters to elicit from experts. Therefore, the first step is to select the most important parameters to be elicited, with the remaining parameters being filled in by the modellers. There are several ways to do this. In some cases, the structure may indicate that some parameters will not (greatly) affect the target variables of interest. These can be dropped from the elicitation, and approximated instead (if required at all). In other cases, modellers may have enough knowledge of the domain to perform an initial parameterisation of the network. From this, a sensitivity analysis can be performed to further screen out variables and identify influential ones. This may work in a wide range of cases, even when the modellers' knowledge is quite limited.

In most cases, the key to a substantial reduction in the number of parameters is to recognise that each node's CPT contains local structure. Just as a joint distribution can usually be described more compactly by a BN structure, so too can a node's CPT usually be described more compactly by a local structure. A node CPT can often be described exactly or approximately by decision trees, logit models, interpolated parameters or equations. Of course, each of these have their own parameters — but there are often many fewer of these than there are parameters in the CPT. As such, identifying the local structure of a node can dramatically reduce the number of parameters that need to be elicited.

In some cases, the type of local structure is not obvious. When this is true, the nature of the local structure should itself be elicited from experts.[2] In other cases, the modellers can determine the local structure, and then only elicit the parameters of that structure from experts. In the simplest case, this involves identifying key rows in the CPT from which other parameters can be interpolated or inferred. For example, commonly, there is an ordering over the states of each node, and the effect on the child is monotonic. In such cases, one can identify the parent combinations that cause the minimum and maximum node values and interpolate other parameters that sit in between.

3.3.2 Elicitation

Once the parameters have been chosen, a list of survey questions is drawn up based on those parameters, and the facilitator then presents experts with the survey. As was the case for structure elicitation, this survey is embedded in a Delphi protocol. In the first round, experts are asked for their estimates in isolation. The results are collected by the facilitator, and then a summary is sent back to the experts. In the second round, experts discuss the results from the first round and revise their estimates in the light of new information presented by other experts.

3.3.3 Process for Parameter Elicitation

1. Modellers select N key parameters from the previously elicited BN structure. This may be based on selecting parameters from node local structures instead of the underlying CPTs, or otherwise on identifying which parameters are the most important.

2. A new round is opened, and M experts are invited to answer questions (and provide comments), but in isolation from other experts.

3. Experts answer the questions. They can change their answers and comments as many times as they wish before the round is over.

[2] Indeed, this ought to be a key part of any future work to automate the elicitation process.

4. Facilitator collects MxN answers and provides experts with feedback. This may or may not include an initial parameterised BN.

5. Facilitator repeats the process from Step 2 to 4, but this time allowing the experts to see each other's answers and comments.

6. Modellers collects the average responses for the parameters, and uses these to parameterise the elicited network. This model may then be subject to further manual refinement.

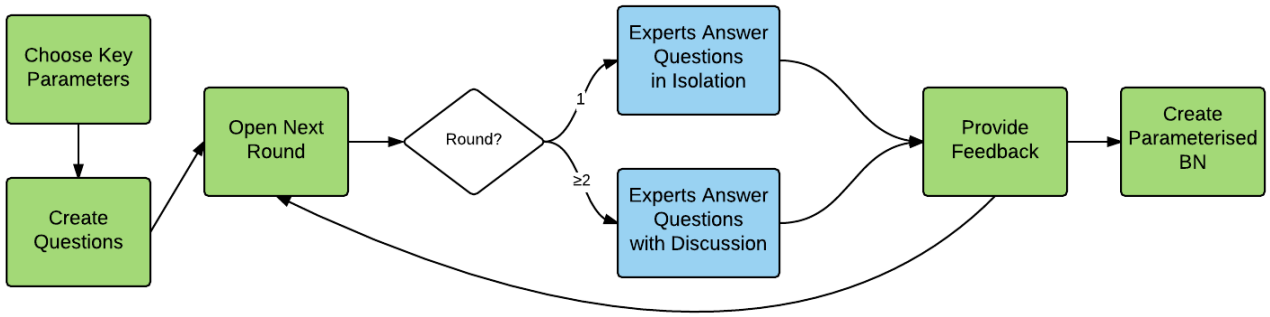


Figure 5: Process for parameter elicitation phase

3.4 Phase 4 - Evaluation

The key aim of the evaluation phase is to ensure that the network behaviour is appropriate. This can be done by evaluating the model's predictions (or conclusions) on data or known cases, or by having experts check the model's conclusions. Given the process here involves experts, it seems natural to validate the model with experts as well. It is also possible that the experts involved in the elicitation do not see the final model until (an iteration of) the elicitation is complete. Hence, it also makes perfect sense to have these very same experts assess model conclusions.

One way to do this is to develop a set of scenarios that covers a range of possibilities, including common scenarios, extreme scenarios (e.g. when the target variables are expected to be at their most extreme values or probabilities), rare scenarios and random scenarios. Experts can then be provided with a survey with these scenarios, and asked to predict the outcomes. Entering these scenarios into the elicited network will yield a prediction that can then be checked against the expert's evaluations, either with simple measures such as predictive accuracy, or preferably measures that take into account probability assessments, such as log scores. Indeed, the very same approach can be used with data that has been derived from measurements.

Of course, Phase 4 is not the end of the process. Evaluation will almost certainly provide much more information about the accuracy and utility of the model. This may lead to modellers refining the elicited model, or perhaps another round of elicitation. Indeed this kind of iteration can occur at any stage (after variable selection, structure elicitation, etc.), because new information can be uncovered at any stage that makes revisiting an earlier stage worthwhile.

While the automated process described here does reduce the expense of the elicitation (in terms of time and resources), the expense is still much more significant than for many other forms of modelling. Due to this expense, it is often better for the modellers to work on the model as much as possible themselves, before performing another iteration with experts. Increasing the amount of automation involved in the elicitation process, along with reducing the burden on experts may change this equation in future.

4 Applying the Method: Bayesian Delphi Elicitation for Tuberculosis Management

To trial the Delphi BN Elicitation process, we wanted a simple case study within a domain for which we had ready access to both domain knowledge and experts. At the beginning of this project, one of the authors of this report was in the process of developing a model for tuberculosis (TB) management using manual elicitation.[3] Through this work, we were able to secure the involvement of 8 health experts who had good knowledge of the epidemiology of tuberculosis. Of these, 7 experts came from Royal Melbourne Hospital and 1 from Monash University. Communication with all experts was conducted entirely online, via email. However, one of the experts also helped us with the development of the problem and surveys, which was done via both email and phone conversations.

4.1 Case Study

Our case study focused on the same problem as the manually elicited TB management model: managing the spread of TB due to immigration. In particular, the intention is to create a model to help decide how testing for TB should be done for those migrating to Australia. TB is not prevalent in Australia, but is in surrounding regions and hence poses a potential threat if the risk is not well managed. The case study was designed to contain dynamic elements (though does not contain all the variables needed for a full DBN), requiring experts to project forward into the future for each migrant arrival to assess their probability of manifesting TB. These assessments would be based on the migrant's background and history, as well as the result of a blood test, if available.

The manually elicited TB management model was mostly completed by the time of our first elicitation experiments. This model (with no evidence entered) can be seen in Figure 18. It captures the relationship between a migrant's probability of having TB now and their probability of having TB in the future, mediated by many different background factors. In addition, various tests can be conducted to shed more light on a migrant's current TB status.

[3] In the following, this pre-existing model will be called the *manually elicited model*, in contrast to the Delphi elicited model that is the product of the current trial.

Jack's wet lawn conundrum. Jack's suburban house has a beautiful front yard with a lush green garden but it lacks a clear pathway to the front door. One must walk through the grass to approach the house. Thus, Jack is often worried about **Overnight Rain** and his **Sprinkler**. Since, either of them going all night would result in **Wet Grass**. Jack hates the grass getting too wet because then there will be **Dog Tracks** in the house. The wet grass also makes his **Guest Slip** in the mud. Jack would like to assess the probability of his guests slipping before he makes any changes to his front yard set up.

It is obvious to Jack that **Rain** directly influences **Wet Grass**. He wants to figure out the direct influences between the remaining variables.

Figure 6: The calibration scenario as it was put to the experts

4.2 Calibration

To perform calibration (that is, to ensure that experts, facilitators and modellers are all speaking a common language), we created a toy problem that nonetheless exhibited the range of common relationships possible within a BN. A virtue of any problem used for calibration is that it does *not* require expert knowledge of any sort; that way, all lessons learned relate solely to the language used and not to any particulars of the problem. Hence, we settled on a very simple problem: wet grass. The scenario, as it was put to expert participants, can be found in Figure 6.

Of course, prior to developing the scenario text, we had also created our solution (Figure 7) to ensure that it captured the relationships that we needed. The model contains the three main types of graphical node relationships: causal chain (e.g. Rain → Wet Grass → Guest Slips), common cause/ancestor (e.g. Wet Grass → Dog Tracks, Guest Slips) and common effect/descendent (e.g. Rain, Sprinkler → Wet Grass). In addition, we created an ambiguity in the problem description, such that, based on one's interpretation, one may or may not draw an arc between Rain and Sprinkler.

The facilitator emailed experts with instructions on how to access and complete the calibration survey, which was in the form of an online survey. After signing into the site, experts were presented with the scenario followed by a list of questions (Figure 9). Given the simplicity of the survey, a short deadline of two days was given. The survey was well received, with experts commonly answering all questions correctly. The experts were taken to a model answers page immediately after submitting their responses (Figure 8). This page was used to further improve the experts' understanding of causality and influence. Additional feedback was provided by the facilitator to all experts (as a group) in an email follow-up.

4.3 Structure Elicitation

4.3.1 Variable Selection

To begin the structure elicitation of the model proper, a list of variables was created in cooperation with one of the experts. Since our aim with this case study was simplicity (while still capturing an interesting set of relationships), these variables corresponded to a subset of the variables from the manually elicited model. The variables chosen included the target variable (Future TB) along with its present day correlate (Current TB), background factors (Age, Region of Origin and Relative with Active TB), variables associated with testing TB (Do Blood Test and Blood Test Result) and also the decision to perform treatment for latent TB (Treat Latent TB). (It is assumed that the unlikely case of a migrant with active TB would always be treated.)

In the manually elicited model, both `Do Blood Test` and `Treat Latent TB` are decision

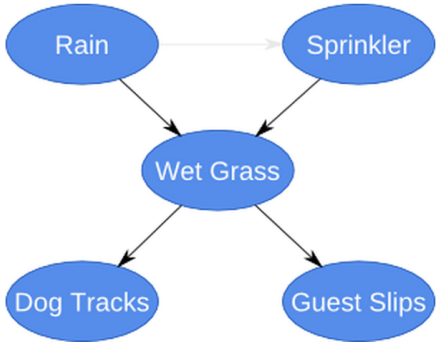


Figure 7: Our solution to the calibration problem.

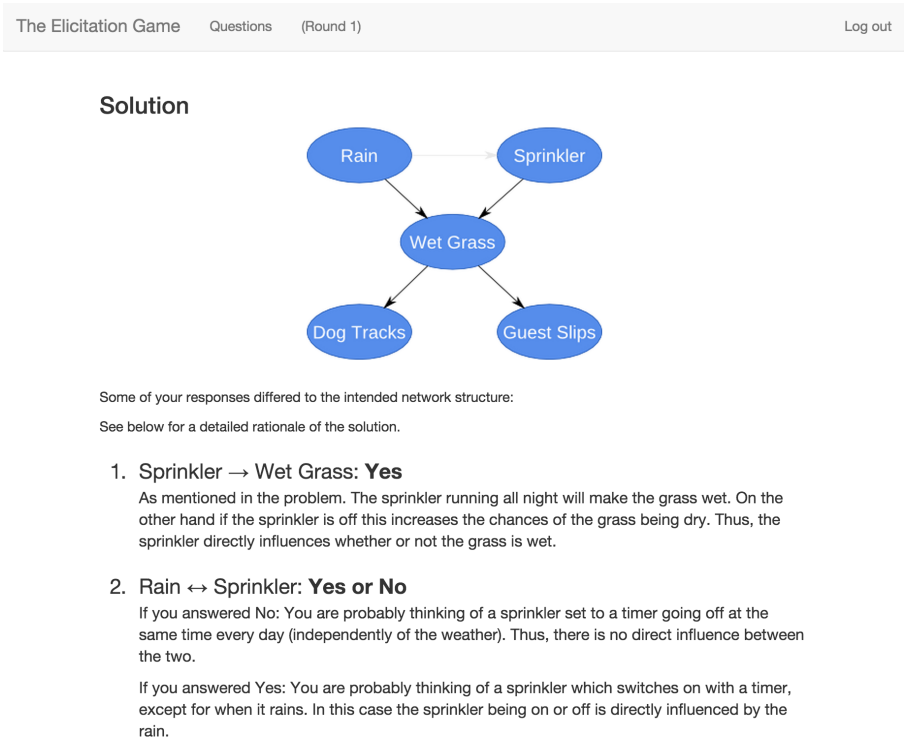


Figure 8: Sample of the feedback page after answering the calibration questions

nodes and are indeed nodes that we would normally have direct control over. They are treated differently during the Delphi elicitation: namely, as ordinary causal nodes for structure elicitation, and omitted altogether for parameter elicitation. This approach allows us to identify what factors the experts believe influence these decisions. In the final Delphi elicited model, the nodes are converted back to decision nodes.

4.3.2 Causal Structure

To perform the structure elicitation, we constructed a survey with questions regarding the directed influence between each of the chosen variables. Had we included all such possible questions, the survey would have contained $N(N-1) = 8 \times 7 = 56$ questions. However, some arcs are clearly invalid; for example, Future TB cannot causally influence Current TB, Age, Region of Origin, etc. We took advantage of this by grouping the chosen variables into 4 tiers as follows:

- Background (Tier 1): Age, Region of Origin, Relative with Active TB
- Current Disease and Tests (Tier 2): Current TB, Do Blood Test, Blood Test Result
- Treatment (Tier 3): Treat Latent TB
- Future Disease (Tier 4): Future TB

The variables within a tier are capable of influencing each other, as well as variables in downstream tiers (but *not* upstream tiers). Splitting the variables into tiers allows the number of questions to be reduced to 34. In addition, for pairs of variables within a tier, an expert is first asked whether there is any direct influence between the two variables. Only if they answer yes, are they asked for the direction. Typically, this saves many additional questions.

With the questions ready, the facilitator emailed experts with instructions on how to access and fill in the online survey. The software for the survey was the same as that used for

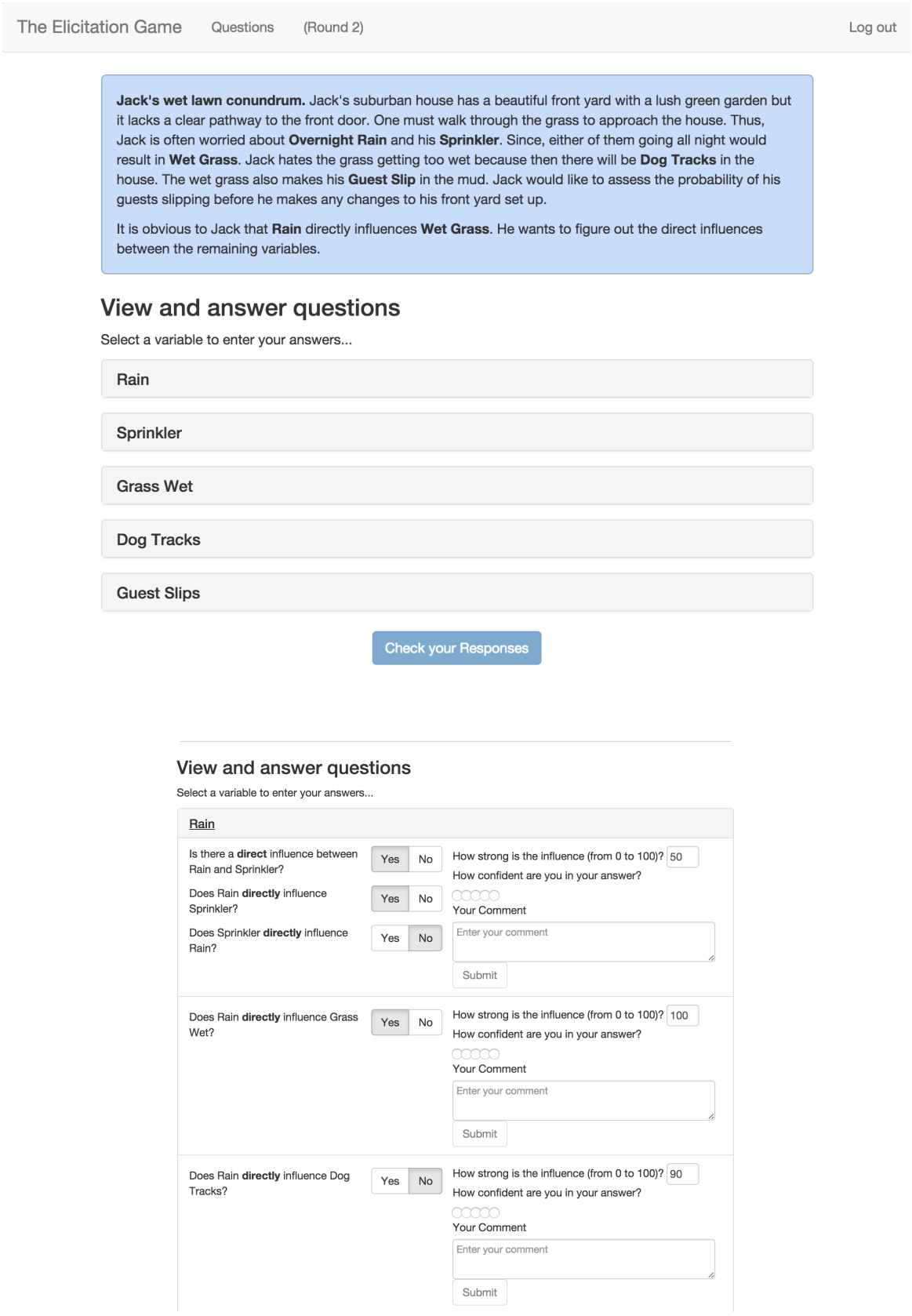


Figure 9: Questions as part of the calibration phase. The top screenshot shows the text, with the relevant variables following. The bottom screenshot shows the 'Rain' section expanded, revealing questions related to the Rain variable.

	Region of Origin	Relative with Active TB	Age	Current TB Infection Status	Blood Test Decision	Blood Test Result	Treatment Decision	Future TB Outcome
Region of Origin		6	-7	7	3	-6	3	-4
Relative with Active TB	-6		-8	6	1	-5	-3	-1
Age	-7	-4		5	-2	6	6	4
Current TB Infection Status	0	0	0		7	7	5	7
Blood Test Decision	0	0	0	-7		-5	-3	-7
Blood Test Result	0	0	0	-7	-5		5	-1
Treatment Decision	0	0	0	0	0	0		6
Future TB Outcome	0	0	0	0	0	0	0	

Table 2: Adjacency frequency matrix after Round 2

the calibration (see Figure 9). In this first round, the site was setup to isolate experts from each other when answering the questions. Experts were permitted (and encouraged) to record comments and explanations for their responses, but these would be kept hidden from other experts during the first round. Experts were initially given three days to submit their answers, but an extension of several days was provided just prior to the deadline. (Extensions were also provided after each subsequent survey.)

After completing the first round, the facilitator analysed the comments and responses, and sent an email to the experts summarising the results. This included pointing out some of the more interesting responses, where some experts answered quite differently (and with different reasoning) to others.

This email also contained instructions for commencing Round 2. In the second round, the answers and comments from all of the experts in the first round were made visible to all. In addition, any further comments and responses made would be made immediately available. Experts were again given three days to complete the survey (which was, again, extended). Once complete, all experts were again emailed with a summary of the results.

The final answers obtained at the end of the structure elicitation were aggregated to obtain an adjacency frequency matrix representing BNs, using the Serwylo counting method described earlier (i.e. +1 for a positive report of arc presence, -1 for a negative report). The adjacency matrix obtained after the trial elicitation is given in Table 2:

We produced several BNs based on different thresholds for arc presence. Many of the positive thresholds produced similar or identical networks. In the end, we settled on a +3 threshold due to it being far enough from 0 to avoid controversial arcs, while not being so severe as to require a unanimous response. The selected structure is shown in Figure 10. We inserted the highlighted arc after the elicitation. This is discussed in Section 4.4.1.

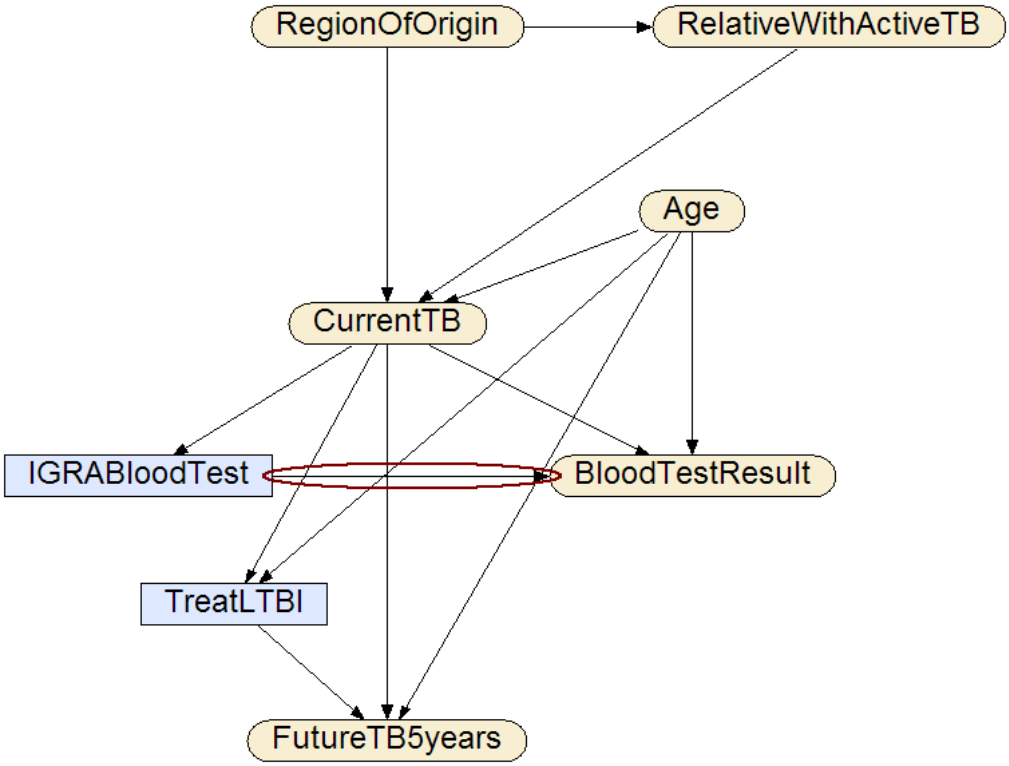


Figure 10: Resultant structure from the structure elicitation phase

4.3.3 Parameter Elicitation

After the structure elicitation was complete, the structure was assessed by the modellers to determine the critical parameters that needed to be elicited in the next phase. These critical parameters were mainly a mix of extreme cases (i.e., worst case and best case) and a few average cases from the CPTs for Current TB, Future TB, Blood Test Result and Relative with Active TB.

All in all, 18 CPT rows (out of a possible 138) were identified as important for elicitation. These rows contained 47 parameters all together (or 29 free parameters). A survey was then created based on the chosen parameters. Rather than asking questions on just the free parameters (which would raise the sometimes difficult decision of which parameters to treat as free), questions were designed to elicit an entire CPT row at once. Users were then free to choose which parameters within the CPT row to focus on. A typical question from this survey can be seen in Figure 11.

With the survey constructed, the parameter elicitation followed a loosely similar approach to the structure elicitation. The facilitator emailed experts, notifying them of the availability of the survey and that the first round had commenced, and provided instructions on how to complete it. (Due to the lack of a calibration session dedicated to the software used in this phase, this included a quick guide to the software, as can be seen in Figure 12.) In Round 1, experts completed the survey in isolation (without seeing answers and comments from other experts). The facilitator then analysed the answers, and emailed participants with a summary, pointing out any notable points of difference between the experts. The facilitator also opened Round 2, at which point experts could see other people's comments and answers, and could revise their own answers in response. Furthermore, discussion was also permitted, and this sometimes led to changes in estimates.

With the elicitation complete, the final estimates for each question were averaged to produce parameters that could be entered into the structure obtained in phase 2. Of course, only 18

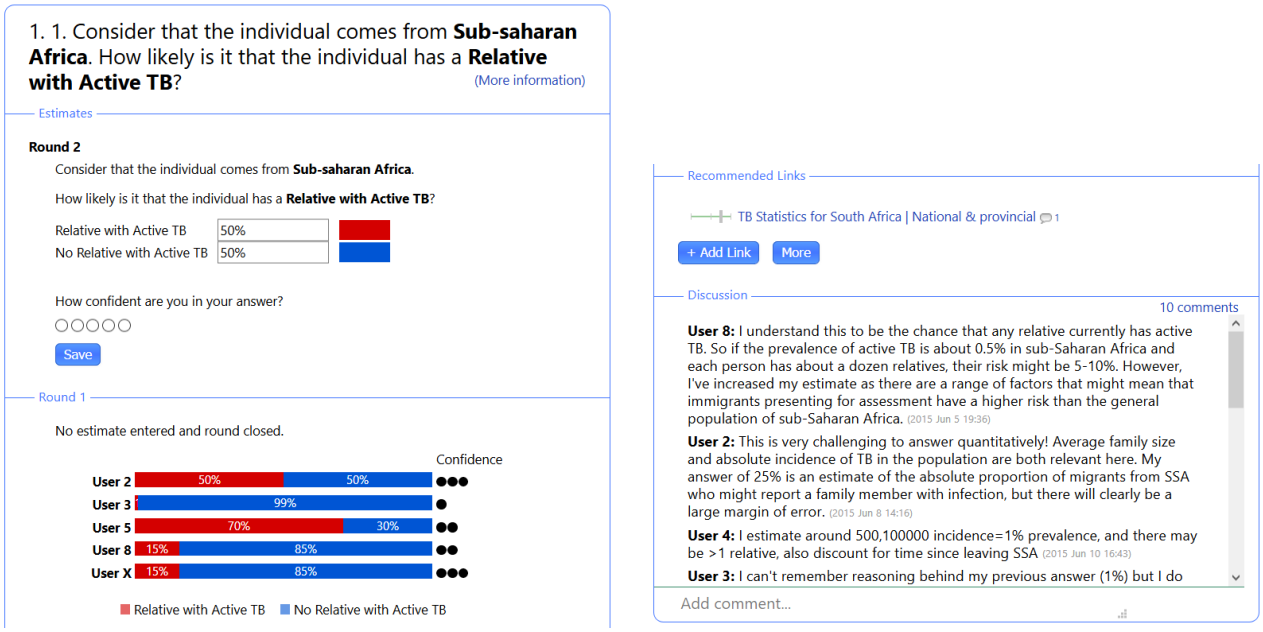


Figure 11: (Left) Interval probability judgments from round 1 are displayed back to each group, (right) links to useful websites are recommended, and discussion takes place below it.

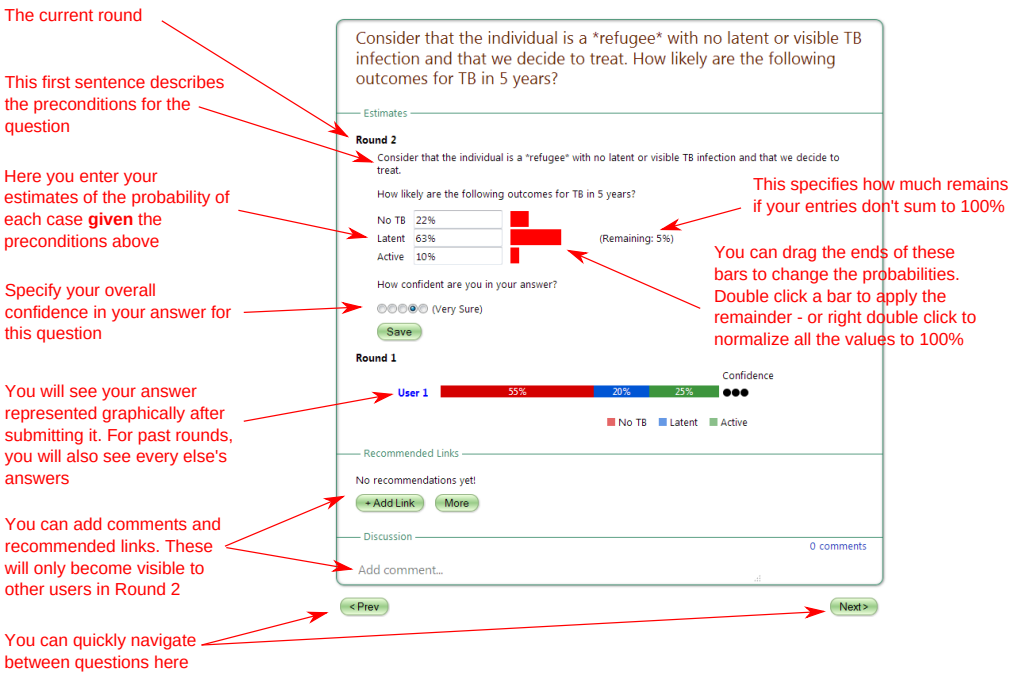


Figure 12: The quick guide to the parameter elicitation software

CPT rows could be filled directly. The modellers had to fill the remaining 120 rows via other means.

Prior probabilities for root nodes were not entered, as they can be entered during the application of the network based on the population being tested. Therefore, CPTs for Age and Region of Origin were left blank, which Netica treats as uniform parameters.

In other cases, the elicited values were the best case, worst case or average case for their CPT. Therefore, the missing values could be generated by interpolation. Interpolation software was used to parameterise Relative with Active TB, Current TB, Future TB and Blood Test

Result.[4] This software takes the probabilities for the best case and worst case, along with weights for each parent state as input. It then provides probability distributions for each case as per the weights and their interaction. The resultant distributions are bound between the best and worst case probabilities entered by the user. (A detailed explanation of the working and usage can be found in the software.) Different weights for the parent variables were applied and the resultant CPT generated by the interpolator was compared to the elicited values available. (For the Current TB node, this included several non-extreme CPT rows.) The generated CPT most similar to the elicited values was used, however the values elicited from the experts were always used in preference to the generated parameters. Figure 13 illustrates the final version of the network with all the necessary parameters filled in.

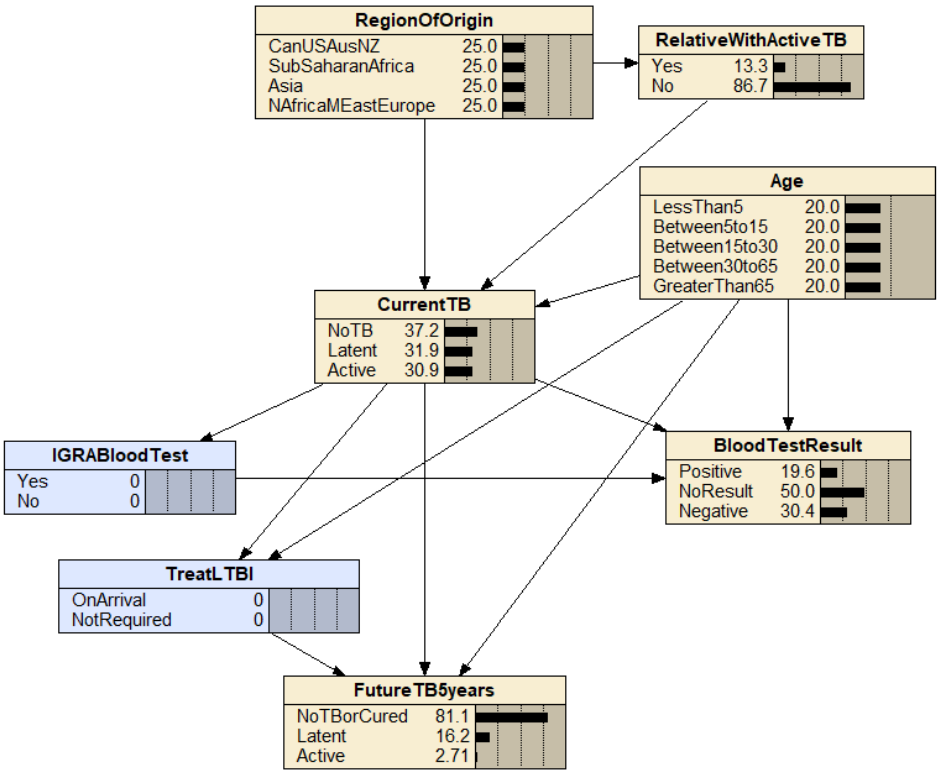


Figure 13: Final network with all parameters filled in

4.3.4 Evaluation

To evaluate the network, the modellers tested scenarios within the elicited network to ensure that there was no obviously incorrect behaviour. This allowed for some iteration with the CPT interpolation, to tweak the generated (but not elicited) probabilities.

Following this, a survey was created with scenarios drawn from the final BN. Experts were presented with the various scenarios, and asked to assess the plausibility of each one. Each scenario specified a definite state for *every* variable from the elicited network, including Current TB and Future TB. The expert's responses were then checked against the probability of these scenarios given by the network (i.e., the probability of the findings when a scenario is entered into the network).

[4]This tool is available at `https://bayesian-intelligence.com/interpolator/`.

4.4 Analysis

The final network accorded well with our expectations, both in terms of the structure and parameters.

4.4.1 Structure

The structure produced by the elicitation process produced no obviously incorrect arcs: as just two examples, Region of Origin was not judged to influence Age and Blood Test Result was not judged to influence Current TB. These questions were asked in the survey due to the variables appearing within the same tier. If we were particularly concerned about the number of questions being asked of experts, we would of course not have included them in the survey at all. However, here they clearly demonstrate that the experts were answering the questions with the correct causal interpretation in mind. As a particular example, it is a common error when creating Bayesian networks to model the direction of reasoning (i.e., Blood Test Result $\rightarrow$ Current TB), rather than the causal direction (i.e., Current TB $\rightarrow$ Blood Test Result). In aggregate, the experts made no errors of this kind.

One arc that was missed by the experts was Do Blood Test $\rightarrow$ Blood Test Result. This was likely due to several experts not having a good understanding of the possible states for Blood Test Result — namely, Positive, Negative and (importantly) No Result. A sample of the experts' comments for this question bears this out:

- "decision doesn't change the outcome, it's like schrodingers cat"
- "I've interpreted this as per user 1 ie. the mere fact that you've done the test doesn't directly affect the result."

Several other experts noted that it is only possible to get a result, if you decide to do the test:

- "Interested in reasons for those answering no. Can't get a result without a decision to do the test."
- "Only those with a Decision are tested."

In the end, the latter were not enough to overcome the negative votes, hence the arc did not appear in the network produced by aggregating the expert responses. As the modellers, we of course included it ourselves.

In addition, the experts identified several arcs that we were not originally expecting. These included Region of Origin $\rightarrow$ Relative with Active TB and Age $\rightarrow$ Blood Test Result.

It is interesting to note that the first is not a direct causal (or ancestral) influence — the mere fact that one comes from a certain location does not *cause* one to have a relative with active TB. However, there *is* a hidden complex network of common causes (e.g., children are born near parents, families tend to stay in physical proximity over time, TB spreads via physical proximity, etc.) that this one arc captures in the form of a correlation. Given that this complex subnetwork is missing from the model, it is perfectly legitimate to capture the important dependency with a (non-causal) arc. In this case, experts possibly chose Region of Origin $\rightarrow$ Relative with Active TB because that is the common direction of reasoning.[5] This seems to be supported by some of the expert comments:

[5]It would be quite unusual for any person, clinician or otherwise, to puzzle out whether a person comes from a particular region, based on whether or not they have a relative with active TB.

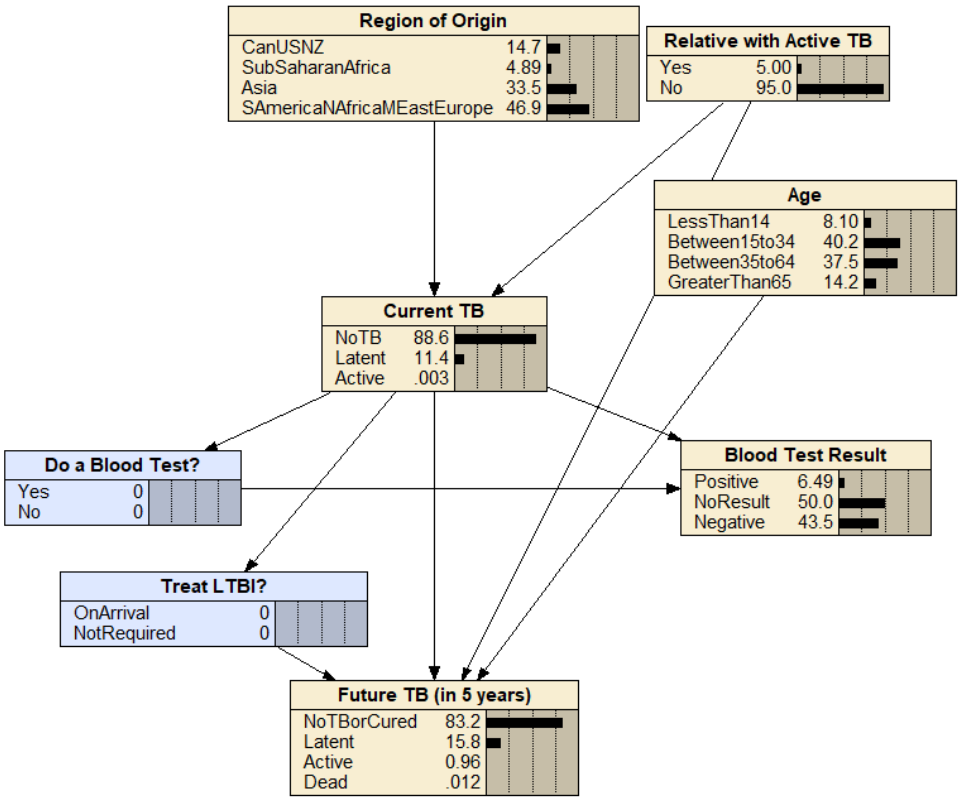


Figure 14: The manually elicited TB model, with differing nodes absorbed

- "people with high risk regions of origin are also more likely to experience exposure at home"
- "The more prevalent TB is in a region, the more likely it is that someone from that region will have a relative with TB."

In the case of Age → Blood Test Result, it was noted by several experts that both younger and older individuals produced less reliable results. This was an example in which experts changed their mind in response to discovering new information arising from the comments in the first round. One expert stayed with their original negative response, but lowered their confidence, while the second expert changed their response.

Comparison to the manually elicited model While the manually elicited model contains several more variables than the elicited model, we can make a direct comparison between the structures of the two models by marginalising out the excess nodes — what Netica calls 'absorbing' — from the manually elicited model. Absorbing nodes allows us to preserve the statistical relationships that are present between the remaining variables in the full variable model, while keeping the new structure as simple and as similar to the original structure as possible.[6] The absorbed version of the manually elicited model can be seen in Figure 14.

There were several differences in structure between the Delphi-elicited model and the manually elicited model. In some cases, these were a consequence of the difference in purpose of the two models, however in others, the Delphi-elicited model has provided an idea of how to

[6]There are some cases where absorbing will not produce the simplest model — for example, when two nodes are connected by counteracting paths, absorbing the intermediate nodes will preserve a connection between the nodes with an arc, despite there being no (or virtually no) relationship between the two remaining variables. This does not occur here, however.

improve the manually elicited model. For example, no direct relationship between Age and Blood Test Result was included in the manually elicited model, however the Delphi-elicited model suggests this may be a relationship that needs to be included. By contrast, the expert identified relationship between Region of Origin and Relative with Active TB may not be so relevant for the manually elicited model, given that that model makes the assumption that all the input (in this case, root) nodes will be specified whenever the model is used.

In the absorbed manually elicited models, there are several nodes that affect the Future TB node, due to the presence of paths in the full model that needed to be preserved in the absorbed model. For example, Region of Origin and Relative with Active TB both influence of Future TB directly, rather than via Current TB. This occurs due to the presence of a common cause of both Current TB and Future TB (namely, a "Refugee" node) in the original manually elicited model. The Delphi-elicited model does not include these connections. It's possible that the connection should be there, but was missed due to the missing common cause. Alternatively, it's possible that the common cause is either incorrect or in practice has limited influence.

We can calculate the edit distance between the Delphi-elicited model and the absorbed manually elicited model to get an idea of the magnitude of the difference between the two network structures. The edit distance calculates how many arcs would need to be either added, deleted or reversed to make the structure of one of the networks identical to that of the other. The edit distance between these two networks is 8, none of which involve arc reversal. Here is the full list of differences from the manually elicited model to the Delphi-elicited model:

1. (missing) Region of Origin → Future TB
2. (missing) Relative with Active TB → Future TB
3. (missing) Do Blood Test → Blood Test Result
4. (added) Region of Origin → Relative with Active TB
5. (added) Age → Current TB
6. (added) Age → Blood Test Result
7. (added) Age → Treat Latent TB
8. (added) Blood Test Result → Treat Latent TB

This is quite a large number considering the average number of arcs across both networks is 12. However, some differences can be accounted for quite easily. For example, the arcs from Region of Origin and Relative with Active TB to Future TB in the manually elicited model would very likely not be included if building the model directly. Furthermore, the decision nodes were interpreted differently in the Delphi-elicited model and the manually elicited model. And as noted earlier, the interpretation of Do Blood Test affecting Blood Test Result was unclear, and as modellers we would immediately notice that this arc needs to be present. Accounting for these easily explained differences leaves us with an edit distance of 3 over an average of 10.5 arcs, with the specific arc differences as follows:

1. (added) Region of Origin → Relative with Active TB
2. (added) Age → Current TB
3. (added) Age → Blood Test Result

Each of these arcs would be worth considering as inclusions into the manually elicited model.

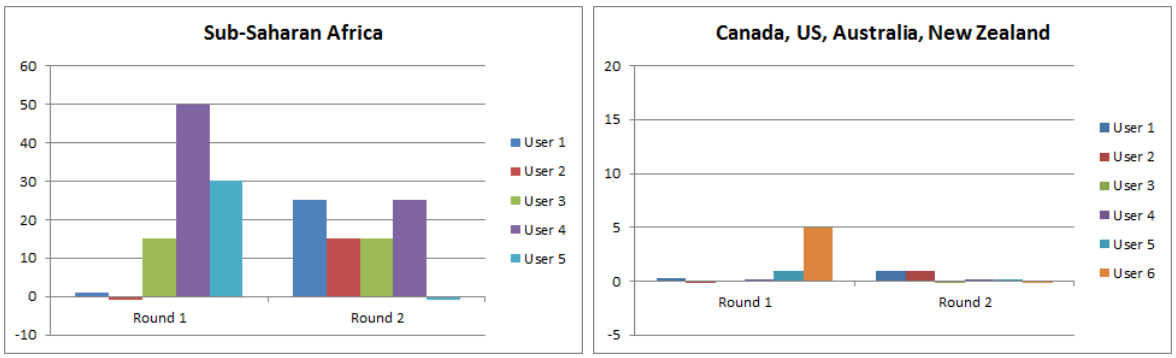


Figure 15: Responses for the two questions regarding the likelihood of having a Relative with Active TB. Only the probabilities for 'Yes' are shown, for both rounds 1 and 2. A negative bar indicates the user did not provide a response for that round.

4.4.2 Parameters

The number of experts actively participating in the elicitation dropped during the parameter elicitation phase (discussed in Section 4.4.3), however the response rate was still healthy. As noted earlier, due to the number of parameters involved, only a small number of key parameters (distributions) were elicited. We will look at how the experts responded to these here.[7]

We looked at the Bhattacharyya distance between all pairs of expert provided distributions for all questions in each round (Table 3).[8] While the aim of the Delphi process is not convergence or consensus, we nonetheless see evidence of that occurring here. Indeed, we see a lower Bhattacharyya distance across every question in Round 2 than in Round 1, except for Question 4.

While we will not look at all the elicited parameters in detail, we will look a little more closely at the parameters for two example nodes, the Relative with Active TB and Future TB nodes. We elicited two distributions for the CPT of Relative with Active TB (Questions 1 and 2). The group responses for these two questions are shown in Figure 15 for both Rounds 1 and 2. All experts responded with a marked difference in their distributions for the two questions. For the question regarding Sub-Saharan Africa, there was a fair amount of variability in responses in the first round, with convergence occurring in the second round. While for the question regarding western countries, there was much more agreement (at least superficially) from the word go. It is interesting to note, however, that these probabilities are very close to 0 or 1. It is often the case that a small difference near 0 or 1 is much more significant than a small difference around (say) 0.5, particularly when large volumes are involved. Hence, it is not especially clear whether the differences in the expert responses here represent genuine disagreement that would lead to notable differences in network predictions.

We elicited 6 distributions for the Future TB CPT. The experts seemed to agree well on all these distributions. First round responses by the experts were extremely similar in almost all cases, while second round estimates showed a very strong convergence. This can be seen in Table 3, which shows the average Bhattacharyya distances for each question and round. Question 13 involves the individual having Latent TB, but not receiving treatment, while questions 14 through 18 involve a mix of Latent TB and Active TB that *do* involve treatment. It seems that expert agreement on the efficacy of treatment under different scenarios is quite

[7]Two experts on a handful of occasions provided responses that were the exact opposite of what we expected. Since these were very clearly incorrect, we took these to be accidental errors, and inverted them.

[8]A value of 0 for the Bhattacharyya distance indicates the distributions are identical, while a value of ∞ indicates entirely different (i.e., incompatible) distributions. Incompatibility in this case means that, for every state, at least one of the distributions assigns a 0 probability. In practice here, the maximum distance between two 2-state distributions is expected to be around 1.6 ([0.99, 0.01] vs [0.01, 0.99]).

	Round 1	Round 2
Q1	0.1752	0.005298
Q2	0.007776	0.001567
Q3	0.0101	0.001278
Q4	0.006149	0.007449
Q5	0.01115	0.005055
Q6	0.02959	0.01908
Q7	0.03942	0.01367
Q8	0.1205	0.02076
Q9	0.09767	0.001305
Q10	0.07479	0.01752
Q11	0.00496	0.001278
Q12	0.1148	0.008548
Q13	0.1632	0
Q14	0.006201	0.002333
Q15	0.01329	0.003735
Q16	0.02373	0.005773
Q17	0.00639	0.004755
Q18	0.007573	0.005817

Table 3: Average Bhattacharyya distance in Round 1 and Round 2 for all questions

strong in all cases. Of these, Question 16 exhibited the greatest variation in responses, followed by Question 15. These were also questions in which the average distribution contained the highest entropy (i.e., were the furthest away from deterministic). In the case of Question 13 (i.e., Latent TB in a young child with no treatment), there was a large amount of variation in the first round which was eliminated in the second round. This suggests that at least some of the experts had a high degree of uncertainty for this, despite the fact that all experts generally indicated medium to high confidence for this question.

Comparison to the manually elicited model distribution An exact comparison of the model distributions cannot be made due to some differences in node states (that occurred during the course of the manually elicited TB project) and the differences in structure that we described earlier. However, an approximate comparison can be made.

First, we can compare the marginal probabilities of the three key comparable nodes, which gives the results in Table 4. (Before performing this comparison, we first remove the CPTs for nodes in the manually elicited model that were not elicited at all.) This shows exceptionally good agreement in most cases, except for Future TB's first two states, which are nonetheless within a similar range. To emphasise, there is no reason to suppose the manually elicited model is in any sense the 'correct' model, but it is nonetheless notable that there is such good agreement. This gives us some reason to have confidence in the automated process.

We can get a very rough sense of how the CPTs for each of these three nodes compare by altering the network (in particular, absorbing nodes) until the two nodes being compared have the same parents. Doing this for Current TB, and then taking the average Bhattacharyya characteristic across all rows in the CPT, gives a value of 0.00575, which indicates good agreement.[9]

[9]We adopted here a version of the Bhattacharyya characteristic that is calculated as $1 - \sum_i \sqrt{p_i q_i}$. This

Node	State	Manually Elicited Marginal	Delphi-Elicited Marginal
Current TB	NoTB	0.885	0.892
	Latent	0.115	0.105
	Active	0.00004	0.0027
Blood Test Result	NoResult	0.5	0.5
	Positive	0.0653	0.0893
	Negative	0.435	0.411
Future TB	NoTBorCured	0.831	0.925
	Latent	0.159	0.0643
	Active	0.0098	0.0102

Table 4: Comparison between the marginal probability distributions for key nodes in the manually elicited and Delphi-elicited networks. (Note that Future TB has the additional state 'Dead' in the manually elicited network, but the probability for this is negligible.)

Similarly, for Blood Test Result, the characteristic is 0.00119. For Future TB, however, the value is 0.28, which is quite a large difference. The marginal probabilities indicate some difference, but perhaps not to this degree. The reason this does not manifest in a larger difference in marginal probabilities is because the the most significant divergences occur when Current TB is 'Active' — which is (for the marginal) an improbable case.

4.4.3 User Activity

There is a wide range of statistics on user activity that can illuminate how the Bayesian Delphi process functions. We will only look at a small set of these statistics here.

Structure Elicitation Since users often enter comments on answering questions, we can take a look at the word frequencies to gain some insight into the thoughts of the experts. Table 5 shows the word frequencies for the top 20 words in Round 1 and Round 2 (left and middle table), along with the top 10 words with the greatest increase (as well as decrease) in relative frequency (right table).[10] It should come as no surprise that 'TB' is the most commonly used word in both rounds. The second most commonly used word in both rounds is 'test', suggesting that there was substantial interest in the impact on and effect of testing in the given scenario. Other common words across both rounds tend to cover both the subject matter (e.g., 'infection', 'region', 'active', etc.) as well as causal modelling matters (e.g., 'risk', 'likely', 'influence').

The change in relative frequencies (rightmost table) shows several points of interest. We see that words like 'influence', 'indirect' and 'direct' have increased in frequency. This is possibly due to the experts focussing more intently on how these terms affect their responses. The change in the frequency of 'age' is quite substantial, which suggests that the experts had cause to focus more on age's role in the scenario. (This matches with the identification as age by some experts as important in blood tests, which was picked up by other experts in Round 2.) Also of note is the drop in the relative frequencies for both 'risk' and 'likely'.

Table 6 shows the average number of times users entered or changed their responses for a question (across both rounds). We can see that most users answered questions around twice on average, which is as expected if answering just once for Round 1 and Round 2. (Keeping in mind that participants could change their answers as often as they liked.) For several questions, users revisit their answer more than once; most of these questions are for cause and effect variables

returns 0 for identical distributions, and 1 for maximally different distributions — i.e., deterministically contradicting states.

[10]The 40 most commonly used English words were first removed from all the comments.

Round 1 Words

Word	Frequency
tb	94
test	51
if	33
more	32
status	30
risk	29
likely	29
infection	27
region	22
active	21
may	21
result	21
high	19
treatment	19
relative	18
decision	17
current	16
via	16
positive	15
blood	15

Round 2 Words

Word	Frequency
tb	73
test	36
influence	35
no	29
age	27
active	26
relative	26
decision	25
indirect	22
if	21
direct	18
via	17
status	17
result	16
risk	15
likely	15
so	14
outcome	13
influences	12
directly	12

Change from Round 1 to Round 2

Word	Change
influence	0.018047614
no	0.016522519
indirect	0.012912996
age	0.010294638
direct	0.008006995
whether	0.007854583
directly	0.007168242
question	0.00714284
patients	0.006431096
if	-0.00770266
prevalence	-0.008159896
status	-0.00849061
high	-0.008770032
risk	-0.009227756
likely	-0.009227756
test	-0.009380656
infection	-0.012125535
region	-0.012252544
tb	-0.012558834
more	-0.014133756

Table 5: Words used in comments during the structure elicitation stage. (Left) Top 20 words in Round 1 and (Middle) Round 2. (Right) Top 10 words with the greatest increase (blue) and decrease (red) in frequency.

Cause - Effect	Changes/user
Age - Relative with Active TB	2.75
Relative with Active TB - Age	2.75
Blood Test Result - Current TB Infection Status	2.375
Current TB Infection Status - Blood Test Result	2.375
Blood Test Decision - Blood Test Result	2.25
Blood Test Result - Blood Test Decision	2.25
Age - Blood Test Decision	2.125
Age - Blood Test Result	2.125
Blood Test Decision - Current TB Infection Status	2.125
Current TB Infection Status - Blood Test Decision	2.125
Age - Region of Origin	2
Age - Treatment Decision	2
Current TB Infection Status - Treatment Decision	2
Region of Origin - Age	2
Region of Origin - Blood Test Decision	2
Region of Origin - Relative with Active TB	2
Region of Origin - Treatment Decision	2
Relative with Active TB - Blood Test Decision	2
Relative with Active TB - Blood Test Result	2
Relative with Active TB - Region of Origin	2
Age - Current TB Infection Status	1.875
Age - Future TB Outcome	1.875
Blood Test Result - Treatment Decision	1.875
Current TB Infection Status - Future TB Outcome	1.875
Region of Origin - Blood Test Result	1.875
Region of Origin - Future TB Outcome	1.875
Relative with Active TB - Current TB Infection Status	1.875
Treatment Decision - Future TB Outcome	1.875
Blood Test Decision - Future TB Outcome	1.75
Blood Test Decision - Treatment Decision	1.75
Region of Origin - Current TB Infection Status	1.75
Relative with Active TB - Future TB Outcome	1.75
Blood Test Result - Future TB Outcome	1.625
Relative with Active TB - Treatment Decision	1.625

Table 6: Average number of times the participants altered their responses for each question

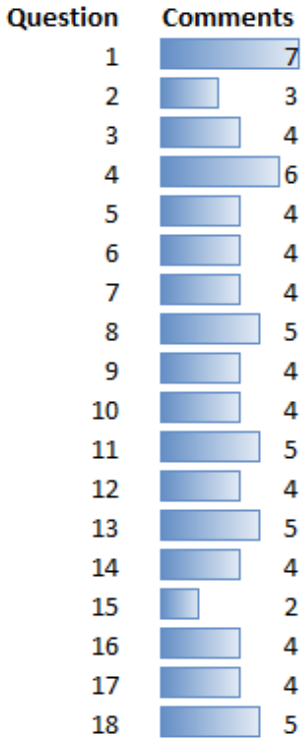


Figure 16: Number of comments per question in the Parameter Elicitation

that belong to the same tier (these are highlighted in yellow in the table). Same tier variables are asked twice (variable A to B, and then variable B to A), however the second time the question is answered, the answer is pre-filled from the first response. So it is likely that some experts are reconsidering their answer upon seeing it a second time. Note that this does not mean answers are reversing the direction of influence; it may just mean that experts are (for example) assigning a different confidence to the answer.

Table 7 shows a break down of changes by user for each question. We can see that some users frequently revisted questions (for example, User 4 and User 8), while some users provided just a single response across both rounds most of the time (for example, User 2 and User 3).

Parameter Elicitation Table 8 shows the top word frequencies for Rounds 1 and 2 of the parameter elicitation, along with the top changing word frequencies across the two rounds. Again, 'TB' is the most common term. However, 'test' no longer ranks as a common word, as it seems most of the concern around testing was solely regarding the structure of the scenario, rather than the parameters. However, the term 'risk' has become very common. The use of 'risk' here is due to the involvement of probability of course, since there are no utilities involved in the elicitation. 'Active' and 'latent' are also quite common, as is 'infection'.

It appears that experts feel much more subjective about the answers provided in the parameter elicitation stage. Hence, words like 'I'm', 'I've', 'estimate', 'assuming', 'probably', 'think' are quite common across both rounds in this stage. The lower level of certainty during parameter elicitation is certainly expected.

'Figure' becomes a much more common word in Round 2. The word is being used as a synonym for 'number' or 'probability', and suggests that people are commenting on the numerical responses from the first round. In support of this, 'answer' is the second most common word. Many of the words that become less common in Round 2 are related to uncertainty (for example, 'chance', 'probably', 'assuming').

In the parameter elicitation the experts had the option of submitting multiple comments

Question	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8
1	3	2	1	2	2	2	2	3
2	3	2	1	2	2	3	2	2
3	3	1	1	2	2	2	2	2
4	2	1	1	2	2	2	2	3
5	3	1	1	2	2	2	2	3
6	4	3	1	2	3	3	2	4
7	3	2	1	2	2	2	2	2
8	2	2	1	3	2	2	2	4
9	2	1	2	2	2	2	2	4
10	2	1	1	2	2	2	2	2
11	2	1	1	2	2	2	2	2
12	2	2	1	3	2	2	2	4
13	3	1	1	4	2	2	2	4
14	1	1	1	2	2	2	2	2
15	2	1	1	2	2	1	2	4
16	2	1	2	2	2	2	2	4
17	3	1	1	4	2	2	2	4
18	2	1	1	3	2	2	2	2
19	2	1	1	3	2	3	2	2
20	3	1	1	2	2	2	2	3
21	2	1	1	4	2	2	2	2
22	2	1	1	2	2	3	2	2
23	2	1	1	2	2	2	2	2
24	2	1	1	3	2	2	2	2
25	3	1	1	2	2	2	2	3
26	2	2	1	2	2	2	2	3
27	4	3	1	2	3	3	2	4
28	2	3	1	2	2	2	2	2
29	3	1	1	3	2	2	2	2
30	2	1	1	3	2	2	2	2
31	2	1	1	2	2	2	2	2
32	3	1	1	2	2	2	2	3
33	2	1	1	2	2	1	2	2
34	2	2	1	2	2	2	2	2

Table 7: Number of times the responses were altered across users and questions

Word	Frequency
tb	33
active	22
risk	16
latent	14
I'm	13
infection	11
estimate	10
however	10
question	9
so	8
assuming	8
infected	8
no	8
group	7
probably	7
individual	7
about	7
here	6
proportio	6
means	6

Word	Frequency
tb	53
risk	25
active	22
figure	19
latent	18
infection	15
answer	13
so	13
I've	12
user	12
which	11
less	11
think	10
estimate	10
dont	10
am	10
here	9
if	9
group	9
than	9

Word	Change
figure	0.010342825
answer	0.006781318
I've	0.006000678
think	0.005561732
which	0.005220038
tb	0.00433689
less	0.004097704
because	0.004000452
recent	0.003561506
up	0.003561506
infected	-0.005075475
means	-0.005172726
proportion	-0.005172726
chance	-0.005953367
probably	-0.006295061
individuals	-0.006295061
assuming	-0.006636755
active	-0.007517275
I'm	-0.009906507
however	-0.010442704

Table 8: Words used in comments during the parameter elicitation stage. (Left) Top 20 words in Round 1 and (Middle) Round 2. (Right) Top 10 words with the greatest increase (blue) and decrease (red) in frequency.

for one question. Figure 16 shows the total number of comments across questions. On average, there were 4.3 comments per question, with most comments occurring (as one might expect) on the first question. The number of comments per question remained stable throughout the elicitation process.

Figure 17 shows the number of estimates made by each user for each question. For the most part, users answered each question just twice (once for each round), however some users (User 3 and 4) answered questions less, while User 5 revisited questions a few more times. Generally, the difficulty of answering parameter questions is much higher than for structure questions, so one expects lower response and revision rates. However, there may be more value in revising parameter questions than structure questions, since structure questions are often just Yes/No (not counting additional questions about confidence), while there are more degrees to work with for parameter questions. In any event, there was no significant difference in the response rate of individual experts who chose to participate in the both elicitation stages.

5 Conclusion

While a number of techniques are used in strategic risk assessment for both qualitative and quantitative analyses, very few are capable of both. BNs (and particularly causal BNs) provide an approach for performing a clear and intuitive qualitative analysis, with the further option of a completely rigorous quantitative analysis. As such, they arguably fulfil the goal of strategic risk assessment better than any other modelling or knowledge representation technique. The most frequently raised point of concern is the difficulty and effort required to build and parameterize these BNs, particularly in a collaborative setting.

Meanwhile, since it was first introduced in the 1950s, the Delphi protocol has proven to be an extremely robust approach to eliciting estimates from collaborating groups of experts. However, the process can take a lot of time and effort — perhaps too much for anything more than a small number of questions. It would therefore seem folly to marry the Delphi protocol

Question	User 1	User 2	User 3	User 4	User 5	Total
1	2	3	0	1	2	8
2	3	2	1	2	2	10
3	2	3	1	1	3	10
4	2	2	2	2	2	10
5	2	2	1	1	2	8
6	2	2	1	1	2	8
7	2	2	1	2	2	9
8	2	2	2	2	3	11
9	2	3	1	2	3	11
10	2	2	1	2	3	10
11	2	2	1	1	3	9
12	2	2	2	2	2	10
13	2	2	1	2	2	9
14	2	2	2	1	2	9
15	2	4	1	1	2	10
16	2	2	2	1	2	9
17	2	2	2	1	2	9
18	2	1	2	1	2	8

Figure 17: Number of estimate changes per user per question in the Parameter Elicitation

with BN engineering.

Here we have seen that this marriage is not only possible, but also practical and fruitful. To be clear, there are many things that are needed to make this marriage work. The right technology is essential, so too careful planning and oversight. A very healthy dose of prior knowledge is also needed to ensure that expert time is used wisely and effectively. But with all this in hand, it is possible to create models and an understanding that go beyond what other techniques can provide.

There is a long way to go before the techniques described here can be made to work for large problems. A key issue to solve is how to focus questions on just those that most require expert involvement. A great deal of simple but extremely useful prior knowledge is often readily available, and which doesn't require expert input. For structure elicitation, this can include temporal order for many of the variables (which might be represented in the form of the tiers used in the modelling here), known variable associations (or non-associations) and logical constraints. For parameter elicitation, this can include logical constraints, correlations and (perhaps most importantly) local structure, which in many common cases can simplify the number of parameters needed dramatically.

To build a model, we need to pass through at least five stages (typically iteratively): select a problem, select the variables, draw the structure (the direct dependencies), define the types of relationships (the local structure), and quantify the relationships. In addition, testing and validation are critical, and appear across all of the other five stages. Experts may need to be involved in all of these different aspects of model building. Of these, we have focused here just on structure and parameterisation, however a robust group-based elicitation method would assist in all these aspects of development. Our work so far suggests that such a method would be a very worthwhile pursuit.

References

Baran, E. and Jantunen, T. (2004). Stakeholder consultation for bayesian decision support systems in environmental management. *Forest*, 27(35.6):31–37.

Bashari, H., Smith, C., and Bosch, O. J. H. (2009). Developing decision support tools for rangeland management by combining state and transition models and bayesian belief networks. *Agricultural Systems*, 99(1):23 – 34.

Bishop, P., Hines, A., and Collins, T. (2007). The current state of scenario development: An overview of techniques. *Foresight*, 9(1):5 – 25.

Brown, B. B. (1968). Delphi process: A methodology used for the elicitation of opinions of experts. DTIC Document.

Cinar, D. and Kayakutlu, G. (2010). Scenario analysis using bayesian networks: A case study in energy sector. *Knowledge-Based Systems*, 23(3):267 – 276.

Clemen, R. T. and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203.

Costa, P. C. G. and Kathryn, B. L. (2006). Multi-entity bayesian networks without multi-tears.

Crance, J. H. (1987). Guidelines for using the Delphi technique to develop habitat suitability index curves. Technical report, US Fish and Wildlife Service.

Deloitte (2013). Exploring Strategic Risk. Technical report.

Frühling, S. (2007). *Managing Strategic Risk: Four Ideal Defence Planning Concepts in Theory and Practice*. PhD, Australian National University, Strategic and Defence Studies Centre, Research School of Pacific and Asian Studies.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701.

Hejblum, G., Ioos, V., Vibert, J.-F., Böelle, P.-Y., Chalumeau-Lemoine, L., Chouaid, C., Valleron, A.-J., and Guidet, B. (2001). A web-based delphi study on the indications of chest radiographs for patients in icus. *Chest*, 133(5):1107–1112.

Helmer, O. and Rescher, N. (1959). On the epistemology of the inexact sciences.

Hodgetts, R. M. (1977). Applying the Delphi technique to management gaming. *Simulation*, 29(1):209–212.

Hsu, C.-C. and Sandford, B. A. (2007). The delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation*, 12(10).

Kelly, D. L. and Smith, C. L. (2009). Bayesian inference in probabilistic risk assessmentthe current state of the art. *Reliability Engineering & System Safety*, 94(2):628 – 643.

Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition.

Kragt, M. (2009). *A Beginners Guide to Bayesian Network Modelling for Integrated Catchment Management*. Landscape logic technical report. Landscape Logic, University of Tasmania. Landscape Logic and Australia. Department of the Environment, Water, Heritage, and the Arts.

Linstone, H. A. and Turoff, M., editors (1975). *The Delphi method : techniques and applications*. Reading, Mass. Addison-Wesley Pub. Co., Advanced Book Program.

Lucas, P. J. F. (2005). Bayesian network modelling through qualitative patterns. *Artificial Intelligence*, 163(2):233 – 263.

MacMillan, D. C. and Marshall, K. (2006). The delphi process an expert-based approach to ecological modelling in data-poor environments. *Animal Conservation*, 9(1):11–19.

Millett, S. M. (2009). Should probabilities be used with scenarios? *Journal of Futures Studies*, 13(4):61–68.

Nadkarni, S. and Shenoy, P. P. (2004). A causal mapping approach to constructing bayesian networks. *Decis. Support Syst.*, 38(2):259–281.

Nicholson, A. E. and Flores, M. J. (2011). Combining state and transition models with dynamic bayesian networks. *Ecological Modelling*, 222(3):555–566.

Ouchi, F. and Bank, W. (2004). *A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis*. Policy Research Working Paper. World Bank.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Fransisco.

Ritchey, T. (2002). General morphological analysis a general method for non-quantified modelling.

Rowe, G. and Wright, G. (2001). Expert opinions in forecasting: The role of the delphi technique. In Armstrong, J. S., editor, *Principles of Forecasting*, volume 30 of *International Series in Operations Research & Management Science*, pages 125–144. Springer US.

Sahal, D. and Yee, K. (1975). Delphi: An investigation from a bayesian viewpoint. *Techonological Forecasting and Social Change*, 16:5–18.

Schultze, T., Mojzisch, A., and Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes*, 118(1):24–36.

Serwylo, P. (2016). *Eliciting Bayesian Networks via Online Surveys:*. PhD thesis, Monash University, Faculty of Information Technology.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday.

Tetlock, P. (2005). *Expert Political Judgment: How Good is It? how Can We Know?* Princeton paperbacks. Princeton University Press.

Tversky, A. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185:1124–1131.

VicHealth (2014). Victorian infectious diseases bulletin.

Wellman, M. P. (1990). Graphical inference in qualitative probabilistic networks. *Networks*, 20(5):687–701.

White, A. L. (2012). Conceptual Models for Victorian Ecosystems. Technical Report 64, Parks Victoria, Melbourne.

Wintle, B., Mascaro, S., Fidler, F., McBride, M., Burgman, M., Flander, L., Saw, G., Twardy, C., Lyon, A., and Manning, B. (2013). The intelligence game: Assessing delphi groups and structured question formats. In *Australian Security and Intelligence Conference*. SRI Security Research Institute, Edith Cowan University, Perth, Western Australia.

Wooldridge, S. (2003). Bayesian belief networks. *CISRO 2003*.

Wright, S. (1934). The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

Zwicky, F. and Zwicky, F. (1969). Discovery, invention, research through the morphological approach.

A Appendix

A.1 Variables

Variables	Description	States
Age	The age of the immigrant.	<5 5-15 15-30 30-65 65+
Region of Origin	Which region or group of countries is the immigrant from?	CAN-US-AUS-NZ Sub Saharan Africa Asia North Africa-Middle East-Europe
Relative with Active TB	Whether the immigrant has had a relative with active TB?	Yes No
Current TB Infection Status	Represents the expected TB state of the immigrant	No TB Latent Active TB
Blood Test Decision	Whether the immigrant should take the blood test for TB?	Yes No
Blood Test Result	Result of the TB blood test taken by the immigrant	Positive (TB) Negative (No TB) No Result (No test taken)
Treatment Decision	Whether the immigrant should be given treatment?	Not Required On Arrival
Future TB Outcome	Represents the expected TB state of the immigrant in 3-5 years	No TB Latent Active TB

A.2 Baseline BN

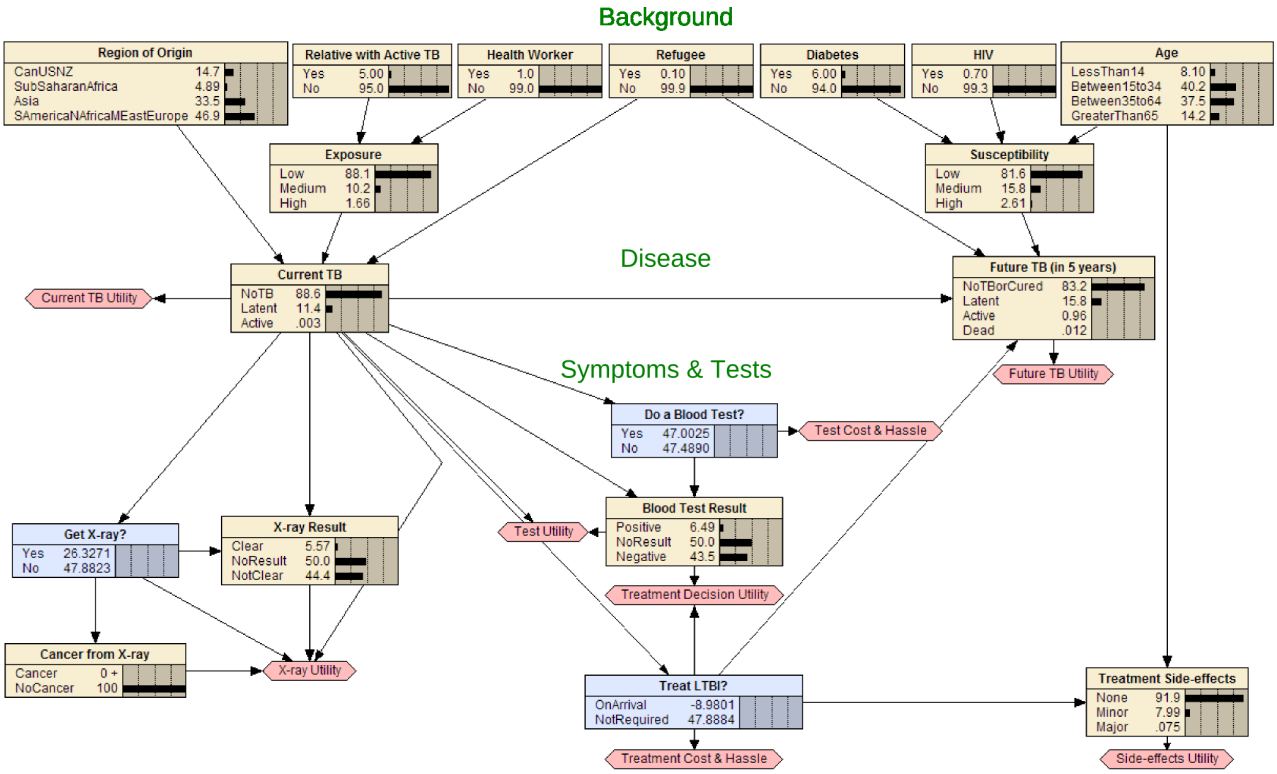


Figure 18: The manually elicited TB model

A.3 Structure Questions

The structure questions are listed by category (e.g. Age) followed by a list of questions associated with the category.

1. Age
    - Is there a direct influence between Age and Relative with Active TB?
    - Is there a direct influence between Age and Region of Origin?
    - Does Age directly influence Current TB Infection Status?
    - Does Age directly influence Blood Test Decision?
    - Does Age directly influence Blood Test Result?
    - Does Age directly influence Treatment Decision?
    - Does Age directly influence Future TB Outcome?
2. Relative with Active TB
    - Is there a direct influence between Relative with Active TB and Age?
    - Is there a direct influence between Relative with Active TB and Region of Origin?
    - Does Relative with Active TB directly influence Current TB Infection Status?
    - Does Relative with Active TB directly influence Blood Test Decision?
    - Does Relative with Active TB directly influence Blood Test Result?
    - Does Relative with Active TB directly influence Treatment Decision?
    - Does Relative with Active TB directly influence Future TB Outcome?
3. Region of Origin
    - Is there a direct influence between Region of Origin and Age?
    - Is there a direct influence between Region of Origin and Relative with Active TB?
    - Does Region of Origin directly influence Current TB Infection Status?
    - Does Region of Origin directly influence Blood Test Decision?
    - Does Region of Origin directly influence Blood Test Result?
    - Does Region of Origin directly influence Treatment Decision?
    - Does Region of Origin directly influence Future TB Outcome?
4. Current TB Infection Status
    - Is there a direct influence between Current TB Infection Status and Blood Test Decision?
    - Is there a direct influence between Current TB Infection Status and Blood Test Result?
    - Does Current TB Infection Status directly influence Treatment Decision?
    - Does Current TB Infection Status directly influence Future TB Outcome?
5. Blood Test Decision
    - Is there a direct influence between Blood Test Decision and Current TB Infection Status?

- Is there a direct influence between Blood Test Decision and Blood Test Result?
- Does Blood Test Decision directly influence Treatment Decision?
- Does Blood Test Decision directly influence Future TB Outcome?

6. Blood Test Result
    - Is there a direct influence between Blood Test Result and Current TB Infection Status?
    - Is there a direct influence between Blood Test Result and Blood Test Decision?
    - Does Blood Test Result directly influence Treatment Decision?
    - Does Blood Test Result directly influence Future TB Outcome?
7. Treatment Decision
    - Does Treatment Decision directly influence Future TB Outcome?

A.4 Parameter Questions

The parameter questions and their associated states are listed below.

1. Consider that the individual comes from Sub-saharan Africa. How likely is it that the individual has a Relative with Active TB?
    - Relative with Active TB
    - No Relative with Active TB
2. Consider that the individual comes from Canada, the US, Australia or New Zealand. How likely is it that the individual has a Relative with Active TB?
    - Relative with Active TB
    - No Relative with Active TB
3. Consider that the individual has No TB and is greater than 65 years old. How likely is it that we Treat for Latent TB Infection on arrival?
    - Treat
    - Don't Treat
4. Consider that the individual has No TB and is less than 5 years old. How likely are the following outcomes for the Blood Test Result?
    - Positive
    - Negative
5. Consider that the individual has No TB and is between 30 and 65 years old. How likely are the following outcomes for the Blood Test Result?
    - Positive
    - Negative
6. Consider that the individual has Latent TB and is between 30 and 65 years old. How likely are the following outcomes for the Blood Test Result?

- Positive
- Negative

7. Consider that the individual has Latent TB and is greater than 65 years old. How likely are the following outcomes for the Blood Test Result?

   - Positive
   - Negative

8. Consider that the individual comes from Sub-saharan Africa, has a Relative with Active TB and is greater than 65 years old. How likely are the following as the individual's Current TB Infection Status?

   - No TB
   - Latent
   - Active

9. Consider that the individual comes from Sub-saharan Africa, has a Relative with Active TB and is between 15 and 30 years old. How likely are the following as the individual's Current TB Infection Status?

   - No TB
   - Latent
   - Active

10. Consider that the individual comes from Sub-saharan Africa, has a Relative with Active TB and is less than 5 years old. How likely are the following as the individual's Current TB Infection Status?

    - No TB
    - Latent
    - Active

11. Consider that the individual comes from Canada, the US, Australia or New Zealand, has no Relative with Active TB and is between 15 and 30 years old. How likely are the following as the individual's Current TB Infection Status?

    - No TB
    - Latent
    - Active

12. Consider that the individual comes from Canada, the US, Australia or New Zealand, has a Relative with Active TB and is between 15 and 30 years old. How likely are the following as the individual's Current TB Infection Status?

    - No TB
    - Latent
    - Active

13. Consider that the individual has Latent TB, has not been Treated and is less than 5 years old. How likely are the following outcomes for the individual's Future TB Status in 5 years time?

    - No TB
    - Latent
    - Active

14. Consider that the individual has Latent TB, has not been Treated and is between 30 and 65 years old. How likely are the following outcomes for the individual's Future TB Status in 5 years time?

    - No TB
    - Latent
    - Active

15. Consider that the individual has Latent TB, has been Treated and is less than 5 years old. How likely are the following outcomes for the individual's Future TB Status in 5 years time?

    - No TB
    - Latent
    - Active

16. Consider that the individual has Latent TB, has been Treated and is greater than 65 years old. How likely are the following outcomes for the individual's Future TB Status in 5 years time?

    - No TB
    - Latent
    - Active

17. Consider that the individual has Active TB, has been Treated and is between 15 and 30 years old. How likely are the following outcomes for the individual's Future TB Status in 5 years time?

    - No TB
    - Latent
    - Active

18. Consider that the individual has Active TB, has been Treated and is greater than 65 years old. How likely are the following outcomes for the individual's Future TB Status in 5 years time?

    - No TB
    - Latent
    - Active

A.5 Participant Communications

Dear XXXX,

On behalf of the Forecasting and Futures group of the Australian Defence Science and Technology Organisation (DSTO) and Monash University, I would like to invite you to participate in an online Delphi experiment for the elicitation of a probabilistic causal network model, called Bayesian networks, for TB risk and management. The experiment is aimed at investigating a methodology for eliciting such models from experts.

Delphi participants Participation is by invitation only and will consist of specialists in TB management. The aims of the experiment are to:

1. Explore the idea of generating Bayesian networks (BNs) through a Delphi process.
2. Explore how or if we can arrive at a consensus model for the structure of the TB management network.
3. Explore how to use experts to capture the probabilistic impact of the relevant factors on the TB management process and outcomes

Guidelines for the Participants Time commitment is approximately 2-4 hours total via electronic communication over a 8-day period, starting in early May. The process consists of:

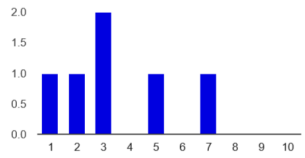
1. A preparatory brief quiz (approx. 12 questions) about an example causal model
2. An elicitation stage aimed at defining the network structure for managing TB where you will be asked to give your opinion regarding causal factors, strength of influence and your confidence in your assessment. This will consist of approx. 18 questions plus any comments you'd like to make.
    - You will be given an opportunity to view other anonymous participants' responses.
    - You will be asked to do a second round of the same questions. You may give the same or change your response.
3. An elicitation stage aimed at defining the network parameter values.
    - You will be given an opportunity to view what other anonymous participants have submitted
    - You will be asked to do a second round of the same questions. You may give the same or change your response.

Please indicate via return email to Emma McBryde (ADD EMAIL). your intention to participate to a delegate:

Yours sincerely,

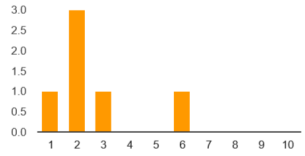
A.6 Evaluation Survey

Region of Origin: Sub Saharan Africa, Relative with Active TB: Yes, Age: 23, Current TB: Latent, Blood Test: Yes, Test Result: Negative, Treatment: Not Required, Future TB: Active



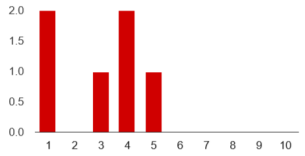
Impossibility: 1	**1**	16.7%
2	**1**	16.7%
3	**2**	33.3%
4	**0**	0%
5	**1**	16.7%
6	**0**	0%
7	**1**	16.7%
8	**0**	0%
9	**0**	0%
Certainty: 10	**0**	0%

Region of Origin: Canada, Relative with Active TB: No, Age: 19, Current TB: Latent, Blood Test: No, Test Result: , Treatment: , Future TB: Active



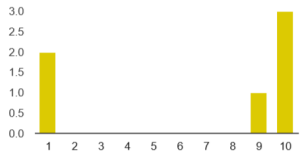
Impossibility: 1	**1**	16.7%
2	**3**	50%
3	**1**	16.7%
4	**0**	0%
5	**0**	0%
6	**1**	16.7%
7	**0**	0%
8	**0**	0%
9	**0**	0%
Certainty: 10	**0**	0%

Region of Origin: India, Relative with Active TB: Yes, Age: 25, Current TB: No TB, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: Active



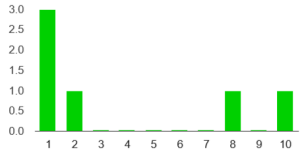
Impossibility: 1	**2**	33.3%
2	**0**	0%
3	**1**	16.7%
4	**2**	33.3%
5	**1**	16.7%
6	**0**	0%
7	**0**	0%
8	**0**	0%
9	**0**	0%
Certainty: 10	**0**	0%

Region of Origin: Europe, Relative with Active TB: No, Age: 35, Current TB: Active, Blood Test: No, Test Result: No Result, Treatment: Not Required, Future TB: Active



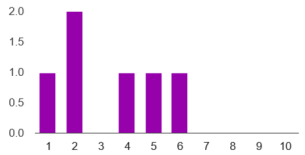
Impossibility: 1	**2**	33.3%
2	**0**	0%
3	**0**	0%
4	**0**	0%
5	**0**	0%
6	**0**	0%
7	**0**	0%
8	**0**	0%
9	**1**	16.7%
Certainty: 10	**3**	50%

Region of Origin: US, Relative with Active TB: Yes, Age: 71, Current TB: Active, Blood Test: Yes, Test Result: Negative, Treatment: Not Required, Future TB: Latent



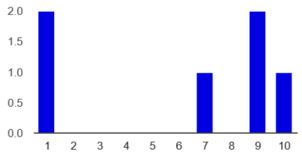
Impossibility: 1	**3**	50%
2	**1**	16.7%
3	**0**	0%
4	**0**	0%
5	**0**	0%
6	**0**	0%
7	**0**	0%
8	**1**	16.7%
9	**0**	0%
Certainty: 10	**1**	16.7%

Region of Origin: Sub Saharan Africa, Relative with Active TB: Yes, Age: 23, Current TB: Latent, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: Active



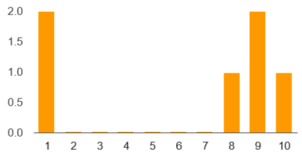
Impossibility: 1	**1**	16.7%
2	**2**	33.3%
3	**0**	0%
4	**1**	16.7%
5	**1**	16.7%
6	**1**	16.7%
7	**0**	0%
8	**0**	0%
9	**0**	0%
Certainty: 10	**0**	0%

Region of Origin: NZ, Relative with Active TB: No, Age: 47, Current TB: No TB, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: No Tb



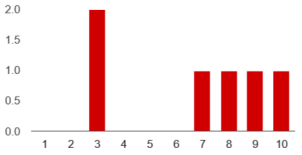
Impossibility: 1	**2**	33.3%
2	**0**	0%
3	**0**	0%
4	**0**	0%
5	**0**	0%
6	**0**	0%
7	**1**	16.7%
8	**0**	0%
9	**2**	33.3%
Certainty: 10	**1**	16.7%

Region of Origin: US, Relative with Active TB: No, Age: 16, Current TB: No TB, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: No TB



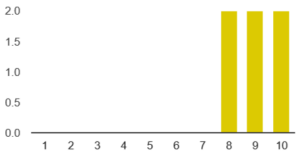
Impossibility: 1	**2**	33.3%
2	**0**	0%
3	**0**	0%
4	**0**	0%
5	**0**	0%
6	**0**	0%
7	**0**	0%
8	**1**	16.7%
9	**2**	33.3%
Certainty: 10	**1**	16.7%

Region of Origin: China, Relative with Active TB: Yes, Age: 20, Current TB: Latent, Blood Test: No, Test Result: No Result, Treatment: Not Required, Future TB: Latent



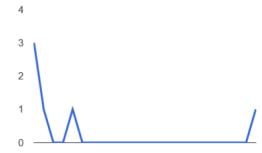
Impossibility: 1	**0**	0%
2	**0**	0%
3	**2**	33.3%
4	**0**	0%
5	**0**	0%
6	**0**	0%
7	**1**	16.7%
8	**1**	16.7%
9	**1**	16.7%
Certainty: 10	**1**	16.7%

Region of Origin: Middle East, Relative with Active TB: No, Age: 55, Current TB: Latent, Blood Test: Yes, Test Result: Positive, Treatment: Not Required, Future TB: Latent



Impossibility: 1	**0**	0%
2	**0**	0%
3	**0**	0%
4	**0**	0%
5	**0**	0%
6	**0**	0%
7	**0**	0%
8	**2**	33.3%
9	**2**	33.3%
Certainty: 10	**2**	33.3%

Number of daily responses



Latent TB Survey Form

We have developed a Bayesian network which models the latent problem in immigrants. The model assists in recognizing the high-risk immigrants and deciding who must be tested for latent TB.

We have created the following survey to test the model. Below are 10 hypothetical scenarios related to latent TB in immigrants. Each scenario represents an immigrant arrived in Australia in the last 6 months.

You are required to assess the probability of each case and respond with the closest option available.The responses of the experts will be compared to the responses of the network to assess its performance.

The variables used for the analysis are as follows:

1) Region of Origin: The country or region of origin of the immigrant being considered.
2) Relative with Active TB (Yes/No): Whether or not the immigrant has ever been in contact with a relative with Active TB
3) Age: The age of the immigrant being considered.
4) Current TB (No TB/Latent/Active): The current TB status of the immigrant in question i.e. the TB status on arrival
5) Blood Test (Yes/No): Was the immigrant tested for TB before arrival in Australia with the IGRA blood test
6) Test Result (Positive/Negative/No Result): The result of the test taken for latent TB. The result will be "No Result" if no test was taken.
7) Treatment (On Arrival/Not Required): Whether the immigrant was treated for TB after arrival in Australia
8) Future TB (No TB/Latent/Active): TB risk in the next 3-5 years.

Given the above variables you must decide the probability of each scenario occurring.

Respondent's Name

Respondent's Email

Region of Origin: Sub Saharan Africa, Relative with Active TB: Yes, Age: 23, Current TB: Latent, Blood Test: Yes, Test Result: Negative, Treatment: Not Required, Future TB: Active

	1	2	3	4	5	6	7	8	9	10	
Impossibility	○	○	○	○	○	○	○	○	○	○	Certainty

Region of Origin: Canada, Relative with Active TB: No, Age: 19, Current TB: Latent, Blood Test: No, Test Result: , Treatment: , Future TB: Active

	1	2	3	4	5	6	7	8	9	10	
Impossibility	○	○	○	○	○	○	○	○	○	○	Certainty

Region of Origin: India, Relative with Active TB: Yes, Age: 25, Current TB: No TB, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: Active

	1	2	3	4	5	6	7	8	9	10	
Impossibility	○	○	○	○	○	○	○	○	○	○	Certainty

Region of Origin: Europe, Relative with Active TB: No, Age: 35, Current TB: Active, Blood Test: No, Test Result: No Result, Treatment: Not Required, Future TB: Active

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Region of Origin: US, Relative with Active TB: Yes, Age: 71, Current TB: Active, Blood Test: Yes, Test Result: Negative, Treatment: Not Required, Future TB: Latent

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Region of Origin: Sub Saharan Africa, Relative with Active TB: Yes, Age: 23, Current TB: Latent, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: Active

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Region of Origin: NZ, Relative with Active TB: No, Age: 47, Current TB: No TB, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: No Tb

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Region of Origin: US, Relative with Active TB: No, Age: 16, Current TB: No TB, Blood Test: Yes, Test Result: Positive, Treatment: On Arrival, Future TB: No TB

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Region of Origin: China, Relative with Active TB: Yes, Age: 20, Current TB: Latent, Blood Test: No, Test Result: No Result, Treatment: Not Required, Future TB: Latent

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Region of Origin: Middle East, Relative with Active TB: No, Age: 55, Current TB: Latent, Blood Test: Yes, Test Result: Positive, Treatment: Not Required, Future TB: Latent

1 2 3 4 5 6 7 8 9 10

Impossibility ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Certainty

Submit

Never submit passwords through Google Forms.