Received: 27 May 2020 | Revised: 13 October 2020 | Accepted: 2 November 2020

DOI: 10.1111/medu.14407

THE CROSS-CUTTING EDGE

The potential use of Bayesian Networks to support committee decisions in programmatic assessment

Nathan Zoanetti[1] | **Jacob Pearce**[2]

[1]Psychometrics and Methodology, Australian Council for Educational Research, Camberwell, Vic., Australia

[2]Tertiary Education (Assessment), Australian Council for Educational Research, Camberwell, Vic., Australia

Correspondence
Nathan Zoanetti, Research Director, Psychometrics and Methodology, Australian Council for Educational Research, 19 Prospect Hill Rd, Camberwell VIC, 3124, Australia.
Email: Nathan.Zoanetti@acer.org

Abstract

Context: The benefits of programmatic assessment are well-established. Evidence from multiple assessment formats is accumulated and triangulated to inform progression committee decisions. Committees are consistently challenged to ensure consistency and fairness in programmatic deliberations. Traditional statistical and psychometric techniques are not well-suited to aggregating different assessment formats accumulated over time. Some of the strengths of programmatic assessment are also vulnerabilities viewed through this lens. While emphasis is often placed on data richness and considered input of qualified experts, committees reasonably wish for practical, defensible solutions to these challenges.

Methods: We draw upon on existing literature regarding Bayesian Networks (BN), noting their utility and application in educational systems. We provide illustrative examples of how they could potentially be used in contexts that embed programmatic principles. We show a simple BN for a knowledge domain before presenting a full-scale 'proof of concept' BN to support committee decisions. We zoom in on one 'node' to demonstrate the capacity of incorporating disparate evidence throughout the network.

Conclusions: Bayesian Networks offer an approach that is theoretically well-supported for programmatic assessment. They can aid committees in managing evidence accumulation, help them make inferences under conditions of uncertainty, and buttress decisions by adding a layer of defensibility to the process. They are a pragmatic tool adding value to the programmatic space by applying a complementary statistical framework. We see four major benefits of BNs in programmatic assessment: BNs allow for visual capturing of evidentiary arguments by committees during decision-making; 'recommendations' from probabilistic pathways can be used by committees to confirm their qualitative judgments; BNs can ensure precedents are maintained and consistency occurs over time; and the imperative to capture data richness is maintained without resorting to questionable methodological strategies such as adding qualitatively different things together. Further research into their feasibility and robustness in practice is warranted.

1 | INTRODUCTION

The benefits of taking a programmatic approach to assessment are now well-established in the medical education literature.[1-4] Increasingly, medical schools and specialist training colleges use programmatic principles as part of their assessment frameworks. One core principle of programmatic assessment is that decision-making regarding student progression or promotion is offset from single assessment moments, which are treated as single data points. Information is aggregated about given domains of content or specific competencies from disparate assessment events, and the information is reviewed in combination rather than in isolation. Evidence from multiple forms of assessment is accumulated and triangulated to inform committee or progression panel decision-making. These committees use the accumulated assessment information (which must remain nuanced, rich and meaningful) to inform their deliberations and ensure that progression decisions are made with reference to a substantive evidence base.

Assessment committees are consistently challenged to ensure consistency and fairness in programmatic deliberations. Indeed, it is the operationalisation of programmatic assessment, and specifically the element of decision-making that requires further research. There is a dearth of research on this topic to date.[5] Although the arsenal of statistical and psychometric techniques for quality assuring traditional assessment formats in medical education is well-established,[6] these tools are not well-suited to aggregating different forms of assessment accumulated over time. Some of the above strengths of programmatic assessment can also be considered as its vulnerabilities when viewed through a lens of reliability and reproducibility. While we acknowledge this re-introduces a post-positivist lens on the problem, we argue that this is an important lens that has been overlooked. After all, one of the drivers of the programmatic model was an argument based on reliability and sampling. Fairness and consistency in assessment remain an important driver of assessment research and innovations in practice, and it is important that the drive does not result in pre-psychometric mindsets.[7,8]

From this psychometric perspective, there are multiple vulnerabilities in operationalising a programmatic approach. First, programmatic approaches promote the aggregation of information about a given domain of content (eg anatomy) or about a specific competency (eg diagnosis) from disparate assessment events. Proponents argue for the meaningful triangulation of rich, nuanced information,[9] but this can be difficult to execute in practice. For example, there are issues with how to triangulate different formats of information where sources of data are captured in different contexts: such as narrative feedback from supervisors in clinical contexts[10]; enstrustability ratings from work-based assessments[11]; and sources of information from higher-stakes assessment tasks (such as traditional hurdle examinations). These sources of evidence cannot readily be added together, yet a programmatic approach encourages them to be considered in combination rather than in isolation. While traditional test theory models provide a useful framework for constructing and quality assuring specific measures within an assessment program, they do not directly address the problem of aggregating conceptually distinct measures. The problem of how best to combine scores from different assessments is well-known but relatively under-researched,[12] with various reliability, validity and error rate trade-offs arising from different combinations of arithmetic and logical rules.[13] Specific shortcomings, such as where weighted sum scores across multiple measures might lead to different decisions being made for equally capable trainees, or vice versa, become more likely with overly simplistic aggregation approaches. Even less commonplace is research into approaches for combining assessment scores with more qualitative, categorical assessment information in a decision-making context, such as might arise in programmatic assessment systems. At certain points in time during a training program, binary decisions pertaining to progression are required. These binary decisions in a programmatic paradigm involve aggregation of a variety of disparate sources of evidence across multiple domains or competencies.

As a second specific challenge, in programmatic assessment, there is an assumption of continuous learning, intentionally supported by more frequent, diagnostic and meaningful feedback.[14,15] This temporal dimension to the development of multiple competencies may complicate attempts to use traditional test theory models to support making inferences, particularly in the context of multifarious assessment evidence. Although numerous longitudinal psychometric approaches are available, these models may not have the flexibility to simultaneously address the evidence accumulation challenge described in the previous paragraph. The feasibility of these models requires further evaluation. These challenges highlight the need for sufficiently flexible yet robust statistical frameworks in a programmatic assessment context.

The challenges above are not unheard of in other assessment contexts.[16] In educational assessment specific to medical education, these challenges have already been echoed by Schuwirth and van der Vleuten in a plea for new, conceptually different psychometric models.[17] In response to these challenges, committees reasonably wish for practical, defensible solutions to ensuring consistency and fairness. This is where numerous tensions arise. The programmatic approach champions the richness of the data and the considered input of suitably qualified committees of experts to marshal these data into an evidentiary argument about whether the trainee is ready to progress. With this kind of approach, it is difficult to ensure that consistent judgements are being made both within a cohort and over time. A further threat may arise as the composition of committees naturally changes over time. To address this issue, we explore the potential for utilising Bayesian Networks (BNs) as a pragmatic tool to inform programmatic deliberations and buttress committee decisions.

Bayesian Networks are probabilistic graphical models which provide a visual and statistical framework for reasoning under uncertainty.[18] They have been applied over several decades in disciplines including medicine,[19] agriculture,[20] finance,[21] ecology[22]

and information technology,[23] where they have fulfilled numerous functions including diagnosis, prediction, classification and decision-making. BNs are widely used because they enable flexible and intuitive modelling of uncertainty and complexity in almost any real-world system where alternative statistical models or rule-based algorithms prove inadequate or intractable.[24] Culbertson[25] provides a recent summary of their application in educational assessment contexts, and notes that, despite their flexibility, BNs have garnered relatively little attention from psychometricians to date. Although Schuwirth and van der Vleuten actually foreshadowed probabilistic and Bayesian statistical approaches to medical education assessment in their seminal 2006 paper,[17] we are not aware of the use of BNs to support implementation of a programmatic approach to assessment in medical education. We contend that BNs are well-suited to accommodating the diversity of evidence under consideration in a programmatic approach, and that they have the potential to address a glaring gap in credible methods to consistently and defensibly aggregate this evidence.

The purpose of this paper is to introduce the concept of BNs and to provide some illustrative examples of how they could potentially be used in medical education assessment contexts that embed programmatic principles. Research into validity frameworks for BNs is limited in educational contexts[26] and more work in this area is warranted, although this is not our focus in this paper. Instead, we outline how BNs can be used as heuristics that can add value to the benefits that come with thinking programmatically, while also drawing upon BNs as a way of overcoming some of the vulnerabilities of a programmatic approach. We argue that BNs can re-introduce an element of fairness and consistency, dealing statistically with data that is collected temporally and from disparate sources. BNs can aid committees with managing the accumulation of evidence, helping make inferences under conditions of uncertainty, and in buttressing decisions by way of adding a layer of defensibility. After providing a brief background to BNs, we proceed by way of providing some illustrative specific examples of how BNs can work in medical education contexts, emphasising their utility.

2 | BAYESIAN NETWORKS: A PRACTICAL SOLUTION FOR REASONING UNDER UNCERTAINTY

A BN is defined as a directed acyclic graph, meaning it consists of a number of nodes that represent random variables. These nodes are connected by edges or arcs which are represented as arrows. The direction of these arrows is commonly specified to indicate the direction of causality. Each node has an underlying conditional probability table or function that specifies how the probability of each possible state within that node (a so-called child node) depends on its so-called parent nodes (connections are directed from parent nodes to child nodes in a BN). The lack of a direct connection between two nodes represents conditional independence assumptions. A sparse set of edges can result in a compact representation of the joint probability distribution for a system that is intuitive to work with and that is computationally tractable. We provide illustrative examples of the probability theory underpinning BNs in the following section.

Much like in classical and modern test theory models, both observed and latent variables can be specified in a BN. For example, latent variables representing attributes of interest can be specified as parent nodes, and observations from assessment tasks can be specified as child nodes that are 'caused by' the parent nodes. Although the assumptions in many measurement models are too strict to work with a programmatic approach, BNs are a flexible statistical model that can accommodate complex relationships between variables, and are thus more applicable for programmatic approaches that embrace the sampling of diverse evidence at different points in time to appraise the development of competence.

2.1 | Simple BN for a knowledge domain in medical education

To make this theoretical introduction and the flexibilities of BNs more concrete, it is instructive to first explore a relatively simple BN showing the structure of the graph and the corresponding conditional probability tables underpinning it. The structures of the BNs demonstrated in this article have been specified using expert judgement and built using the Netica software package.[27] We begin by showing a simple BN for the Pathology knowledge domain in Figure 1. The adequacy of a trainee's Pathology knowledge is specified as a latent variable with two categories: adequate and inadequate. There are three pieces of observable assessment evidence including outcomes from the relevant components of a multiple-choice and a short-answer question examination, and a rating from a work-based assessment focusing on Pathology knowledge. The evidence in this example is of two different kinds, with examination score performance categories (though examination scores could be used) in addition to a more qualitative rating derived from a work-based assessment. We note here that a BN does not preclude the simultaneous use of particular assessment methods or test theory models within a programmatic system, for example using Item Response Theory to scale and equate examinations.

A BN allows disparate assessment information to be incorporated without resorting to methodologically questionable strategies such as adding qualitatively different things together.[28] These observable variables are modelled as being dependent on the latent Pathology knowledge variable and conditionally independent of each other. The structure of a BN can be specified by domain experts or, in some cases, it can be learned from a database using appropriate software. BNs can be employed either as confirmatory or exploratory tools.

Much like the structure of a BN, the conditional probabilities that underpin the network can be determined in different ways. They can be determined via theory or expert judgement, they can be learned from available assessment data (ie a database), or they

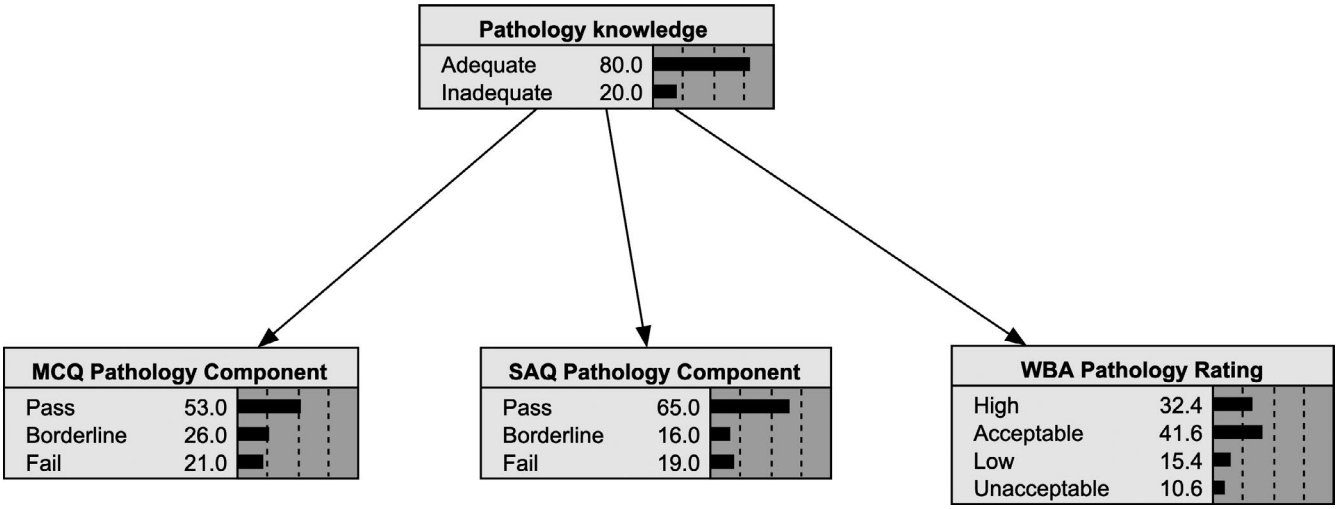

FIGURE 1 Simple BN for inferring Pathology domain knowledge from three assessments

can be continually refined using a combination of these approaches (for instance being updated as new data become available).[29] To produce the BN shown in Figure 1, expert judgement was applied to initially specify the network structure and to subsequently specify the conditional probabilities. First, we specified prior probabilities for the parent node (Pathology knowledge), where we specified an 80% likelihood that a trainee for whom no assessment evidence was yet available would have 'Adequate' Pathology knowledge. Second, we specified conditional probabilities for each of the child nodes. For example, the conditional probabilities that we specified for the WBA (Work-based assessment) Pathology Rating node are shown in Table 1.[26] These probabilities represent the likelihoods of the WBA Pathology Rating being observed in each of the possible states of 'High', 'Acceptable', 'Low' or 'Unacceptable' given that a trainee has an assumed level of 'Pathology knowledge' that is either 'Adequate' or 'Inadequate' in turn. In this example, we specified that a trainee with 'Adequate' Pathology knowledge would have a 40% likelihood or 0.40 probability of having an observed rating of 'High' on this particular WBA and only a 2% likelihood of having an observed rating of 'Unacceptable', and so on.

Once these probabilities have been specified, the BN can be instantiated. This results in the prior probabilities being calculated for each of the categories within each child node in the network. This calculation step applies the law of total probability and is completed automatically by BN software packages. For example, the value of 32.4% in Figure 1 for a WBA rating of 'High' is calculated by summing over the product of each Pathology Knowledge (PK) category probability with the conditional probability of a 'High' WBA rating given that PK category. This is shown in Equation 1.

$$\begin{aligned}\sum_{PK} P[\text{WBA}=\text{'High'}] &= (P[\text{WBA}=\text{'High'}|\text{PK}=\text{'Adequate'}]\times P[\text{PK}=\text{'Adequate'}]) \\ &\quad +(P[\text{WBA}=\text{'High'}|\text{PK}=\text{'Inadequate'}]\times P[\text{PK}=\text{'Inadequate'}]) \\ \sum_{PK} P\left[\text{WBA}=\text{'High'}\right] &= (0.4\times 0.8)+(0.02\times 0.2)=0.324\end{aligned} \tag{1}$$

Figure 2 illustrates an important concept. As observations are entered into a BN, a process referred to as 'belief propagation' occurs. In this process, the probabilities of other nodes in the network are updated to reflect the current belief about the most likely state of the system. More technically, this process updates the posterior probability distribution of variables in the network. In this example, our prior belief about a trainee for whom we have no assessment information is that there is 80% likelihood that they have adequate knowledge of pathology. After one unfavourable observation on the WBA, our belief that their knowledge is adequate reduces to 41.6% likelihood. This calculation, which is an application of Bayes' Theorem, is shown in Equation 2.

$$\begin{aligned}P[\text{PK}=\text{'Adequate'}|\text{WBA}=\text{'Low'}] &= \frac{P[\text{WBA}=\text{'Low'}|\text{PK}=\text{'Adequate'}]\times P[\text{PK}=\text{'Adequate'}]}{P[\text{WBA}=\text{'Low'}]} \\ P[\text{PK}=\text{'Adequate'}|\text{WBA}=\text{'Low'}] &= \frac{0.08\times 0.80}{0.154}=0.416\end{aligned} \tag{2}$$

The belief propagation process then updates the likelihoods in all other nodes for which an observation has not yet been recorded. This process again applies the law of total probability, but now uses the updated likelihood of 'Adequate' Pathology knowledge of 41.6% as opposed to the 80% value that was used in Equation 1. Readers interested in further technical details beyond those covered in this article are encouraged to refer to Pearl's seminal text[18] or Charniak.[30]

Pathology knowledge	High	Acceptable	Low	Unacceptable
Adequate	0.40	0.50	0.08	0.02
Inadequate	0.02	0.08	0.45	0.45

TABLE 1 Conditional probabilities for the WBA Pathology Rating node given each possible state of Pathology knowledge

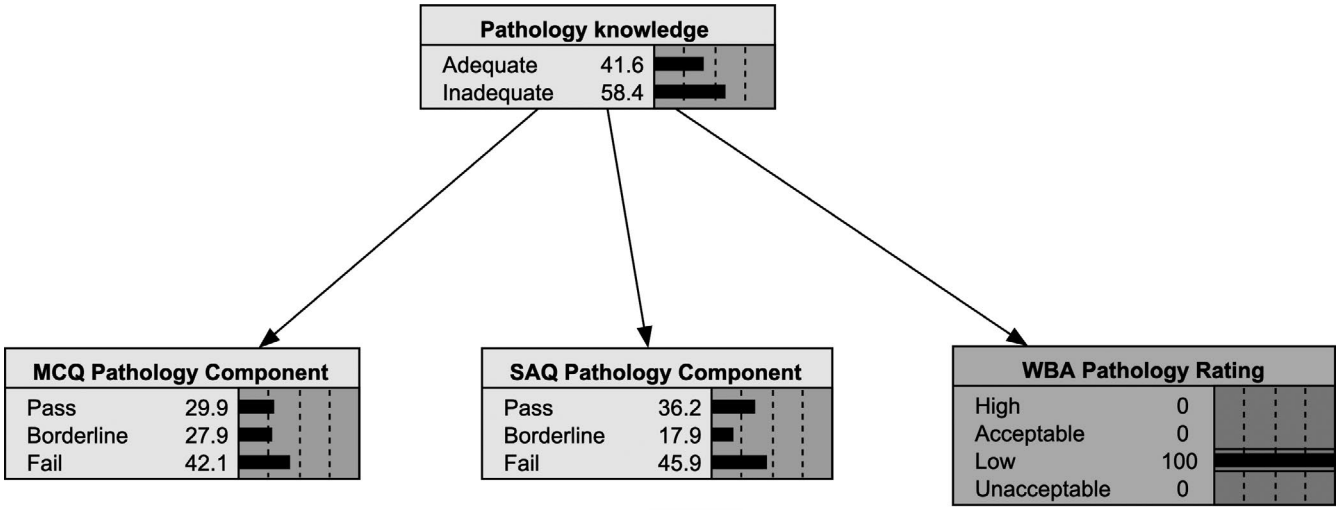

FIGURE 2 Simple BN with one observation instantiated showing belief propagation through the network

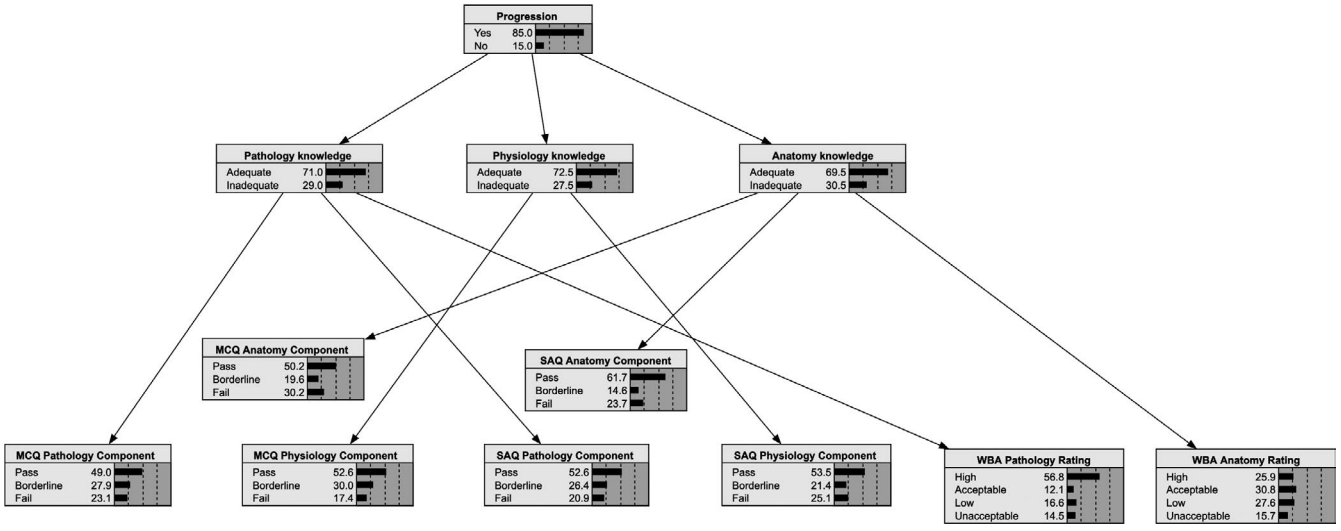

FIGURE 3 Possible BN for an overall progression decision relating to knowledge of three basic sciences domains

As more observations are entered into a BN, the uncertainty around the most likely state of the trainee's pathology knowledge reduces. This is analogous to score reliability increasing with increased sampling of questions or cases and can provide insight into the incremental value of additional observations in the decision-making process. This also provides an opportunity for experts to review the extent to which the reasoning captured in the BN and the relative weight given to individual observations accords with their expectations. This process of BN evaluation and refinement can be undertaken both qualitatively and via formal statistical model criticism techniques.[31]

An important feature of the first example BN shown in Figure 1 and Figure 2 is that it can be treated as what is known as a BN fragment. This means that this small BN can be combined with other BN fragments, for example analogous fragments for anatomy and physiology, thus providing a full BN that could be applied in the context of progression decisions relating to basic sciences. This extensibility is illustrated in Figure 3, which shows a BN for a decision-making point in specialty training progression. The binary decision is whether the trainee is safe to progress to the next phase of clinical training. The committee must take into account the trainee's performance in three knowledge domains (pathology, physiology and anatomy), each of which is comprised of assessment information from different formats (both examination components and WBA ratings).

3 | ASSEMBLING BNS TO ADDRESS CHALLENGES IN PROGRAMMATIC ASSESSMENT

Now that several of the main building blocks of BNs have been presented, it is worth briefly summarising how these can be potentially assembled to address situations and challenges that may arise in a programmatic approach to assessment. The following provides an explanation of BN affordances that are aligned to characteristics

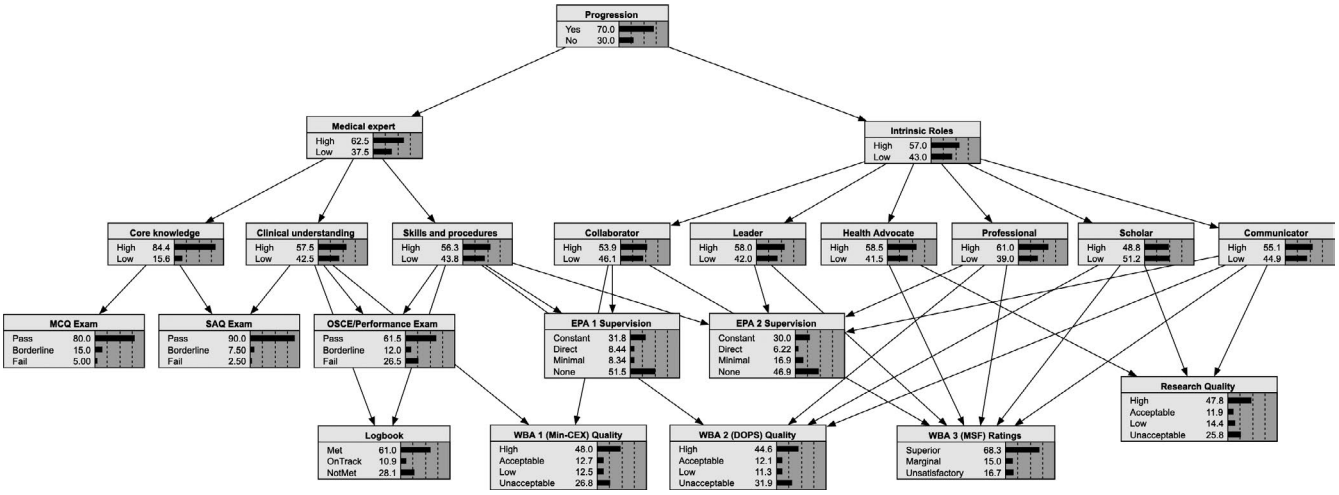

FIGURE 4 Possible BN for an overall fellowship progression decision in a programmatic assessment system

of a programmatic approach: multidimensional skills and attributes; complex dependencies between skills, attributes and evidence; inter-dependencies between multiple scores or judgements derived from the same task; missing observations; and learning or improvement over time.

A BN approach is well-suited to dealing with the need to consider data from multiple sources. Programmatic approaches to assessment generally aim to consider a breadth and diversity of knowledge, skills and attributes, such as those described in the CanMEDs framework. This breadth and diversity implies that the accompanying program of assessment and associated processes ought to be conceptualised as multidimensional, with, from a measurement perspective, multiple latent variables under consideration simultaneously. The examples so far show that this is readily accommodated in BNs, with incorporation of multiple parent nodes, and optionally with dependencies between parent nodes specified too.

Bayesian Networks can deal with complex dependencies between skills, attributes and evidence. Programmatic committees must account for multiple, disparate attributes and observations with the additional challenge of properly capturing the nature of the relationships between these variables. Furthermore, there may be a requirement to impose certain policies that dictate hard rules or deterministic relationships between these variables. Fortunately, BNs provide considerable flexibility in these respects. For example, Almond[32] explains that the relationship between variables and their influence on assessment observations can be modelled as conjunctive, disjunctive or compensatory, among other possibilities. Finally, it is possible to fix certain probabilities in the conditional probability tables to be zero or one,[18] meaning that a deterministic rule or policy can be instituted in the BN (eg if a candidate fails a specific task, then they cannot progress).

BNs can handle inter-dependencies between multiple scores or judgements derived from the same task. In assessment situations where a single task or stimulus contributes multiple observations about a trainee, it is important to consider whether strong dependencies between the observations have been introduced. In such cases, it may not be appropriate to count each observation separately towards a final score or outcome as if each contributed independent information about the attributes being assessed. Instead, the shared variation in scores across the observations should be accounted for.[33] Like several other test theory models, these dependencies can be explicitly modelled in a BN.[34]

Missing data or assessment evidence that has gaps (due to trainee rotations, for example) can be managed by BNs. This is because BNs do not require complete data in order to support the probabilistic inference processes described so far. This is an attractive feature in that the current belief about a trainee's proficiencies and attributes can be updated progressively and in real-time as new assessment information becomes available. Also, for programs looking to learn conditional probabilities from historical data, parameter estimation processes that can accommodate missing data are available in most if not all popular BN software tools.

Finally, BNs can effectively track learning progressions over time. BNs afford the possibility of making inferences about the same collection of skills and attributes at separate points in time in a medical education program (eg years or stages). Versions of BNs that are sometimes referred to as Dynamic Bayes Nets (DBNs) (and the process of Bayesian Knowledge Tracing (BKT) more generally) are prominent in the Intelligent Tutoring System (ITS) and Learning Analytics research communities.[35] While we do not present an example of a DBN or BKT in this paper, we note their potential applicability for training programs that may seek to explicitly model changes in skills and attributes over time.[36]

3.1 | BNs to support decision-making in a programmatic system

Figure 4 presents a full-scale BN model for a programmatic assessment system as a proof of concept. A high-stakes decision—in this example, progression to Fellowship in a specialty training program—is made based on a constellation of evidentiary information from

disparate sources across the spectrum of assessment approaches. How might BNs support progression committees in their decision-making in such truly programmatic contexts?

The progression committee reviews assessment information and makes a series of informed, collective judgments about the trainee. These judgments are resolved into different categorical scales, depending on the assessment. In this example, Work-based assessment tasks and research components use an 'acceptability scale'; EPAs use an 'entrustability scale' relating to the amount of supervision required; the Logbook uses a scale relating to progress; and the examination components are resolved into broad performance level ratings. The committee needs to make a series of judgments about whether the trainee's learning, at this point in time, has reached the appropriate level to grant fellowship status.

Blueprinting and constructive alignment principles ensure that each assessment moment is aligned to one of the CanMEDs competencies, either Medical Expert or the six 'Intrinsic Roles'. In this example, Medical Expert is broken down further into 'Core Knowledge', 'Clinical Understanding' and 'Skills and Procedures'. These domains along with the six Intrinsic Roles allow for nine overall categorisations of performance. The final decision for the committee is a binary one—does this trainee meet the requirements for progression to fellowship? This process is a typical one followed in programmatic assessment committees. The development of a BN to visually capture this process can occur independently. Thus, the BN can become a powerful statistical tool based on a probabilistic framework to confirm whether the final judgment by the committee is consistent with precedent and mitigate against potential bias.

3.2 | Incorporating rich evidence into nodes

None of the judgments above remove the detailed or 'rich' information that may have been collected—WBAs might be populated with detailed narrative comments; Multi-Source Feedback (MSF) may have specific and nuanced information that a trainee used to inform their learning. This more nuanced information can actually be incorporated into the BN nodes, coded as a qualitative judgment based on a rubric, without the need for potentially arbitrary numerical operations. The raw data can be maintained separately in its more meaningful format for review by the committee if required. The more detailed 'fragment' in Figure 5 can be added to the full programmatic network in Figure 4 and feed into the final progression decision.

4 | DISCUSSION: BENEFITS OF UTILISING BNS AS A HEURISTIC TOOL

The methodological gap we are addressing is the current lack of theoretically appropriate statistical methods for accumulating different observations and measures to support decision-making in programmatic assessment systems. Bayesian Networks are theoretically well-suited to support this process of substantive reasoning under conditions of uncertainty. The examples shown are all 'proof of concept'—other instantiations of networks could be readily developed for different medical education assessment contexts with particular specificities and idiosyncrasies. However, we argue that BNs have immense potential in the programmatic space as they add a complementary probabilistic framework (one

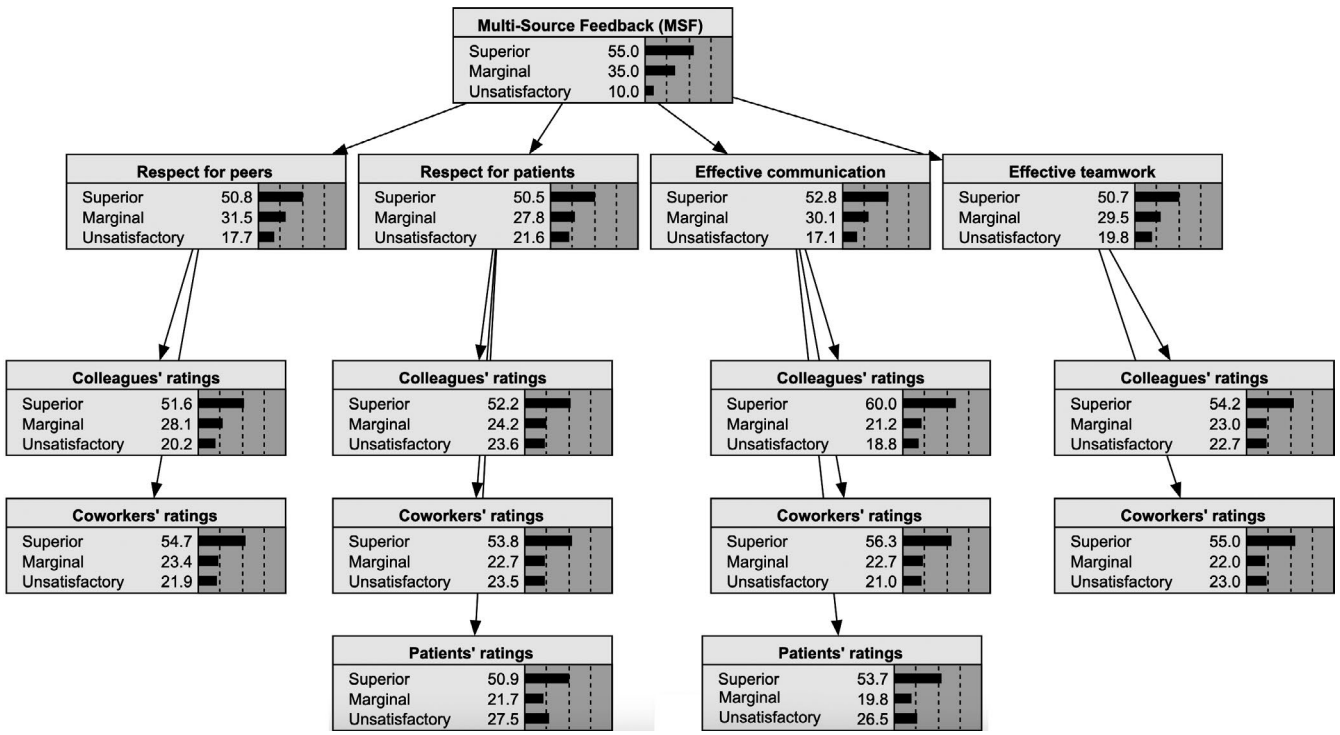

FIGURE 5 Possible BN drilling down on MSF ratings

even foreshadowed by Schuwirth and van der Vleuten)[17] for guiding or reviewing committee decisions. There is the matter that some medical education programs attract small candidatures, and therefore, the feasibility of BNs needs to be proven in programs that differ in size and other respects. This is a limitation worth considering. Similarly, we see a need for further research into the feasibility and robustness of BNs specified with different types of latent variables, at different levels of granularity, and incorporating observations that are of different levels of quality. Nevertheless, we see four major potential benefits of BNs in programmatic assessment.

First, BNs allow for the visual capturing of evidentiary arguments in decision-making processes. That is, at their very simplest, a BN allows for the recording of committee decisions in context. Programmatic committees must marshal a vast array of disparate forms of evidence of trainee progression. BNs articulate the interconnectivity of assessments across time, explicitly show the relationships between certain assessment moments and framework components (such as competencies), and visually demonstrate how decisions propagate and impact on final progression decisions. In one trial instantiation with a specialist medical college, we triangulated assessment data across assessment formats and aggregated by curriculum modules and proficiency domains. In the candidate review process, the decision to award/not award fellowship incorporated rich evidence across assessment formats via detailed interrogation of performance. A BN was built to visually capture the process and highlight transparency of decisions.

Second, committees can confirm their qualitative judgments against quantitative 'recommendations' from the probabilistic pathways that emerge from the BN. BNs provide not only a record of the evidentiary trail that led to final decisions; BNs can be fed with real data over time. Data from previous committee reviews can be fed in, allowing for probability calculations to emerge regarding the likelihood of Pass/Fail outcomes based on multiple decision-making points. The BN is thus a form of an 'expert system' that recommends a final outcome. This does not imply that BNs should become a type of 'black box' or artificial intelligence machine that replaces the process of committee decision-making. Instead, we see the BN as another tool in the arsenal of evidence available to the panel to add a layer of defensibility to the process. Transparency regarding how the BN is built, and justification in how it is employed remains crucial. Here, we echo the call for the judicious use of psychometric information in assessment (rather than the blind acceptance of it), and comprising part of an evidentiary framework.[8] Further, as there are established statistical techniques[37] available for evaluating and refining BNs based on real data, there is scope for critiquing the quality of assessment evidence included in each node; reviewing the reasonableness of conditional independence assumptions in the network structure; checking the stability of the network structure and conditional probabilities over time; and, for validating and continually improving a programmatic system using BNs. Such evaluative information will feed back into the whole assessment approach and ultimately enhance the decision-making process over time.

Third, BNs can be used to ensure that precedents are maintained. In other words, the same constellation of evidence would lead to the same decision, either for different trainees or at different points in time. Consistency of decisions over time is a major vulnerability of the programmatic approach, with potential for bias from committee members and even changing committee membership over time. Does the BN recommendation align with the decision of this committee? If yes, then a further check that precedent is being maintained and consistency (in the form of fairness to trainees) is occurring over time. If no, then something may have gone awry somewhere in the process and further interrogation of the data by the committee is warranted before finalising a decision. The way these processes flow will depend on context-specific guidelines for decision-making relating to the alignment of the BN outcomes.

Fourth, BNs avoid artificial and arbitrary assignment of numbers to ratings so that data can be more easily aggregated from disparate sources. Instead, different qualitative judgments from different rating scales can be coded according to a rubric, while still being maintained in a pure form elsewhere. Such coding of nodes can retain qualitatively distinct categories that reflect granular performance categories and do not require reducing the rich information from different assessment strategies to numbers. This is particularly relevant for programs that are collecting large volumes of assessment evidence across years (such as residency programs in the United States). If data collection is centralised across specialities, the burden of data accumulation is effectively managed by the BN and potentially distilled in a way that is helpful to educators. Using BNs maintains the imperative in programmatic assessment to capture the richness of the data and, more importantly, the decision-making points that are enacted by committees.

5 | CONCLUSION

There are multiple potential benefits of utilising BNs as a pragmatic tool where programmatic assessment is implemented. We have drawn upon BNs as a means for overcoming several of the perceived vulnerabilities in implementing a programmatic approach to assessment. BNs are able to deal statistically with disparate forms of data collected over time. And yet BNs do not reduce the rich information from different assessment strategies to numbers. Instead, they retain much of the richness of the data and, more importantly, the decision-making points that are enacted by committees. Evidentiary arguments for final Pass/Fail decisions are visually captured; panels can confirm their qualitative judgments against quantitative 'recommendations' from the probabilistic pathways; precedents from previous years can be maintained, ensuring fairness to cohorts of trainees; and the artificial and arbitrary assignment of numbers or weightings to assessment data is avoided.

Although BNs offer an approach that is theoretically well-supported for programmatic assessment, further research into their feasibility and robustness in practice is warranted. Many questions emerge from thinking about their application to programmatic assessment. For instance: How much data is needed to reach saturation to make BNs useful? How much trialling across cohorts is required before the approach is defensible? How are repeated measurements and change over time optimally modelled using BNs or DBNs? What is the impact of curricular or assessment changes on accumulated data sets? We call for specific research in programmatic assessment contexts on these, and other, operational issues. Prior to full-scale studies, simulation-based research approaches may be particularly useful.

Overall, BNs have the potential to become a powerful tool to add a layer of defensibility to the process of decision-making in programmatic assessment. Systematic approaches to ensuring consistency, transparency, fairness and ultimately, defensibility of decision-making in programmatic assessment can be readily applied. Although further research into their feasibility is required, BNs have the potential to aid programmatic assessment committees with managing the accumulation of evidence, supporting the process of making inferences under conditions of uncertainty, and in enhancing the robustness of a programmatic approach to assessment in medical education.

ACKNOWLEDGEMENTS

We would like to thank Dr Kate Reid and Dr Ling Tan at ACER, and Medical Education editors and reviewers for helpful comments on earlier versions of this manuscript.

AUTHOR CONTRIBUTIONS

Both authors contributed substantially to the conception, research and development of this piece of work. Both authors drafted and edited the piece. Both approved the final version. Both authors agree to be accountable for all aspects of the work.

ORCID

Nathan Zoanetti https://orcid.org/0000-0002-0296-8424
Jacob Pearce https://orcid.org/0000-0003-3081-9132

REFERENCES

1. van der Vleuten C, Heeneman S, Schuwirth L. Programmatic Assessment. In: Dent J, Harden R, Hunt D, eds. *A Practical Guide for Medical Teachers*. Edinburgh, UK: Elsevier; 2017:295-303.
2. van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39:309-317.
3. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve Tips for programmatic assessment. *Med Teach*. 2015;37(7):641-646.
4. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214.
5. Tweed M, Wilkinson T. Student progress decision-making in programmatic assessment: can we extrapolate from clinical decision-making and jury decision-making? *BMC Med Educ*. 2019;19(1). https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-019-1583-1
6. Schauber SK, Hecht M, Nouns ZM. Why assessment in medical education needs a solid foundation in modern test theory. *Adv Health Sci Educ*. 2018;23(1):217-232.
7. ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94(3):333-337.
8. Pearce J. In defence of constructivist, utility-driven psychometrics for the 'post-psychometric era'. *Med Educ*. 2020;54(2):99-102.
9. Uijtdehaage S, Schuwirth LWT. Assuring the quality of programmatic assessment: Moving beyond psychometrics. *Perspect Med Educ*. 2018;7(6):350-351.
10. Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. *Acad Med*. 2017;92(11):1617-1621.
11. Cate O. When I say … entrustability. *Med Educ*. 2020;54(2):103-104.
12. McBee MT, Peters SJ, Waterman C. Combining scores in multiple-criteria assessment systems: the impact of combination rule. *Gifted Child Q*. 2014;58(1):69.
13. Kane M, Case SM. The reliability and validity of weighted composite scores. *Appl Measur Educ*. 2004;17(3):221-240.
14. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-485.
15. Ajjawi R, Regehr G. When I say … feedback. *Med Educ*. 2019;53(7):652-654.
16. Bennett R, Persky H, Weiss A, Jenkins F. *Problem Solving in Technology-Rich Environments: A Report From the NAEP Technology-Based Assessment Project (NCES 2007–466)*. Washington, DC: U.S. Department of Education. National Center for Education Statistics; 2007.
17. Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ*. 2006;40(4):296-300.
18. Pearl J. *Morgan Kaufmann Series in Representation and Reasoning. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann; 1988.
19. Heckerman D. *Probabilistic Similarity Networks*. Cambridge, MA: MIT Press; 1991.
20. Drury B, Valverde-Rebaza J, Moura M-F, de Andrade LA. A survey of the applications of Bayesian networks in agriculture. *Eng Appl Artif Intell*. 2017;65:29-42.
21. Neapolitan RE, Jiang X. *Probabilistic Methods for Financial and Marketing Informatics*. Elsevier Inc.; 2007.
22. Carriger JF, Yee SH, Fisher WS. An introduction to Bayesian networks as assessment and decision support tools for managing coral reef ecosystem services. *Ocean Coast Manag*. 2019;177:188-199.
23. Horvitz E, Breese J, Heckerman D, Hovel D, Rommelse K. The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence; 1998.
24. Pourret O, Naim P, Marcot B, eds. *Bayesian Networks: A Practical Guide to Applications*. John Wiley & Sons Ltd; 2008.
25. Culbertson MJ. Bayesian networks in educational assessment: the state of the field. *Appl Psychol Meas*. 2016;40(1):3-21.
26. Pitchforth J, Mengersen K. A proposed validation framework for expert elicited Bayesian Networks. *Expert Syst Appl*. 2013;40(1):162-167.
27. Netica Application for Belief Networks and Influence Diagrams 1995–2020 [Internet]. Norsys Software Corp. www.norsys.com. Accessed 14 May 2020.
28. Schuwirth L, van der Vleuten C. How to design a useful test: the principles of assessment. In: Swanwick T, Forrest K, O'Brien B. eds. *Understanding Medical Education: Evidence, Theory and Practice*. Oxford: The Association for the Study of Medical Education; 2018:275-289. https://doi.org/10.1002/9781119373780.ch20

29. Almond RG, Mislevy RJ, Steinberg LS, Yan D, Williamson DM. *Bayesian Networks in Educational Assessment*. Springer; 2015.
30. Charniak E. Bayesian networks without tears. *AI Magazine*. 1991;12:4.
31. Sinharay S. Model diagnostics for bayesian networks. *J Educ Behav Stat*. 2006;31(1):1-32.
32. Almond RG. I can name that Bayesian network in two matrixes!. *Int J Approx Reason*. 2010;51(2):167-178.
33. Yen W. Scaling performance assessments: strategies for managing local item dependence. *J Educ Measure*. 1993;30(3):187-213.
34. Almond RG, Mulder J, Hemat LA, Yan D. Bayesian network models for local dependence among observable outcome variables. *J Educ Behav Stat*. 2009;34:491-521.
35. Pardos ZA, Bergner Y, Seaton D, Pritchard D. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. In: Conference: 6th International Conference on Educational Data Mining [Internet]. 2013. http://educationaldatamining.org/EDM2013/proceedings/paper_20.pdf. Accessed 9 May 2020.
36. Reichenberg R. Dynamic bayesian networks in educational measurement: reviewing and advancing the state of the field. *Appl Measur Educ*. 2018;4:335-350.
37. Marcot B. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecol Model*. 2012;230:50-62.

How to cite this article: Zoanetti N, Pearce J. The potential use of Bayesian Networks to support committee decisions in programmatic assessment. *Med Educ*. 2020;00:1–10. https://doi.org/10.1111/medu.14407