

Decision Support using Bayesian Networks for Clinical Decision Making

Oluwole Victor Ogunsanya



University of London

Thesis submitted for the degree of Doctor of Philosophy
at Queen Mary, University of London

April 2012

Declaration of originality

I certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

.....

Oluwole V Ogunsanya

.....

Date

Abstract

This thesis investigates the use of Bayesian Networks (BNs), augmented by the Dynamic Discretization Algorithm, to model a variety of clinical problems. In particular, the thesis demonstrates four novel applications of BN and dynamic discretization to clinical problems.

Firstly, it demonstrates the flexibility of the Dynamic Discretization Algorithm in modeling existing medical knowledge using appropriate statistical distributions. Many practical applications of BNs use the relative frequency approach while translating existing medical knowledge to a prior distribution in a BN model. This approach does not capture the full uncertainty surrounding the prior knowledge.

Secondly, it demonstrates a novel use of the multinomial BN formulation in learning parameters of categorical variables. The traditional approach requires fixed number of parameters during the learning process but this framework allows an analyst to generate a multinomial BN model based on the number of parameters required.

Thirdly, it presents a novel application of the multinomial BN formulation and dynamic discretization to learning causal relations between variables. The idea is to consider competing causal relations between variables as hypotheses and use data to identify the best hypothesis. The result shows that BN models can provide an alternative to the conventional causal learning techniques.

The fourth novel application is the use of Hierarchical Bayesian Network (HBN) models, augmented by dynamic discretization technique, to meta-analysis of clinical data. The result shows that BN models can provide an alternative to classical meta analysis techniques.

The thesis presents two clinical case studies to demonstrate these novel applications of BN models. The first case study uses data from a multi-disciplinary team at the Royal London hospital to demonstrate the flexibility of the multinomial BN framework in learning parameters of a clinical model. The second case study demonstrates the use of BN and dynamic discretization to solving decision problem. In summary, the combination of the Junction Tree Algorithm and Dynamic Discretization Algorithm provide a unified modeling framework for solving interesting clinical problems.

To my loving wife Oluwatoyin and my children Precious & Emmanuel

Contents

1	Introduction	18
1.1	Research Hypotheses	22
1.2	Structure of the Thesis	25
2	Bayes and Decision Analysis	26
2.1	Background	26
2.1.1	Notation	27
2.1.2	Bayes' Theorem	27
2.1.3	Bayesian & Frequentist approach to Data Analysis	29
2.1.4	Bayesian Inference	30
2.2	Decision Trees	32
2.3	Bayesian Networks	34
2.3.1	Conditional Independence Assertion	38
2.3.2	Knowledge representation in BNs	40
2.3.3	Constructing a BN	44
2.3.4	Variable selection	44
2.3.5	Types of variables commonly used in a BN model	46
2.3.6	Construct the DAG (qualitative part)	47
2.3.7	Assess the conditional probabilities (quantitative part)	47
2.4	Inference in Bayesian Networks	47
2.4.1	Junction Tree Algorithm	50
2.4.2	Potential function	51
2.4.3	Message Propagation on Junction Tree	51
2.4.4	Projection of Message from $V_1 \rightarrow V_2$	52
2.4.5	Absorption of Message into V_2	53
2.5	Learning parameters & structures of probabilistic Networks	54
2.6	Software Packages for Graphical Models	55

2.7	Application of BN models to clinical decision making	57
2.7.1	Bayesian classification model	59
2.7.2	A BN model to support clinical decision for diagnosing pyloric stenosis	60
2.8	Summary	61
3	Dynamic Discretization Algorithm	63
3.1	Introduction	63
3.2	Hybrid Bayesian Networks (HBNs)	64
3.3	Analytical Solutions	64
3.4	Static Uniform Discretization	66
3.5	Dynamic Discretization	67
3.6	Beta-Binomial formulation	68
3.7	Multinomial formulation	71
3.8	Multinomial BN models	74
3.9	Hierarchical Bayesian Network (HRBNs)	76
3.10	Hierarchical Bayesian Network: Beta-Binomial	77
3.11	Summary	78
4	Clinical applications of Multinomial BN models	80
4.1	Modeling Subjective opinions with statistical distributions	80
4.2	Model Specification	89
4.2.1	Binomial Likelihood	91
4.2.2	Normal Likelihood	94
4.3	Parameter Learning using dynamic discretization	97
4.4	Scoring causal relations between variables using dynamic discretization	100
4.5	Summary	105
5	Meta Analysis of clinical data	106
5.1	Introduction	107
5.2	Conventional approach to Meta Analysis	111
5.3	Bayesian approach to Meta Analysis	113
5.3.1	Empirical justification for the exchangeability assumption	115
5.3.2	Application of HRBN to meta analysis	117

5.4	Beta-blockers clinical trials	119
5.4.1	Data from a meta-analysis on Beta-Blocker conducted by Marcelo <i>et al.</i> [143]	119
5.5	Magnesium trials in myocardial infarction	125
5.5.1	Summary of data on magnesium trials in myocardial infarction	126
5.5.2	Fixed effect Model	128
5.5.3	Random effect Model	130
5.5.4	Analysis with a constrained τ	132
5.6	The prevalence of intracranial saccular aneurysms	134
5.7	Data from a meta-analysis on the prevalence of intracranial saccular aneurysms [190]	134
5.7.1	Exploring the data for heterogeneity	135
5.7.2	Summary Estimate of the prevalence of Aneurysms: Fixed effect assumption	139
5.7.3	Summary Estimate of the prevalence of Aneurysms: Random effect as- sumption	141
5.7.4	Fixed parameters for retrospective and prospective studies	144
5.7.5	Fixed parameters for autopsy and angiography	144
5.8	Summary	147
6	BN Decision Making Models	149
6.1	Introduction	149
6.2	Evaluating the impact of multidisciplinary team meetings on treatment selection .	151
6.2.1	Description of Data	152
6.2.2	Summary of the MDT Data	154
6.2.3	Estimating Conditional probabilities from data	156
6.2.4	Summary of results	162
6.2.5	Direction of future work on the MDT analysis	166
6.3	Comparing the risk of Magnetic Resonance Imagine and Catheter Angiogram . .	167
6.3.1	Medical Background	167
6.3.2	Background to a clinical negligence case study	170
6.3.3	BN model for comparing risk of two screening procedures for intracranial aneurysm	170

6.3.4	Screening for intracranial aneurysms	172
6.4	Summary	182
7	Summary and Conclusions	183
7.1	Summary	183
7.1.1	Modeling existing medical knowledge with plausible statistical distributions (H_1)	184
7.1.2	Parameter learning with multinomial BN models (H_2)	184
7.1.3	Causal learning with BN models (H_3)	185
7.1.4	Solving meta analysis with BN models (H_4)	186
7.2	Conclusions	188
7.3	Future work	188

List of Figures

1.1	Model for evidence based decisions making process adapted from [201].	19
2.1	Decision tree representation for needle-stick injury example	33
2.2	A simple Bayesian Network model	36
2.3	A simple Bayesian Network model with node probability tables	37
2.4	Sample Graphical representation of probabilistic model with chain rule.	38
2.5	Sample BN representation of the problem	39
2.6	The Three primitive Connection Types.	41
2.7	BN combining the three primitive connection types in the undirected path between R and Q	42
2.8	Chest clinic Bayesian Network model	43
2.9	A simple BN model to predict whether the grass is wet or not.	46
2.10	Singly connected and multiply connected BN	48
2.11	Simple Join Tree from BNs	51
2.12	A Bayesian classifier for thyroid nodules [247]	60
2.13	A Bayesian decision network for diagnosing pyloric stenosis, [2]	61
3.1	A Beta-Binomial BN model	69
3.2	Posterior Densities of θ from dynamic discretization with 25 iteration and static discretization with 25 uniform intervals	70
3.3	A BN fragment for summing seven parameter nodes	73
3.4	A BN fragment for summing seven parameter nodes with binary factorization ap- proach	73
3.5	Hierarchical BN model	76
3.6	Hierarchical BN model	77
3.7	Hierarchical BN model	78
4.1	Parameter taking a fixed value on the parameter space.	83
4.2	Representing prior information with a discrete node	83

4.3	Representing prior information about the sensitivity of a test with Triangle distribution.	84
4.4	Representing prior information about the sensitivity of a test with a) Uniform b) Truncated Normal and c) Beta distributions.	85
4.5	Representing enthusiastic prior information using dynamic discretization.	85
4.6	Representing enthusiastic prior information using dynamic discretization.	86
4.7	Enthusiastic prior specified with the Triangle distribution.	87
4.8	Enthusiastic prior specified with the Uniform distribution.	87
4.9	Enthusiastic prior specified with the Truncated Normal distribution.	88
4.10	Enthusiastic prior specified with the Beta distribution.	88
4.11	Prior mean =0.5, sample proportion $y/n = 0.446$ and posterior statistics for different models ($\alpha + \beta$ values)	91
4.12	Prior marginal distributions of the Beta-Binomial Example after incorporating subjective information	92
4.13	Posterior marginal distributions of the Beta-Binomial Example after incorporating subjective information	93
4.14	Posterior marginal distributions of parameter θ using the reference, Sceptical and enthusiastic priors.	93
4.15	BN representation of the Normal likelihood example	95
4.16	BN representation of the Normal likelihood example	96
4.17	BN representation of the Normal likelihood example	96
4.18	A Multinomial BN model	99
4.19	Three competing causal relations between variables “Sex” and “Year”: a) Sex and Year are independent; b) Sex is dependent on Year and c) Year is dependent on Sex	102
4.20	Beta Binomial model to calculate the probability of the observed data (x_i) given H_1	102
4.21	Beta Binomial model to calculate the probability of the observed data (x_i) given H_2	104
5.1	Plot showing the number of published meta-analysis studies since 1980	108
5.2	Graphs showing the likelihood densities of two hypothetical studies (X,Y) and the posterior density of the combined effect	116
5.3	Posterior densities of the combined treatment effect based on fixed effect ($\mu y_1,y_2$) and random effect ($\mu y_1,y_2,\tau = 10$)	118

5.4	Hierachical BN model for Beta-Blocker Trial	123
5.5	Exploratory analysis of the data from the Magnesium trial.	127
5.6	Posterior distribution of the common effect (Fixed effect model)	129
5.7	Posterior distributions of true effect based on a) 1991 data analysis b) additional data found in 1995 and c) additional data found in 1995 with result of 1991 analysis as the prior	130
5.8	Posterior distribution of the common effect (Random effect model)	131
5.9	Posterior means conditional on the constraints on τ	133
5.10	Visual display showing the degree of overlap between two posterior densities . .	136
5.11	Posterior densities from du Boulay 1965 and other studies.	136
5.12	Posterior densities from du Jakubowski 1978 and other studies.	137
5.13	Fixed effect model for prevalence of intra-cranial aneurysms	140
5.14	Fixed Effect Model parameter from Winbugs	141
5.15	Posterior predictive effect from the two models a) θ_i is assumed to be normally distributed and b) θ_i is assumed to have beta distribution	142
5.16	Posterior means and lower and upper quartiles.	144
5.17	Hierarchical BN model for summarizing the prevalence of aneurysms by study designs.	145
5.18	Hierarchical BN model for summarizing the prevalence of aneurysms by diagnostic techniques.	146
6.1	BN modeling of treatment recommendations at a multi-disciplinary setting . . .	152
6.2	Multinomial BN model for learning parameters in Conditional table of “MDT Recommendation”	158
6.3	BN fragment with two categorical nodes: Parameters were learnt with multinomial BN models	162
6.4	Treatment recommendations made by the HPB multi-disciplinary team from 2005 to 2009.	163
6.5	A graph showing the posterior means of treatment recommendation for benign diagnoses.	164
6.6	A graph showing the posterior means of treatment recommendation for benign diagnoses.	165

6.7	A clinical model for accessing risks of two screening procedures for intracranial aneurysms.	171
6.8	BN model to compare risk of the MRA and CA techniques, Adapted from Fenton <i>et. al's</i> model, [62].	174
6.9	The prior Marginal distributions for all nodes	178
6.10	The posterior Marginal distributions of BN nodes given MRA test and large aneurysm	178
6.11	The posterior Marginal distributions of BN nodes given CA test and large aneurysm	179
6.12	The posterior Marginal distributions of BN nodes given (No test) and large aneurysm	180

List of Tables

2.1	Reduction in NPT size based on BNs representation	40
2.2	Joint probability distribution from different Connections	40
2.3	Comparing features of AgenaRisk, Hugin Expert and Winbugs packages	57
3.1	Conjugate prior distribution and analytical expression for posterior inference	65
3.2	Summary of posterior distributions from static and dynamic discretization	71
3.3	Summary statistics for the parameters (θ) and Tail ($1 - \theta$) from a hierarchical BN model	78
4.1	Statistical distributions to model prior knowledge about the sensitivity of a diagnostic procedure.	84
4.2	NPT to constrain a parameter space to a range between 92 and 98.	86
4.3	Different Beta-binomial models for the hypothetical rape victim data	90
4.4	Three mathematical representations of prior information	91
4.5	Table showing ages of twelve patients from a hypothetical population	94
4.6	Summary means for the parameters θ_1 , θ_2 and θ_3	99
4.7	Frequency distributions of patients treated between 2005 and 2009	101
5.1	2-by-2 Presentation of dichotomous data	110
5.2	Effects of beta-blockers from 22 clinical Trials, extracted from Table 5.4 of [73]	120
5.3	Effects of beta-blockers from 22 clinical Trials, extracted from Table 5.4 of [73]	121
5.4	Summary of posterior inference of beta-blockers	124
5.5	Summary data from 15 randomized trials of intravenous magnesium for acute myocardial infarction (data from Sterne <i>et al.</i> [218])	126
5.6	Summary of the Combined effects based on Fixed effect	128
5.7	Statistics of true effect based on data 1991 and 1995	130
5.8	Summary of the Combined effects based on Random effect	131
5.9	Constraints introduced on the heterogeneity parameter	133
5.10	Prevalence of aneurysms (Rinkel, <i>et al's.</i> [190])	135

5.11	Estimates of the KL-Distance between posterior densities θy_i and θy_j	138
5.12	Summary result of prevalence of intracranial aneurysm based on fixed effect assumption (excluding evidence from Housepain 1958 and McCormick 1965)	140
5.13	Summary result of prevalence of intracranial aneurysm based on fixed effect assumption	141
5.14	Posterior prediction of prevalence	142
5.15	Posterior Summary of study specific prevalence	143
5.16	Lower and Upper posterior percentiles of prevalence of aneurysm per 1000 patients in each study design	145
6.1	Descriptions of possible diagnosis options	153
6.2	Descriptions of possible treatment options	154
6.3	Number of cases per Diagnosis	154
6.4	Number of cases per Diagnosis by Year of discussion	155
6.5	Treatment recommendation by diagnosis	155
6.6	Treatment recommendation by diagnosis	155
6.7	Treatment recommendations by diagnoses for the five years period	156
6.8	Parameters of the conditional probability $p(D Y)$ for “Diagnoses”	157
6.9	Posterior mean for each parameters p_i in conditional Table $p(D Y)$	159
6.10	Parameters of the conditional probability for “MDT recommendation”, $p(R Y,D)$	159
6.11	Number of Parameters in table $p(R Y,D)$ by diagnosis.	160
6.12	Posterior mean for each parameters p_i in conditional Table $p(R Y,D)$	161
6.13	Variables in the model	171
6.14	The conditional probability distribution $p(Stroke tp)$	175
6.15	The conditional probability distribution $p(ta tp, cause)$	176
6.16	The conditional probability distribution $p(Outcome Rupture, Stroke?)$	177
6.17	Posterior distributions of outcome for different scenarios	179

List of Abbreviations

BN	Bayesian Network. 19
CSP	Cavernous Sinus Pathology. 149
DAG	Directed Acyclic Graph. 33
DDA	Dynamic Discretization Algorithm. 58
DT	Decision Tree. 30
EBM	Evidence Based Medicine. 17
GP	General Practitioners. 133
HBN	Hybrid Bayesian Network. 58
HIV	Human immunodeficiency virus. 30
HPB	HepatoPancreaticoBiliary. 133
HRBN	Hierarchical Bayesian Network. 20
ID	Influence Diagram. 19
JT	Join Tree. 47
JTA	Junction Tree. 47
KL	Kullback-Leibler Distance. 106
MCMC	Monte Carlo Markov Chain. 24
MDT	Multidisciplinary Team. 133
MRA	Magnetic Resonance Angiography. 30

MTE	Mixture of Truncated Exponential. 58
NPT	Node Probability Table. 45
OR	Odd Ration. 100
RD	Risk Difference. 100
RR	Relative Risk. 100

Acknowledgements

I want to thank those who have contributed to the success of this work. Firstly, I like to acknowledge the support and helpful guidance of my supervisors Professor Martin Neil and Professor Norman Fenton. Secondly, I want to thank the EPSRC and Agena Ltd for providing sponsorship for this work. Also, I like to thank my family for their unflinching support and sacrifice during the period of my study. Above all, I want to thank God almighty for guiding me during my research.

Chapter 1

Introduction

In the past clinical decisions were based entirely on the opinion of medical experts and based on the clinical experience of the expert, [81, 206]. Often, the most senior clinician of a clinical group uses personal experience, clinical intuition and anecdotal evidence¹ for making decisions. This approach is based on the following assumptions, [81]:

1. The authority of an opinion is directly related to the volume of clinical experience;
2. Clinical experience and judgement enable clinicians to evaluate new tests and procedures.

Contingent on this, a patient's involvement in the decision-making-process was restricted to giving or withholding consent to a given diagnostic test or treatment.

Evidence-Based Medicine (EBM) emerged in the early 1990s and has since become a major influence on many national health-care organizations, [81]. EBM advocates the use of up-to-date scientific evidence from health care research in the process of making medical decisions [201]. The popularity of EBM stems from the shortcomings of the opinion-based approach, the most prominent being the wide variabilities and inconsistencies of opinions between different experts on the same subject [81]. The following sources of clinical information are recognized by the EBM:

- patient's clinical state and circumstances.

¹Evidence based on casual observations rather than scientific analysis

- patient's preferences and actions.
- research evidence.

A model describing evidence based decision-making process is presented in Figure 1.1 whereby different sources of information are recognized in the decision-making process.

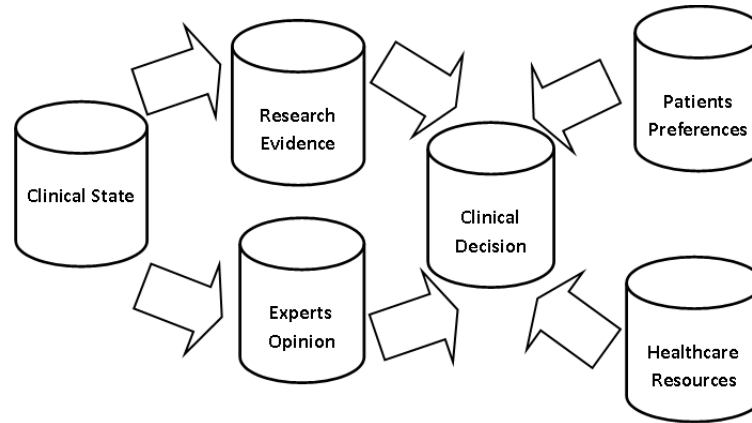


Figure 1.1: Model for evidence based decisions making process adapted from [201].

Meta analysis² is often used to synthesise clinical evidence in order to generate knowledge for clinical decisions. As indicated in Figure 1.1, clinicians combine relevant information from different sources, in addition to their own expertise, in order to choose the best clinical strategy for a given patient. Also, since patients may want to be actively involved in the process of making a decision regarding their condition, [37], they need to be well-informed about risks and benefits of the available clinical options. Clinical information will guide their perception of risks and benefits of various clinical alternatives. Consequently, the role of clinicians is no longer restricted to administering treatment but also communicating risk to their patients,[81] . This is particularly challenging because clinicians deal with a great deal of unavoidable uncertainty, arising from a number of sources. For example, clinical decisions about a diagnostic test, therapy, or prognosis are made under uncertainty, [81, 198–200, 202]. Laboratory test results and various diagnostic tools such as imaging techniques help a clinician to determine the cause and severity of an illness, predict its clinical course, and evaluate the patient's response to a treatment. However, these are themselves agents of uncertainty, [40], since tests are

²Meta-analysis is a technique for summarizing the findings of related studies in order to generate a generalizable conclusion about a treatment effect from a larger pool of data [47].

prone to systematic and random error, both in the conduct of the test and analysis of the results [84, 105, 197, 198]. The practice of EBM requires practitioners to be aware of how susceptible any source of information is to random and systematic error [40].

Although EBM criticizes other sources of clinical evidence such as non-controlled observations, experience and experts' opinions and also rejects practice standards based on clinical intuition, [45, 81, 83, 84, 198–200, 202], clinicians frequently encounter situations where there is no relevant evidence from either basic or applied research [197]. Therefore a technique that allows the combination of clinical evidence and other sources of information is central to clinical decision analysis. Bayesian Networks provide an inference mechanism that combines experts' opinion with available clinical evidence thereby making EBN attractive to the sceptics.

Bayesian Networks (BN) and Influence Diagrams (IDs) have been used in several studies,[1, 2, 8, 11, 23, 24, 35, 38, 41, 86, 86, 87, 87, 108, 112, 113, 115, 118, 128, 129, 137, 137, 138, 148, 151, 162, 169, 171, 183, 183, 204, 205, 233, 236, 238] for creating models to support clinical decisions.

There have been numerous attempts to apply decision support tools in medicine and despite the recent popularity and success in the use of Bayesian techniques a number of barriers to their systematic and routine use are evident. Firstly, there is a clear split between those involved in Bayesian statistics, mainly used to learn parameters from data and for prediction, and Bayesian network methods, where the emphasis has been on modelling causal interdependencies, diagnosis and classification. Around these two approaches two separate communities have evolved: a Bayesian statistics community mainly interested in experimental trials, meta-analysis and population studies and a causal modelling community, focused on decision making about the individual, optimal treatment identification and the use of causal explanations.

This thesis is part of a more general realignment of research efforts, most prominently applied to the medical domain, to provide an approach that integrates both of these approaches. Buchan, Winn and Bishop [22] have termed this realignment the 4th paradigm and emphasise the role of Bayesian methods, both statistical and causal, in unifying different sources of information available to clinicians when they make decisions about patient care and carry out research to identify optimal treatments. They, rightly, perceive both

approaches as two sides to the same coin and have presented a vision of a future approach where graphical models are the lingua franca for all stages in the decision support process including prediction, hypothesis testing, inference and diagnosis. However, this 4th paradigm lacks a unifying toolset, methodology or indeed a core algorithm for carrying out the necessary computations needed to transform decision support in medicine.

This thesis attempts to provide some concrete support to the 4th paradigm approach, starting by identifying a core computational algorithm that can be used for hybrid situations involving both causal reasoning and statistics. The algorithm is called dynamic discretization (DD) and it results in a class of Bayesian models called hybrid Bayesian Networks (HBNs), which can model the discrete variables, needed in casual modelling, and the continuous variables, often used in Bayesian statistics together. By demonstrating how DD is used we can show how it generalises and scales up to wider medical problems irrespective of whether it involved medical trials or individual patient diagnosis.

Methodologically, there also is a need to provide more flexibility to incorporate domain knowledge into this 4th paradigm process and this need highlights two gaps in current practice (though there may be more). Firstly, there is the perennial problem of prior distributions. To date conjugate (informative or otherwise) have been favoured, whether these are used for parameter learning in the diagnostic context (E.g. Dirichlet), or for learning hyper-parameters in meta analysis. However the use of conjugate priors has tended to be used irrespective of whether these conjugate priors actually match those that experts might wish to assert. To some extent conjugacy is a by product of the limitations of popular Bayesian statistical algorithms (and in particular Monte Carlo Markov Chain (MCMC) and in closed form analysis), therefore the choice, in this thesis, has been on an approach to prior formulation that makes no assumptions about what prior the expert can express. The approach taken provides considerable flexibility. Higgins and Spiegelhalter's work on sceptical priors [94] is a good example where researchers seek out to test differing strengths of priors, in this case for meta analysis, but all of the priors investigated are conjugate.

The second problem we need to address is the, perhaps obvious with hindsight, problem of encoding real world deterministic constraints into a causal model in a way that reflects expert choice, treatment protocols or some other non-probabilistic mechanism at

work. This seems critical but has received scant attention in the literature. This issue, combined with the inability of current algorithms to support non-conjugate expert priors are both major barriers to the complete representation of expert knowledge in Bayesian decision support in many application areas, and in the application to the medical domain in particular. Antal et al [3, 4] and Lucas et al [184] are good examples where soft data is used, either from literature or from experts, to inform the structure of a Bayesian model, where deterministic constraints are not identified and so have been neglected. For instance in [184] the decisions taken by clinicians are clearly not entirely probabilistic but are treated as such.

In addition to the questions surrounding these computational and methodological barriers to the use of the 4th paradigm this thesis has attempted to address a number of subsidiary questions relating to how we can use knowledge gained, computationally using DD, from the Bayesian statistical domain into a Bayesian causal model. The thesis focuses on using Bayesian statistical models, from clinical studies on populations, to generate probabilistic estimates for use within causal models at the individual patient level. There was insufficient time and resources to do this systematically covering model generation, refinement and application implied by the 4th paradigm approach, so instead we test some hypotheses relating to the use of DD for testing hypothetical causal relations between variables, and to show how the methodological and computational approach can be used to improve meta analysis. This is, of course, neither exhaustive nor the selection of these particular problems, over others, completely scientific but gives some insight into how the 4th paradigm might operate in practice.

1.1 Research Hypotheses

Dynamic discretization has been around for a few decades, however, it has not been applied to a number of important classes of problems described in this thesis. Specifically, this thesis investigates applications of BN models augmented to explore the following interesting hypotheses:

1. (H_1): Can we use plausible statistical distribution to model prior medical knowledge on a BN model with dynamic discretization? Modeling subjective prior is the most controversial aspect of bayesian analysis. A great deal of research effort has

been spent in developing techniques for formulating prior distributions. In the clinical domain, it is common to have a situation whereby a clinician can make some subjective assessments about some interesting parameters based on their clinical intuition and experience. A commonly used approach for translating such subjective assessment into a mathematical equivalence that can be used in the analysis is the relative frequency approach. This approach assumes that the prior value of the parameter takes a fixed value on the parameter space. Modeling a subjective prior with a plausible statistical distribution would allow the analyst to incorporate a degree of uncertainty around the subjective value.

2. (H_2): Can we use a multinomial BN augmented by dynamic discretization to learn parameters of a BN model? Parameter learning is an important aspect of modeling domain problems. Various parameter learning algorithms have been proposed over the past three decades. Chapter 2 presents a literature review of learning algorithms. In contrast to modeling the subjective opinion of an expert with statistical distributions, parameter learning is the use of the available data to learn parameter in a model. In the case, we use a subjective prior on the parameter and use the data to revise the parameter. With the second hypothesis (H_2), we want to explore the use of the multinomial BN formulation in learning parameters of categorical variables. This approach is very similar to the parameter learning algorithms implemented in the Deal package. The Deal package, [20] requires equal number of parameters in a node probability table and therefore cannot distinguish between zero observations³ and impossible observations⁴. We interpret observation zeros resulting from violations of health-care policies as logically impossible and then dynamically adjust the number of parameters to learn.
3. (H_3): Can we use the BN framework coupled with the Dynamic Discretization Algorithm to score hypotheses about causal relations between variables? Another important aspect of modeling with Bayesian network is the causal learning i.e. learning the structure of the network using available data. Many algorithms have been developed for learning causal relations entirely from data. This thesis applies the multi-

³lack of data or missing record

⁴logically impossible observation that violates health-care policies

nomial BN formulation and dynamic discretization to scoring different hypotheses about causal relations between variables. Our approach is not a full fledged causal discovery mechanism but a technique for scoring different causal hypotheses proposed by domain experts. We consider competing causal relations between variables as hypotheses and use data to identify the best hypothesis given the available data. This approach however requires substantive knowledge of the domain to formulate useful hypotheses.

4. (H_4): Can we use BNs models and dynamic discretization to generate summary estimates from meta-analysis based on fixed and random effect assumptions? Meta analysis has gained a tremendous popularity in the clinical domain over the past few decades. It is now the most widely used method of combining clinical evidence on the same subject. A range of classical statistical techniques have been developed for solving meta-analysis problems. In the recent times, meta-analysis has also generated a tremendous interest in the bayesian community. Meta-analysis is generally adjudged as the strongest source of evidence for generating clinical knowledge about some parameters. The fourth hypothesis (H_4) explores the possibility of using the Bayesian network framework for solving meta-analysis problems. Solving a meta-analysis problem with a BN model implies that the full posterior distribution of a summary estimate can be used as prior in clinical decision models. This approach not only captures the full uncertainty surrounding the summary estimate but also allows an analyst to check the sensitivity of a decision to the underlying studies.

Two clinical case studies are described in chapter 6 to assess the clinical useability of the models we have develop in the cause of testing hypotheses H_1 - H_4 listed above. The first case study uses a BN model to evaluate the impact of multi-disciplinary team meetings on treatment selection for patients with cancer. The second case study described clinical decision models that recommend optimal screening strategies for the management of patients with an unruptured intracranial aneurysm. In summary, we show that the powerful combination of the Dynamic Discretization Algorithm and the Junction Algorithm provides a unified modeling framework for creating clinical decision models that

include meta-analysis, parameter learning and decision support models.

1.2 Structure of the Thesis

The rest of the thesis is organized as follows. Chapter 2 provides a background to Bayes' probability theorem and decision analysis and its suitability in the clinical domain. The chapter reviews the Bayesian Network framework and a popular inference algorithm (the Junction Tree Algorithm) for performing inference on BN models with only discrete variables. Chapter 3 reviews the Dynamic Discretization Algorithm. This algorithm works seamlessly with the Junction Tree Algorithm thereby facilitating the use of BN models in solving problems with both discrete and continuous variables. Chapter 4 describes novel applications of BN models, augmented by dynamic discretization, to three important classes of problem. These include modeling subjective opinions from clinical experts with plausible statistical distributions; learning parameters of a model using multinomial BN models; and scoring different hypotheses about causal relations between clinical variables. The chapter addresses hypotheses H_1 , H_2 and H_3 . Chapter 5 describes a novel application BN models to meta-analysis. This chapter addresses hypotheses H_1 and H_4 . Finally, two clinical case studies are presented in chapter 6 to demonstrate these novel applications of BN models. The first case study uses data from a multi-disciplinary team at the Royal London hospital to demonstrate the flexibility of the multinomial BN framework in learning parameters of a clinical model. The second case study describes BN models to compare risks of two diagnostic techniques for intracranial aneurysms, namely: the Magnetic resonance imaging test and Catheter Angiogram test. The summary of the thesis and conclusion is discussed in chapter 7.

Chapter 2

Bayes and Decision Analysis

There are various sources of uncertainty in the process of making clinical decisions; these include diagnostic and treatment uncertainties among others. The main techniques used in this thesis for handling clinical uncertainty is the Bayesian Networks (BNs) framework. This chapter reviews this technique and the associated theoretical and technical issues relating to its development, uses and limitations.

2.1 Background

The origin of the term “Bayesian” goes back to 1763, when the work of Rev Thomas Bayes was posthumously published, [10]. This work contained the first detailed description of Bayes’ theorem derived from elementary probability theory. Central to Bayes theorem is the idea of “inverse probability” (where parameters are inferred from data). During the first few decades of the twentieth century, an alternative approach, based on Fisher’s work on statistical inference [65], was developed named as “*frequentist*” approach. The name “*frequentist*” suggests its interpretation of probability as the “long-run expected frequency of occurrence. $P(A) = n/N$, where n is the number of times event A occurs in N opportunities”. The method of *hypothesis testing* and *confidence intervals* were later developed. These developments revolutionized both the theory and practice of statistics and the application of frequentist methods quickly spread to diverse areas and replaced inverse probability in the first half of the twentieth century [54].

In the 1950s, there was a renewed interest in statistical decision theory that recognized the role of *subjective probability* in scientific inference and decision-making. Despite the computational challenges posed by the Bayesian approach, the application of Bayesian statistics grew in the following decades as evidenced by the number of published papers. The advent of Monte Carlo Markov Chain (MCMC) methods in the late 1980s made Bayesian computations possible for realistic problems. Since then Bayesian methods have been used in many different domains for reasoning under uncertainties.

2.1.1 Notation

In the rest of this chapter, we will use capital letters to denote random variables (e.g. A , B and C) and lower case letters for the realization or instance of a random variable (e.g. $A = a$). We will denote a set of variables (e.g. $\mathbf{V} = \{A, B, C\}$) by a bold capital letter and set of realization e.g. ($\mathbf{e} = \{e_1, e_2, e_3\}$) by a bold lower letter. Variable names are emphasized in italics.

2.1.2 Bayes' Theorem

Bayes' theorem shows how one conditional probability (such as the probability of a hypothesis given observed evidence) depends on its inverse (the probability of evidence given the hypothesis). A good and intuitive explanation of Bayes' theorem (taken from the Economist article [46]) is given below:

“The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sun rises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (i.e., the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sunrise the next day, the child adds an-

other white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on. Gradually, the initial belief that the sun is just as likely as not to rise each morning is modified to become a near certainty that the sun will always rise.”

Definition 2.1. Bayes’ theorem expresses the posterior probability $p(H|E)$ of the hypothesis after observing the evidence in terms of the prior probability $p(H)$, probability of the evidence $p(E)$ and the conditional probability of the evidence given the hypothesis $p(E|H)$.

Formally, Bayes’ theorem is stated as:

$$p(H|E) = \frac{p(H, E)}{p(E)} = \frac{p(E|H)p(H)}{p(E)}, p(E) \neq 0 \quad (2.1)$$

The prior $p(H)$ represents the prior belief about the hypothesis before observing any evidence. The probability of observing the evidence given the hypothesis $p(E|H)$ is called the *likelihood function*. The numerator $p(E)$ is called the *marginal probability* of evidence (E). This is the probability of witnessing new evidence E under all possible hypotheses. It can be calculated as:

$$p(E) = p(E|H)p(H) + p(E|\neg H)p(\neg H) \quad (2.2)$$

Where $\neg H$ is the complement of H . In general, given n mutually exclusive and exhaustive hypotheses $H_1, H_2 \dots H_n$ such that $p(H_i) \neq 0$ for $i = 1, 2 \dots n$ the Bayes’ theorem is given as Equation 2.3.

$$p(H_i|E) = \frac{p(E|H_i)p(H_i)}{\sum_j^n p(E|H_j)p(H_j)} \quad (2.3)$$

We can also apply this theorem to problem involving continuous variable. Let θ represents a parameter of interest, measurable on a continuous scale. For example the probability of an adverse effect following a particular clinical intervention. We can express a prior probability on this parameter as $\pi(\theta)$ and then apply the Bayes’ theorem (Equation 2.4) to revise this probability after observing y cases of adverse effect from N patients who received this intervention.

$$f(\theta|y) = \frac{l(y|\theta)\pi(\theta)}{\int l(y|\theta)\pi(\theta)d\theta} \quad (2.4)$$

In Equation 2.4, $f(\theta|y)$ is the posterior density function of θ given y observations.

2.1.3 Bayesian & Frequentist approach to Data Analysis

As opposed to the point estimators, such as means and variances, used by *frequentist* statistics (often referred to as the *classical approach*), *Bayesian* statistics is concerned with the idea of generating a posterior distribution for the unknown parameter given both the data and a prior density for the parameter. The Bayesian approach, therefore, provides a much more complete picture of the uncertainty surrounding the estimates of the unknown parameters.

In the classical approach, parameters are treated as fixed variables while data are considered as random variables. Probabilities are expressed in terms of relative frequency while inference is based on confidence interval techniques [66]. In contrast, the Bayesian approach treats the parameters as random variables and allows subjective interpretation of probability (i.e. the probability of an event is the degree to which you believe that the event is true). Bayesian inference is based on Bayes' theorem which allows inference about the parameter given the observed data.

Suppose we are interested in estimating a clinical parameter θ from data $y = (y_1, \dots, y_n)$. The classical approach proceeds by estimating θ with its best unbiased estimator and then constructs a confidence interval for this unknown parameter. In the Bayesian approach, θ cannot be determined exactly with a point estimator because of its inherent uncertainties. These are expressed through probability statements and distributions. We can express the prior knowledge, for example, as a normal distribution with mean 0 and variance 1 (i.e. $\theta \sim N(0, 1)$). This expresses a subjective belief and the uncertainty associated with it.

The frequentist approach for measuring uncertainty requires information about many past instances of an event under identical condition (i.e. repeated trials). The subjective approach, on the other hand, is based on the existing body of knowledge. This can be (for example) experts' subjective opinions about the parameter of interest. In reality, many uncertain events of interest do not have sufficient historical data to satisfy the condition for frequentist statistics. Therefore, the Bayesian approach is a feasible option for tackling

many practical problems. For example, if we are interested in the probability that England wins the next world cup, the identical assumption of the frequentist approach will break down but the Bayesian approach can express subjective opinions regarding the chances of the English team winning the next world cup [116].

2.1.4 Bayesian Inference

Conceptually, Bayesian inference can be thought of as a collection of basic ideas. Suppose we are uncertain about a parameter θ but we can quantify this uncertainty using a subjective probability $p(\theta)$. We can then express the conditional probability for the observations x given this parameter with the appropriate model (likelihood function) ($p(x|\theta)$). When data is available, we can then revise the prior probability to a new probability distribution (posterior distribution, $p(\theta|x)$) using Bayes' theorem, [64]. This enables us to combine the prior distribution and the likelihood as previously described in Equations 2.3 and 2.4.

The popularity of Bayesian analysis (in the clinical domain) stems from the following benefits.

- **Prior Specification:** Bayesian analysis provides a natural way of combining prior information with data, within a solid decision-theoretical framework, [101]. We can use the historical data to learn a prior parameter in the current analysis. When new observations become available, the posterior distribution from the current analysis can then be used as a new prior for the subsequent analysis. This temporal and dynamic inference capability of Bayesian analysis makes it attractive for decision analysis in the clinical domain. Clinical decision analysis often involves management of a patient's condition over time.
- Possibility of valid inference for a small sample.
- It also provides a convenient setting for a wide range of models including hierarchical models and missing data problems among others.

However, there are some challenges to Bayesian analysis. These challenges include the following among others:

- Bayesian analysis often requires skills to formulate a mathematically tractable prior distribution from the subjective beliefs provided by an expert;
- The posterior distribution can be heavily influenced by the prior distribution especially where there is insufficient data;
- High computational cost may be incurred, especially in a model with a large number of parameters.

Example 2.1. In a study presented by Casscells and Graboys [26], the following problem was put to a number of students and staff at Harvard Medical School.

“One in a thousand people has a prevalence for a particular heart disease. There is a test to detect this disease. The test is 100% accurate for people who have the disease and is 95% accurate for those who don’t (i.e. 5% of people who do not have the disease will be wrongly diagnosed as having it). If a randomly selected person tests positive, what is the probability that the person actually has the disease?”

About 50% of the respondents gave a posterior probability of 0.95, whereas, the posterior probability of the heart disease given a positive test result is 0.0196 as shown below:

$$\begin{aligned}
 p(D|+ve) &= \frac{p(+ve|D)p(D)}{p(+ve|D)p(D) + p(+ve|\neg D)p(\neg D)} \\
 &= \frac{(1)(0.001)}{(1)(0.001) + (0.05)(0.999)} = 0.01962
 \end{aligned}$$

When people give a high answer such as 0.95, they are falling victim to a very common fallacy known as the ‘base-rate neglect’ fallacy, [62]. People often fail to take cognizance of the low probability of having the disease. Also, the probability of a false positive test is high (5% is the same as 50 in a thousand, whereas there is only a one in a thousand chance of having the disease). There is much debate about why intelligent people can get such important (and not too difficult) mathematical reasoning completely wrong and the ramifications of it [62]. Fenton *et al.*, [62] presented an informal but intuitive graphical explanation to this problem which is easier to understand and appreciate but visible only for a simple inference problem.

Inference problems involving only two variables, such as the one in the above example, is fairly simple. However, computation becomes much more complex when several variables are involved with complex conditional dependencies between them. To address this, various graphical models have been developed for complex probabilistic inference. Indeed, the idea of using graphical models to represent uncertainty can be traced back to the 1920's when Sewal Wright, [245] developed a method for measuring the degree to which an observed variation of a given effect is determined by the causes [135]. During the 1960's, there was a renewed interest in research around methods for representing and reasoning with problems characterized by uncertainty. As research evolved, there was a growing need for efficient mechanisms for representing, encoding, storing and manipulating probabilistic problems. Consequently, graphical models such as Decision Trees, Bayesian Networks and Influence Diagram emerged.

2.2 Decision Trees

Decision trees (DT) are expressive graphical models consisting of chance nodes representing random variables and decision nodes representing the decision to be made. The decision is a set of mutually exclusive actions, alternatives or options that the decision maker can take. For example, a clinical action could be '*diagnose with Magnetic Resonance Angiography MRA*' for patients with a high risk of developing aneurysms. The actions performed will lead to a set of possible uncertain outcomes (for instance an MRA scan can result in a false negative). Decision trees can be used to quantify the potential risks of an action in relation to possible costs or benefits, in the face of uncertainties. They are useful in determining the most favorable outcome from the several alternatives [32].

Example 2.2. The needle-stick injury example, [32] shown in Figure 2.1 is a decision problem involving a health officer. This health officer suffered a needle-stick injury while administering a treatment to a potentially human immunodeficiency virus (HIV) -positive drug user. Suppose the prevalence of HIV within the population of drug users is 0.15 and that 5 in every 1000 needle-stick injuries with HIV+ blood seroconverts ¹ to HIV. To put in another form, the conditional probability of developing HIV given an exposure to HIV

¹The development of detectable antibodies in the blood directed against an infectious agent after an exposure to the agent. Following seroconversion, a person tests positive in tests based on the presence of antibodies [152].

is 0.005. Let B denote the HIV status of the drug user such that $B = b_1$ if the drug user is HIV+ and $B = b_2$ if not. Let also denote the rate of conversion into HIV by C i.e. $C = c_1$ and $C = c_2$ respectively for seroconversion and not seroconversion to HIV.

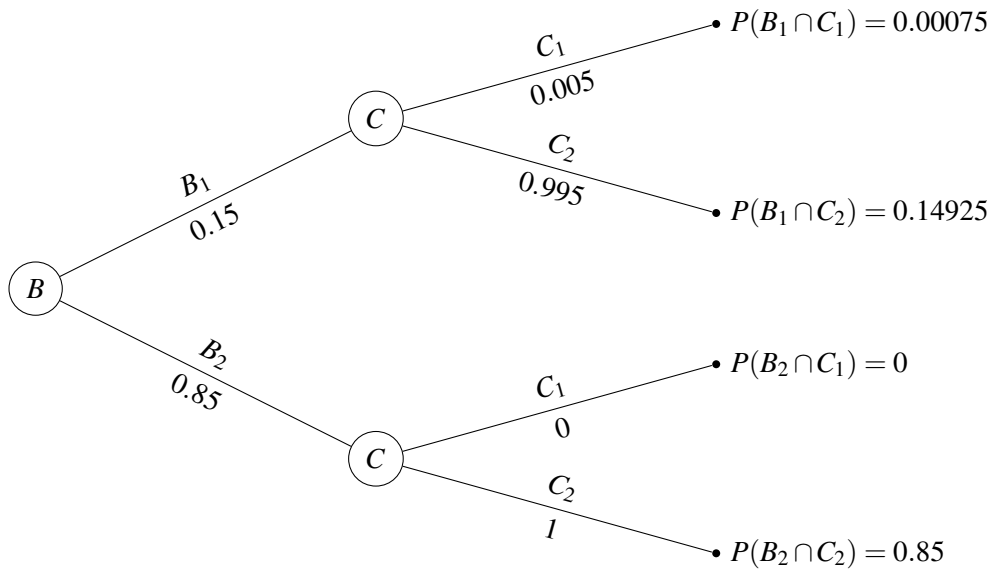


Figure 2.1: Decision tree representation for needle-stick injury example

The simple Decision tree in Figure 2.1 has four paths and the joint probability along each part is shown at the end of the path. These joint probabilities are computed by simply multiplying all the probabilities along a path. For instance, 0.00075 is the product of two probabilities $P(B = B_1)$ and $P(C = C_1|B = B_1)$. In order to compute the likelihood of an outcome, say the likelihood of developing HIV $P(C_1)$, we will simply combine the joints of all the branches of the tree involving C_1 . This can be formally expressed as given in Equation 2.5.

$$P(C_1) = P(B = B_1)P(C = C_1|B = B_1) + P(B = B_2)P(C = C_1|B = B_2) \quad (2.5)$$

Therefore, $P(C_1)$ and $P(C_2)$ are respectively $(0.15)(0.005) + (0.85)(0) = 0.00075$ and $(0.15)(0.005) + (0.85)(1) = 0.99925$ from the tree. We can use the information on the decision tree to compute the posterior probabilities of HIV status of the drug user if we have a test result confirming the HIV status of the health user. We have assumed, in this example, that the only way the health officer could have contracted HIV is through the exposure resulting from the needle stick injury and the diagnostic test for detecting HIV

is 100% accurate for detecting HIV and non HIV cases.

We can compute a posterior probability of HIV+ of the drug user having observed a positive test result for the health officer i.e. $P(B = B_1|C = C_1)$.

$$p(B = B_1|C = C_1) = \frac{p(B_1, C_1)}{p(C_1)} = \frac{0.00075}{0.00075} = 1$$

This posterior value has an intuitive explanation. A positive test result for the injured health officer implies that the drug user is, certainly, HIV+. This is only possible when there are no other means by which the health officer could have contracted HIV and the diagnostic tool is 100% accurate. We may also want to compute the posterior probability that the drug user is HIV+ having observed a negative test result for the health officer $p(B = B_1|C = C_2)$.

$$p(B = B_1|C = C_2) = \frac{p(B_1, C_2)}{p(C_2)} = \frac{0.14925}{0.99925} = 0.149362$$

With a negative test result, the revised probability that the drug user is HIV+ is now lower than the prevalence of HIV in the population of drug users.

Decision trees have traditionally been used in decision making to choose an optimal decision from a finite set of choices. The value being optimized is the utility function which is expressed for each outcome. However there are a number of problems with decision trees, the main ones being:

1. The order of decision nodes in a tree is arbitrary regardless of the condition and information relationships that exist in the real world (causal knowledge)
2. The number of state combinations grows in size exponentially as the number of decisions and outcomes increase

Consequently, many practical decision problems are tackled by a more sophisticated decision framework such as Bayesian networks.

2.3 Bayesian Networks

Similar to Decision trees, Bayesian networks (BNs) also provide a framework for reasoning, with graphical models, about problems involving uncertainty [179]. The Bayesian

Networks formalism offers a natural way to represent the uncertainties in clinical decision. It helps us quantify the uncertainty in a clinical parameters such as *diagnostic test results, the prevalence of a disease* etc, with prior distributions. It also allows us to specify an appropriate likelihood function for the observable variables, such as the number of deaths resulting from a particular clinical procedure. The computational mechanism in BNs is based on Bayes' theorem, described earlier. This computational mechanism computes posterior distributions for unobserved variables given observations on one or more observable variables. The beauty of a BN stems from its capacity for intuitive representation and effective computation of the joint probability distribution over a set of random variables [179].

Formally, a Bayesian Network (BN) is a graphical model $\Lambda = \{D, \Phi\}$ consisting of qualitative part D and quantitative part Φ . The qualitative part is a directed acyclic graph (DAG) with a set of nodes and edges. The nodes² represent a random variable in the domain of study. Clinical variables such as *family history of aneurysm* will be represented as a node on the BNs. Edges are used to represent the relationship between nodes.

By convention, BN nodes are represented as circles labeled by the names of the variable they represent while the edges between nodes are represented by arrows connecting them. The direction of an edge is from the parent to the child node. If there is a directed arc from node X_i to node X_j , then we say X_i is a parent of X_j and, as stated earlier, the edge encodes probabilistic dependencies between X_i and X_j . We can denote the set of parent nodes of X_i by $pa(X_i)$.

The idea of “inverse probability” makes BNs versatile and able to carry out bi-directional inference. This makes BNs ideal for modeling clinical decision problems where reasoning is, sometimes, from causes to effects; in some other times, it could be backward reasoning from effects to causes.

Example 2.3. Figure 2.2 shows three nodes, each of which represents a clinical variable. There are also two directed edges encoding causal relationship between these variables. The first directed edge between ‘*family history of aneurysms (H)*’ and ‘*third nerve palsy (TNP)*’ encodes a dependence relation between them i.e. developing a third-nerve palsy may be influenced by the *family history of aneurysms* node. This link only captures a

²We do not distinguish between a node and the variable represented, the two are used interchangeably.

positive association between them and does not have a causal interpretation. The second edge connecting ‘*third nerve palsy*’ with ‘*Ptosis (P)*’ (i.e. drooping of the upper eyelid) encodes the causal relationship between them as ptosis is one of the clinical manifestations of third nerve palsy.

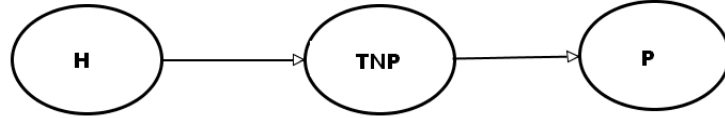


Figure 2.2: A simple Bayesian Network model

The causal relationship depicted in Figure 2.2 above is uncertain. For example, we cannot say with certainty that having a third nerve palsy will always show ptosis symptom. However, clinicians can use their expert knowledge to assign probabilities to the relationship. For example they can use a probability to express the degree to which they believe ptosis will manifest in a patient with a third nerve palsy. The quantitative part Φ enables us to quantify the strength of these causal relationships by specifying the conditional probabilities on each node given the configuration of the parent nodes. These conditional probabilities are stored in Node probability Tables (NPT).

Example 2.4. Suppose we are studying a hypothetical population of aneurysms susceptible patients in which half of the population have a family history of aneurysm (h). Cases of patients diagnosed with third nerve palsy (tnp) were recorded over a period of 48 months. From the group with a family history of aneurysm, 20% cases were reported while only (5%) were recorded in the group without family history. 90% of patients with tnp developed *ptosis* and 1% in the group without tnp also developed ptosis. We can summarize the information provided into priors and conditional probabilities of the corresponding variables. We will start by specifying the prior probability on family history of aneurysm (h).

$$\begin{aligned} P(h) &= 0.5 \\ P(\neg h) &= 0.5 \end{aligned}$$

Then, we can specify the conditional probability of tnp given the family history of aneurysm (h).

$$\begin{aligned}
 P(\text{tnp}|\text{h}) &= 0.2 \\
 P(\text{tnp}|\neg\text{h}) &= 0.05 \\
 P(\neg\text{tnp}|\text{h}) &= 0.8 \\
 P(\neg\text{tnp}|\neg\text{h}) &= 0.95
 \end{aligned}$$

Finally, we need to specify the conditional probability for *ptosis* given *tnp* as follows:

$$\begin{aligned}
 P(\text{ptosis}|\text{tnp}) &= 0.9 \\
 P(\text{ptosis}|\neg\text{tnp}) &= 0.01 \\
 P(\neg\text{ptosis}|\text{tnp}) &= 0.1 \\
 P(\neg\text{ptosis}|\neg\text{tnp}) &= 0.99
 \end{aligned}$$

Thus, in order to specify the probability distribution of a BN, one must provide the prior probabilities for all root nodes³ and conditional probabilities for all other nodes, given all possible combinations of their parent nodes. Figure 2.3 shows both qualitative and quantitative parts of a fully specified BN model for the problem.

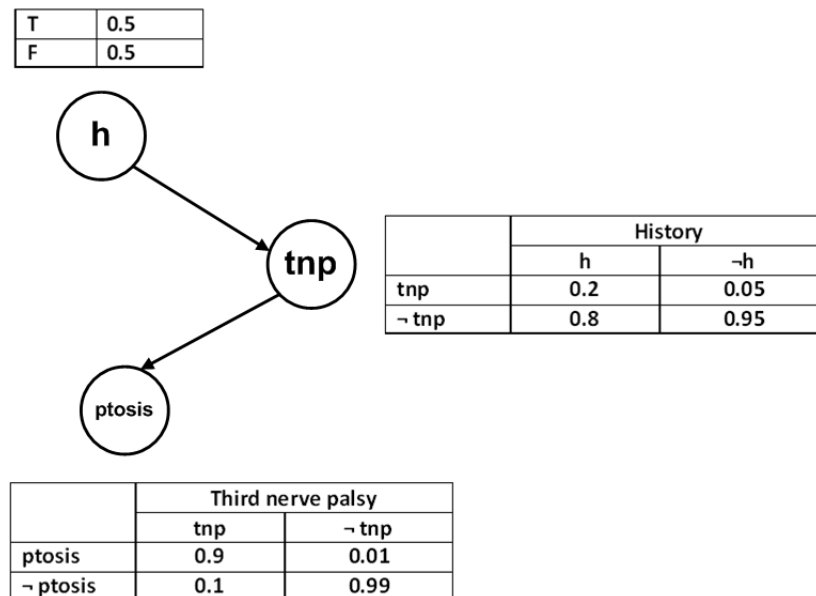


Figure 2.3: A simple Bayesian Network model with node probability tables

This model facilitates making inference about the probable cause of ptosis and allows analysts to predict the probability of ptosis.

³These are nodes with no parent parent nodes.

2.3.1 Conditional Independence Assertion

The conditional independence assumption embedded in the BN framework makes BNs attractive for probabilistic reasoning. The traditional approach, i.e. the chain rule, requires a full specification of the probability distributions. This may introduce a high number of probability entries that will be difficult or impossible to encode in a node probability table. In this traditional inference approach, a probabilistic model with n random variables, each with a binary outcome, requires 2^n entries for the largest probability table. For instance, a probabilistic problem involving six variables $\{A, B, C, D, E$ and $F\}$, the full joint probability distribution is as follows (Equation 2.6):

$$P(A, B, C, D, E, F) = P(A).P(B|A).P(C|A, B).P(D|A, B, C).P(E|A, B, C, D).P(F|A, B, C, D, E) \quad (2.6)$$

Let us now use a graphical model to represent this chained dependency structure. The order of events follows the same order in which the variables appears in Equation 2.6. That is, the event labeled A is known first, followed by event B . Therefore, event B only depends event A . By the same reason, event C depends on both events A and B and so on for other events (Figure 2.4).

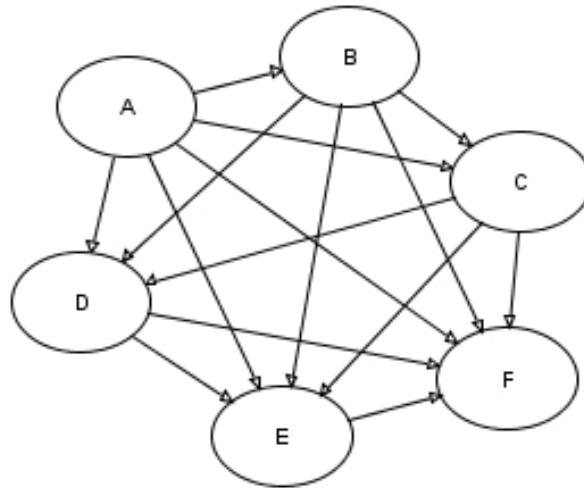


Figure 2.4: Sample Graphical representation of probabilistic model with chain rule.

In this example, we need 64 entries to specify the probability table for the last event (F). In contrast, a BN model exploits the independence between variables to specify a more compact joint probability distribution. To illustrate this we can use the existing

knowledge in the problem domain and specify dependence structures as appropriate. Suppose we have the following knowledge about the problem domain; 1) variables B and A are independent causal factors of variable C ; 2) both variables D and E depend only on variable C and 3) variables D and E are independent causal factors of variable F . Armed with this piece of information, we can create another graphical model shown Figure 2.5. This is a reduced version of the graphical model in Figure 2.4.

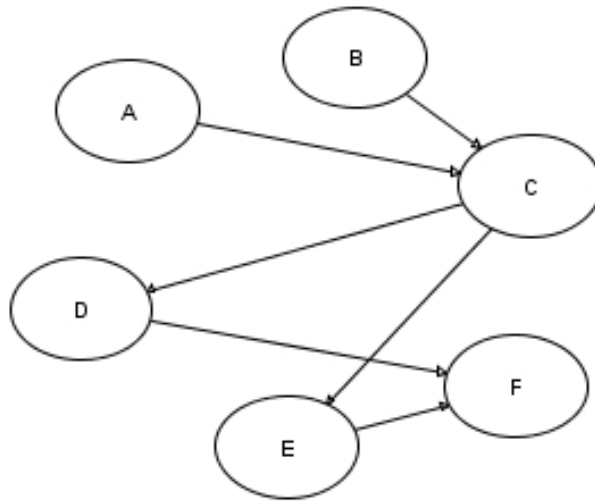


Figure 2.5: Sample BN representation of the problem

Note that the joint probability distribution for the BN model in Equation 2.7 is a reduced version of the full probability distribution presented earlier.

$$P(A, B, C, D, E, F) = P(A).P(B).P(C|A, B).P(D|C).P(E|C).P(F|D, E) \quad (2.7)$$

As stated earlier, the largest probability table from the fully specified model requires 2^6 i.e. (64). In contrast, the largest table in the BN model, requires just 8 entries. Table 2.1 shows the size of conditional probability tables required for all the variables. The model with the full joint probability distribution requires a total of 126 probability entries, whereas the total entries required for the BN model is only 28 probabilities.

The full probability distribution approach is more complex and maybe counter intuitive to the heuristic governing human reasoning, [180] pp 78. People base probabilistic judgements on a small number of relevant propositions, especially conditional statements such as the likelihood of a disease given a set of symptoms, as opposed to using a complex

Table 2.1: Reduction in NPT size based on BNs representation

Variables	Full Specification		BNs Model	
	Probability	Size	Probability	Size
F	$P(F A, B, C, D, E)$	64	$P(F D, E)$	8
E	$P(E A, B, C, D)$	32	$P(E C)$	4
D	$P(D A, B, C)$	16	$P(D C)$	4
C	$P(C A, B)$	8	$P(C A, B)$	8
B	$P(B A)$	4	$P(B)$	2
A	$P(A)$	2	$P(A)$	2

conjunction of all possible propositions, [135].

The conditional independence assumptions reduce the complexity inherent in the full joint probability distribution by reducing the number of probabilities and the time complexity of an inference. For example, a problem with n binary variables requires $O(2^n)$ complexity for the fully specified model i.e. using the full joint probability distribution. Whereas, the complexity of a BN model, for the same problem, is $O(n2^k)$, where k represents the maximum number of parent nodes in the model, [158].

2.3.2 Knowledge representation in BNs

The structure of a BN model can be composed from three primitive fragments, which are called serial, diverging and converging connections (Figure 2.6). With these three fragments, we can capture all the possible ways in which variables can become dependent or independent given evidence. In the linear/serial fragment, node B is between nodes A and C . The neighboring BN fragment (converging) shows both A and C as parents of B and finally an example of diverging fragment with B as a parent node to both A and C .

The joint probability distribution for these connections are provided in Table 2.2

Table 2.2: Joint probability distribution from different Connections

Connection Type	$P(A, B, C)$
Linear/Serial	$P(A).P(B A).P(C B)$
Converging	$P(A).P(C).P(B A, C)$
Diverging	$P(B).P(A B).P(C B)$

Definition 2.2. The path between any two nodes P and Q in a BN is d-connected with respect to the evidence nodes \mathbf{E} if every interior node N on the path between P and Q is either

1. linear or diverging and not belong to \mathbf{E} or

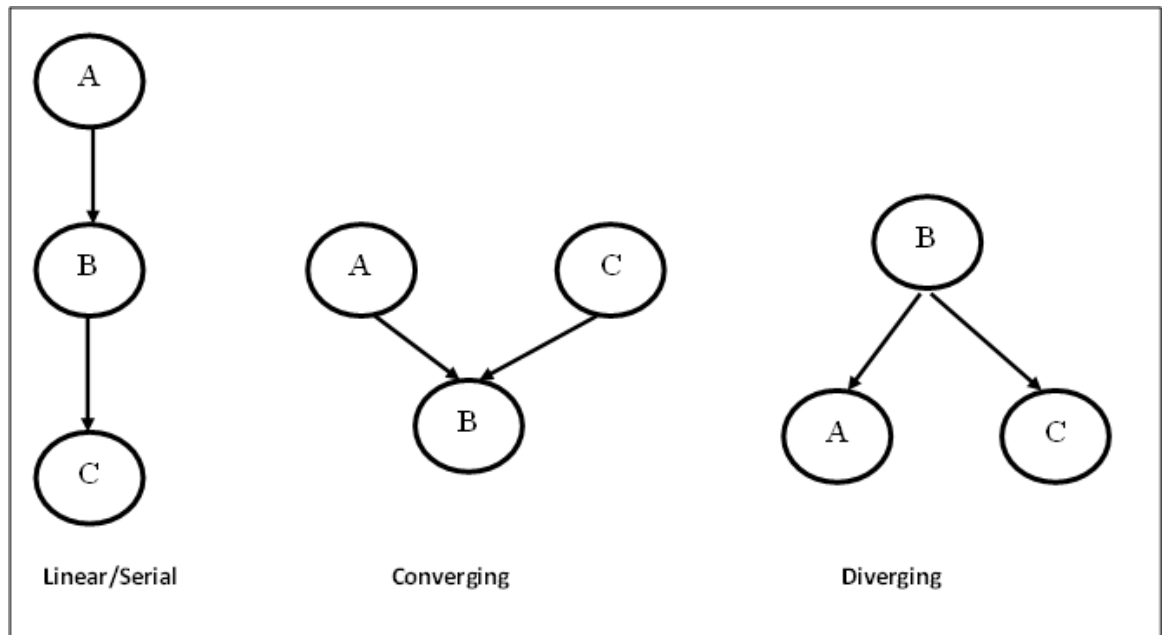


Figure 2.6: The Three primitive Connection Types.

2. converging and either N or one of its descendants (if any) belong to \mathbf{E} [57],

where \mathbf{E} is a set of nodes with observations, this can be empty if there is no observation on any node. \mathbf{N} is a set of interior nodes linked on the path between P and Q . Therefore, it is possible to know whether any two variables are dependent given evidence.

As shown in the example in Figure 2.7 there are 3 interior nodes i.e. $\mathbf{N} = \{A, B, C\}$ on the “undirected path”⁴ between R and Q . The connection is converging on A , diverging on B and serial on C . Any evidence entered on R would not be propagated to B unless there is evidence on A or its descendants (if any). That is, evidence supplied on A or its descendants (if any) renders R and B d-connected. Evidence on B or C will block any evidence coming from R , hence, R and Q are d-connected if A belongs to evidence node E and B and C does not.

In a BN, a variable (A) is independent of another variable (B) given the evidence (e) if there is no d-connecting path from A to B given \mathbf{e} (where e is the instantiations of evidence nodes \mathbf{E}). Formally, A is independent of B given e if $P(A|B, \mathbf{e}) = P(A|\mathbf{e})$.

At this point, we can generalize the earlier joint probability distributions for n random variables $X_1, X_2, X_3, \dots, X_n$. Equation 2.8 shows the joint probability distribution model using the chain rule.

⁴The undirected path is the graph obtained by dropping the direction of the arcs on a BN.

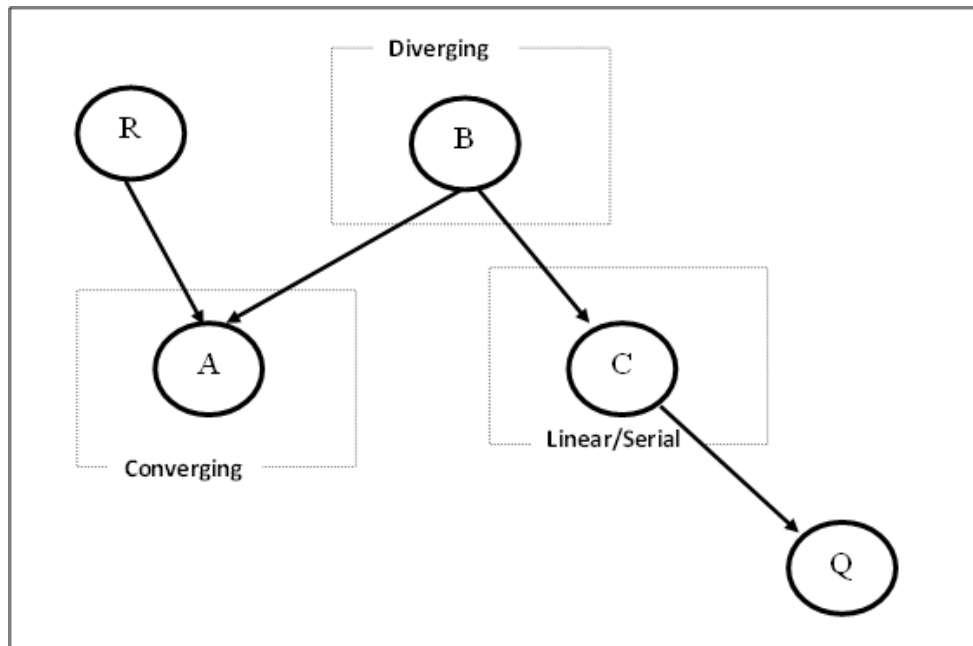


Figure 2.7: BN combining the three primitive connection types in the undirected path between R and Q

$$P(X_1, X_2, \dots, X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdots P(X_n|X_1, X_2, \dots, X_{n-1}) \quad (2.8)$$

Again, a BN specification reduces this joint probability distribution to Equation 2.9.

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | pa(X_i)) \quad (2.9)$$

In general, a BN variable is independent of its more remote ancestors given the values of its immediate parents in the graph.

Example 2.5. The example considered here is the Chest Clinic BN model for medical diagnosis of shortness of breath [125]. ‘*Tuberculosis*’ and ‘*lung cancer*’ cause an ‘*abnormality in the chest*’ which can result in *shortness of breath* (‘*dyspnea*’). ‘*Abnormality in the chest*’, when present can be diagnosed by a ‘*positive X-ray result*’. Another potential cause of ‘*dyspnea*’ is ‘*bronchitis*’ which is common among smokers. ‘*Smoking*’ is considered a risk factor for both ‘*lung cancer*’ and ‘*bronchitis*’, while a ‘*visit to Asia*’ sometimes increases the chance of contracting ‘*tuberculosis*’. The joint probability distribution for the BN with variables V, S, T, L, B, A, X and D can be expressed as given in Equation 2.10.

$$P(V,S,\dots,D) = P(X|A).P(D|A,B).P(A|T,L).P(B|S).P(L|S).P(T|V).P(V).P(S) \quad (2.10)$$

Figure 2.8 depicts causal relationships where each node on the graph represents a random variable with two possible values {true or false}.

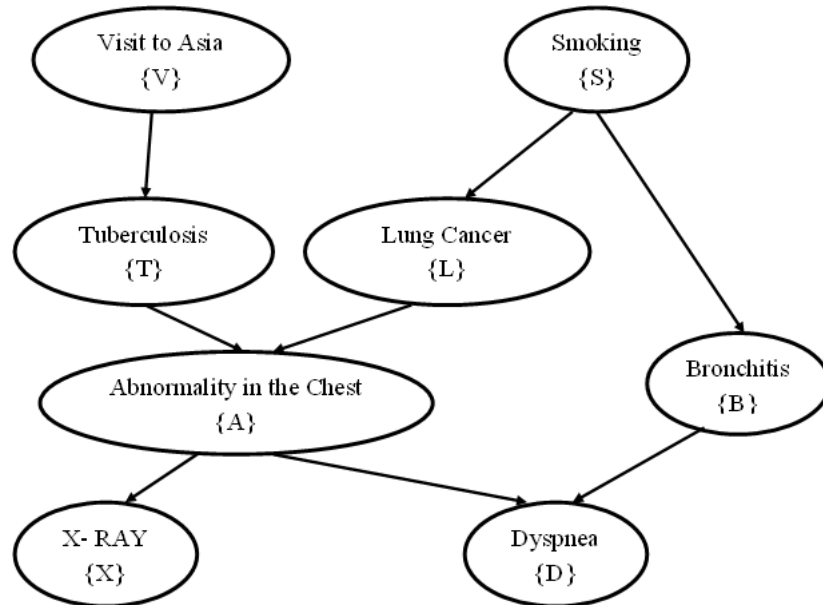


Figure 2.8: Chest clinic Bayesian Network model

The arc between the ‘*bronchitis*’ and ‘*dyspnea*’ represents the dependencies between them, that is to say, the former influences the later. On the other hand, lack of an arc between ‘*smoking*’ and ‘*visit to Asia*’ shows that there is no such direct dependency between them. However, evidence on ‘*abnormality in the chest*’ or any of its descendants ‘*positive X-ray result*’ and ‘*dyspnea*’ d-connects ‘*smoking*’ and ‘*visit to Asia*’ provided there is no evidence on ‘*tuberculosis*’ or ‘*lung cancer*’.

‘*Smoking*’ increases the chance of developing ‘*lung cancer*’ which in turn increases the chance of ‘*abnormality in the chest*’. Once we know that a patient has ‘*lung cancer*’, then we know the cause of the ‘*abnormality in the chest*’, any additional information regarding the ‘*smoking*’ habit will no longer change the probability of developing ‘*abnormality in the chest*’. Therefore, evidence on ‘*lung cancer*’ d-separates ‘*Smoking*’ and ‘*abnormality in the chest*’.

2.3.3 Constructing a BN

Having described the BN graphical language in the preceding sections, we will now describe the steps in constructing a BN model. The benefits of using BNs for modeling problems characterized by uncertainty are well documented [106, 137, 179]. However, constructing a BN model requires gathering complex domain specific information from disparate sources into an intuitive, coherent and easy to understand form. This nontrivial and time consuming task requires a structured approach for capturing relevant knowledge for the qualitative and quantitative parts, [124].

In practice, it is common to use the expert's knowledge while constructing a BN model. It is also possible to use the existing data from the problem domain to construct a BN. In both cases, domain experts must be involved in the construction of BNs models. The role of domain experts is to supply the analyst with specialist domain knowledge. This helps the analyst to synthesize domain knowledge and translates this information into the BN components such as nodes, causal relations and probabilities. The process of building a BN model has been classified into three sequential steps:

- Selection of variables.
- Creating the qualitative part (DAG) using expert's knowledge or historical data.
- Assessing the conditional probabilities (quantitative part) using experts' knowledge or historical data.

In principle, these steps should be performed sequentially. However, building a BN requires a careful trade-off between accuracy and complexity. Therefore, these steps are repeated until a desired BN model is achieved. Let us look at these three steps in turns.

2.3.4 Variable selection

The process of identifying variables is not always straightforward. Heckerman [91] suggests the use of the following guidelines to identify variables:

- The objectives of the model must be identified (e.g. prediction, explanation or exploration).
- Possible observations that may be relevant to the problem must be identified.

- Determine the subset of the identified observations that is worthwhile, considering the complexity of the network.
- For each node, define a set of mutually exclusive states and collectively exhaustive states.

The variables gathered from the problem domain are directly mapped onto nodes in a BN model, [135]. Therefore, nodes and variables are used interchangeably as they refer to the same phenomenon. A node on a BN can represent a continuous variable as well as a discrete variable. A continuous variable takes any value between any two points on a scale. These points may be finite or infinite. In contrast, a discrete variable only takes a finite number of discrete values. An example of discrete node in a BN might be the outcome of a test result. This node can take two values *+ve* for a positive result and *-ve* for a negative result. Jensen [106] suggests three types of variables when building a BN model.

- Hypothesis variables: These are the outcome variables that provide values of interest. These nodes are often referred to as query or target nodes which are not observable. Identifying these variables is the primary task in building a BN model.
- Information variables: Variables whose values are observable and relevant to the hypothesis events. These are the so called evidence nodes whose state can be observed.
- Mediating variables: These are introduced for a special purpose. For example, a mediating variable can be introduced to simplify the conditional probabilities tables.

Figure 2.9 shows a BN model to illustrate roles played by each node with respect to the suggestions, [91] above.

This simple BN model predicts the condition of the grass, wet or dry, given observations of the weather. The root node labeled “Weather” is a categorical node with three possible values *Sunny*, *Cloudy* and *Raining*. This is an observation (evidence) node representing the current observable weather.

There are two intermediate nodes “Sprinkler” and “Rain” and one query node “Wet grass”. All these nodes are boolean nodes with values $\{On, Off\}$, $\{True, False\}$ and $\{Wet, Dry\}$ respectively.

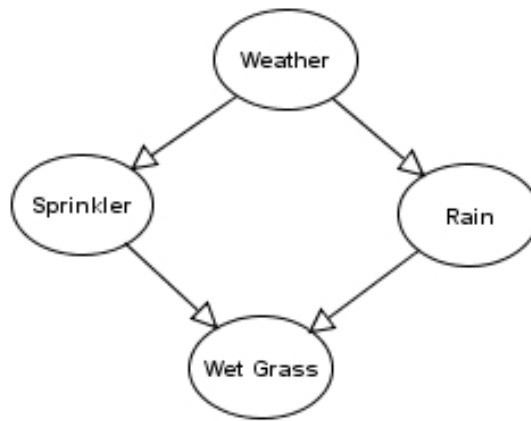


Figure 2.9: A simple BN model to predict whether the grass is wet or not.

2.3.5 Types of variables commonly used in a BN model

A BN node can represent different data type such as binary and categorical. A binary node can take on binary values. Categorical or multinomial nodes takes on more than two values, for example, *small, medium, large*. A discrete integer node takes on possible values between two finite integer values. For example, a node that represent the number of days before treatment which normally takes between one and sixty days.

The values defined for a node should adequately represent the level of detail required by the model, [135]. However, it is important to balance granularity with efficiency. The values⁵ of discrete variables must be mutually exclusive [23] and collectively exhaustive. This means that there must not be overlapping of states and all states must account for all possibilities. For example, let us consider a possible initial design to encode a query node labeled “Treatment” representing different treatment options for cancer patients. Let us consider four treatment options namely: *Chemotherapy; Radiotherapy; Surgery* and *Intervention Radiology*. Suppose we use these four options as the possible states of the treatment node. This initial modeling choice does not satisfy the exhaustive property because other possible strategies such as *watchful waiting* are excluded. The exhaustive condition can be satisfied by adding another state *Other Treatment* to encode all other possible treatments. In some cases, cancer patients are given more than one treatment strategy. For example, a patient with cancer of the Liver can be treated with a surgical

⁵Values and states are used interchangeably to represent possible configuration of a discrete or categorical variable.

intervention followed by chemotherapy. Hence, a patient who has received both surgery and chemotherapy violates the exclusivity property. To address this, we might introduce a further state (*Combination therapy*) to encode information about treatments combination.

2.3.6 Construct the DAG (qualitative part)

Once we have identified all the variables of a problem domain, the next stage is to describe the relationships between variables. One commonly used approach is the *causal relationship analysis* [135] by the domain experts. The expert must identify variables that cause another variable to take a value or prevent it from taking a given value. In addition to using experts' opinion, the dependency structure of a BN can be learnt for the historical data.

2.3.7 Assess the conditional probabilities (quantitative part)

Finally, we define a node probability table (NPT) for each node. The NPT quantifies the strength of the causal relations. This might be a discrete probability table for a discrete node or a continuous probability distribution for a continuous node.

We can estimate NPT subjectively using experts' opinion or from historical data. The most common sources of information for probabilities are from the literature and domain experts. In data-rich application domains, statistical data might be used to estimate probabilities. However, there is scarcity of data in many application domains. Therefore, the knowledge and experience of experts in the domain is the main source of probabilistic information, [116]. A number of formal methods have been developed for estimating probabilities, the most commonly used being the structured interviews with experts, [153, 189, 230]. However, given the extremely tight schedule of a domain expert, it might be difficult, if not impossible, to apply this technique to a real-life problem [51].

2.4 Inference in Bayesian Networks

The main goal of constructing a BN is to perform inference about unobserved variables given the observed variables. The BN inference algorithm computes the posterior marginal distributions for all unobserved nodes given the evidence nodes. The computational complexity of the inference depends on the structure of the BN model. The structure

of a BN can be singly connected or multiply connected depending on the causal relations between variables.

A singly connected network is one in which the underlying undirected graph has not more than one path between any two nodes. A multiply connected network, on the other hand, is a network with two or more paths between any two nodes in the underlying undirected graph. Figure 2.10 shows an example of a singly connected BN (on the left) and a multiply connected BN (on the right).

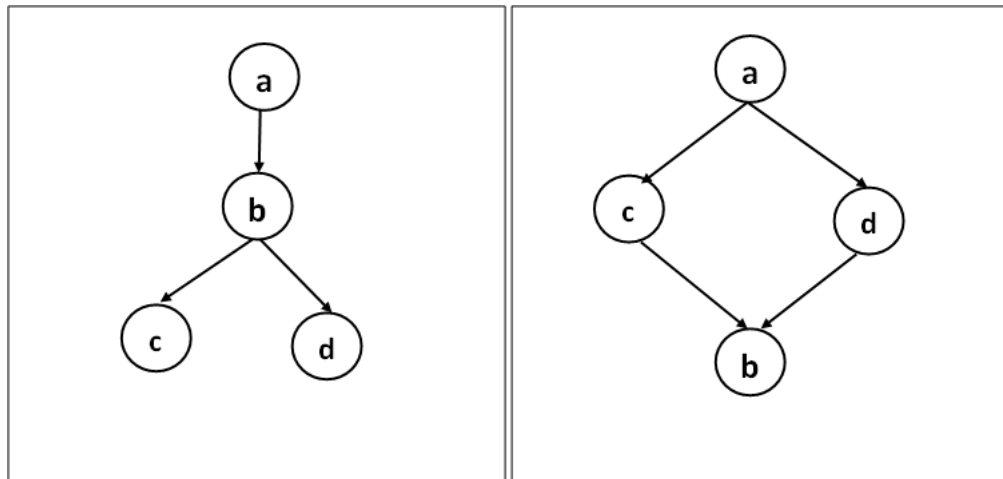


Figure 2.10: Singly connected and multiply connected BN

Several exact inference algorithms have been developed for performing inference on singly-connected networks [117, 130, 177, 178, 208–210, 212, 250]. However, exact inference in multiply connected network is known to be NP-hard⁶ [34, 42].

In the early 1980s, Pearl developed message propagation inference algorithm for a singly connected network [117, 177, 178]. This is an exact algorithm which has polynomial complexity in the number of nodes. Also in the 1980s, Shachter published an arc-reversal algorithm,[208] for inference on a singly connected network. This algorithm reverses the links using Bayes' rule by applying a sequence of operators to the network. Zhang and Poole developed a variable elimination algorithm that eliminates other variables one by one by summing them out [250]. In the 1990s, Lauritzen and Spiegelhalter published the junction tree algorithm [125], also called the clique-tree propagation algorithm or “clustering” algorithm. This is a very popular exact BNs inference algorithm.

Probabilistic inference on a BN is the process of computing the posterior marginal

⁶A problem that could not be solved in polynomial time.

distribution given the evidence nodes.

$$p(X|E = e)$$

Therefore a version of Equation 2.9 that takes into account evidence E in the BN is given in Equation 2.11.

$$P(X_1, X_2, \dots, X_n | \mathbf{E}) = \prod_{i=1}^n P(X_i | pa(X_i), \mathbf{E}) \quad (2.11)$$

In many cases, clinical decisions are made under uncertainty, [81, 198–200, 202]. A laboratory test result or an image from an imaging technique are normally used to determine the cause and severity of illness, predict its clinical course and evaluate the patient's response to a treatment. However, these are themselves agents of uncertainty, [40], since diagnostic tests are prone to systematic and random error, both in the conduct of the test and in the analysis of the results [84, 105, 197, 198]. More so, the technological advances, over the years, means that the search space for such diagnostic tests or imaging techniques or even treatment options has also increased. For example, aneurysm can now be treated effectively by surgical clipping and embolization coiling, whereas the only available option was surgical clipping a few decades ago. The objective of a clinical decision is to maximize the probability of the desired health outcomes. Clinical decision analysis is an objective and explicit use of model to represent clinical decision problems [211]. The CDA is the vehicle driving the application of evidence based medicine [100]. The CDA is a mathematical approach for analyzing difficult decisions faced by clinicians, patients and everyone involved in the clinical decision making process. This thesis proposes the use of Bayesian Network technique as a unified framework for integrating clinical evidence and also for analyzing clinical decisions. Medical knowledge is growing at an ever increasing rate and so is the use of Bayesian network models in the medical domain. Bayesian network models can assist clinicians in making decisions by integrating numerous probabilities that may overwhelm the human mind.

2.4.1 Junction Tree Algorithm

The Junction Tree Algorithm (JTA), proposed in [125] and refined by [107], translates a given BN into another tree called a Join Tree (JT). This Join Tree organizes the joint distribution so that computation of marginal distributions is handled efficiently. Similar to the BNs, a Join Tree also has both quantitative and qualitative parts. At the qualitative level, we have clusters of variables often referred to as the cliques in the literature. A clique on the Junction tree is represented by a circle. Unlike the BN graph, a connection between any two cliques is established through their intersecting variables often called *sepsets*. A sepset is conventionally drawn as a rectangle on the Join Tree. A Join Tree must satisfy the following properties, [125]:

1. each variable must appear in at least one cluster involving all its parent;
2. if a variable appears in any two clusters, say V_1 and V_2 , then it must also appear in every cluster on a path between V_1 and V_2 .

At the quantitative level, each clique on a Join tree has a corresponding potential function (ψ). JTA determines the cliques membership and assigns a variable to the clique containing its parents. The transformation of a BN into a joint tree follows a set carefully defined procedures. These procedures are summarized into three steps.

- Moralization: This involves linking pairs of nodes that have a common child, and converts a directed graph into an undirected graph.
- Triangulation: The moral graph is triangulated such that there are no cordless cycles in the graph.
- Then construct a JT from the triangulated graph.

For a detailed discussion see [125, 178] or see [97] for a procedural guide. As an illustration, suppose we have a BN model, shown on the left of Figure 2.11, with four nodes A , B , C and D . The Node labeled C is the only parent node to D and nodes A and B are the parents to C .

The joint probability distribution for this BN model is shown in Equation 2.12.

$$P(A, B, C, D) = P(A).P(B).P(C|A, B).P(D|C) \quad (2.12)$$

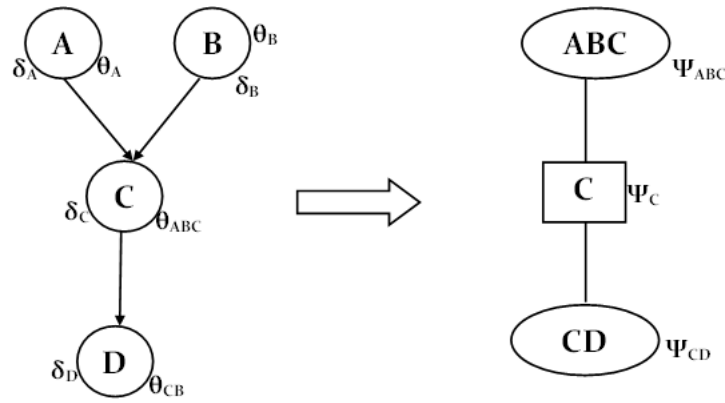


Figure 2.11: Simple Join Tree from BNs

From Figure 2.11, θ_A , θ_B , θ_{ABC} and θ_{CD} represent the NPTs defined respectively over A, B, C and D while δ_A , δ_B , δ_C and δ_D represent their respective evidence.

The resulting Join Tree from this BN is given on the right of Figure 2.11 with two cliques drawn in circle and a sepset containing common variable C to both cliques. Let us label these cliques as V_1 for clique containing ABC and V_2 for clique made of variables C and D . The sepset is represented by $V_1 \cap V_2$.

2.4.2 Potential function

Just like we defined a node probability table for each node on a BN, we need to define a potential function on every node on a Join Tree. Let us denote the potentials on V_1 and V_2 as Ψ_{ABC} and Ψ_{CD} respectively while Ψ_C denotes the potential function on the sepset C . These potentials are initialized as follows:

$$\Psi_{ABC} = \theta_A \cdot \delta_A \cdot \theta_B \cdot \delta_B \cdot \theta_{ABC} \cdot \delta_C \quad (2.13)$$

$$\Psi_{CD} = \theta_{CD} \cdot \delta_D \quad (2.14)$$

$$\Psi_C = 1 \quad (2.15)$$

2.4.3 Message Propagation on Junction Tree

The main goal of inference is to compute the posterior marginal distribution of some query nodes given observations from some other nodes. Message passing is the process through which the JT attains local consistency, that is a state such that the revised potential on each clique is the marginalization of the joint probability distribution to variables in the domain of the clique. Messages are passed between two neighboring cliques via

their intervening sepset, message passed from clique $V_1 = \{A, B, C\}$ to clique $V_2 = \{C, D\}$ forces the intervening sepset $V_1 \cap V_2$ to be consistent with V_1 . Global propagation induces every cluster to pass a message to all its neighbors in such a way that each message pass will preserve the consistency introduced by previous message passes. When global propagation is completed, each cluster-sepset pair is consistent, and the Join Tree is locally consistent [97]. From this point we will use the term “node” to represent a clique on the junction tree just like it was used to represent a variable in the BNs.

There are two nodes in the Join Tree in Figure 2.11 represented by, V_1 and V_2 . One of these nodes is chosen as the root node, the choice of the root node is normally any node in the JT whose members include the query variables in the underlying DAG. For instance if we are interested in the posterior marginal distribution of node A in the underlying BN, then V_1 can be chosen as the root node. Message passing between two adjacent nodes occurs in two steps (projection and absorption) [97]. Projection involves marginalizing out variables that are not in the intervening sepset, (i.e. variables in the *sending clique* that are not in the *receiving clique*). Projection ensures that the dimension of the incoming message is the same as the dimension of potential table defined over the intervening sepset. Absorption, on the other hand enables us to combine incoming message with the table in the receiving clique. Following these steps, we can describe message passing on the JT.

2.4.4 Projection of Message from $V_1 \rightarrow V_2$

V_1 computes its message by summing over local variables A, B (Equation 2.16).

$$\Psi_{ABC}^{\downarrow(C)} = \sum_{AB} \Psi_{ABC} \quad (2.16)$$

$$(2.17)$$

This message is assigned to the intervening sepset $V_1 \cap V_2$. Borrowing the same notations used by Huang and Darwiche [97] to denote the old and new table on the sepset, then we can store the potential function for sepset C and assign a new function as given in Equation 2.18 and Equation 2.19 respectively.

$$\Psi_C^{old} \leftarrow \Psi_C \quad (2.18)$$

$$\Psi_C^{new} \leftarrow \Psi_{ABC}^{\downarrow(C)} \quad (2.19)$$

2.4.5 Absorption of Message into V_2

The second leg of the message passing is to combine the incoming message from V_1 with the potential (ψ_{CD}) of V_2 . This step updates the potential with the incoming message. A new potential is assigned to V_2 as given in Equation 2.21.

$$\Psi_{CD} \leftarrow \Psi_{CD} \frac{\psi_C^{new}}{\psi_C^{old}} \quad (2.20)$$

$$(2.21)$$

This is a simple example where the receiving clique has just one neighbor. However this same process is merely repeated in cases involving more than one neighboring cliques. It is important to note that once a receiving clique has received messages from all of its neighbors, the clique is said to be locally consistent. Clique V_2 becomes locally consistent after receiving messages from V_1 and we can compute marginal distributions for C and D . However, if we are interested in the marginal distributions for A and/or B which can only be computed from V_1 , we need to make V_1 consistent by sending a message based on the revised potential for V_2 back to V_1 in the same way. Once we have attained a local consistency state in all the cliques, we can compute the marginal for any node in the network as follows. Assuming we are interested in the marginal for node B i.e. $p(B)$, we can compute this in the following steps:

- Identify a cluster (or sepset) that contains node B i.e. clique V_1 . The updated potential function for V_1 after receiving message from V_2 is Ψ_{ABC}^* .
- Compute $p(B, e)$ by marginalizing Ψ_{ABC}^* as shown in Equation 2.22.

$$p(B, e) = \sum_{AC} \Psi_{ABC}^* \quad (2.22)$$

Where Ψ_{ABC}^* is the updated potential of V_1 based on the message received from V_2 .

- Having computed $P(B, e)$ where e represent the evidence entered into in the model, we can

now compute $P(B|e)$ by normalizing $p(B, e)$ as shown in Equation 2.23:

$$p(B|e) = \frac{p(B, e)}{p(e)} = \frac{\sum_{AC} \Psi_{ABC}^*}{\sum_B p(B, e)} \quad (2.23)$$

2.5 Learning parameters & structures of probabilistic Networks

So far we have been dealing with examples of BN models in which both the qualitative and quantitative parts are known or provided by experts. Indeed, we can come up with the parameter and structure of a BN by using the knowledge gathered from domain experts. A number of techniques has been developed to assist analyst in the elicitation of probabilities. Two well cited references in this context are Good [78] and Winkler [239]. However, these techniques may be extremely time consuming for a realistic clinical network involving many parameters [122]. Druzdzel and van der Gaag [52] also offer an important technique for eliciting the parameters of a BN model from the experts. Because of the very tight schedule of experts (especially clinicians), analysts often resort to learning parameters or/and causal relations from the available data when data is available.

The most straightforward learning situation is to learn parameters of BN models with known structures from the available data. In some cases, there could an adequate knowledge of the domain to define relations between variables. Thus the causal relations are taken as given and then used to learn parameters in the network. However, some situations require the use of the data to learn both the structure and then parameters. A wide range of techniques have been developed for learning parameters of Bayesian networks. Heckerman *et al.* [91] described a method for learning networks with only discrete variables while Geiger and Heckerman [71] described an approach for learning parameters in Gaussian networks. Btcher and Dethlefsen, [20] generalized these methods by describing a technique for learning the parameters and structure of a Bayesian network with discrete and continuous variables. They defined the local probability distributions such that the joint distribution over a set of random variables is a conditional Gaussian (CG) distribution. Therefore discrete variables are not allowed to have continuous parents, so the network can factorize into a discrete part and a mixed part [21]. Their technique has been implemented into a free causal learning software (*Deal* [20]), written in R [99]. This software provides methods for learning both parameter and structure of Bayesian networks. The package also includes procedures for defining priors, calculating network scores, performing heuristic search and a procedure for simulating data sets from a given dependency structure. One assumption common to all these

techniques is the parameter independence assumption. This means that the parameters of a node can be learned independently of the parameters of the other nodes [20].

A major limitation of these algorithms is that they require complete dataset. In particular, they assume zero value for any missing observation and then perform the analysis as if zero observation was observed. To address this limitation, a number of learning techniques have been described [194–196] to deal with learning problems with incomplete dataset. A survey of various learning techniques can be found in [122].

In a situation where the structure of a BN is unknown, we can learn both the structure and parameters from data. Here we will stick with the use of *structure* as against *causality* because data driven approach would require biological explanation to justify the use of the term *causality*, especially in the clinical context. The Deal package *deal* [20] is an example of algorithm for learning structure and parameters from data. The central goal of all structure learning algorithms is to identify structure that provide the best approximation for the data. In other words, they tend to identify a BN model that best fit the data. The problem is that the search space for all possible structures is more than exponential. Hence exploring the entire search space for a large network is an intractable problem.

The question now is, how do we identify the structure that provide the best fit? A widely used general strategy is to combine the scoring criterion with a search algorithm. Conceptually, learning structures involves two ideas namely scoring and searching. A search algorithm determines structures within the search space to be scored while the scoring algorithms provides an empirical score for each structure. Each of the structure identified can be considered as an hypothesis competing with other hypotheses.

The Bayesian Dirichlet (BD) criterion is one scoring criteria commonly used for the categorical data. Madigan and Raferty [140] proposed a scoring metric that uses the idea of relative posterior probability in conjunction with heuristics based on the principle of Occams Razor. Others include the Akaike information criterion (AIC) [85], the Bayesian information criterion (BIC) [69] and minimum description length [191].

2.6 Software Packages for Graphical Models

The JTA has been implemented in a number of popular commercial BNs software tools including Hugin [98] and AgenaRisk [131] among others. The growing interest in Bayesian network research, over the past few decades, has been accompanied by the proliferation of BN software tools

developed to support research efforts and applications of BNs to a wide range of domains. There are several commercial [36, 98, 131, 133, 136, 220] and noncommercial [14, 20, 33, 43, 44, 58, 79, 88, 110, 126, 134, 139, 149, 154, 155, 157, 159, 161, 175, 182, 186, 193, 207, 215, 217, 221, 227–229, 240, 246] software tools for developing BN models. These tools provide a graphical editor for building the BN and inference mechanism for evidence propagation.

In this thesis, all BN models are built using the AgenaRisk toolset [131] and the posterior distributions, in some cases, are compared with those from either Hugin [98] or Winbugs [217]. Table 2.3, adapted from the work of Kelvin Murphy [160], compares features of these three software packages and their suitability to this research work. For comparison with other Bayesian software package, see [160].

Table 2.3: Comparing features of AgenaRisk, Hugin Expert and Winbugs packages

Features	AgenaRisk	Hugin Expert	Winbugs
Availability of Source Code	Yes, Java source code available to research students.	No	No
Application Programming interface	Java	Java, C, C++ C# (.Net) Visual Basic	No API
Handling of Continuous variables	By Dynamic discretization and inference calculation with Junction Tree. A wide range of built-in statistical distributions and expressions for constructing NPTs.	Restricts distribution on the continuous variables to conditional Gaussian distribution	simulation using Gibbs sampling
Allows parameter learning	Yes	Yes	Yes
Support for Utility node	No	Yes	No
Supports for Dynamic BN models (DBN)	uses the concept of Object oriented BN models.	DBN by creating different BN forests	Yes
Structure learning	No	By conditional independence test	No
Allows discrete child of continuous parent	Yes	No	No

We cannot consider anyone of these three Bayesian tools as the overall best, each of them has some strengths over the remaining two softwares. Therefore, a single software tool may not provide all the functionalities required, so it is common to use a combination of software tools in research projects. A crucial consideration in the choice of a tool for this project is the availability of the source code. With the source code, a researcher can implement any functionality that are not available in the official release of a software. For example, one of the hypotheses of the thesis is to confirm whether the combination of the Junction tree and dynamic discretization can be used for scoring causal relations. This functionality is not available in the official release of AgenaRisk software but was implemented for the purpose of the research using the source code. To this end, we will use the AgenaRisk software in this work and wherever possible, we will compare the result with those obtained using Winbugs or HuginExpert.

2.7 Application of BN models to clinical decision making

Bayesian Networks models have been widely used in the medical domain since algorithmic breakthroughs in the late 1980s [125, 178]. This meant that large-scale BN models could be efficiently calculated. Clinical decision-support systems based on BNs were first developed in the late 1980s [11, 92]. Since then hundreds of BN papers have been published within the medical do-

main. For example, Bayesian network models have been developed for the diagnosis of diseases, [1, 2, 41, 87, 112, 113, 115, 137, 138, 151, 169, 236, 238]. For example, Bayesian network models have been developed for diagnosing pyloric stenosis [2], breast cancer [236] and pneumonia [5]. Also, a BN model for the diagnosis and procedure selection for patients suspected to have gallbladder disease was described in [86]. Kline *et al.*, [118] also applied a Bayesian network model to predict the pretest probability of venous thromboembolism. Lukas *et al* [183] described a decision-theoretic model for the management of gastric non Hodgkin lymphoma (NHL). BN models have also been used in predicting the risk of diseases [24, 118, 204, 205] and the risk of specific medical outcomes [75, 95, 114, 203, 234]. Other clinical applications of BN include models for monitoring patients in intensive care [11], radiology [23], therapeutics [171] and biomedical informatics [233] and clinical models for improved medical procedures [8, 86, 108, 128, 129, 148].

General guidelines on using BNs in medical applications can be found in [141, 156, 170], while comparisons of BNs with alternative approaches in the medical context can be found in [5, 49].

Despite the headway that has been made in the application of BNs in the medical domain they have yet to become mainstream. Whilst it is outside of the scope of this thesis to offer definitive explanations for this a number of factors seem to be at play, each of which might limit the extent to which technology transfer has been and might be successful in future. Of course, these factors are just as relevant to this thesis as it is to past research.

In addition to purely economic considerations there appear to be three broadly intertwined reasons for successful technology transfer: organisational, cultural and technological. Bayesian modelling, like all mathematical disciplines, is a craft industry where each model is designed around the domain and context, hand in hand with domain experts. This leads to the production of bespoke models most useful in the particular context they were designed for. If the context changes, or is difficult to generalise, or even communicate to those outside of the immediate situation any benefits of the model, as they appear to others, may seem limited. Also, although Bayesian ideas have been around for a long time, the gestation period for a new technology, to move from laboratory into the field can be lengthy. Frequentist statistics took decades to move into the mainstream and their use is relatively pain-free, in terms of up front time investment, compared to Bayesian methods (but with less gain). Lastly, the issue of incentives looms large in determining whether a successful BN model is used beyond the domain within which it was developed: researchers can have little vested interest in technology transfer since it does not involve the generation of new

knowledge, similarly, but perhaps much less so, for clinicians sheer pressure of work means that technology transfer takes time and is subject to changing priorities. Having discussed this with my supervisors we agreed that any attempt to tackle the point in more detail would extend the scope of the thesis beyond that defined and constrained by the hypotheses. There are many reasons for the lack of any technology to “cross the chasm” into routine use, amongst these organisational, cultural and resource constraints.

What this thesis does is explicitly recognised the issue, the constraints in time etc that prevents it being systematically addressed as part of my research and the limitations that naturally flow from this.

2.7.1 Bayesian classification model

A Bayesian classification, which is a form of Bayesian network, is a commonly used model in the clinical domain. Just like other classification techniques, a BN classification model can classify an object into one of the c mutually exclusive classes. The simplest form of classification is a situation with two possible classes ($c = 2$) such as “Yes” or “No” indicating absence or presence of a disease. For example a clinician can use such model to classify a patient’s case as benign or malignant given the clinical observations on the patient. Bayesian classifiers have been used in different areas in the clinical domain. For example, Jian *et al.*, [109] used Bayesian classifier to develop a computer-aided diagnosis of Cerebral Aneurysm. They process the cerebral vessel image and used the geometrical shape characteristics extracted from a suspected lesion area to construct feature vector. Burnside *et al* [23] also used a Bayesian classifier model to predict breast cancer. In the work of Yueyi *et al* [247], they apply bayesian classification model to differentiate benign versus malignant thyroid Nodules using sonographic Features. They included all sonographic features known to be predictors of malignancy in their model. Also included are variables that significantly influence the probability of a nodule being malignant such as age and gender. They discretized age into two crude states i.e. < 50 (which accounts for 75% of the population) and ≥ 50 (25% of the population). Figure 2.12 shows the resulting BN model described by Yueyi *et al* [247].

Although, discretizing age in this way reduces the number of parameters required in their model but this simplicity comes at the expense of accuracy of the result. Any observation entered on node age is not precise. The model treats a 20 years old patient and a 49 years old patient alike and makes no distinction between a 50 year old patient and 89 year old patient.

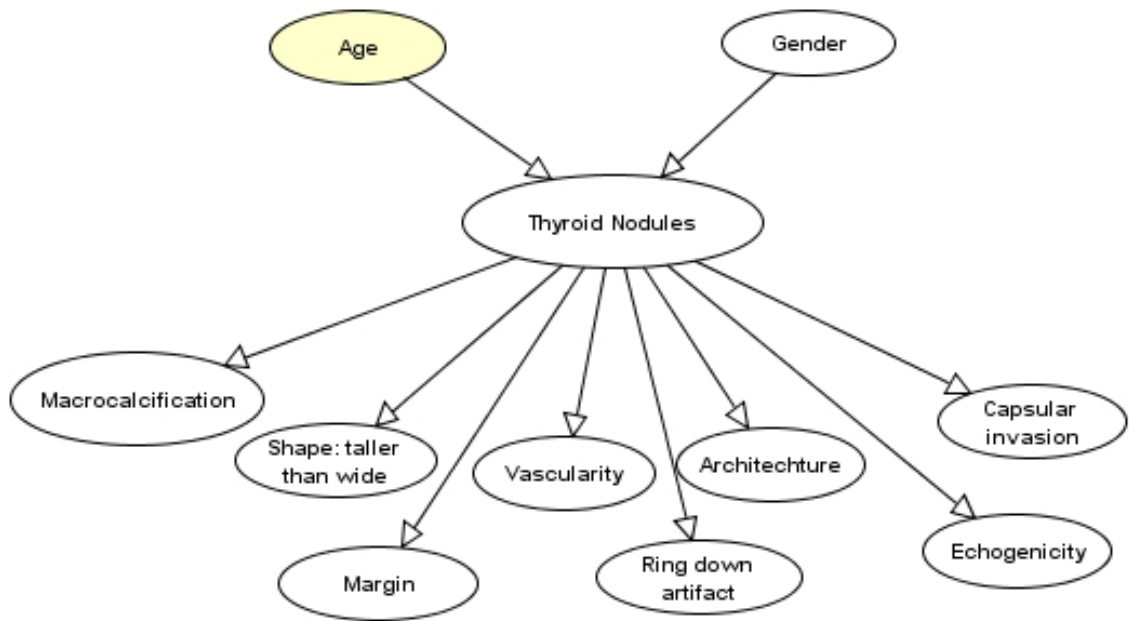


Figure 2.12: A Bayesian classifier for thyroid nodules [247]

2.7.2 A BN model to support clinical decision for diagnosing pyloric stenosis

This section reviews a BN model to support decision for diagnosing pyloric stenosis. When evaluating an infant with suspected pyloric stenosis, a clinician uses available information to estimate the likelihood of the disease. This likelihood is updated as additional information is obtained. Sonia *et al.*, [2] conducted a study to evaluate the feasibility of using a Bayesian network to improve the accuracy of diagnosing pyloric stenosis. The potential factors of pyloric stenosis identified include *age, sex, race, family history of pyloric stenosis, number of days of vomiting, projectile vomiting, increasing vomiting, diarrhea, weight loss* while laboratory tests include (*sodium, chloride, potassium, serum bicarbonate, and total bilirubin*), [174, 213, 231].

Figure 2.13 shows their BN model with the node for pyloric stenosis dependent on risk factors and the nodes for the clinical features such as symptoms and laboratory findings dependent on the node for pyloric stenosis.

In this analysis, two continuous variables were manually discretized. The node *age* was discretized with a 2-week age intervals until age 16 weeks after which all ages were grouped together. For the node representing the *number of days of vomiting*, 1-day increments were used until 1 week and 1-week intervals until 7 weeks after which all time intervals were grouped. Clearly, the analysts knew a priori the high density region in the *number of days of vomiting* variable. The first week is partitioned into days with 7 partitions whereas other weeks, from week 2 to week 7, were grouped together. In essence, we can say that pyloric stenosis may last for several weeks but ma-

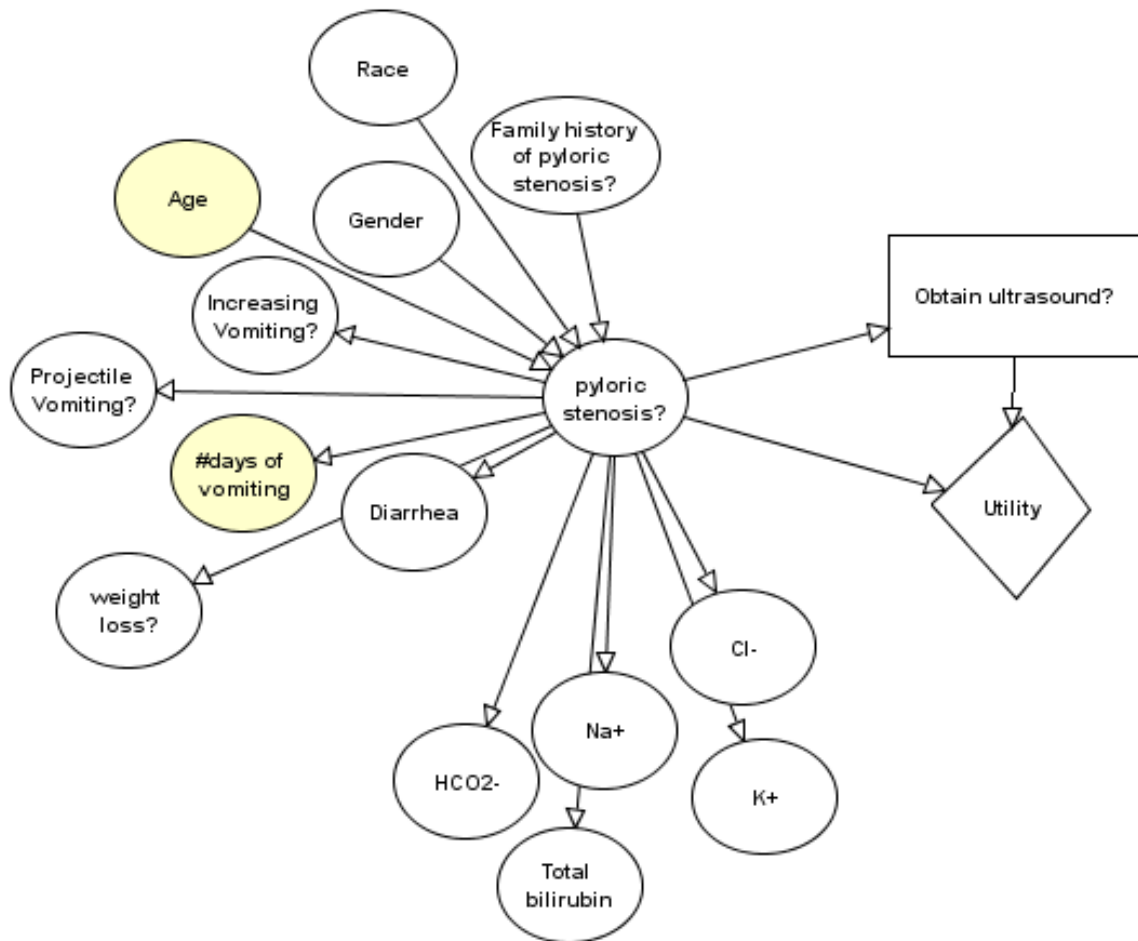


Figure 2.13: A Bayesian decision network for diagnosing pyloric stenosis, [2]

majority of the affected children may recover after one week. The static discretization approach relies on the expert to know the high density region before the analysis.

2.8 Summary

In this chapter, we have described the core concepts of Bayesian networks and Bayes' theorem. The chapter presented a basic overview of the BN framework and described the steps in constructing a BN model. The suitability of the BN approach to model clinical problems was emphasized by drawing the reader's attention to two important features of the BN framework. These are the compact representation of probabilistic problems for efficient probabilistic inference and capability for forward and backward reasoning. This chapter also provided a brief description of the Junction Tree Algorithm and message passing procedure on the Join Tree. A detailed description can be found in Huang and Darwiche [97].

The chapter also described two BN models applied to solving problems in the clinical domain.

These models show the approach used in tackling BN models with continuous variables. Although the approach is convenient but the simplicity comes at a cost of sacrificing accuracy of the result. In the next chapter, we will review an existing technique for discretizing continuous variables within the BN framework.

Chapter 3

Dynamic Discretization Algorithm

The Junction Tree Algorithm described in the previous chapter performs inference on BNs with discrete nodes. This chapter reviews the concept of Hybrid Bayesian Network and also describes an algorithm for approximating continuous distribution functions (the Dynamic Discretization Algorithm).

3.1 Introduction

So far, we have described the concept of Bayesian Network and the Junction Tree Algorithm for solving inference problems. Often, analysts resort to the static discretization approach in handling BN models with both discrete and continuous variables. For example, an analyst may decide on using three partitions say $[0 - 36.1]$, $[36.1 - 37.5]$ and $[37.5 - 43]$ to model the body temperature. Suppose this analyst later realizes that three values does not adequately model this situation, and then decides on five values or seven values or even more. The greater the number of partitions, the slower the processing time. Even worse, the analyst cannot say apriori which range of temperature would be the high-density region given some evidence. Therefore, the analyst might consider treating the temperature as a continuous variable. We will now describe a Bayesian network with both discrete and continuous variables.

3.2 Hybrid Bayesian Networks (HBNs)

Hybrid Bayesian Networks (HBNs) contains both discrete and continuous variables. Theoretically the Junction Tree Algorithm is extendable to provide solutions to HBNs by merely replacing summation with integration when computing marginal distributions. However performing multidimensional integration is computationally expensive. Hence actual implementations of algorithms for HBNs perform some kind of approximations to the continuous probability density functions which cannot be integrated in closed form [31]. Some of the techniques for approximating continuous functions include stochastic simulation known as Markov Chain Monte Carlo (MCMC) [74, 77], Dynamic Discretization Algorithm (DDA) [120, 168] and Mixture of Truncated Exponential (MTE) [188] among others. Indeed analysts can manually discretize continuous variables and then treat as a discrete model using static discretization.

3.3 Analytical Solutions

The concept of “conjugate prior” applies to a prior distribution $p(\theta)$ that yields a posterior distribution $p(\theta|x)$ from the same distribution family as the prior probability distribution $p(\theta)$. In this case, the prior $p(\theta)$ is a conjugate prior for the likelihood $f(x|\theta)$. Table 3.3 shows the analytical posterior distributions for normal, binomial and multinomial likelihood distributions.

Table 3.1: Conjugate prior distribution and analytical expression for posterior inference

Conjugate family	Prior	Likelihood	Posterior
Beta distribution	$\theta \sim \text{Beta}(\alpha, \beta)$ $f(\theta \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ $E(\theta) = \frac{\alpha}{\alpha+\beta}; V(\theta) = \frac{E(\theta)(1-E(\theta))}{\alpha+\beta+1}$	$y \sim \text{Binomial}(n, \theta)$ $f(y \theta) \propto \theta^y(1-\theta)^{n-y}$ $E(y) = n\theta; V(y) = n\theta(1-\theta)$	$(\theta y) \sim \text{Beta}(\alpha+y, \beta+n-y)$ $f(\theta y) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$ $E(\theta y) = \frac{\alpha+y}{\alpha+\beta+n}; V(\theta y) = \frac{E(\theta y)(1-E(\theta y))}{\alpha+\beta+n+1}$
Dirichlet distribution	$p(\theta_i) \sim \text{Dirichlet}(\alpha^a)$ $f(\theta \alpha) = \frac{\Gamma(\alpha_1+\dots+\alpha_k)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$ $E(\theta_j) = \frac{\alpha_j}{\sum_i^k (\alpha_i+x_i)}; V(\theta_j) = \frac{\theta_j(\sum_i^k (\alpha_i+x_i) - \theta_j)}{[\sum_i^k (\alpha_i+x_i)]^2}$	$y \sim \text{Binomial}(n, \theta)$ $f(y n, \theta) = \frac{n!}{y! \dots y_k!} \theta_1^{y_1} \dots \theta_k^{y_k}$ $E(y) = n\theta_j; V(y) = n\theta_j(1-\theta_j)$	$(\theta y, n) \sim \text{Dirichlet}(\alpha^{*b})$ $f(\theta n, y) \propto \theta_1^{\alpha_1+y_1-1} \dots \theta_k^{\alpha_k+y_k-1}$ $E(\theta_j y, n) = \frac{\alpha_j+y_j}{\sum_i^k (\alpha_i+y_i)}; V(\theta_j) = \frac{\alpha_j(\sum_i^k \alpha_i - \alpha_j)}{[\sum_i^k \alpha_i]^2 (\sum_i^k \alpha_i + 1)}$
Normal distribution	$\theta \sim \text{Normal}(\mu_0, \tau_0^2)$ $f(\theta) \propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2}$ $E(\theta) = \mu_0; V(\theta) = \tau_0^2$	$y \theta \sim \text{Normal}(\theta, \sigma^2)^c$ $f(y \theta) \propto e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$ $E(y \theta) = \theta; V(y \theta) = \sigma^2$	$\theta y \sim \text{Normal}(\mu_1, \tau_1^2)^d$ $f(\theta y) \propto e^{-\frac{1}{2}\{\frac{y-\theta}{\sigma^2} + \frac{(\theta-\mu_0)}{\tau_0^2}\}^2}$ $E(\theta y) = \mu_1; V(\theta) = \tau_1^2$ see note ^c

^a $\alpha = \alpha_1, \dots, \alpha_k$ ^b $\alpha^* = \alpha_1 + y_1, \dots, \alpha_k + y_k$ ^cVariance σ^2 is known^d $\mu_1 = (\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y) / (\frac{1}{\tau_0^2} + \frac{1}{\sigma^2})$ and $\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$

The use of a conjugate prior is computationally convenient as the posterior distribution can be expressed (analytically) in a closed-form thereby eliminating the challenge of difficult numerical integrations. The next two sections review two approximation techniques, static and dynamic discretization, to calculate posterior distributions in models with non conjugate priors.

3.4 Static Uniform Discretization

Static discretization is a relatively easy approximation approach, it is obtained by subdividing an interval into sub-intervals such that the density for each of these sub-intervals can be approximated by a piecewise constant distribution. We can select a point, in each sub-interval, and assign a constant probability mass to this point. The combination of the selected points and the assigned probabilities provide the approximation for a discrete distribution. Two widely used methods of static discretization are the extended Pearson-Tukey method [50] and the bracket median method [30].

The extended Pearson-Tukey method is a three-point approximation of a continuous distribution in which a continuous distribution is approximated by a discrete distribution with probabilities 0.185, 0.63, and 0.185 assigned to the 0.05, 0.5, and 0.95 fractiles of the continuous distribution. The *fractile* is similar to *percentile*, except that it is expressed as a fraction rather than a percentage thus a 0.05 fractile is the same as 5th percentile.

In the bracket median method, a continuous distribution is discretized into n equally likely intervals. The median of each interval is used as the value representing the interval and each point is assigned a probability $\frac{1}{n}$.

The static discretization approach is flawed for two reasons. Firstly, it requires that an analyst understands the posterior high density region of a continuous distribution and must guess the state ranges before performing inference, thus presupposing that they know the resulting probability distribution of the results beforehand. This may be quite easy, in simple cases, but it would be difficult to guess the posterior density region of a continuous node with several observable child nodes. Secondly, evidence on a poorly discretized node would be less precise. Consequently, unreliable summary statistics may result from a poor discretization of continuous distributions.

3.5 Dynamic Discretization

In contrast to the uniform discretization, the dynamic discretization technique discretizes a continuous density function such that the region whereby the density is changing rapidly (conditional on the evidence) are discretized more than other regions. This thesis focuses on the application of the Dynamic Discretization Algorithm (DDA) for approximating continuous distribution functions. The technique was proposed by Kozlov and Koller [121] for the purpose of supporting inference in HBNs described in section 3.2. Kozlov and Koller [121] described an iterative scheme for partitioning multivariate continuous functions in a non-uniform way. Their approach used the relative entropy or Kullback-Leibler (KL) distance [123] as a metric for measuring the error introduced by discretizing continuous density functions. Formally, let $f(x)$ be a continuous probability density function and $g(x)$ be an approximation to $f(x)$, the KL distance measures the divergence between $f(x)$ and $g(x)$ and is given in equation 3.1

$$D(f(X)||g(X)) = \int (f(x) \log \frac{f(x)}{g(x)}) dx \quad (3.1)$$

Since exact computation of the relative entropy error is computationally expensive for a general function, they proposed a bound on the KL distance between the continuous function $f(x)$ and its discretization $g(x)$ based on the function mean \bar{f} , function maximum f_{max} and function minimum f_{min} in the given subregion w_j .

$$\int_{w_j} f \log \left(\frac{f}{\bar{f}} \right) d\Omega \leq \left\{ \frac{f_{max} - \bar{f}}{f_{max} - f_{min}} f_{min} \log \frac{f_{min}}{\bar{f}} + \frac{\bar{f} - f_{min}}{f_{max} - f_{min}} f_{max} \log \frac{f_{max}}{\bar{f}} \right\} |w_j|$$

where $|w_j|$ denotes the volume of a discretization subregion w_j and Ω denotes the entire region. Influenced by their work, Neil *et al* [168] describe a univariate discretization approach that works seamlessly with the Junction Tree Algorithm, [125]. The algorithm presented in Neil *et al* [168] for discretizing continuous function is presented below:

Algorithm 1 Dynamic Discretization

Require: BNs model with continuous variable $\mathbf{V} = V_1, \dots, V_n$
 Error threshold for each continuous variable $\xi_1 \dots \xi_n$
 Maximum iteration *maxit*
 Error threshold for the model ξ^v
 initialize partition on \mathbf{V} to $W = w_1, \dots, w_n$

for $i = 1$ to *maxit* **do**
 for $k = 1$ to n **do**
 Calculate NPT_k given w_k
 Compute posterior marginal distributions given current discretization w_k and evidence
 Query the BN to get posterior marginal distribution
 Compute E_k from Equation 3.2 $E_k \Leftarrow \sum_{j=1}^{|w_k|} E_j$
 if $E_k > \xi_k$ **then**
 Create a new discretization w'_k for the node by splitting interval with highest entropy error
 $w_k \Leftarrow w'_k$
 end if
 Compute global error $\xi_i^v = \sum_k^n E_k$
 end for
 if $\xi_i^v \leq \xi^v$ **then**
 stop
 end if
end for

Several Hybrid BN models have been constructed using the univariate Dynamic Discretization Algorithm. These include reliability modeling [147, 166, 167], BN models for solving dynamic fault trees problems [145, 146], operational risk models [164, 165], BN models for predicting software defects [59–61] and BN models for software project risk assessment [63, 90].

3.6 Beta-Binomial formulation

Section 3.3 describes an analytical expression for the posterior distribution of a parameter (θ) in a beta-binomial model. The Beta distribution is a conjugate prior distribution to the binomial likelihood. Therefore, the posterior distribution of (θ) can be expressed as given by Equation 3.2. This section shows the use of the Dynamic Discretization Algorithm to approximate the posterior marginal distribution of parameter θ in a beta-binomial BN model.

Example 3.1. To illustrate the strength of the Dynamic Discretization Algorithm, let us consider a simple inference problem involving two random variables x and θ . Variable y denotes the number of Heads from n repeated tosses of a coin and variable θ represents

the parameter i.e. probability of a Head from a single toss of the coin. We can proceed by assuming that these repeated tosses are exchangeable i.e the probability of a head (θ) is constant throughout the experiment. A convenient formulation of this problem is the beta-binomial formulation i.e. a beta distribution for the parameter θ and binomial distribution for the random variable y . Therefore, the posterior distribution for θ after observing y can be expressed analytically using the Bayes' Theorem as follows:

$$p(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \quad (3.2)$$

with mean and variance given as Equations 3.3 and 3.4 respectively.

$$E(\theta|x) = \frac{\alpha+x}{\alpha+\beta+n} \text{ and variance given as} \quad (3.3)$$

$$V(\theta|x) = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{E(\theta|x)[1-E(\theta|x)]}{\alpha+\beta+n+1} \quad (3.4)$$

Figure 3.1 shows a Bayesian Network representation of a Beta-binomial problem.

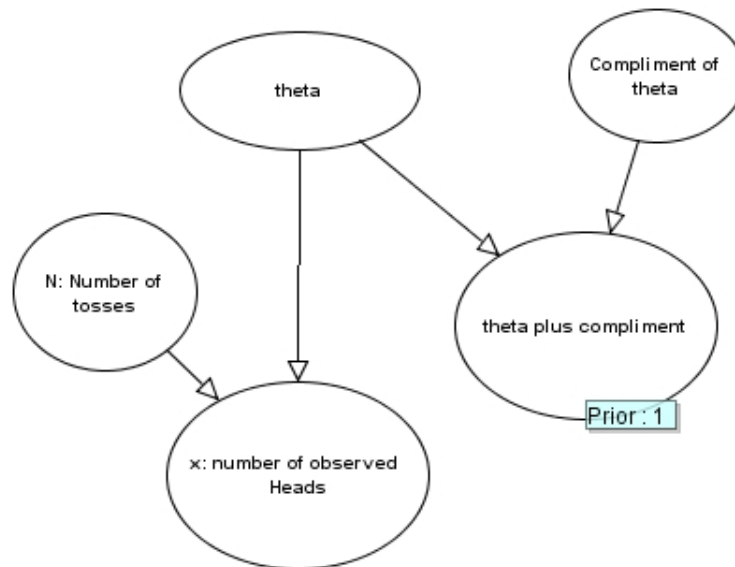


Figure 3.1: A Beta-Binomial BN model

In Figure 3.1, the parameter of interest (θ), is represented by a node labeled “theta” and the observable variable x , representing the number of “Heads” from n exchangeable tosses, is represented by a node labeled x . We also introduce an additional node labeled “compliment of theta” i.e. $(1 - \theta)$. This is achieved by ensuring that the sum of the

two posterior distributions add up to one i.e. $p(\text{Head}) + p(\text{Tail}) = 1$. The sum of these posterior probabilities is stored as another BN node “theta plus compliment”. We then enter hard evidence 1 on the “theta plus compliment” node to constrain the sum of two posterior distributions to one. Also we need to specify appropriate distributions for parameter θ and observation x . We can use both conjugate and non conjugate prior but here we will stick with the conjugate so that we can compare the result obtained from dynamic discretization with those from the analytical solution. We can assume the following prior for parameter θ :

$$\theta \sim \text{Beta}(1, 1) \quad (3.5)$$

This Beta distribution in Equation 3.5 is a non informative prior and a special case of the uniform distribution. The second stage is to specify a Binomial distribution for the observable data x .

$$x \sim \text{Bin}(n, \theta) \quad (3.6)$$

If after 100 tosses we observe 1 head then we can obtain a posterior mean and variance for parameter θ using Equations 3.3 and 3.4 as 0.01961 and 0.00019 respectively.

Figure 3.2 presents the posterior densities obtained using dynamic discretization with 25 iterations and those from static discretization with 25 uniform intervals.

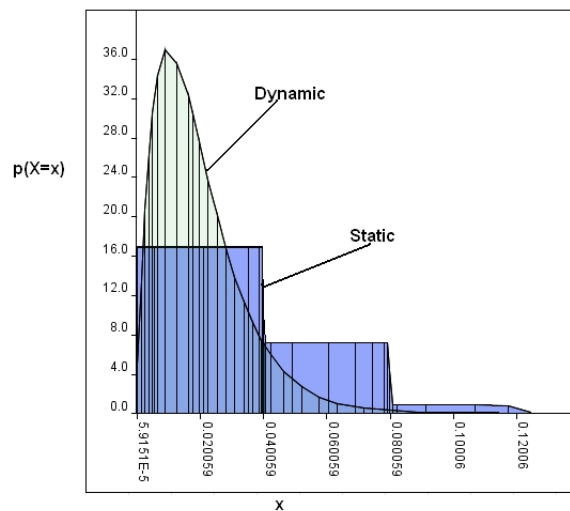


Figure 3.2: Posterior Densities of θ from dynamic discretization with 25 iteration and static discretization with 25 uniform intervals

We can repeat this calculation with 100 iterations for the model with dynamic discretization and 100 uniform partitions for the model with static discretization. Table 3.2 shows the posterior summary statistics, mean and variance, from static discretization with 25 and 100 uniform partitions and dynamic discretization based on 25 and 100 iterations.

Table 3.2: Summary of posterior distributions from static and dynamic discretization

Parameters	Analytical	25 iterations		100 iteration	
		Static	Dynamic	Static	Dynamic
Mean	0.01961	0.034499	0.01987	0.034499	0.019634
Variance	0.00019	0.0005	0.00021	0.00062	0.000194

Clearly, dynamic discretization achieves better accuracy than a static uniform discretization. More so, discretization can be adjusted any time in response to new evidence and achieve greater accuracy.

3.7 Multinomial formulation

The multinomial distribution is a generalization of the binomial distribution. The binomial distribution gives the probability of “successes” and “failures” in n independent trials of a two-outcome process. The probability of “success” and “failure” in any one trial is given by the fixed probabilities θ and $1 - \theta$ respectively. The multinomial distribution gives the probability of each combination of outcomes in n independent trials of a k -outcome process. The probability of each outcome in any one trial is given by the fixed probabilities $\theta_1, \dots, \theta_k$. The probability density function (pdf) of a multinomial distribution is given by Equation 3.7.

$$p(x|n, \theta) = \frac{n!}{x_1! \dots x_k!} \theta_1^{x_1} \dots \theta_k^{x_k} \quad (3.7)$$

where $x = (x_1, \dots, x_k)$ gives the number of each of k outcomes in n trials of a process with fixed probabilities $\theta = (\theta_1, \dots, \theta_k)$. The vector x has non-negative integer components and $\sum_i^k x_i = n$. Where $0 \leq \theta_i \leq 1$, $i = 1, \dots, k$ and $\sum_i^k \theta_i = 1$. The expected value and variance of outcome i are $n\theta_i$ and $n\theta_i(1 - \theta_i)$ respectively. We can treat parameters θ as a random parameter vector drawn from Dirichlet distributions i.e. $\theta \sim \text{Dirichlet}(\alpha)$, where $\alpha = (\alpha_1, \dots, \alpha_k)$ is a vector of hyper parameters. The pdf for a Dirichlet distribution is given by Equation 3.8.

$$p(\theta|\alpha) = \frac{\Gamma(\alpha_1 + \dots + \alpha_K)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \quad (3.8)$$

The Dirichlet distribution is a conjugate prior distribution for a multinomial distribution therefore the posterior distribution of $p(\theta|x)$ can also be computed analytically. Equation 3.9 shows the analytical expression for the posterior distribution $p(\theta|n,x)$.

$$p(\theta|n,x) \propto \theta_1^{\alpha_1+x_1-1} \dots \theta_k^{\alpha_k+x_k-1} \quad (3.9)$$

The posterior mean $E(\theta_j|x)$ and variance $V(\theta_j|x)$ are given by the Equations 3.10 and 3.11

$$E(\theta_j) = \frac{\alpha_j + x_j}{\sum_i^k (\alpha_i + x_i)} \quad (3.10)$$

$$V(\theta_j) = \frac{\alpha_j (\sum_i^k \alpha_i - \alpha_j)}{[\sum_i^k \alpha_i]^2 (\sum_i^k \alpha_i + 1)} \quad (3.11)$$

A multinomial problem with just three parameters can be solved by merely extending the Beta-Binomial model. However, with several parameters, it will be impossible to use a node to sum all the parameters because the clique size will become so large that the computation becomes infeasible. For example, seven parameters implies that the summation node would have seven dynamically discretizable parents. The Dynamic Discretization Algorithm might consume too much computer resources or even break down completely when a node has many discretizable parents. To address this problem, Neil *et al.*, [163] proposed a binary factorization approach for optimizing the calculation of conditional probability tables in Hybrid Bayesian Networks. For example, let us consider a multinomial BN with seven parameters. The node that sums these parameters will have seven parents as shown in Figure 3.3.

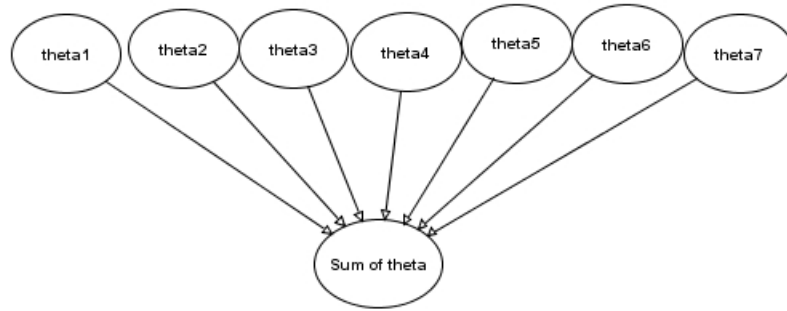


Figure 3.3: A BN fragment for summing seven parameter nodes

We can borrow the idea of binary factorization to reduce the complexity of the calculation. Rather than summing up all the parameters at once, we can recursively sum the parameters as shown in Figure 3.4.

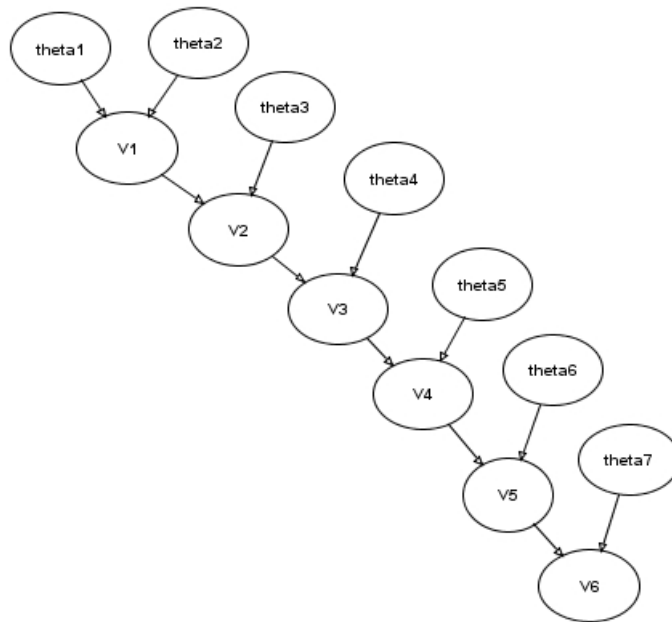


Figure 3.4: A BN fragment for summing seven parameter nodes with binary factorization approach

In Figure 3.4, the ‘summation’ node has been replaced by six nodes (V_1, \dots, V_6). These V_i s nodes recursively sum the parameters as shown below:

$$\begin{aligned}
 V_1 &= \theta_1 + \theta_2 \\
 V_2 &= \theta_1 + \theta_2 + \theta_3 \\
 V_3 &= \theta_1 + \theta_2 + \theta_3 + \theta_4 \\
 V_4 &= \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 \\
 V_5 &= \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6 \\
 V_6 &= \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6 + \theta_7
 \end{aligned}$$

Although the node V_6 has two parents, the node actually represents the sum of all the parameter nodes.

3.8 Multinomial BN models

This section describes the steps for creating a multinomial BN model. The observed data $x = (x_1, \dots, x_k)$ can be treated as random variables from a multinomial distribution i.e. $x \sim \text{multi}(\theta, n)$.

The marginal distribution of single x_i is binomial. We can simulate a multivariate draw for the i^{th} configuration in AgenaRisk by using a sequence of binomial draws as discussed in [73], page 583. This is as follows:

- Draw x_1 from a $\text{Binomial}(n, \theta_1)$ distribution.
- Then draw x_2, \dots, x_{k-1} in order as follows. For $j = 2, \dots, k-1$, draw x_j using Equation 3.12.

$$\text{Bin}\left(n - \sum_{i=1}^{j-1} x_i, \frac{\theta_j}{\sum_{i=j}^k \theta_i}\right) \quad (3.12)$$

If at any time during the simulation the binomial sample size equals zero, we can use the convention that a $\text{Bin}(0, \theta)$ variable is identically zero, [73].

- Finally, set $x_k = n - \sum_{i=1}^{k-1} x_i$

Unlike the beta-binomial formulation, constructing a multinomial BN model is not straightforward. The following steps are taken while constructing a multinomial BN model with k parameters.

- Create a set of k continuous nodes for all parameters $\theta_1, \dots, \theta_k$. Define any distribution on each parameter θ_i . This can be informative or non informative, conjugate or non conjugate. Here we stick with the Dirichlet distribution.
- Introduce $k-1$ continuous nodes (v_1, \dots, v_{k-1}) to perform pairwise summations of parameter nodes as follows:

$$\begin{aligned}
v_1 &= \theta_k + \theta_{k-1} \\
v_2 &= v_1 + \theta_{k-2} \\
&\vdots \\
&\vdots \\
v_{k-1} &= v_{k-2} + \theta_1
\end{aligned}$$

- Create an integer node to represent the total number of independent draws n and also introduce a set of $k - 1$ integer nodes (n_1, \dots, n_{k-1}) for sample size. Define arithmetic expression on the $k - 1$ nodes as follows:

$$\begin{aligned}
n_1 &= n - x_1 \\
n_2 &= n - (x_1 + x_2) \\
&\vdots \\
&\vdots \\
n_{k-1} &= n - \sum_{i=1}^{k-2} x_i
\end{aligned}$$

- Create a set of k integer nodes for observation x . Define NPT for these nodes using statistical expressions as follows:

$$\begin{aligned}
x_1 &\sim \text{Bin}(n, \theta_1) \\
x_2 &\sim \text{Bin}(n_1, \frac{\theta_2}{V_{k-2}}) \\
&\vdots \\
&\vdots \\
x_{k-1} &\sim \text{Bin}(n_{k-2}, \frac{\theta}{V_1}) \\
x_k &\sim \text{Bin}(n - \sum_{i=1}^{k-1} x_i, 1)
\end{aligned}$$

An example of a multinomial BN model with three parameters is shown in Figure 3.5.

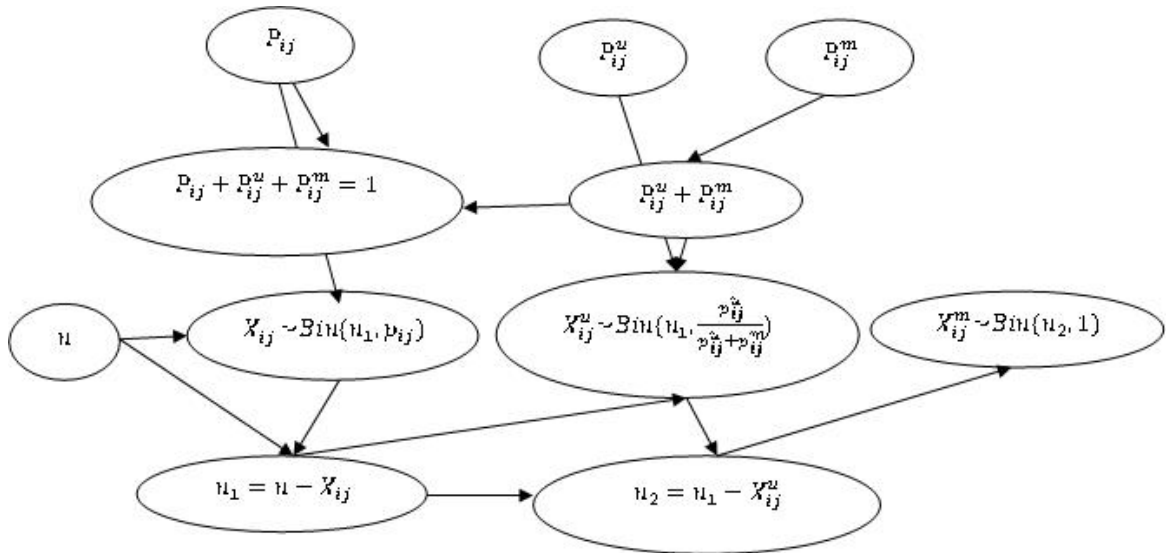


Figure 3.5: Hierarchical BN model

3.9 Hierarchical Bayesian Network (HRBNs)

Apart from being able to incorporate both discrete and continuous variables in a BN model, many practical applications of BNs require more than a simple structure of prior, likelihood and posterior distributions [73]. Hierarchical models allow us to represent complex problems by assuming appropriate parametric distributions for uncertain parameters in the model. Hierarchical Bayesian Networks (HRBNs) can be regarded as Bayesian Networks which consist of parameters and hyper parameters. The HRBNs are very useful in learning parameters from data. An example of Hierarchical Bayesian model is of the form

$$p(y|\theta) \sim \text{Bin}(n, \theta)$$

$$p(\theta) \sim \text{Beta}(\alpha, \beta)$$

Where α and β are the hyper-parameters and θ is an uncertain parameter. In general, hierarchical models are characterized by parameter θ and hyper parameters (ϕ) which is a vector consisting of $\{\alpha, \beta\}$ in the model above.

An HRBN model to represent this example is shown in Figure 3.6.

One challenge in hierarchical modeling is choosing an appropriate prior distributions for the hyper parameters. This task requires substantive knowledge about the parameters,

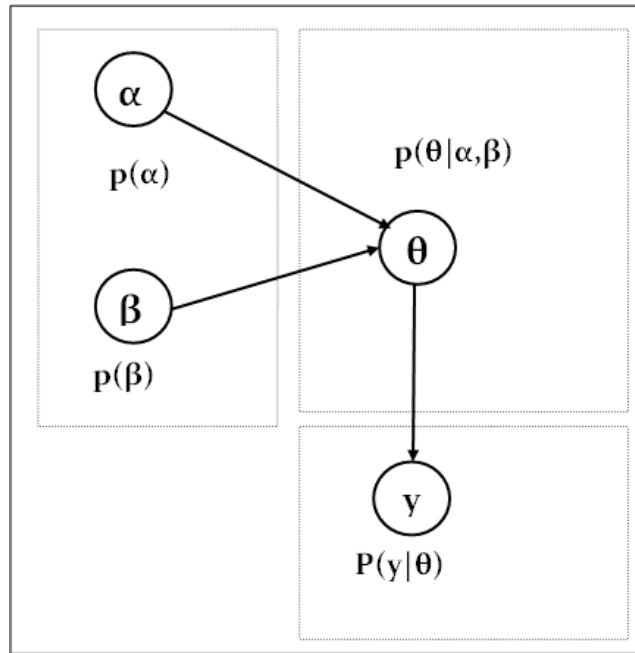


Figure 3.6: Hierarchical BN model

at least sufficient to be able to confine them into a finite region, [73]. In many cases the information required to specify these prior distributions precisely is not available. Consequently, non-informative prior (diffuse prior) distributions are often used for the hyper parameters.

3.10 Hierarchical Bayesian Network: Beta-Binomial

Here we will use a hierarchical model to analyze the Beta-binomial example described in Example 3.1. The prior distribution assumed for the parameter (θ) was a simple prior distribution i.e. $\theta \sim \text{Beta}(1, 1)$. We can assume a complex structure so that the parameter (θ) depends on hyper parameters α and β as given by Equation 3.13.

$$\theta \sim \text{Beta}(\alpha, \beta) \quad (3.13)$$

We can specify uniform distributions on the hyper parameters (Equation 3.14).

$$\alpha, \beta \sim U(0, 1) \quad (3.14)$$

A hierarchical BN model showing the posterior marginal distributions for parameter θ and hyper parameters α and β is shown in Figure 3.7. This is based on the realization

of 105 “Heads” from the 200 repeated tosses of the coin.

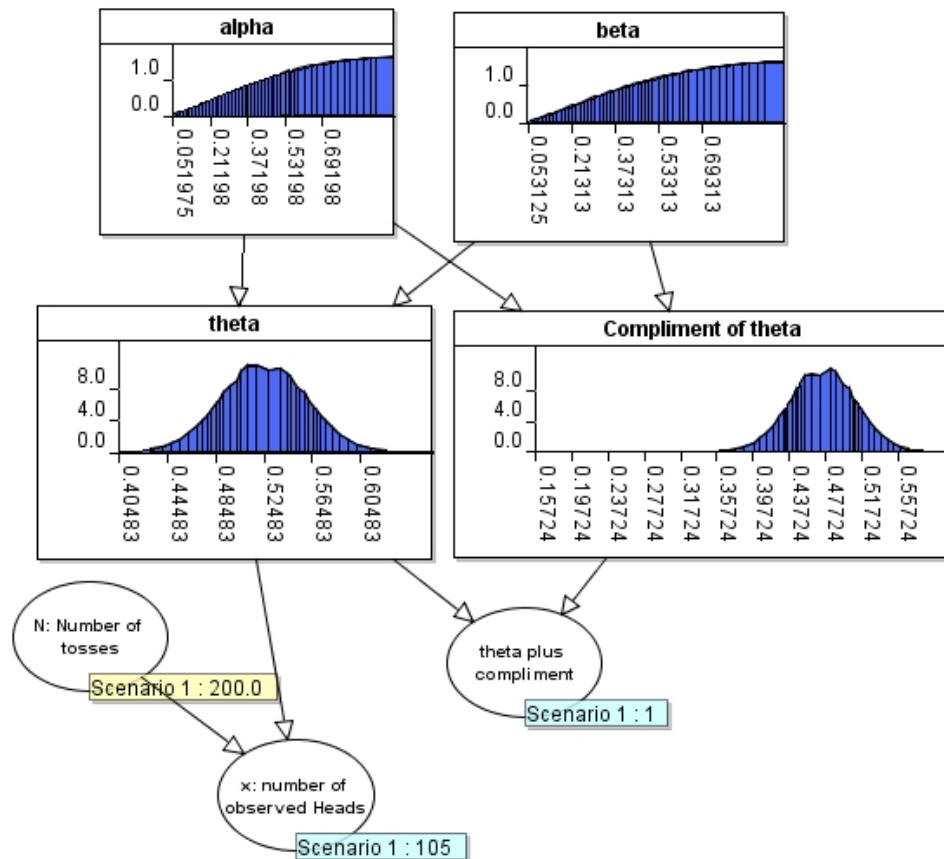


Figure 3.7: Hierarchical BN model

These result in shown in Table 3.3.

Table 3.3: Summary statistics for the parameters (θ) and Tail ($1 - \theta$) from a hierarchical BN model

Parameters	Summary Statistics			
	Mean	Median	Lower quartile	Upper quartile
θ : Head	0.52493	0.52479	0.50007	0.54635
θ' : Tail	0.47469	0.47515	0.45162	0.49977

3.11 Summary

This chapter reviewed the concept of Hybrid Bayesian Network and also described the Dynamic Discretization Algorithm. This version of the Dynamic Discretization Algorithm works seamlessly with the Junction Tree Algorithm, [125]. Also, the chapter described the beta-binomial and multinomial BN formulations. A simple Beta-binomial BN model was introduced to compare the result obtained from dynamic discretization

with those from static discretization. The chapter demonstrated the potential of the Junction Tree Algorithm and dynamic discretization in performing general inference using a Bayesian Network model. This powerful combination facilitates the use of a BN model to solve other interesting modeling problems such as learning parameters and summarizing clinical evidence in meta-analysis.

Chapter 4

Clinical applications of Multinomial BN models

This chapter presents three novel applications of Bayesian Networks augmented by dynamic discretization to solve some important modeling problems in the clinical domain. These classes of problems are frequently encountered in the clinical domain. These are:

- modeling subjective opinion from clinical experts with statistical distributions;
- learning parameters of a model using multinomial BN models; and
- scoring different hypotheses about causal relations between clinical variables.

This chapter describes the approach to solve these problems using the concept of Hybrid Bayesian Networks.

4.1 Modeling Subjective opinions with statistical distributions

One attractive feature of Bayesian analysis is the ability to incorporate subjective knowledge in the analysis. In a real life application of BN models, domain experts are normally involved in the analysis to provide subjective opinion about some interesting clinical parameters. Gelman *et. al* [73] provided two intuitive interpretations of the prior information. These are population interpretation and subjective interpretation. From the population interpretation point of view, a prior distribution represents a population of possible parameter values within which the parameter of interest has been drawn. The subjective

interpretation, on the other hand, represents our knowledge about the parameter of interest and the degree of uncertainty associated with it. Whichever interpretation we assume, the prior distribution should cover all plausible values of the parameter but needs not to be concentrated around the true value [73]. Different types of prior distributions has been proposed by different authors. Gelman *et. al* [73] decribed three categories of prior information namely: noninformative; highly informative and moderately informative [18] hierarchical prior distributions. The use of noninformative prior distribution is common when there is not information about the parameter of interest. The use of non-informative prior distributions has been addressed by a number of authors. For example see Bernardo and Smith [12], Carlin [25] and Gelman *et. al* [73] for full discussions on the theoretical principles for prior distributions. Gelman discussed the hierarchical prior distribution for the Physiological pharmacokinetic analysis in [72]. A commonly used approach for representing a informative prior in a BN model is to assume that the parameter takes a specific value in the parameter space. This is by far the most convenient approach which is compatible with many inference algorithms. However this approach fails to encode uncertainty about the subjective information given by the experts.

Using informative prior distributions allows the incorporation of information gathered from clinicians in light of their experience. However, this may lead to problems because of the subjective nature of the information provided by the experts [67]. Once an expert has provide the subjective values about a parameter, a Bayesian analyst needs to translate the information into a mathematical form that can be incorporated into the analysis. An important question therefore is how should analysts incorporate subjective information in the model? Should the analyst take this subjective value as the only possible value or as a candidate prior from a pool of potential values? In practice, two experts can give two different subjective values about the same parameter. Therefore, it is important to introduce uncertainty in the subjective value. For example Fenton *et al.*, [62] reported a case study where two clinicians disagreed on the sensitivity of the Magnetic Resonance Angiogram technique.

In this work, rather than using this *relative frequency approach*,¹, we will borrow the idea used in Spiegelhalter, *et al.* [216] to model subjective values received from the

¹This approach normally assume that the parameter of interest takes a specific value on the parameter space, i.e. The prior value of the parameter is set to the subjective estimate provided by the experts.

expert. Spiegelhalter, *et al.* [216] recommended a set of prior families that can be used to specify prior belief about a treatment effect, these are: the reference prior; Sceptical prior; and enthusiastic prior.

The reference prior is the least subjective which represents a minimal prior information. An analysis based on this prior can be used as a baseline against which to compare analysis using other priors. We can specify this prior with a uniform distribution or use a conjugate parametric family that includes the uniform distribution as a special case. This specification implies that we do not have any confidence in the value we received from the expert. This may be applicable when there is no substantive knowledge from domain experts. In this case we can define a prior that assume equal probability to every value on the parameter space.

In the context of this work, we can use the skeptical prior to model the degree to which we believe the parameter is likely to take other values on the parameter space from the value specified by the expert. For example an expert may say that the sensitivity of a test is about 90% and then we may want to specify a prior distribution on this parameter that assigns a very low probability to the proposition that the sensitivity is not equal to the value specified by the expert. To specify this prior, we will use a statistical distribution that assigns probability to all possible values in the parameter space. For a parameter with a bounded space, we can use the Triangle distribution, Truncated Normal distribution and Beta distribution to specify the Sceptical prior.

In contrast to the Sceptical prior, we may rather want to use a prior distribution that assign a very high probability to a proposition that the parameter is within an interval containing the subjective value specified by the expert. We can use the enthusiastic prior to model the degree to which we believe in the subjective value provided by the domain experts. To specify the enthusiastic prior, we will use appropriate statistical distributions and also introduce constraints on the parameter. We will use the constraints to assign a probability (ϕ) to a proposition that the parameter is within a range ($[a,b]$) on the parameter space i.e. $p(\theta \in [a,b] = \phi)$. For example if a clinician says there is 5-10% risk of complication (θ) from a clinical procedure, analysts can use the enthusiastic prior to encode a degree of belief (ϕ) in this proposition, that is $p(\theta \in [0.05, 0.1] = \phi)$.

Let us start with an example of the sensitivity of a diagnostic procedure for intracranial

aneurysm. Suppose a clinician says that the sensitivity of this procedure is 92%. Therefore we can assume a prior mean of 0.92 for a parameter, θ , representing the propensity of this procedure to detect cases with intracranial aneurysms. It is a common practice to use a summary value, such as the mean, to represent the prior probability. This representation implies that the parameter takes a specific value on the parameter space as shown in Figure 4.1.

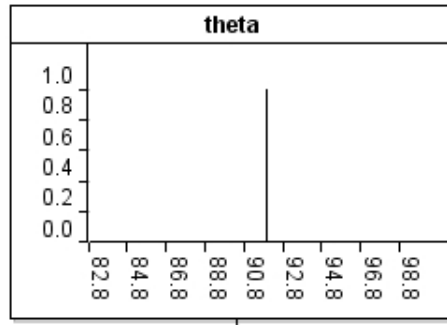


Figure 4.1: Parameter taking a fixed value on the parameter space.

Figure 4.2 shows a Boolean node representing the prior distribution for parameter θ .

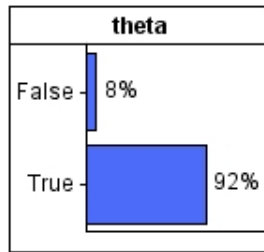


Figure 4.2: Representing prior information with a discrete node

Representing prior information in this way is very convenient and easy to achieve but it does not incorporate the uncertainty associated with θ . In order to treat θ as a random parameter, we can use statistical distributions to model our subjective belief about θ . Since sensitivity takes values between 0 and 100, the parameter space for θ is bounded to a range of values between 0 and 100 i.e. $[0, 100]$. We can use the Triangle distribution to model θ by specifying the lower, middle and upper values for θ . In this example the lower value is 0, upper is 100 and the middle value is 92. Figure 4.3 shows the prior marginal distribution for θ using the Triangle distribution. The second node labeled “ $p(\theta > 92)$ ” represents the probability that θ is higher than 92 i.e. $p(\theta > 92)$.

The prior marginal distribution of θ shows that all values, in the the parameter space,

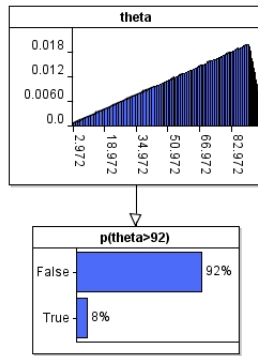


Figure 4.3: Representing prior information about the sensitivity of a test with Triangle distribution.

are likely but the region around the middle value (i.e. 92%) have higher probability. The mean and median of θ , based on the Triangle distribution (Figure 4.3), are 63.998 and 67.822 respectively. The prior marginal distribution for the discrete node “ $p(\theta > 92)$ ” shows a probability 8% for “True” and 92% for “False”. This means that there is 0.08 probability that the sensitivity of this screening procedure is higher than 92%. In addition to the Triangle distribution, we can also model this prior knowledge using other statistical distributions shown in Table 4.1.

Table 4.1: Statistical distributions to model prior knowledge about the sensitivity of a diagnostic procedure.

Prior	Mathematical formulation
Uniform	$U(0, 100)$
Truncated Normal	$N(92, 0.01, 0, 100)$
Beta	$Beta(92, 8)$

Figure 4.4 shows three representations of the medical knowledge using statistical distributions in Table 4.1.

Unlike the Triangle distribution the prior distribution for θ based on the Uniform distribution shows that all values in the parameter space are equally likely with mean and median equal 50. Also, this model gives 0.08 probability that the sensitivity of this screening procedure is higher than 92%. Unlike the duo (Triangle and Uniform), modeling the prior with the Truncated Normal and Beta distributions concentrate around the subjective value 92%. Some ranges of values in the parameter space, for example [0 - 90] and [93 - 100] have zero probability. The mean and median of θ from the Truncated Normal and Beta distributions are approximately 92. These models give 0.55 probability that the sensitivity of this screening procedure is higher than 92%.

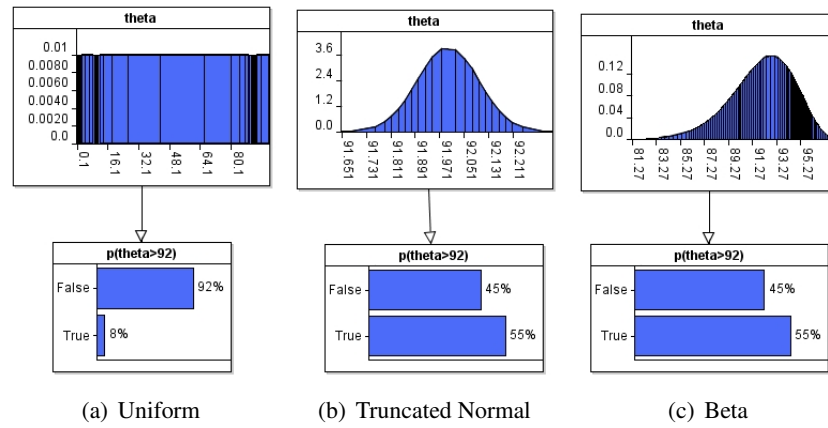


Figure 4.4: Representing prior information about the sensitivity of a test with a) Uniform b) Truncated Normal and c) Beta distributions.

So far we have used the Uniform distribution to specify a reference prior and we have also used the Beta, Truncated Normal and Triangle distribution to specify a Sceptical prior. We can say that the degree of skepticism is higher in the Triangle distribution than using a Truncated Normal distribution with a low variance. We can change the parameters of each of these distributions to reflect the degree of our skepticism about the existing medical knowledge. Let us now specify enthusiastic priors on parameter θ . To do this, let us assume that a clinician says that the sensitivity of this diagnostic test is between 92% and 98%. The intention is to model this prior information so that we can assign a probability to a proposition that $\theta \in [87, 98]$.

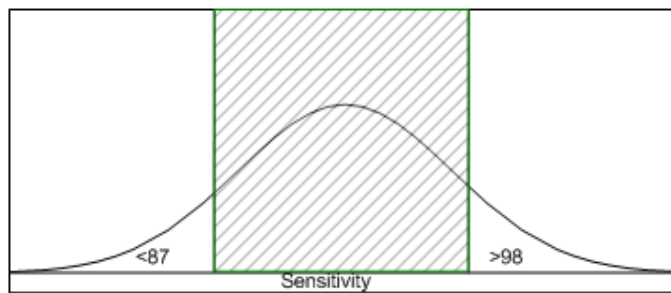


Figure 4.5: Representing enthusiastic prior information using dynamic discretization.

To achieve this with dynamic discretization, we will introduce three Boolean nodes in addition to the parameter node as shown in Figure 4.6.

For the parameter node, we can use an appropriate statistical distribution. Let us use the priors defined in Table 4.1 for θ . For each of these distributions, we will constrain θ so that we will have 90% probability that the prior distribution for θ is defined between

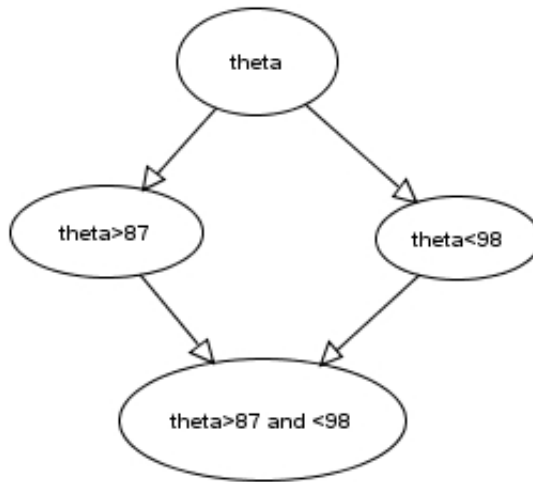


Figure 4.6: Representing enthusiastic prior information using dynamic discretization.

92 and 98. We will define comparative expressions in the NPT for nodes labeled “theta >87” and “theta <98”. For “theta >87” we define a comparative expression {if($\theta > 87$, “True”, “False”)} and {if($\theta < 98$, “True”, “False”)} for the second node “theta >98”. These expressions calculate the percentages of the areas specified relative to the whole distribution. For instance the node “theta > 87” calculates the percentage of the prior distribution of θ that is greater than 87. Lastly, we define the NPT for the third Boolean node “theta <87 and <98” as given in Table 4.2.

Table 4.2: NPT to constrain a parameter space to a range between 92 and 98.

“theta > 87”	“theta < 98”	“theta > 87 and < 98”	
		True	False
True	True	1	0
True	False	0	1
False	True	0	1
False	False	0	1

Figure 4.7 shows the prior marginal distributions for θ and nodes used to enforce the constraint. This shows that about 17.96% of the distribution of θ is greater than 87 and 99.5% of the distribution of θ is below 98. The percentage in the region 87 to 98 is 17.46%.

Using the Uniform distribution shows that 11% of distribution of θ falls in the interval [87, 98] with 13% higher than 87 and 98% below 98 (Figure 4.8).

For the Truncated normal distribution, the entire distribution of θ falls in the interval [87, 98] as shown in Figure 4.9.

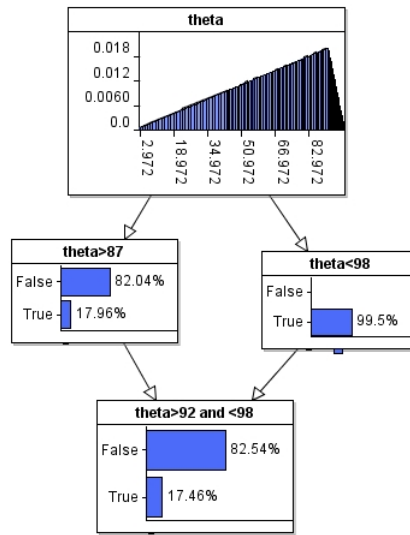


Figure 4.7: Enthusiastic prior specified with the Triangle distribution.

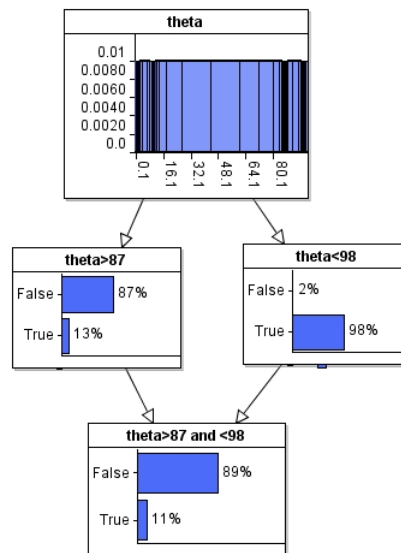


Figure 4.8: Enthusiastic prior specified with the Uniform distribution.

Finally Figure 4.10 shows the prior marginal distribution obtained by using the Beta distribution. In this model, about 95% of the distribution of θ falls in the interval $[87, 98]$.

Even without entering evidence on the node “theta > 87 and < 98”, more than 95% of the distribution of θ is already within the interval $[87, 98]$ for the models specified with the Truncated Normal and Beta distributions. Hence we can use the Truncated Normal distribution with a low variance to model our ‘enthusiasm’ about a hypothesis i.e. $H : E(\theta) \in [a, b]$.

The discussion here has focused on the mechanics of using a number of priors to reflect different expert opinions about the parameters of interest and labels, such as sceptical

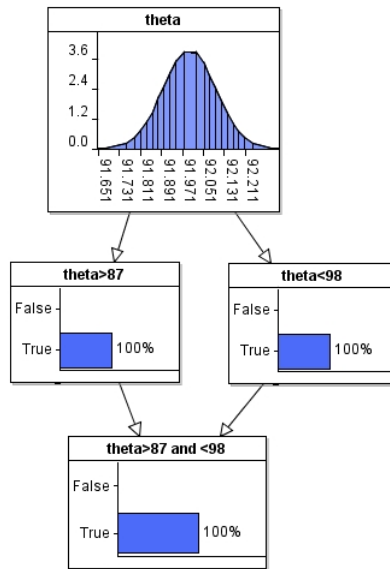


Figure 4.9: Enthusiastic prior specified with the Truncated Normal distribution.

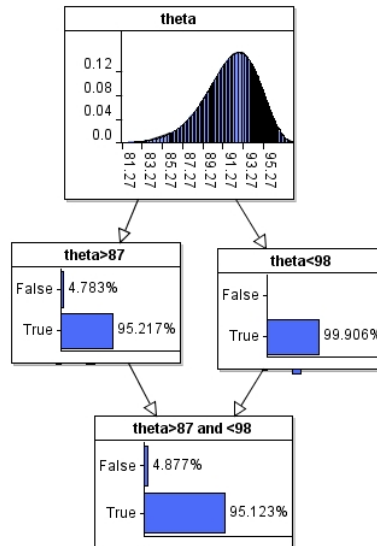


Figure 4.10: Enthusiastic prior specified with the Beta distribution.

or optimistic, have been used to characterise these opinions as priors, mostly for convenience. However we must realise that this is simply the most visible manifestation of Bayesian scientific methodology insofar as the opinions/positions being formulated and evaluated using the data can be considered hypotheses to be tested. Therefore, it would be a mistake to take Spiegelhalter's, or anyone else's, labelled priors as simple opinions when, in fact, they are hypotheses in the sense that they express predictions about real phenomena that we can test with real evidence (as an aside this Bayesian approach to the scientific method is similar to Popper's notion of falsification and provides a rich vein of

research in philosophy of science). Of course in practice we face a multitude of possible hypotheses that could describe the situation at hand but to formulate and test all of these might be infeasible given resource constraints. Given this we should realise that the choice of labels for the priors, and indeed the number of prior hypotheses, used here is used so for illustration only. It is not meant to be exhaustive but rather to reveal an underlying continuous spectrum and select points on the spectrum where expert opinion might gravitate and used these points, in a dispassionate and objective way, as our hypotheses. Of course there is danger here: we are not advocating the use of a standard 3-point (reference, sceptical or enthusiastic) or 5-point scheme for priors. Neither are we saying that the prior formulation, as expressed on the parameters, must follow some dogma: users of Bayesian methodology should instead give careful thought to the number of priors and be careful to use as many as appears right to capture differences in opinion but be realistic about the cost benefit trade off incurred by doing so. Spiegelhalter's prior labelling scheme therefore provides a useful touchstone when applying hypothesis testing but provides nothing more.

4.2 Model Specification

This section demonstrates, with examples, the influence of prior distribution on the posterior inference. BN models can differ in terms of the prior, sampling distributions, causal relations or even in terms of variables included in the DAG, [73]. Because the posterior distribution of a parameter may change with respect to the prior distribution, it is a good practice to consider more than one candidate model in the analysis in order to analyze the impact of the priors on the posterior inference. This helps to check (for example) if making a convenient distributional assumption has an impact on the posterior inference of parameters.

Example 4.1. In a hypothetical experiment involving 980 (n) rape victims, 437 (y) were tested positive for HIV six months after the incident. The sensitivity and specificity of the diagnostic test used are 100%, i.e. the test will accurately identify cases of HIV infection or non infection with certainty. Suppose 50% of convicted rapists are HIV positive, we can then assume a prior probability for the parameter θ representing the prevalence of HIV in the population of convicted rapist. We can approach this problem with the following

beta-binomial model:

$$\begin{aligned} y &\sim \text{Bin}(n, \theta) \\ \theta &\sim \text{Beta}(\alpha, \beta) \\ \alpha, \beta &\sim U(1, 1000) \end{aligned}$$

In order to complete a model's formulation, we need to translate the prior information into a mathematical equivalence that we can easily incorporate in the analysis. In a beta-binomial formulation, the expected value of the parameter θ is given by the mean of the beta distribution $E(\theta) = \frac{\alpha}{\alpha + \beta}$. The historical sample size $(\alpha + \beta)$ represents the value of the prior information. The higher the value, the more we believe in the historical data. We can vary the values of $\alpha + \beta$ to arrive at different models for this problem. Table 4.1 shows eight candidate models (M_i) differing in the degree of belief in the historical information.

Table 4.3: Different Beta-binomial models for the hypothetical rape victim data

Models (M_i)	$E(\theta)$	$\alpha + \beta$
1	0.5	2
2	0.5	50
3	0.5	100
4	0.5	200
5	0.5	500
6	0.5	1000
7	0.5	10000
8	0.5	100000

The posterior means and variances for θ , from different models, are presented in Figure 4.11

Figures 4.11(a) & 4.11(b) depict the posterior mean and variance for parameter θ from different models. Two crucial observations are obvious from these Figures. 1) The posterior mean lies between sampling proportion and the prior. 2) The variance of the posterior distribution decreases with increasing values of $\alpha + \beta$. Models with a small $\alpha + \beta$ value have the posterior distribution concentrated around the sampling proportion y/n thereby suggesting that the analysis was driven mainly by the likelihood. In particular, model (M_1) with $\alpha + \beta = 2$ is a special case equivalent to specifying a uniform prior distribution for θ . However, the posterior means move towards the prior (with increasing

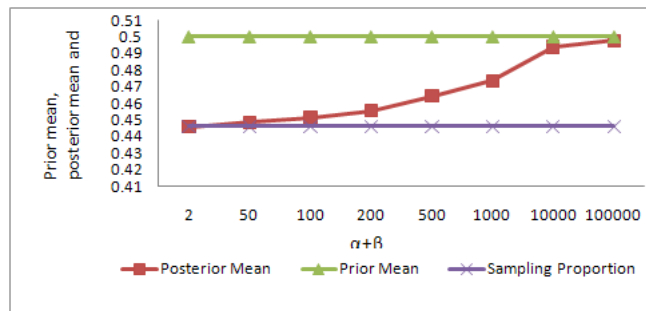
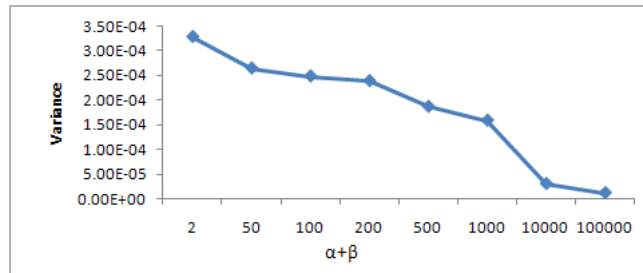
(a) Posterior mean of θ (b) Posterior variance of θ

Figure 4.11: Prior mean = 0.5, sample proportion $y/n = 0.446$ and posterior statistics for different models ($\alpha + \beta$ values)

precision) as $\alpha + \beta$ increases. In essence, we can encode a subjective judgement and also use the value of $\alpha + \beta$ to express the degree of belief in this view.

4.2.1 Binomial Likelihood

In Section 4.2 we used the conjugate prior for a binomial likelihood. Let us now analyze Example 4.1 (Binomial likelihood) with three different, non conjugate prior distributions as shown in Table 4.4.

Table 4.4: Three mathematical representations of prior information

Prior	Mathematical formulation
Enthusiastic	$N(0.92, 0.01, 0, 1)$
Sceptical	$Triangle(0, 1, 0.92)$
Reference	$U(0, 1)$

In order to incorporate these three prior distributions in a single BN model, we introduce an additional node labeled “Priors” as a parent node to the parameter node (“theta”). This node is a categorical node with three states {Reference, Sceptical, Enthusiastic}. The probability table for this node is vacuous i.e. the three prior distributions are equally likely.

$$\begin{aligned} p(\textit{Reference}) &= 1/3 \\ p(\textit{Sceptical}) &= 1/3 \\ p(\textit{Enthusiastic}) &= 1/3 \end{aligned}$$

The conditional table for the parameter node (“theta”) can then be specified using the following expressions.

$$\theta \sim \begin{cases} N(0.92, 0.01, 0, 1) & \text{if Enthusiastic} \\ \textit{Triangle}(0, 1, 0.92) & \text{if Sceptical} \\ U(0, 1) & \text{if Reference} \end{cases}$$

Figure 4.12 shows the marginal prior distributions of “theta” and “Priors” before observing evidence.

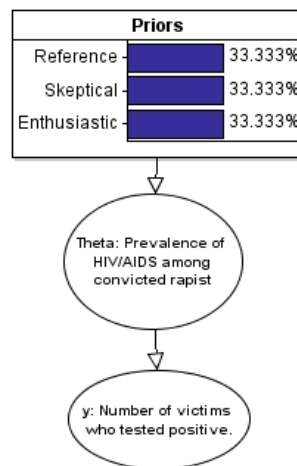


Figure 4.12: Prior marginal distributions of the Beta-Binomial Example after incorporating subjective information

It is obvious from Figure 4.12 that before observing any data, the three prior distributions are equally likely. At this point we do not know if any of these three prior distributions is better than others, given the observed data. After entering evidence the posterior distribution of the node “Prior” shows that the Sceptical prior is more consistent with observed data than the reference and enthusiastic priors. The enthusiastic prior is the least consistent with observed data (Figure 4.13).

In fact we can visually inspect posterior distributions of the parameter node (θ) resulting from these three prior distributions (Figure 4.14).

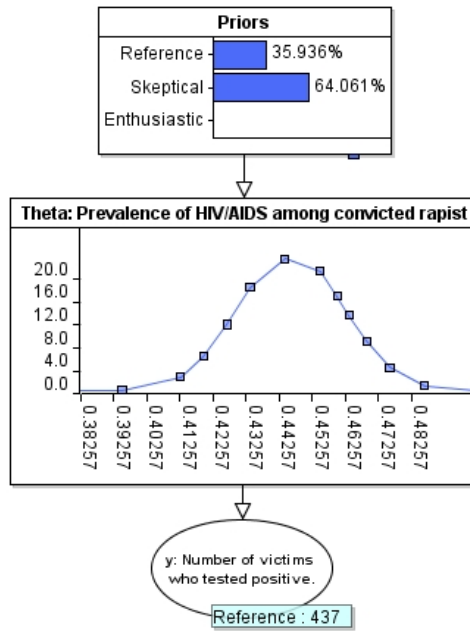


Figure 4.13: Posterior marginal distributions of the Beta-Binomial Example after incorporating subjective information

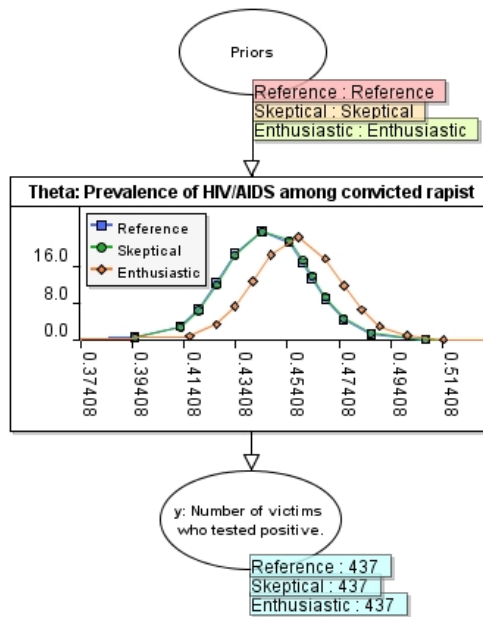


Figure 4.14: Posterior marginal distributions of parameter θ using the reference, Sceptical and enthusiastic priors.

Clearly as shown in Figure 4.14, the posterior inference about θ is different for the enthusiastic prior. This is because we have expressed ‘enthusiasm’ about a hypothesis, $H : E(\theta) = 0.92$, which is inconsistent with the observed data.

4.2.2 Normal Likelihood

Let us now consider a normal likelihood for a variable x representing the age of patients with a particular condition:

$$x_i \sim \text{Normal}(\mu, \tau^2) \quad (4.1)$$

The parameter of interest is the μ representing the average age of this population and we intend to use the BN approach to estimate this parameter from data. Just like before, we will specify three prior distributions on μ to encode the existing knowledge about the parameter. Suppose a clinician believes that the average age of patients is 80 years and that the age of the youngest and oldest patients ever admitted are 30 and 120 years respectively. Therefore we can translate this information into a mathematical representation as follows:

$$\mu \sim \begin{cases} N(80, 100, 30, 120) & \text{if enthusiastic} \\ \text{Triangle}(30, 80, 120) & \text{if Sceptical} \\ U(30, 120) & \text{if reference} \end{cases}$$

For the second parameter (τ) in Equation 4.1, we can use a non informative prior distribution i.e. $\tau \sim U(0, 1000)$. Table 4.5 presents the ages (x_i) of twelve patients randomly selected from this population.

Table 4.5: Table showing ages of twelve patients from a hypothetical population

index (i)	Age (x_i)
1	89
2	31
3	49
4	83
5	77
6	77
7	80
8	90
9	40
10	89
11	70
12	81

Figure 4.15 shows a BN representation of this problem. The observations in Table 4.5 are represented by nodes labeled x_1, \dots, x_{12} in addition to parameter nodes “mu” and “Tau”. We can assume a vacuous probability distribution for the node “Prior” and define

NPT for the observation nodes as follows:

$$x_i \sim TNormal(\mu, \tau, 0, 120)$$

The prior marginal distribution (without observation) shows that the three priors are equally likely as shown in Figure 4.15.

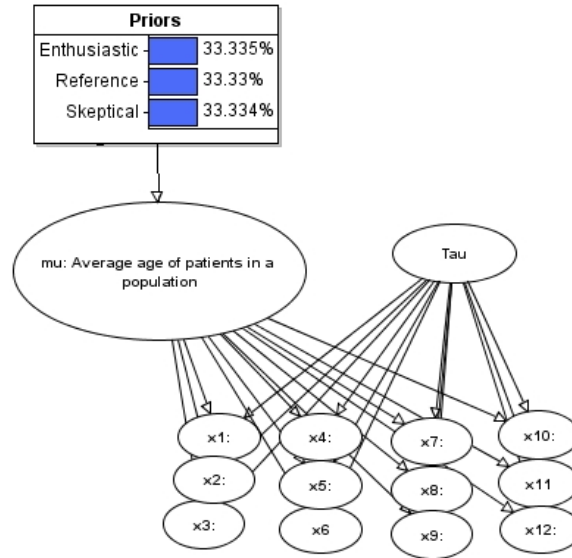


Figure 4.15: BN representation of the Normal likelihood example

However, as shown in Figure 4.16, the three priors specified are now having different distributions. The enthusiastic prior is the most consistent with the observed data while the Sceptical prior is the least consistent. The posterior probability distribution for these three priors are 0.544 for the enthusiastic, 0.176 for reference and 0.280 for sceptical priors.

Figure 4.17 shows a plot of the posterior distributions of μ based on the reference, Sceptical and enthusiastic priors.

We have used conjugate and non conjugate prior distributions in a BN model and presented the posterior distribution based on the dynamic discretization. We have shown how analyst can incorporate a degree of uncertainty about the prior information. It must be noted however that the choice of prior distributions may affect the posterior distributions of parameters of interest. In the worse case scenario, a poorly formulated prior distribution may lead to an *improper posterior distribution*,² thus raising a question about

²An improper prior distribution may lead to improper posterior distribution in which case, the posterior distribution is infinite for some observable values.

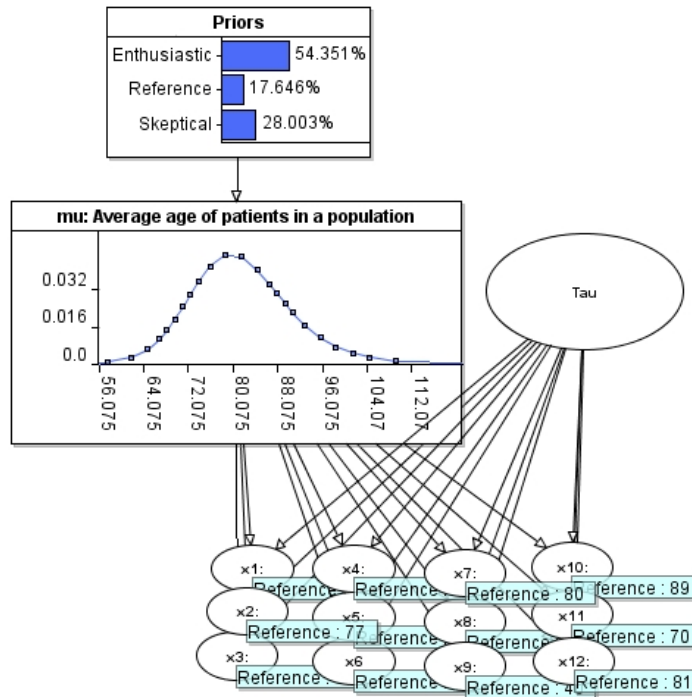


Figure 4.16: BN representation of the Normal likelihood example

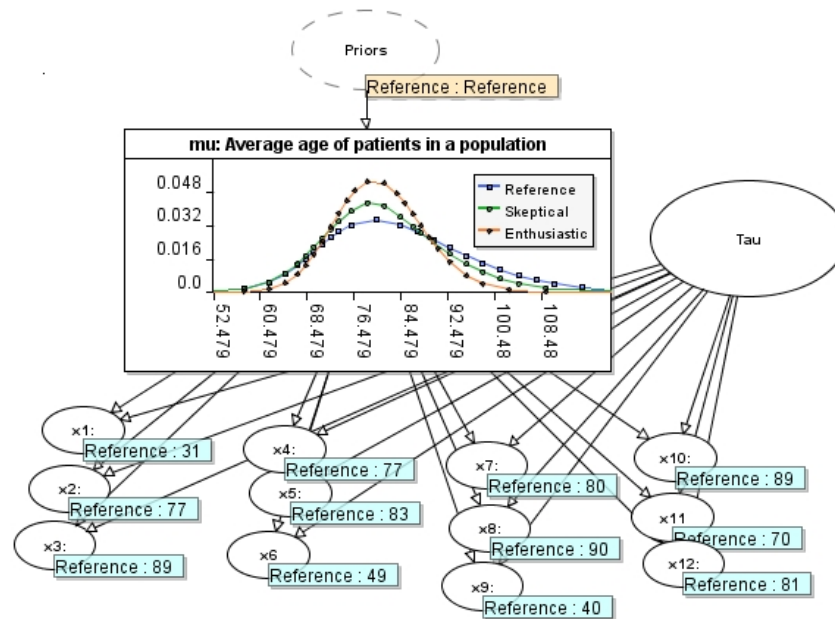


Figure 4.17: BN representation of the Normal likelihood example

the reliability of the result obtained from bayesian analysis. A commonly used approach to access the stability of the posterior distributions to the choice of prior distributions is the sensitivity analysis. Many author have partly addressed the problems of using non informative priors in bayesian analysis. Jeffreys [104] and Hartigan [89] discuss invariance principle for non informative prior distributions. Box and Tiao [19] presented a

straight-forward discussion and practical applications of non informative prior distributions. Jaynes [103] discusses the idea of objectively constructing prior distributions based on invariance principle and maximum entropy. Indeed, this is still an active research area in the bayesian community.

At this point it is worth reminding ourselves of a key criticism of standard Bayesian statistical methodology: this is that the expert is only allowed to express their expertise in the form of conjugate priors (and these mainly for mathematical convenience). The preceding discussion and examples should hopefully have illustrated that the approach taken using DD is much more flexible, allowing expertise to be specified on sub-ranges of parameters, using non-conjugate priors and using mixtures of prior distributions - one for each expert/hypothesis. This provides great flexibility but there is some cost involved in this. Firstly, the meaning and effect of a prior on the prediction can be difficult to foresee, so some experimentation is often needed. If this was challenging in a constrained setup, as is the case with conjugate priors, the challenges are doubled with non-conjugate priors and more responsibility is placed on the expert to formulate these (rightly so, but this may be unwelcome all the same). However, as we show the flexibility in terms of the hypotheses representation in choice is wider than before and thus enhances our ability to properly represent a decision maker's prior choices and the effect of these on posterior results and predictions.

4.3 Parameter Learning using dynamic discretization

Parameter learning is a crucial step in modeling real life problems. In both classical and bayesian analysis, parameter values are used either to predict or to explain some real life phenomenons. In modeling clinical problems, analysts can use the available data to estimate the value of some clinical parameters. In the previous section we have described how we can model a subjective value from analyst, now we can use the same framework to estimate parameters based on empirical values from the literatures or from various experts.

In this section, we will use the dynamic discretization approach to learn parameters of BN nodes. Indeed, the analysis in sections 4.2.1 and 4.2.2 are parameter learning models. Section 4.2.1 presents a parameter learning model with the binomial likelihood. In this

case, we specify three prior distributions and use dynamic discretization to calculate the posterior distributions. Section 4.2.2 is a normal likelihood example to learn average age of a population. Let us now consider an example of learning parameter of a categorical variable using the multinomial BN formulation.

Example 4.2. Let us assume that three treatment options (surgery, chemotherapy and radiotherapy) are available for curing a clinical condition and from a retrospective study of 287 exchangeable patients³ 50 were treated with surgery, 200 with chemotherapy and the remaining 37 with radiotherapy. We can assume three parameters θ_1 , θ_2 and θ_3 as the chance process for surgery, chemotherapy and radiotherapy respectively. In practice, these three options may not be mutually exclusive i.e. a patient may receive more than one treatments. For the purpose of this illustration, let us assume that these three treatment options are mutually exclusive such that $\sum_{i=1}^3 \theta_i = 1$. For these parameters, we can specify a non informative prior distribution is given by Equation 4.2.

$$p(\theta_i) \sim \text{Dirchlet}(1, 1, 1) \quad (4.2)$$

The observation variables are x_1 , x_2 and x_3 respectively for the number of observed patients treated with surgery, chemotherapy and radiotherapy. Equation 4.3 shows the multinomial likelihood function for the observed number of patients in each treatment category.

$$p(x_i|\theta_i) \sim \text{Multinomial}(\theta_i) \quad (4.3)$$

Figure 4.18 is a multinomial BN model showing the posterior marginal distributions for the parameters θ_1 , θ_2 and θ_3 .

The analytical posterior means for these three parameters along with the approximate results from the Dynamic Discretization Algorithm with 25 iterations are shown in Table 4.6.

This relatively simple learning approach provides a desirable benefit to learning clinical parameters subject to some clinical guidelines. Since the learning procedure is based

³That is we do not have any information a priori to distinguish one patient from another based on their responses to these treatment options.

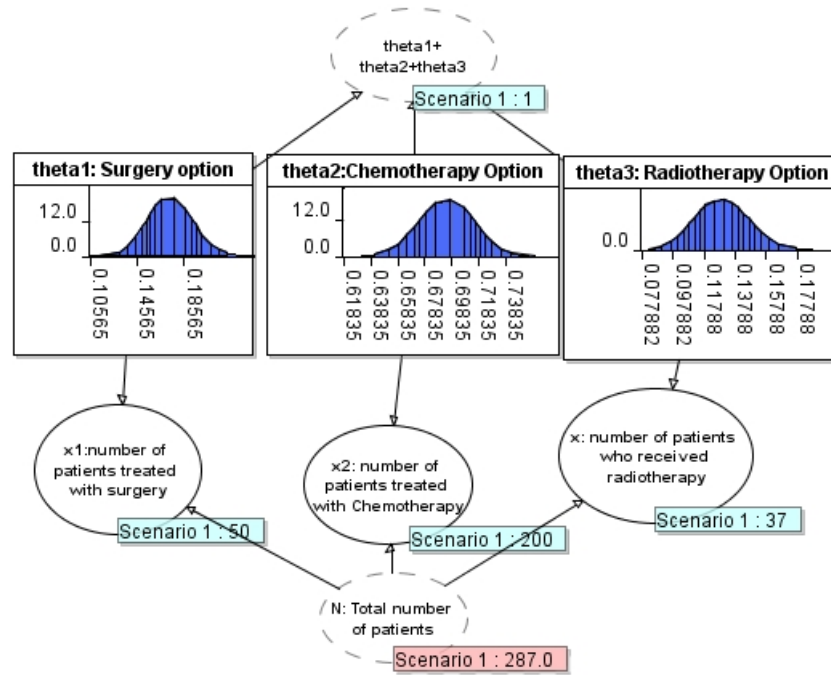


Figure 4.18: A Multinomial BN model

Table 4.6: Summary means for the parameters θ_1 , θ_2 and θ_3

Parameters	Analytical	Dynamic discretisation
θ_1 : Surgery	0.1762	0.17501
θ_2 : Chemotherapy	0.6931	0.69485
θ_3 : Radiotherapy	0.1307	0.13028

on the Bayesian Network formulation, it is easy to introduce constraints on some parameters during the learning process. The idea is to use these constraints to implement clinical guidelines during the learning process. In other words, we can learn the parameters subject to the specified constraints. We will return to a practical application of this technique in chapter 6 where medical policies are incorporated in the learning process.

Antal et al [3, 4] and Lucas et al [184] are good examples where *soft* data is used, either from literature or from experts, to inform the structure of a Bayesian model, where deterministic constraints are not identified and so have been neglected. For instance in [184] the decisions taken by clinicians are clearly not entirely probabilistic but are treated as such. In these studies [3, 4, 184] Dirichlet style learning for multinomial distributions was done without any consideration of the existence of deterministic constraints between subsets of states. Similarly the Deal software package [20] and others assume that all relationships are probabilistic. Indeed there is no mention in any of the Bayesian literature surveyed of the possibility of such constraints existing in theory or practice and as a result

standard statistical methodology runs the risk of being applied blindly. By using the techniques described in Chapter 2 we can overcome this impediment whilst also learning parameters in the multinomial case.

4.4 Scoring causal relations between variables using dynamic discretization

Similar to the parameter learning, another interesting and widely researched area in Bayesian analysis is the causal discovery. In Chapter 2, we have presented a literature review of state of the art techniques for learning causal models. A well known problem about these techniques is the search space for candidate models which is more than exponential in the number of variables. As such, a number of heuristics search algorithms have been developed to handle problems with large scale BN models. Our approach here is to use the BN framework to score different causal hypotheses and choose the one with the highest score. We consider competing models as some forms of hypotheses about a problems and each hypothesis corresponds to a causal explanation of the problem and our approach is to use the available data to score each of these hypotheses. A major distinction between our approach and those described in chapter 2 is that we impose a restriction on the search space before the analysis. In other words, we assume that the search space for competing causal explanations is known a priori. This may not be entirely correct, but we consider this as a crucial step in building realistic clinical models based on clinical knowledge. The problem of gathering the knowledge to impose this restriction on the search space is still an open research question.

We can consider competing causal explanations between variables as a form of hypothesis testing and then assign probabilities or scores to hypotheses H_i , based on the available data, x , using Bayes Theorem:

$$p(H_i|x) \propto p(x|H_i)p(H_i)$$

We can refer to each of the causal explanations as a candidate model and proceed by fitting the parameters (θ_i) of a model using the available data by Bayes Theorem:

$$p(\theta_i|x) = \frac{(p(x|\theta_i, H_i)p(\theta_i|H_i))}{p(x|H_i)}$$

Our idea of causal learning is to determine the hypothesis that best fits the data. In essence, we need to estimate how well a hypothesis explains the data, in competition with other hypotheses. To do this we need to compute the probability of the observed data given the hypothesis and parameters to get a Bayesian score:

$$p(x|H_i, \theta) = \sum_{\theta_i} p(x|\theta_i, H_i)p(\theta_i|H_i)$$

Suppose we want to learn a causal relation between two variables “Year (Y)” and “Sex (S)” using the data presented in Table 4.7.

Table 4.7: Frequency distributions of patients treated between 2005 and 2009

Year	Sex	
	Female (x_i)	Total (N_i)
2005	177	374
2006	88	183
2007	209	399
2008	222	470
2009	296	619
Total	992	2045

There are three possible relations between these two variables. Firstly, they can be independent as shown in Figure 4.19(a). We can treat this as the first hypothesis (H_1 : Sex is independent on Year). Secondly, Figure 4.19(b) shows a dependency of variable “Sex” on the other variable (“Year”). This hypothetical link does not have a causal interpretation as one cannot argue that changes in the variable “Sex” is ‘caused’ by changes ‘Year’. However, we may have reasons to believe that there are changes in the distribution of “Sex” over the years. In this case, the “Year” can be thought of as a surrogate for other (unknown) variables that might have truly caused the changes in the distribution of “Sex”. To this end, we can add this link as the second hypothesis i.e. H_2 : Sex is dependent on the Year. Finally in Figure 4.19(c) we have a hypothesis that reverses the link in the second hypothesis (H_2). This link does not make sense as variable “Sex” cannot influence “Year”.

As a result, the search space for the causal explanations consists of hypotheses H_1 and H_2 . In other words, we have two hypothetical BN models and we can now take them in turns and compute a score for each model. The model with higher score gives better explanation of the data.

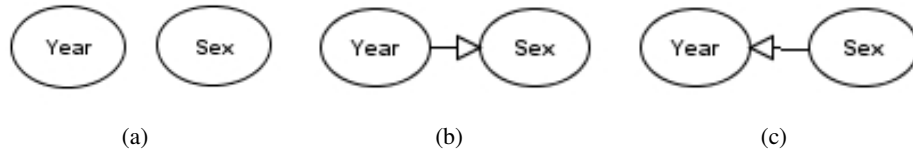


Figure 4.19: Three competing causal relations between variables “Sex” and “Year”: a) Sex and Year are independent; b) Sex is dependent on Year and c) Year is dependent on Sex

H_1 : Sex is independent on Year

To proceed, let us denote the proportion of females by θ and treat the number of female (x) in Table 4.7 as binomial distribution,

$$x_i \sim Bin(\theta, N_i)$$

Using a uniform prior on the parameter θ , we can then apply Bayes Theorem as stated earlier to obtain a posterior distribution $\theta|x$. Considering the first hypothesis (corresponding to BN structure in Figure 4.19(a)), we can create a BN model based on the assumption that the proportion of females is constant over the period i.e. the parameter θ is constant. The BN model showing the posterior marginal distribution θ (based on H_1) is presented in Figure 4.20.

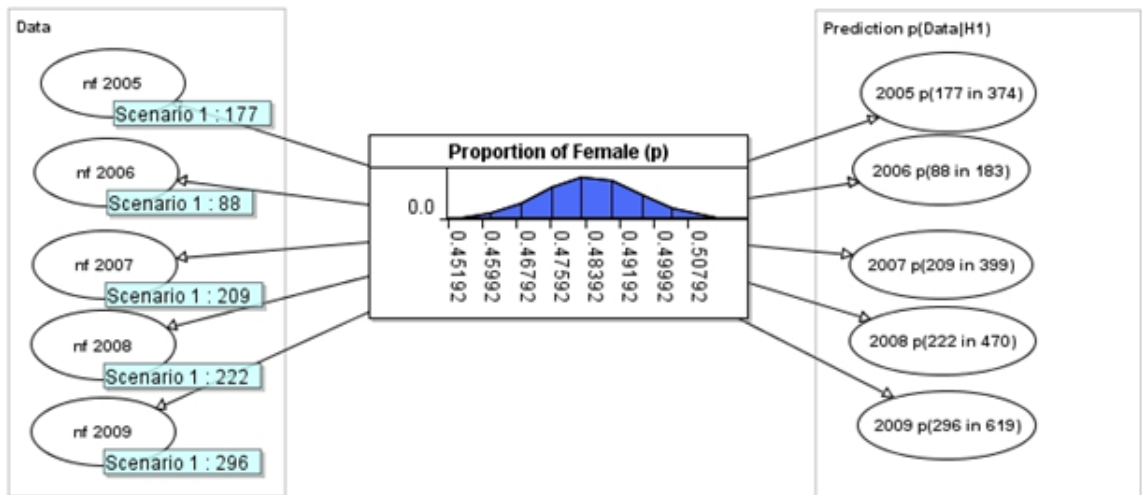


Figure 4.20: Beta Binomial model to calculate the probability of the observed data (x_i) given H_1

On the left of Figure 4.20, we have five “data nodes”, one each for the observed number of females in each year. For example, the node labeled “nf 2005” is the observed

number of females in 2005, “nf 2006” is the observation in 2006 and so on. The node at the center is the parameter node (θ). The nodes on the right are used to compute the probability of the observed data using the posterior marginal distribution of parameter (θ). We refer to these nodes as ‘prediction’ nodes. To compute the probability of the data, we ensure that the partitions of each of the ‘prediction’ nodes contains a point representing the observed data. The posterior marginal distribution of this point is the probability of the observed data.

For example, we are interested in the probability of observing 177 out of 374 females in 2005. To compute this probability, we use three static partitions (0-176, 177, 178-2049) for the node, “2005 p(177 in374)”. The probability of observing 177 females out of 374 is 0.034641. Hence, the probabilities of the observed data $p_i(x_i|H_1, \theta)$ ($i = 1, \dots, 5$) are given below;

$$\begin{aligned} p_1(177 \text{ in } 374) &= 0.035 \\ p_2(88 \text{ in } 183) &= 0.056 \\ p_3(209 \text{ in } 399) &= 0.014 \\ p_4(222 \text{ in } 470) &= 0.029 \\ p_5(296 \text{ in } 619) &= 0.029 \end{aligned}$$

Therefore,

$$S_1 = \prod_{i=1}^5 p_i(x_i|H_1) = 2.05109E - 08$$

Where S_1 is the score of the BN model based on hypothesis H_1 . We can now repeat this process for the second hypothesis (H_2).

H_2 : Sex is dependent on Year

For this hypothesis, there is a link connecting the node labeled “Year” to the second node “Sex”. Therefore the parameter θ is no longer constant for each year i.e. each year has different proportions (θ_i) of females. Figure 4.21 shows the posterior distributions of parameters (θ_i) for each year.

We can estimate the probabilities of the observed data using the posterior distributions ($\theta_i|x_i, H_2$). The probabilities $p_i(x_i|H_2, \theta_i)$ ($i = 1, \dots, 5$) are given below;

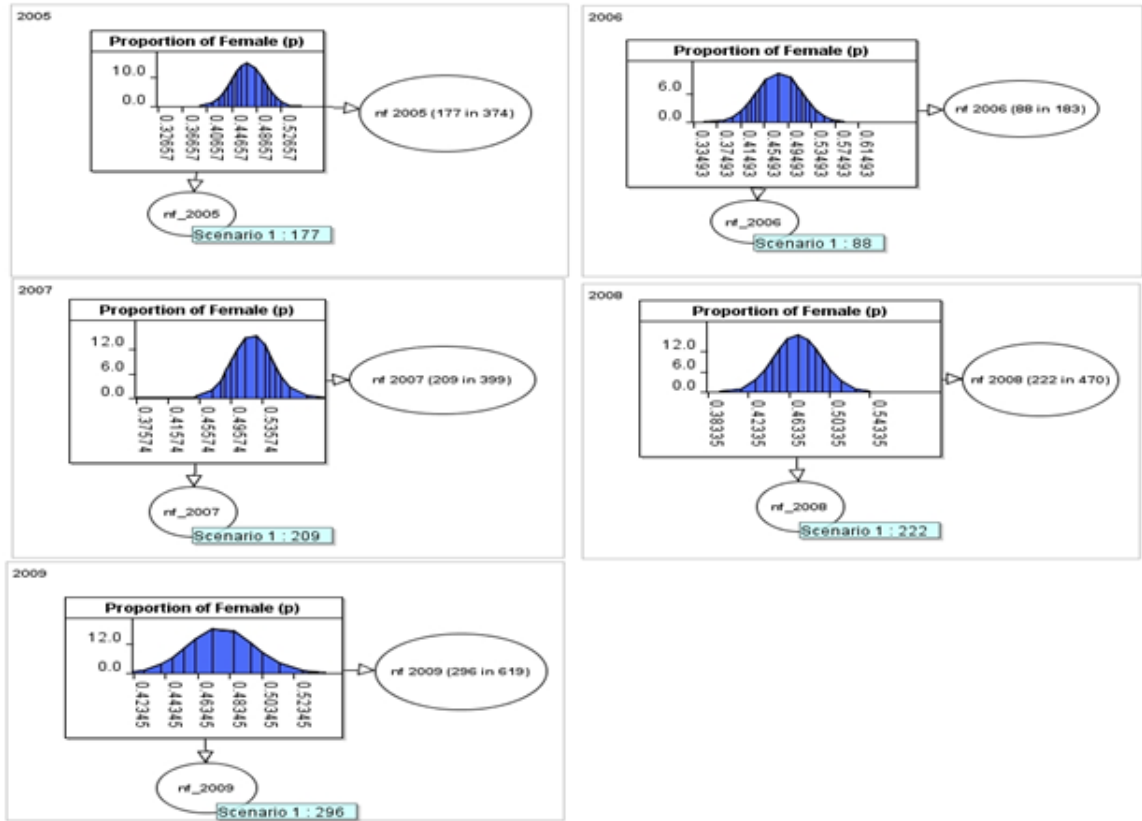


Figure 4.21: Beta Binomial model to calculate the probability of the observed data (x_i) given H_2

$$\begin{aligned}
 p_1(177 \text{ in } 374) &= 0.02891 \\
 p_2(88 \text{ in } 183) &= 0.04148 \\
 p_3(209 \text{ in } 399) &= 0.028048 \\
 p_4(222 \text{ in } 470) &= 0.02567 \\
 p_5(296 \text{ in } 619) &= 0.022404
 \end{aligned}$$

Therefore,

$$S_2 = \prod_{i=1}^5 p_i(x_i|H_2) = 1.93437E - 08$$

In this example, the score for the hypothesis H_1 is slightly higher than the score for H_2 , therefore, the observed data supports hypothesis H_1 better than H_2 . Hence, variables “Sex” and “Year” are independent.

4.5 Summary

In practical applications of BN models, analysts often resort to using experts' opinion in order to generate knowledge about a clinical variable. This chapter demonstrates the flexibility of dynamic discretization in modeling subjective information using different statistical distributions. Three ways of specifying prior knowledge were considered namely: the reference, Sceptical and enthusiastic priors. The chapter described the use of the uniform distribution to model reference priors, Triangle distributions for skeptical priors and the Truncated normal distribution to specify enthusiastic priors. We showed how the Dynamic discretization algorithm makes it possible to specify conjugate and non conjugate prior distributions.

The chapter also demonstrated a novel use of the multinomial BN formulation in learning parameters of a clinical model. In this case, rather than modeling a subjective prior from an experts, we can accumulate data from the literature or from many experts about a clinical parameter. The bayesian approach to learning parameter produced a full posterior distribution of parameters within the BN framework. In theory, these full posterior distributions obtained for the parameters of interest can be used within the bayesian network but the approach to actualize this is still an active research area.

We have also described a simple approach for learning causal relation based on data and existing knowledge of the domain. We use the Bayesian network framework to assigns scores to candidate models based on the knowledge of the domain and recommend the model with the highest scores. We argue that building a very large model requires substantive knowledge of the domain in order to considerably reduce the search space. Here we have fixed the search space for possible causal models to those with possible clinical explanations. In essence, we only consider candidate models that can be justified using knowledge of the domain. However gathering such information to reduce the search space is, on its own, an interesting research topic that require further investigation.

The analysis so far presented have shown the potential of the use of BN model subjective prior information, parameter learning and causal learning data and existing knowledge of the domain. We apply these technique to clinical case studies in chapter 6.

Chapter 5

Meta Analysis of clinical data

This chapter describes a novel application of Bayesian networks in solving meta-analysis problems. Three meta-analysis problems from the literature are analyzed using the Bayesian Network framework. These are:

- The beta-blockers clinical trial;
- The magnesium trials in myocardial infarction; and
- The prevalence of intracranial aneurysms.

Ideally when comparing new methods against established techniques one should conduct the comparative analysis in as systematic and objective a way as possible. There are thousands of meta analyses published every year and unfortunately, due to time and resource constraints it has not been possible to provide an exhaustive review of meta analysis in the clinical domain as one would have liked. Instead a snap shot is provided along with caveats and limitations on what might be learned from this alone. This restriction is also, partly, influenced by availability of data and the clarity of reported analysis needed to enable replication. Fortunately, for the studies reported data was available and a complete description of the modelling approach was available to enable the analysis to be done using DD and HBNs.

5.1 Introduction

In the medical domain it is common to find several studies designed to answer similar questions about the effectiveness of a clinical intervention. These disparate studies provide evidence about clinical effectiveness but the scope of their designs may be too limited to come to a generalizable conclusion. The idea of pooling results from disparate studies can be traced back to the work of Karl Pearson in 1904, [181]. In his account on the preventive effect of serum inoculations against enteric fever, Pearson reported the use of formal techniques to combine data from different samples. The same reasoning that motivated Pearson for pooling studies is still one of the main rationales for conducting meta-analysis today: “Many of the groups are far too small to allow of any definite opinion being formed at all, having regard to the size of the probable error involved.” [181].

Meta analysis, often called evidence synthesis, is a technique for summarizing and integrating findings of related studies in order to generate a generalizable conclusion about the effectiveness of a clinical strategy from a larger pool of data [47]. The most commonly used clinical effects in meta-analysis include the odds ratios, risk ratios and risk differences among others. Over the past few decades, there has been a tremendous increase in the use of meta-analysis for summarizing clinical evidence. In the light of this pervasive use, meta analysis has now become a hallmark of evidence-based medicine [47]. Figure 5.1 shows a plot of the number of published studies on meta analysis. We searched three publicly available databases (Pubmed, web of knowledge and Zetoc) to access the number studies on meta analysis published by year since 1980 up till 2010 at an interval of five years i.e. published studies in 1980, 1985 ... 2010. These databases are the official repositories for published medical studies. Figure 5.1 shows a plot of the number of published meta analysis studies retrieved from each of these three databases by year.

Clearly, the number of papers published work on meta-analysis in the medical domain has increased sharply in the past 30 years. More and more meta analysis studies are published to summarize different *medical effects*. The benefits of meta-analysis have been well recognized [150]:

1. Meta-analysis allows for a more objective appraisal of the clinical evidence which

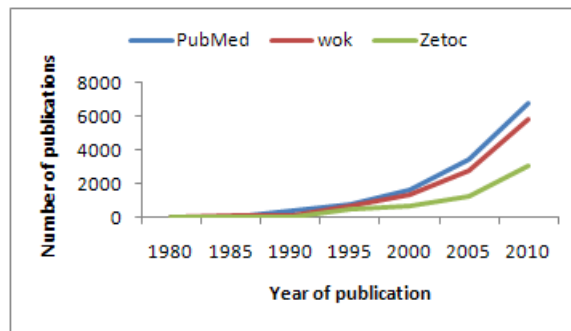


Figure 5.1: Plot showing the number of published meta-analysis studies since 1980

may reduce the uncertainty about a clinical effect;

2. Meta-analysis may reduce the probability of false negative or false positive results, thus preventing undue delays in the introduction of effective treatments into clinical practice;
3. Meta-analysis allows testing (a priori) of hypotheses regarding treatment effects in the study population or subgroups thereof;
4. Meta-analysis allows us to explore heterogeneity between study results which are, in some cases, explainable;
5. In meta-analysis, promising research questions that can be addressed in future studies may be generated.

Recently, meta-analysis has also gained an increasing popularity in the Bayesian community. Many authors have addressed meta-analysis problems from the Bayesian perspective. Examples of Bayesian meta-analyses include modeling the effectiveness of statins in preventing death after an initial myocardial Infarction [9], summarizing the incremental benefit of histamine dihydrochloride when added to interleukin-2 in treating acute myeloid leukemia [13], modeling random effects in a meta-analysis in urinary tract infections [214], modeling heterogeneity in relation to underlying risk [226] and incorporating external evidence on heterogeneity in a trial in cirrhosis [93].

Often the Bayesian approach to meta-analysis incorporates prior distributions which represent subjective opinion about a treatment effect [93]. Ashby described a list of questions researchers may wish to frame that can help in the process of formulating prior

distributions, these include [7]: “What do we think about the relative benefits of the treatments before knowing the results from this trial?”; “What information can be gained from the results of this trial?” and “Considering the results of this trial in the light of previous understanding, what do we now think about the relative benefits of the treatments?” In contrast, Higgins and Whitehead [93] described an approach to incorporate real data from previous studies using Bayesian techniques. Their approach demonstrate the readiness of the Bayesian approach for creating *dynamic meta analysis models*¹. The work of Higgins and Whitehead [93] has set a very good foundation for the work presented in this chapter.

This chapter demonstrates how hierarchical BNs, augmented by dynamic discretization can bridge the methodological gaps between meta-analysis and decision analysis. It is a common practice to use a summary estimate from a meta-analysis as an input in a decision analytical model. Indeed, Ashby and Smith [6] argue that the cardinal focus of evidence-based medicine is about making decisions and that the Bayesian approach is the natural framework to adopt. Using the BN framework to solve meta-analysis problems facilitates the use of a posterior distribution of a “summary effect” obtained from a previous meta analysis as a prior distribution in a subsequent meta analysis. More importantly, the posterior distribution from a meta-analysis can serve as an input (prior) in a decision-model. In addition, the chapter also describe how a meta-analysis can be conducted with a constraint on the posterior distribution of the heterogeneity parameter thereby allowing analysts to make a probabilistic statement about the heterogeneity parameter. Essentially, this approach will recognize the presence of heterogeneity between studies but analyst can decide a priori on the magnitude of the *heterogeneity* to be allowed in the analysis.

Dichotomous data

This section reviews the concepts of Dichotomous data used in this chapter. Dichotomous data are normally presented in a 2-by-2 table where each cell contains the number of patients in treatment or control categories (Table 5.1). Observation y_T and y_C are the number of events (for instance number of deaths following a clinical intervention) recorded in the treatment and the control group respectively while y'_T and y'_C are the

¹By *dynamic meta-analysis models* we mean meta-analysis models that will continue to adapt by updating the *summary effect* when data from a new meta-analysis on the same subject arrives.

number of non events recorded in the two groups.

Table 5.1: 2-by-2 Presentation of dichotomous data

	Treatment	Control
Event	y_T	y_C
No Event	y_T'	y_C'
Total	n_T	n_C

$$y_T' = n_T - y_T, y_C' = n_C - y_C$$

Risk ratio (RR), odd ratio (OR) and risk difference (RD)

Risk ratio (RR), odd ratio (OR) and risk difference (RD) are clinical effects commonly used in practice to summarize the effectiveness of clinical intervention from dichotomous data. For example, if we denote the risk and the odds of a clinical strategy by θ and ϕ respectively, then the risks in the treatment and control groups are given by Equation 5.1 and Equation 5.2 respectively while the odds are respectively given by Equation 5.3 and Equation 5.4. The risk in this context may be the probability of an adverse outcome accompanying a clinical intervention.

$$\theta_T = \frac{y_T}{n_T} \quad (5.1)$$

$$\theta_C = \frac{y_C}{n_C} \quad (5.2)$$

$$\phi_T = \frac{y_T}{y_T'} \quad (5.3)$$

$$\phi_C = \frac{y_C}{y_C'} \quad (5.4)$$

Risk ratio (RR) is simply the ratio of risk in the treatment group divided by the risk in the control group (Equation 5.5).

$$\frac{\theta_T}{\theta_C} \quad (5.5)$$

The odds ratio (OR) is the odds of an event in the treatment group divided by the odds in the control group expressed formally in Equation 5.6.

$$\frac{\phi_T}{\phi_C} \quad (5.6)$$

Both odds ratio and risk ratio are based on relative comparison of risks. On the other hand, risk difference (RD) differs by comparing risks in absolute terms. This is the risk

in the treatment group subtracted from the risk in the control group (Equation 5.7).

$$\theta_T - \theta_C \tag{5.7}$$

In the clinical domain, relative risk (rather than the odds ratio) is often the parameter of greatest interest [173]. The odds ratio, however, has interesting properties such as symmetry and the asymptotic normality of its natural logarithm [173]. Both odd ratio and risk ratio give the same result for a rare event [173]. A rare event is characterized by a very low prevalence. However the odds ratio can overestimate relative risk when dealing with a common event with high prevalence [173, 192, 235, 249].

5.2 Conventional approach to Meta Analysis

The idea of meta-analysis is to generate a combined effect by polling effects from different studies and calculating a weighted average of the selected studies. Suppose all studies included in a meta-analysis are equally precise then we could simply compute the average of the effect size. However, if we believe that some studies are more precise than others then we would assign more weight to those studies that carry more information. Various classical statistical techniques have been developed for computing weighted means. We can broadly categorize these techniques into two classes based on the underlying assumption about the treatment effect. Two different assumptions are made in meta-analysis giving rise to classes of models namely: fixed and random effect models.

In a fixed effect model, we assume that all studies are estimating the same *effect size* and so we want to assign a weight to each of these studies based entirely on the amount of information captured by the study. In this case, a large study would be given a higher weight than a small study. Under this assumption, a large study may dominate the analysis so much so that the relative impact of a small study is almost ignored [17]. By contrast the random effects model allows the true effect to vary from study to study. In other words, the random effect model allows the true *effect size*, in each study, to have a distribution and the combined effect represents the mean of the population of true effects [17].

Theoretically, meta-analytic models are special cases of the general linear mixed-effects model [242]. Let S_i ($i = 1 \dots k$), be k independent studies on the same subject

with empirical evidence y_i about a treatment effect. For a fixed effect model, these observations y_i are determined by the common effect μ plus the within-study error ξ_i . Formally, for any observed effect y_i ,

$$y_i = \mu + \xi_i$$

In the fixed-effect meta-analysis, the fitted model provides an estimate of the weighted average of the true effects in the set of k studies as follows:

$$\frac{\sum_{i=1}^k w_i y_i}{\sum_{i=1}^k w_i}$$

where

$$w_i = \frac{1}{\sigma_i^2}$$

where σ_i^2 is the variability within studies and the variance of the combined effect is $\frac{1}{w_i}$. The computation of the weighted average under the fixed effect assumption is straightforward but the challenge is how to compute the weights. Classical techniques for fixed effect meta-analyses include the Mantel-Haenszel method [142], the Peto method [248], general variance based methods [241] and the confidence interval methods [80, 187].

Under the random effect analysis, we can denote the true *treatment effect* from each study by μ_i . Unlike the fixed effect model, the observations y_i are determined by the study specific effect μ_i plus the within-study error ξ_i .

$$\begin{aligned} y_i &= \mu_i + \xi_i \\ \mu_i &= \mu + e_i \end{aligned}$$

Therefore

$$y_i = \mu + \xi_i + e_i$$

Note that the total error in the random effect model has two components, variability within studies (σ_i^2) and variability between studies² (τ^2). Just like the fixed effect model, the interest is also to compute a weighted mean such that the weight (w_i^*) for each study

²The value τ^2 is often referred to as the “amount of heterogeneity” in the population of studies or simply heterogeneity parameter. Hence we will use these terms interchangeably in the rest of the thesis.

is a function of these two error components.

$$\frac{\sum_{i=1}^k w_i^* \dot{y}_i}{\sum_{i=1}^k w_i^*}$$

where

$$w_i^* = \frac{1}{\sigma_i^2 + \tau^2}$$

The question is how do we estimate the heterogeneity parameter, DerSimonian and Laird [48] proposed a simple, non-iterative method to estimate the inter-study variance of treatment effects.

In general, methods based on the fixed effect or random effect will yield the same result if the studies are homogenous i.e. the variability between studies is very low (close to zero). However, the choice of the most plausible model is challenging while dealing with *non homogenous* studies. In an extreme non homogeneous situation, methods based on the fixed effect assumption can lead to conflicting results in comparison to analysis based on the random effect assumption on the same data [185].

Since the classical techniques make use of test statistics they require a large number to justify the asymptotic assumption. Generally, they are not suitable for meta-analysis with very few studies. In the clinical domain, it is very common to have meta-analyses with less than ten studies summarizing *treatment effects* of interest. The validity of a result from such meta analysis, based on the classical techniques is questionable and should be interpreted with caution. This limitation has generated interest in the use of Bayesian meta analysis. The next section describes the Bayesian approach to meta-analyses.

5.3 Bayesian approach to Meta Analysis

Meta-analysis is closely connected to the hierarchical Bayesian Networks described in the preceding chapter. In essence, observations from each study are considered to be a “unit” of analysis within a hierarchical structure. Let us revisit a BN representation of the hierarchical model presented in Figure 3.6. To represent a meta-analysis problem in the BN framework, we represent the observed effects y_i and associated parameters θ_i by nodes in a BN model. Suppose we are summarizing an effect reported as the log of odds ratio, we can specify a normal likelihood for the effect nodes $f(y_i|\theta_i) \sim Normal(\theta_i, s_i^2)$

where s is the standard error. The basic idea in meta-analysis is the computation of a summary estimate θ which is a weighted sum θ_i .

In a meta-analysis with random effect assumptions, we can treat the effect (θ_i) from each study as a random quantity drawn from a population distribution with some parameters. This is based on the assumption that the underlying studies are exchangeable [216]. It is common to assume a Gaussian distribution with parameter μ and τ for this population. In this case, the posterior values of μ and τ are respectively the “summary effect” and heterogeneity parameter. The heterogeneity parameter captures the variability between the underlying studies. If the underlying studies are homogeneous, the posterior distribution of τ should concentrate around zero. Technically, setting the posterior of τ to zero or constraining it to peak around zero would yield a summary estimate (μ) similar to the estimate from a model with fixed effect assumptions.

Although modeling y_i with a Gaussian distribution is computationally convenient, it may be inappropriate when summarizing outcomes that are not normally distributed. The approach described in this thesis allows us to specify any reasonable distribution for the summary effect. We can also model the likelihood function by using any plausible statistical distribution. Indeed, this framework supports a wider range of prior distributions i.e. both conjugate and non conjugate prior distributions. This novel modeling approach to meta-analysis makes use of Bayesian network models augmented by the dynamic discretization algorithm (described in chapter 3) to compute posterior marginal distributions for the treatment effect. We apply BN models to tackle various meta-analysis problems and demonstrate how the benefits of the Bayesian approach [219] are realized. Sutton *et al* [219] summarizes these benefits as follow:

1. **Unified modeling.** One of the challenges in meta-analysis is the model assumption (fixed effect or random effect). The statistical methods used to combine the study results vary according to the model assumptions [185]. The Bayesian approach provides a unified modeling framework that allows us to model the variability between and within trials. The posterior distribution of the τ (described earlier) in the random effect analysis can be fixed or constrained to achieve an estimate based on a fixed effect assumption.

2. **Borrowing strength.** The exchangeability³ assumption leads to weaker studies (such as studies that recruited too few patients in the trial) to *borrow* information from the other stronger studies (such as clinical trials with many patients), leading to a shrinkage of the estimate towards the overall mean, and a reduction in the width of the interval estimate. However, the degree of pooling depends on the empirical similarity of the estimates from the individual studies.
3. **Allowing for parameter uncertainty.** The parameters in a model may depend on other parameters called hyper parameters. In Figure 3.6, an unobservable parameter θ governing the likelihood of an estimate y was treated as a random variable depending on another parameters μ and τ .
4. **Allowing for other sources of evidence.** One attractive feature of the Bayesian framework is the incorporation of prior information in the analysis. This prior can be based on previous research or a subjective opinion of an expert. This framework allows the use of a current meta-analysis result as prior information in a future study.

5.3.1 Empirical justification for the exchangeability assumption

The underlying assumption in Bayesian data analysis is the exchangeability assumption. Gelman [73] declared that this assumption may be inappropriate in the face of extreme heterogeneity. In reality, clinical studies differ at several fundamental levels including population, measurement error and design methods among others. These are potential sources of heterogeneity in meta-analysis data. In this study, we first explore the data for outlying observations. This is done by plotting the graph of the posterior density of the unknown parameter given the observations i.e. estimates of the treatment effect from each study. Let X be the posterior density of the unknown treatment effect given the observation from a hypothetical study x and Y be the posterior density given observation from another hypothetical study y . We have provided two possible graphs showing proximity of the two posterior densities.

Figures 5.2(a) & 5.2(b) above show the likelihood densities of two hypothetical studies estimating the same effect and the posterior density of the true effect. In the Figure 5.2(a),

³The term exchangeability is used to express the lack of information to distinguish the true effect of one study from another. Although these true effect are not identical but they are not so different that it makes sense to combine them.

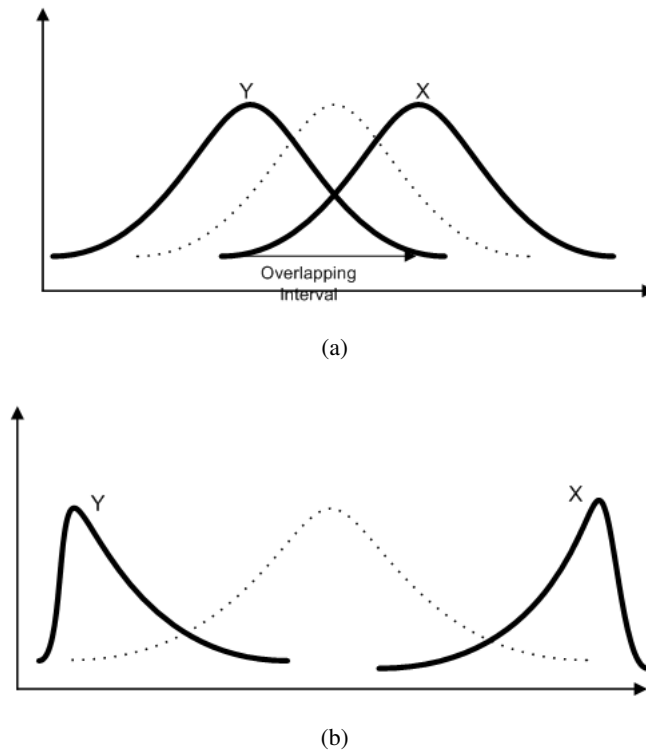


Figure 5.2: Graphs showing the likelihood densities of two hypothetical studies (X, Y) and the posterior density of the combined effect

the likelihood densities of the two studies are very close but these are wide apart in Figure 5.2(b). The fundamental question here is how far apart can we allow these two densities to be before concluding that the exchangeability assumption is questionable.

The method of probability calculation carried out in AgenaRisk aims at making the overall risk model consistent with the assumptions made and the observations entered. If the observations entered on two different nodes result in a very small joint probability of occurring together ($< 10^{-38}$), then we treat these observations as inconsistent with each other [132].

Suppose we have observations x_i , the estimates of a treatment effect θ , from n exchangeable studies. Let the posterior densities $p(\theta|x_i)$ and $p(\theta|x_j)$ be $f(\theta)$ and $g(\theta)$ respectively. The Kullback-Leibler (KL) distance between these two posterior densities is given by Equation 5.8:

$$D(f(\theta)||g(\theta)) = \int p(\theta|x_i) \log \frac{p(\theta|x_i)}{g(\theta|x_j)} d\theta \geq 0 \quad (5.8)$$

The KL distance between the two posterior densities $p(\theta|x_i)$ and $p(\theta|x_j)$ will be low

if observations x_i and x_j are consistent estimates of θ . A very low KL value provides an empirical justification for the exchangeability assumption. The KL metric will be used later in this chapter to explore data on the prevalence of aneurysm for inconsistent observations.

5.3.2 Application of HRBN to meta analysis

This section builds on the methodological steps described earlier and applies a hierarchical BN model to a meta-analysis problem. Let us extend Table 5.1 so that it allows entries for more than one study. These entries can be indexed by $i = 1, \dots, n$ (i.e. the index of the studies that produced them). For each study i , the observation y_{Ti} and y_{Ci} are the number of events recorded in the treatment and the control group respectively while y'_{Ti} and y'_{Ci} represent the number of non events recorded in the two groups. Hence the OR for the i^{th} study is given by Equation 5.9.

$$OR_i = \frac{\theta_{Ti}(1 - \theta_{Ci})}{\theta_{Ci}(1 - \theta_{Ti})} \quad (5.9)$$

Suppose y_1 and y_2 are two observations about a treatment effect from two exchangeable studies S_1 and S_2 . Let us assume that y_1 and y_2 are normally distributed random variables. The Bayesian model for meta-analysis, as described by DuMouchel, [53] is of the form $y_i \sim Normal(\mu, [\sigma_i^2 + \tau^2])$, where σ_i^2 is the estimate of the variance of the i^{th} study and τ is a measure of variabilities between studies. The fixed effect meta-analysis assumes that $\tau = 0$. The posterior mean ($E(\mu|y_1, y_2)$), based on the fixed effect assumption is given as Equation 5.10 with variance given as Equation 5.11.

$$E(\mu|y_1, y_2) = \frac{\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)}y_1 + \frac{\sigma_1^2}{(\sigma_1^2 + \sigma_2^2)}y_2 \quad (5.10)$$

$$\sigma_\mu^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (5.11)$$

On the other hand, the random effect analysis provides an estimate for τ and used this value in estimating a posterior distribution of μ and study specific treatment effect θ_i . The posterior mean $E(\mu|y_1, y_2, \tau)$, based on the random effect is given as Equation 5.12 with variance given as Equation 5.13.

$$E(\mu|y_1, y_2, \tau) = \frac{\tau^2 + \sigma_2^2}{(\tau^2 + \sigma_1^2) + (\tau^2 + \sigma_2^2)} y_1 + \frac{\tau^2 + \sigma_1^2}{(\tau^2 + \sigma_1^2) + (\tau^2 + \sigma_2^2)} y_2 \quad (5.12)$$

$$\sigma_{\mu|\tau}^2 = \frac{(\tau^2 + \sigma_1^2)(\tau^2 + \sigma_2^2)}{(\tau^2 + \sigma_1^2) + (\tau^2 + \sigma_2^2)} \quad (5.13)$$

Let $y_1 = 10$ and $y_2 = 60$ with sampling variances $\sigma_1^2 = 20$ and $\sigma_2^2 = 40$. If we assume that $\tau = 0$, then from Equations 5.10 and 5.11, the posterior mean and variance based on the fixed effect assumption can be calculated analytically as 27 and $\sigma_{\mu}^2 = 13$. Suppose we use $\tau = 10$, then the posterior mean and variance based on the random effect assumption can be computed from Equations 5.12 and 5.13 as 33 and 70 respectively. Using the Dynamic discretization algorithm makes it possible to solve this problem with a Bayesian network model. Figure 5.3 shows the posterior distributions of the summary effect μ based on the fixed and random effect assumptions. The posterior mean and variance obtained from the BN based on the fixed effect assumption are respectively $E(\mu|y_1, y_2) = 27$ and $\sigma_{\mu}^2 = 15$. From the BN model based on the random effect assumption, we obtained a posterior mean $E(\mu|y_1, y_2, \tau = 10) = 33$ and variance $\sigma_{\mu|\tau}^2 = 65$.

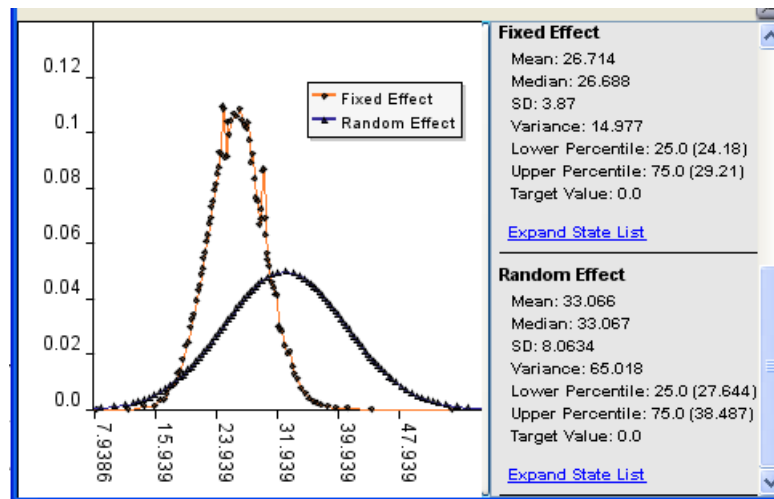


Figure 5.3: Posterior densities of the combined treatment effect based on fixed effect ($\mu|y_1, y_2$) and random effect ($\mu|y_1, y_2, \tau = 10$)

The rest of this chapter presents different applications of this new approach to practical meta-analysis problems. Each of the meta-analysis problems serves a specific purpose.

The Beta-blocker clinical example is intended to compare the results the BN approach with those from a previous analysis using MCMC simulation. The goals of the second example (Magnesium trial) is to show that the BN approach allows us to do something that the conventional approaches cannot do. The BN approach allows us to constrain the posterior distribution of the heterogeneity parameter τ and then perform inference, subject to this constraint. The third example demonstrates the use of a BN model to summarize data from a non-experimental design. The prevalence of aneurysm data shows wide variability, therefore we use the KL metric, describe earlier, to explore the data for extreme heterogeneity.

5.4 Beta-blockers clinical trials

Heart failure is one of the conditions that carry a high burden of mortality and morbidity [143]. A great deal of research effort has been devoted into developing potent and effective drugs to alleviate the problem of heart failure condition, and is still an active research area. The beta-blockers⁴ treatment was introduced a few decades ago for heart failure and since then several randomized trials have been conducted of their effectiveness. One systematic review conducted by Marcelo *et al.* [143] identified 22 randomized controlled trials with a total of 10480 patients and an average of 11 months of treatment. They excluded studies with severe heart failure and used the classical fixed effect and random effect model to summarize the effectiveness of beta-blockers.

In this section, we will describe HRBN models to summarize the effectiveness of the beta-blockers treatment based on the same data [143]. Hence, the scope of this work does not cover the systematic approach employed by Marcelo *et al.* [143] to select the studies. The focus is to use a BN model to summarize data from existing meta-analysis study and compare our result, based on the dynamic discretization algorithm, with estimates produced using Gibbs sampling technique.

5.4.1 Data from a meta-analysis on Beta-Blocker conducted by Marcelo *et al.* [143]

The meta-analysis study conducted by Marcelo *et al.* [143] synthesized evidence from 22 studies. The empirical evidence on mortality after myocardia infarction (heart attack) in

⁴ A family of drugs that affect the central nervous system and can relax the heart muscle

these 22 exchangeable trials is presented in Table 5.2. In each trial, heart attack patients were randomly allocated to receive (treatment group) or not receive (control group) the beta-blockers treatment. The raw data from the study are y_{Ti} , n_{Ti} , y_{Ci} , n_{Ci} , where n_{Ci} and n_{Ti} are respectively number of subjects in the control and treatment groups giving rise to y_{Ci} and y_{Ti} deaths respectively for the i^{th} study.

Table 5.2: Effects of beta-blockers from 22 clinical Trials, extracted from Table 5.4 of [73]

Study i	n_{Ci}	y_{Ci}	n_{Ti}	y_{Ti}
1	39	3	38	3
2	116	14	114	7
3	93	11	69	5
4	1520	127	1533	102
5	365	27	355	28
6	52	6	59	4
7	939	152	945	98
8	471	48	632	60
9	282	37	278	25
10	1921	188	1916	138
11	583	52	873	64
12	266	47	263	45
13	293	16	291	9
14	883	45	858	57
15	147	31	154	25
16	213	38	207	33
17	122	12	251	28
18	154	6	151	8
19	134	3	174	6
20	218	40	209	32
21	364	43	391	27
22	674	39	680	22

Suppose we are interested in summarizing the log odd ratio \hat{OR}_i of the treatment effect. Various methods have been described for estimating the odds ratio from data (see [172] for a detailed discussion). In this example, we can use the method based on the empirical logit to compute a point estimate and standard error for \hat{OR}_i . For each study i , one can estimate \hat{OR}_i by its point estimator y_i given in Equation 5.14 and variance given by Equation 5.15.

$$y_i = \log(y_{Ti}) + \log(y'_{Ti}) - \log(y'_{Ci}) - \log(y_{Ci}) \quad (5.14)$$

$$\sigma_i^2 = \frac{1}{y_{Ti}} + \frac{1}{y'_{Ti}} + \frac{1}{y_{Ci}} + \frac{1}{y'_{Ci}} \quad (5.15)$$

Table 5.3 shows the estimate (y_i) and variance σ_i^2 for the 22 studies.

Table 5.3: Effects of beta-blockers from 22 clinical Trials, extracted from Table 5.4 of [73]

Study i	Effect y_i	sd (σ_i)	var (σ_i^2)
1	0.028	0.850	0.723
2	-0.741	0.483	0.233
3	-0.541	0.565	0.319
4	-0.246	0.138	0.019
5	0.069	0.281	0.079
6	-0.584	0.676	0.457
7	-0.512	0.139	0.019
8	-0.079	0.204	0.042
9	-0.424	0.274	0.075
10	-0.335	0.117	0.014
11	-0.213	0.195	0.038
12	-0.039	0.229	0.052
13	-0.593	0.425	0.181
14	0.282	0.205	0.042
15	-0.321	0.298	0.089
16	-0.135	0.261	0.068
17	0.141	0.364	0.132
18	0.322	0.553	0.306
19	0.444	0.717	0.514
20	-0.218	0.260	0.068
21	-0.591	0.257	0.066
22	-0.608	0.272	0.074

The estimates y_i are treated as normally distributed observations with mean $\hat{O}R_i$ and variances (σ_i^2). The values for y_i and σ_i^2 are indicated in Table 5.3. The normal likelihood models for the observations y_i is given by Equation 5.16. This is the “*observation level*” of a hierarchical model.

$$y_j | \hat{O}R_i, \sigma_i^2 \sim N(\hat{O}R_i, \sigma_i^2) \quad (5.16)$$

At the second level of the hierarchy, we can specify an exchangeable normal prior distribution with mean μ and variance τ for $\hat{O}R_i$. Therefore, μ and τ are the hyper parameters in the hierarchical model (Equation 5.17).

$$\hat{O}R_i \sim N(\mu, \tau) \quad (5.17)$$

Let us consider two set of non-informative priors. Firstly, we can specify a uniform distribution for the hyperparameters i.e. $\mu \sim U(-10, 10)$ and $\tau \sim U(0, 1000)$. The HRBN

model in Figure 5.4, obtained from the AgenaRisk software [131] is based dynamic discretization with 50 iterations.

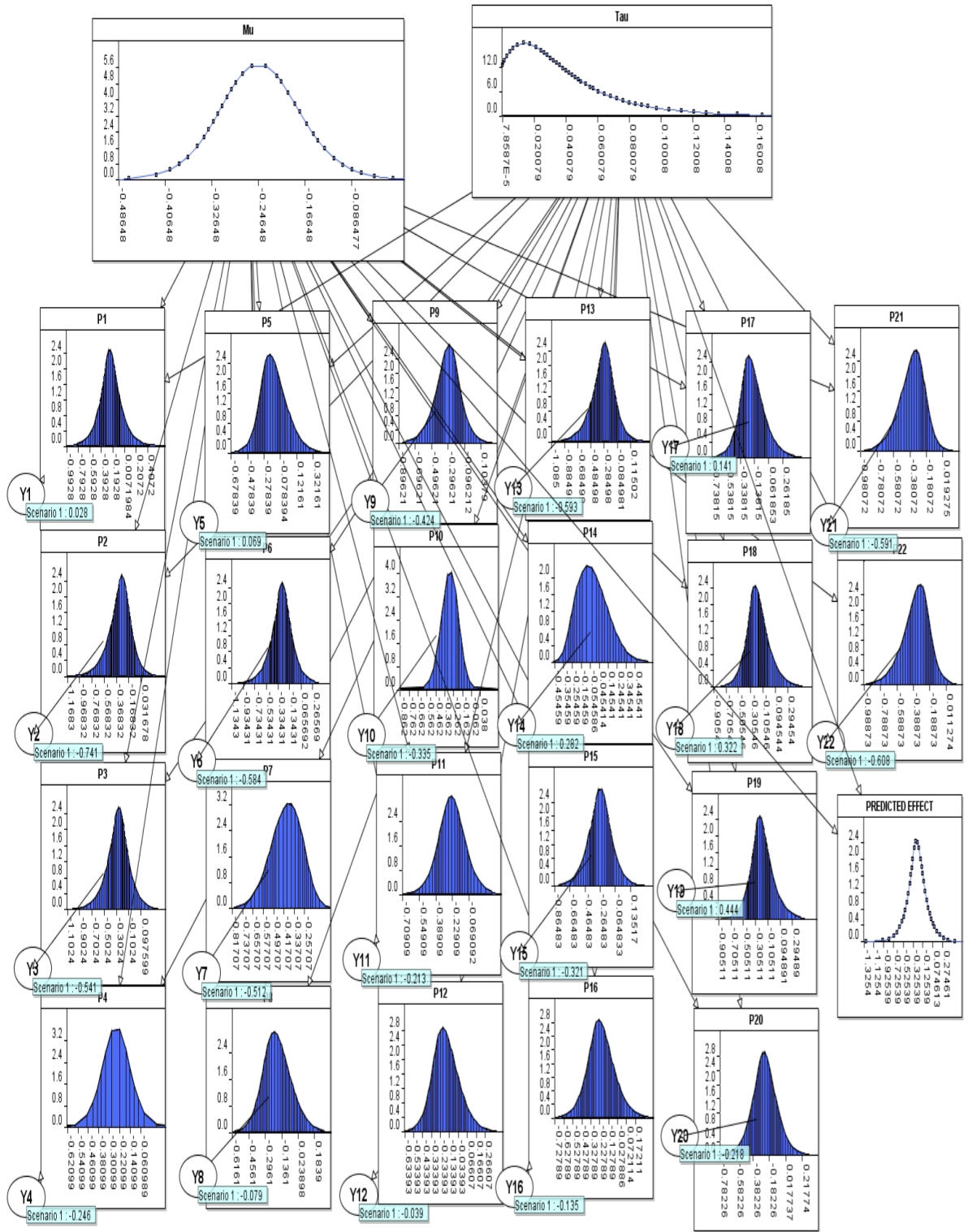


Figure 5.4: Hierarchical BN model for Beta-Blocker Trial

Table 5.4 shows the summary statistics for the posterior distributions of the combined effect (μ), heterogeneity parameter (τ) and a predicted effect of a future study ($y_{23}|y_1, \dots, y_{22}$). These results compare favorably with those achieved using Gibbs Sam-

pling approximation in [73] (estimates produced using Gibbs sampling in parenthesis).

Table 5.4: Summary of posterior inference of beta-blockers

Parameter	Posterior quantiles				
	2.5%	25%	Median	75%	97.5%
Mean (μ)	-0.38 (-0.37)	-0.29 (-0.29)	-0.24 (-0.25)	-0.20 (-0.20)	-0.10 (-0.11)
Variance (τ)	0.002 (4×10^{-4})	0.015 (6.4×10^{-3})	0.001 (0.0169)	0.053 (0.0324)	0.134 (0.0961)
Predicted Effect	-0.68 (-0.58)	-0.36 (-0.34)	-0.25 (-0.25)	-0.13 (-0.17)	0.20 (0.11)

The second non-informative prior distribution assumes truncated normal distributions for the hyper parameters. Both μ and τ have truncated normal distributions with mean 0 and variance (10^9). Formally, $\mu \sim N(0, 10^9)$ and $\tau \sim N(0, 10^9)$ where μ and τ are bounded by the intervals ($\mu \in [-10, 10]$ and $\tau \in [0, 10]$). The posterior distributions for all parameters are identical to those obtained using the uniform distributions (non-informative).

In this analysis, we have also identified a significant effect (odds ratio) for the beta-blocker treatment suggesting that the treatment was effective. That is, the rate of mortality following heart failure condition was higher in the control group than those who received the beta-blocker treatment. The good thing in this analysis is that we are able to express different prior beliefs on the treatment effect even before the analysis, these can be subjective or empirical. We have demonstrated this with two theoretical non-informative prior distributions for lack of information. Also, the result from this analysis can be used to formulate an informative prior distribution in a future meta-analysis study to summarize the effect of beta-blocker treatment.

In terms of comparing our approach, using HBNs and DD, against Gibb's sampling this analysis has shown confirmation that, when the same prior distributions are used very similar results for parameter distributions are provided by both analyses. Thus our new method is as reliable as the Gibb's sampling method. Furthermore the flexibility of changing the priors to non-conjugate values has been illustrated, something that the Gibb's approach cannot accommodate. However one weakness of our technique is that the variances on the estimates appear to be larger under our approach than Gibb's: whether this is a systematic or a local issue needs further investigation.

5.5 Magnesium trials in myocardial infarction

The relevance of magnesium to both the incidence and the management of ischaemic heart disease has been well studied [76, 127, 224, 243]. A study of the geographical comparisons of some regions in South Africa conducted by Leary *et al* showed that death rates from ischaemic heart disease tend to be higher where magnesium concentrations in soil and water are low [127]. Also a case-control study showed that the concentration of magnesium tends to be lower in those who die of ischaemic heart disease than in those who die of other causes [55]. A research has shown that several actions of the magnesium ion could contribute towards some cardioprotective effects [243]. In a study conducted by Chang *et al* [27], magnesium has been shown to limit infarct size in dogs. Ghani and Rabah [76] showed that an infusion of magnesium in animal increases the threshold for electrical excitation of myocardial cells, thereby reducing the likelihood that a current will create a damage near the ischaemic or infarcted tissue [224].

There has been a debate in the literature regarding the benefits of magnesium in patients with acute myocardial infarctions. This is due to the conflicting results from meta-analysis and mega trial to investigate the benefit of magnesium after myocardial infarction. While the initial meta-analysis conducted by Teo *et al* [224] reported a significant treatment effect on the use of magnesium following myocardial infarctions, a multi-center mega trial conducted by ISIS-4 collaborative Group [82] reported a non significant result. Our interest is not to join or extend this debate but to use a Bayesian network model for summarizing treatment of magnesium on patients with myocardial infarctions.

This analysis was motivated by the work of Higgins and Spiegelhalter [94] on the same subject. Higgins and Spiegelhalter [94] used the Bayesian approach to summarize data [218] from 15 randomized trials of intravenous magnesium for acute myocardial infarction conducted. They show how scepticism can be formally incorporated into an analysis as a Bayesian prior distribution. They also show how Bayesian meta-analysis models allow appropriate exploration of different hypotheses about the treatment. In particular they explore an hypothesis that a treatment effect depends on the size of the trial or on the risk in the control group.

Our main goal is to investigate how Bayesian network can be used to summarize treatment effects of intravenous magnesium for acute myocardial infarction. In particular

we want to address a question “how might a BN model for meta-analysis be used to make a probabilistic statement about the distribution of heterogeneity parameter τ ?”

5.5.1 Summary of data on magnesium trials in myocardial infarction

The section presents data (Table 5.5) from 15 randomized trials of intravenous magnesium for acute myocardial infarction. This data consists of published evidence selected in two meta-analysis studies and data from two mega-trial. The first seven studies in Table 5.5 are those included in the meta-analysis conducted by Teo *et al* [224] in 1991. This meta-analysis reported a 55% reduction in odds of mortality. The second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2) [244] is a large trial conducted in 1992 which also reported a 24% reduction in mortality. An updated meta-analysis conducted in 1995 by Teo *et al* [224] identified five additional studies on this subject in addition to a mega-trial conducted by ISIS-4 [82]. However, this mega-trial reported a non significant treatment effect for intravenous magnesium following myocardial infarction.

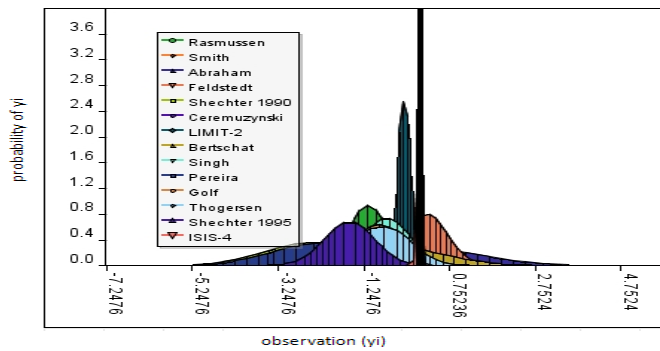
Table 5.5: Summary data from 15 randomized trials of intravenous magnesium for acute myocardial infarction (data from Sterne *et al.* [218])

Study identifier	treatment		Control		Estimate from Data	
	n_{Ti}	y_{Ti}	n_{Ci}	y_{Ci}	y_i	σ_i^2
Morton (1)	40	1	36	2	-0.8303	1.5551
Rasmussen (2)	145	9	135	23	-1.1324	0.1709
Smith (3)	200	2	200	7	-1.2783	0.6531
Abraham (4)	48	1	46	1	-0.0435	2.0435
Feldstedt (5)	150	10	148	8	0.22311	0.2393
Shechter 1990 (6)	59	1	56	9	-2.4075	1.1496
Ceremuzynski (7)	25	1	23	3	-1.2809	1.425
LIMIT-2 (8)	1159	90	1157	118	-0.2993	0.0215
Bertschat (9)	22	1	21	2	-0.7932	1.6003
Singh (10)	76	6	75	11	-0.6957	0.2875
Pereira (11)	27	1	27	7	-2.2083	1.2313
Golf (12)	23	5	33	13	-0.8502	0.3825
Thogersen (13)	130	4	122	8	-0.7932	0.3917
Shechter 1995 (14)	107	4	108	17	-1.5708	0.3295
ISIS-4 (15)	29011	2216	29039	2103	0.05761	0.0012

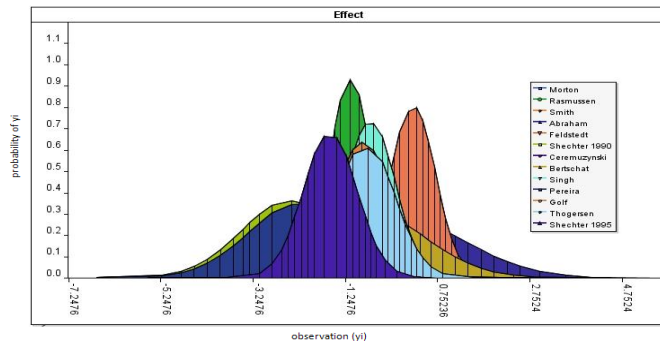
The observations n_{Ti} , y_{Ti} , n_{Ci} , y_{Ci} and the estimates y_i and σ_i^2 are as described in the previous section. Unlike the Beta-blockers clinical trial example where there is no study with substantially higher number of patients than others, both LIMIT-2 and ISIS-4 have considerably higher number of patients than the rest. Therefore, it is reasonable to

explore the data for the validity of the exchangeability assumption. We will now follow our methodological steps described earlier to explore the data and empirically assess if all the observations are indeed combinable.

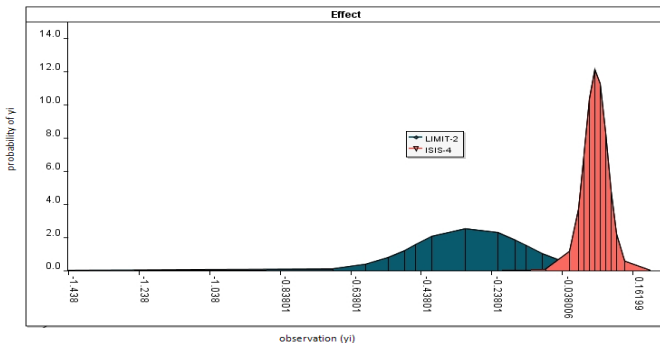
In this example, we will use the plots of likelihood densities to check if observations from a study is inconsistent with those from other studies. A normal likelihood density was assumed for the estimate $y_j | \hat{O}R_i, \sigma_i^2 \sim N(\hat{O}R_i, \sigma_i^2)$ and uniform prior for the treatment effect (log odd ratio) $\hat{O}R_i \sim N(-10, 10)$. We consider visualizing the posterior density of the true effect of each group. For each study i , we obtained the density $f(\hat{O}R_i | y_i)$ as shown in the Figure 5.5.



(a) All studies included the meta analysis



(b) excluding ISIS-4 and LIMIT-2



(c) Two large studies (ISIS-4 and LIMIT-2)

Figure 5.5: Exploratory analysis of the data from the Magnesium trial.

The likelihood graphs of the treatment effect, for all studies, concentrate on the left of 1 with many of them overlapping with each other. However, the densities of the two large studies have slightly different shapes compared to the rest of the studies. Therefore, it is not surprising that [94] found evidence of dependence of treatment effect on the size of the study. Let us now proceed by using the fixed effect assumption to summarize the effect of intravenous magnesium for acute myocardial infarction.

5.5.2 Fixed effect Model

The fixed effect model corresponds to the assumption that all underlying treatment effects are identical such that $\sigma_i^2 = \sigma^2$ [94]. That is, a common odds ratio underlies each and every trial. We can use the usual normal likelihood function for the estimate y_i and a uniform prior distribution for the common effect as given below:

$$\begin{aligned} y_j &\sim N(\hat{OR}, \sigma_i^2) \\ \hat{OR} &\sim U(-10, 10) \end{aligned}$$

The BN model showing the posterior distribution for the combined effect based on the fixed effect assumption is presented in Figure 5.6.

We compare our result with the Peto method [248], a conventional fixed effect approach. Table 5.6 shows the result obtained from the BN model and those from the Peto's method. The BN Result presented here was based on the dynamic discretization algorithm with 40 iterations.

Table 5.6: Summary of the Combined effects based on Fixed effect

Estimand	Peto Method	BN Result
\hat{OR} (log scale)	0.02	0.02
$var(\hat{OR})$ (log scale)	0.031	0.035
$exp^{\hat{OR}}$	1.02	1.02

In practice, prior information may be available (for instance) from previous analysis or even subjective views of experts. It is reasonable to incorporate the existing knowledge in the analysis (if available). Let us now revisit the magnesium trials example by incorporating prior information by grouping the data into two subgroups similar to the grouping made by Higgins and Spiegelhalter [94]. The first group consists of the first eight trials

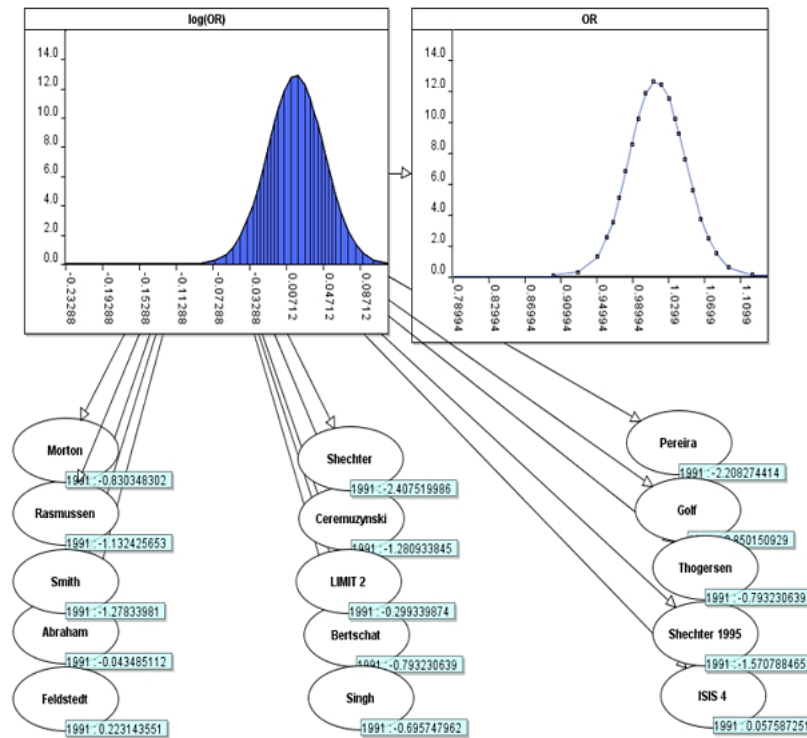


Figure 5.6: Posterior distribution of the common effect (Fixed effect model)

in Table 5.5 corresponding to seven studies included in the 1991 meta-analysis conducted by Teo *et al.* [224] and the LIMIT-2 trial [244]. The remaining seven trials consisting of six additional studies that were identified in the updated meta-analysis and the result of the ISIS-4 mega trial. We can summarize the data in each of these groups using the BN modeling approach. Also, we can leverage on the capability of the BN framework to use the posterior distribution of the 1991 meta analysis as a prior distribution in the 1995 meta analysis. The posterior distributions of the summary effect from these three analyses are shown in Figure 5.7.

Clearly, the posterior distributions of the summary effect for these subgroup lead to conflicting conclusions. The posterior density of the summary effect for the first group concentrates on the left of “1” (odd ratio) thereby suggesting a significant treatment effect. Whereas, the posterior distribution of the *summary effect* for the second group includes “1” (odd ratio) suggesting a non significant treatment effect. More so, using the posterior distribution of the first group as the prior in the second group does not affect the conclusion i.e. a non significant treatment effect. Table 5.7 shows the posterior means for each group.

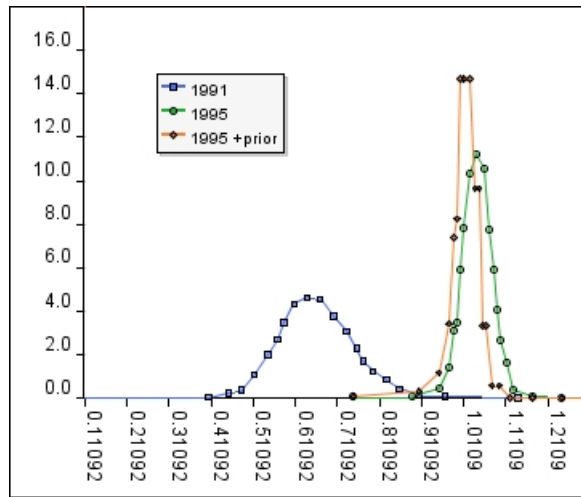


Figure 5.7: Posterior distributions of true effect based on a) 1991 data analysis b) additional data found in 1995 and c) additional data found in 1995 with result of 1991 analysis as the prior

Table 5.7: Statistics of true effect based on data 1991 and 1995

Estimand	Group 1 (1991)	Group 2 (1995)	Group 2 (1991 Prior)
\hat{OR}	0.662	1.04	1.02
$var(\hat{OR})$	0.0081	0.0015	0.0012

5.5.3 Random effect Model

Unlike the fixed effect model, the random effect assumption incorporates study specific variance of the treatment effect, i.e. a treatment effect can vary between the studies. The model is formalized as follows:

$$\begin{aligned}
 y_j &\sim N(\hat{OR}_i, \sigma_i^2) \\
 \hat{OR}_i &\sim N(\mu, \tau^2) \\
 \mu &\sim U(-100, 100)
 \end{aligned}$$

We considered two set of non informative prior distributions for parameter τ . We can assume a uniform prior distribution on τ ($\tau^2 \sim U(0, 100000)$), this is similar to the approach adopted in Julian *et al.* [94]. Also, we can use the inverse gamma distribution (i.e. the conjugate prior) to express the prior $\tau^2 \sim invGamma(0.001, 1000)$. The HRBN models for this problem is shown in Figure 5.8.

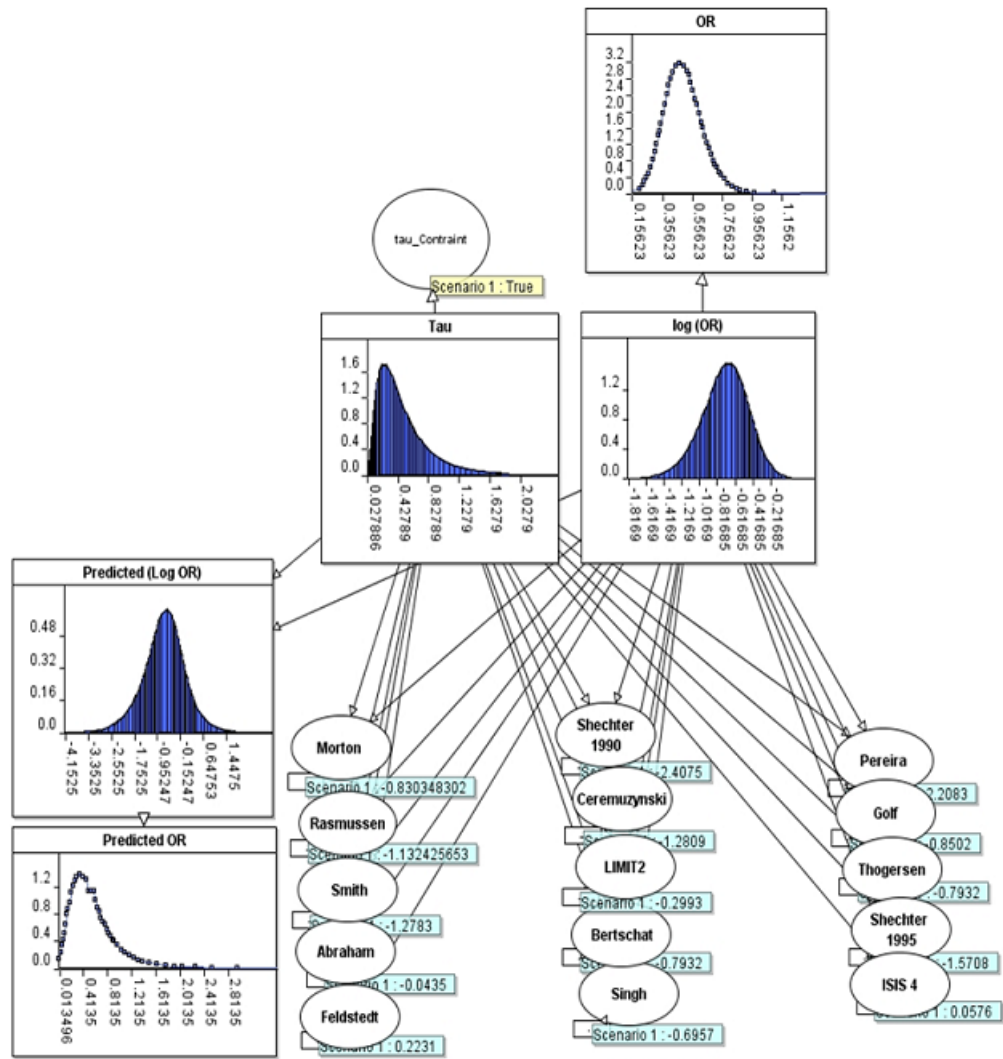


Figure 5.8: Posterior distribution of the common effect (Random effect model)

Table 5.8 shows the posterior means of summary effect based on the two prior assumptions.

Table 5.8: Summary of the Combined effects based on Random effect

Estimand	Prior Distribution	
	Uniform	Inverse Gamma
μ	0.48	0.52
τ	0.227	0.056

With the random effect assumption, the posterior distribution of the *summary effect* suggest a non significant treatment effect. This conclusion conflicts with the result obtained in Section 5.5.2. This is a classical example of meta-analysis problem that echoes the dichotomy between the two opposing philosophies of meta-analysis. Thus, a compro-

mise has to be made in the analysis. What if we are able to adjust the level of variabilities that we are willing to accommodate in our analysis? The BN modeling approach allows an analyst to carry out an inference with a constrained parameter. We will repeat this analysis in the next section with different constraints on the heterogeneity parameter (τ).

5.5.4 Analysis with a constrained τ

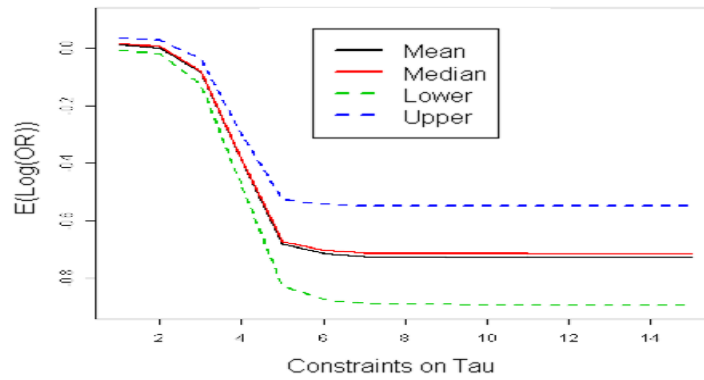
Although the *summary effect* in meta-analysis is often the parameter of interest, the reliability of such an estimate depends on the posterior value of the heterogeneity parameter (τ). While summarizing clinical evidence, analysts might decide a-priori on a threshold γ for τ . Such analysis would be a true compromise between the fixed effect that assumes homogeneity of studies and the random effect that recognizes heterogeneity in data. Analyzing meta-analysis with a constraint on the heterogeneity parameter implies that analyst can make a probabilistic statement such as “there is 0.5% probability that $\tau < \gamma$ i.e. $p(\tau < \gamma) = 0.005$ ”. Therefore, the computation of the summary estimate is then conditional on such constraints. This section describes how we can achieve this in the BN framework.

Using the HBNS and dynamic discretization we can, very easily, introduce a constraint on the posterior density of the heterogeneity parameter. Theoretically, setting the value of $\tau = 0$ implies that the treatment effect is constant between studies i.e. a fixed effect. The random effect, on the other hand allows τ to assume any positive value. With a low value of τ , studies can “borrow strength” from each other thereby shifting the posterior means of all the studies towards the posterior mean of the parameter. DuMouchel [53] suggested computation of the posterior means of the *summary effect* conditional on τ , and plot them as functions of τ , for τ varying in $(0; 2)$. However, constraining the posterior distribution of τ and performing conditional inference is a more plausible assumption. Let us now revisit the random effect analysis of the magnesium trial in Section 5.5.3 by specifying the constraints in Table 5.9 on the heterogeneity parameter.

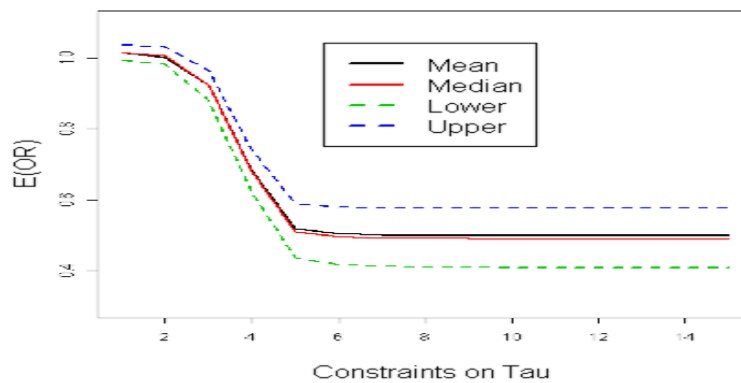
Figures 5.9(a) and 5.9(b) respectively show the graph of posterior statistics (Mean, Median, Lower and Upper quartile) of densities $p(\mu|y_1, \dots, y_n, \tau)$ and $p(\mu'|y_1, \dots, y_n, \tau)$ obtained with $\tau \leq \gamma$ for different values of γ as shown in Table 5.9.

Table 5.9: Constraints introduced on the heterogeneity parameter

Index (i)	γ
1	0.00005
2	0.0005
3	0.005
4	0.05
5	0.5
6	1
7	1.5
8	2
9	2.5
10	3
11	3.5
12	50
13	500
14	1000
15	5000



(a) $\mu(\log OR)$



(b) $\mu'(OR)$

Figure 5.9: Posterior means conditional on the constraints on τ

It is obvious from Figure 5.9 that setting a very low value for τ gives a nonsignificant result whereas higher values of τ will lead to a significant result. Analysts can decide

before the analysis on the size of heterogeneity to be allowed in the analysis and used the value as a cutoff threshold for τ . The resulting posterior distribution for *summary effect*, conditional on this cutoff point can be taken as the summary effect. Similar to a prior probability, the choice of the cutoff can be a subject point or an empirical estimate from a previous analysis.

5.6 The prevalence of intracranial saccular aneurysms

Intracranial saccular aneurysms is a cerebrovascular disorder in which weakness in the wall of a cerebral artery or vein causes a localized dilation of the blood vessel. It is sometimes referred to as cerebral or brain aneurysms. Research has shown that between 3.6 to 6% of the population harbour an unruptured intracranial aneurysm, [237] and many people with brain aneurysms are usually unaware of their presence. The risk factors for the formation of brain aneurysms include a family history of aneurysm and inherited disorders, age greater than 50 years, female gender, behavioral factors such as smoking and cocaine use [29].

Rinkel, *et al's*. [190] conducted a systematic review of data on the prevalence of intracranial aneurysms. They classified the data according to study design, diagnostic methods and study. This review includes studies involving two different diagnostic techniques (Autopsy or angiography techniques) and different designs (retrospectively or prospectively). This section describes different BN models to summarize the prevalence of intracranial using the data for a systematic review conducted by Rinkel, *et al's*. [190] (Table 5.10).

5.7 Data from a meta-analysis on the prevalence of intracranial saccular aneurysms [190]

So far we have been summarizing treatment effects reported in 2-by-2 tables. In practice, it often the case to have meta analysis data without the control group. Table 5.10 shows the data from the systematic review of prevalence of intracranial saccular aneurysm.

Table 5.10: Prevalence of aneurysms (Rinkel, *et al*'s. [190])

Studies	Method	Design	n_{Ti}	y_{Ti}
Housepian-1958	À	R	8663	14
McCormick-1965	À	R	13058	26
Romy-1973	À	R	11696	67
Inagawa-1990	À	R	10259	84
Cohen-1955	À	P	580	9
Chason-1958	À	P	2731	80
Stehbens-1963	À	P	1013	43
McCormick-1970	À	P	1619	82
du Boulay-1965	Á	R	161	0
Jakubowski-1978	Á	R	183	11
Wakai-1979	Á	R	95	7
Atkinson-1989	Á	R	278	3
Ujile-1993	Á	R	1612	44
Sugai-1994	Á	R	605	43
Wakabayeshi-1983	Á	P	17	7
Iwata-1991	Á	P	72	4
Chapman-1992	Á	P	92	4
Nagashima-1993	Á	P	2540	127
Nakagawa-1994	Á	P	400	26
Ruggieri-1994	Á	P	93	10
Leblanc-1995	Á	P	41	1
Ronkainen-1995	Á	P	396	37
Griffiths-1996	Á	P	100	9

^aMethod: Á=Angiographic Study and À= Autopsy study

^bDesign: R=retrospective and P for prospective design.

^c n_{Ti} =Total number of patients investigated in the i^{th} study;

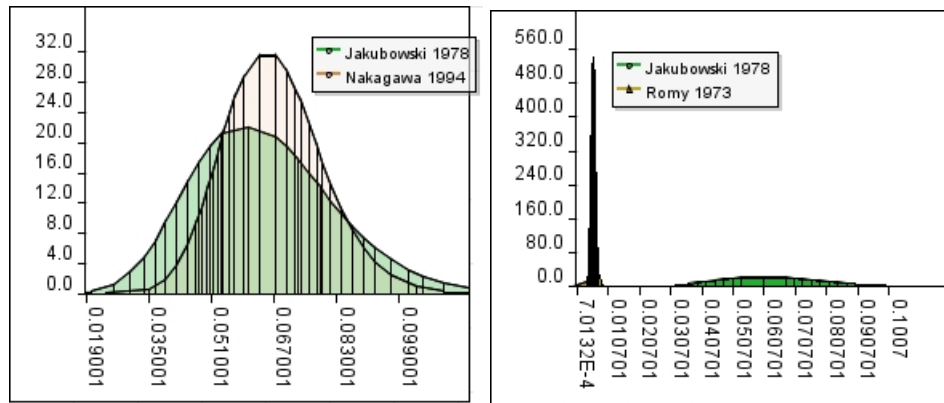
^d y_{Ti} =number of patients with aneurysms in the i^{th} study

5.7.1 Exploring the data for heterogeneity

Rinkel, *et al*[190] identified various sources of heterogeneity in data on the prevalence of aneurysm . Specifically, variation exists between data from autopsy studies and those from angiographies studies. Likewise, data obtained from prospective study vary from those obtained retrospectively. With all these sources of heterogeneity, it is important to examine the degree of variations between these studies before combining them.

Assuming a binomial likelihood function for the observed number of patients with aneurysm (y_{Ti}) from the n_{Ti} patients in the i^{th} study. We can use each of these studies to provide a partial inference about the summary estimate of the prevalence of aneurysm θ . A plot of the posterior densities of $(\theta|y_i)$ and $(\theta|y_j)$ shows the level of consistency expected by combining observations y_i and y_j . We can visually inspect the degree to which these densities $(\theta|y_i)$ and $(\theta|y_j)$ overlap with each other. For instance Figure

5.10(a) shows the posterior densities of Jakubowski 1978 & that of Nakagawa 1994.



(a) Jakubowski 1978 & Nakagawa 1994

(b) Jakubowski 1978 & Romy 1973

Figure 5.10: Visual display showing the degree of overlap between two posterior densities

In Figure 5.10(a), the posterior density of Jakubowski 1978 overlaps greatly with Nakagawa 1994 but poorly with the posterior density of Romy 1973. Also, the study from du Boulay 1965 does not to overlap with 7 other studies as shown in Figure 5.11.

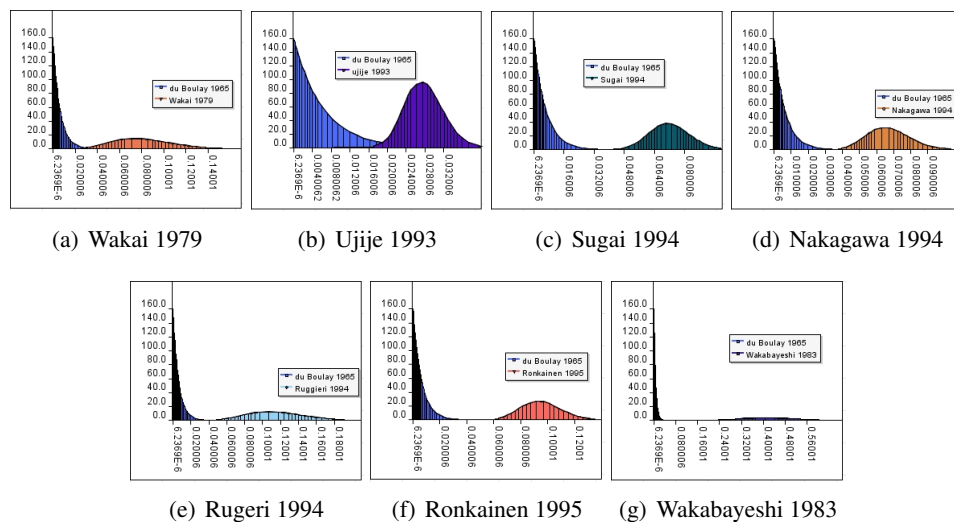


Figure 5.11: Posterior densities from du Boulay 1965 and other studies.

Clearly, the data from the study conducted by Wakabayeshi 1983 is the least consistent with those from du Boulay 1965. Figure 5.12 compares the data from Jakubowski 1978 with nine least consistent studies.

The study from Housepain 1958 has a relatively higher number of patients (13058) than other studies. The data from this study appears to be inconsistent with empirical observations from 16 other studies.

Although the density plots reveal potentially inconsistent studies, they do not quan-

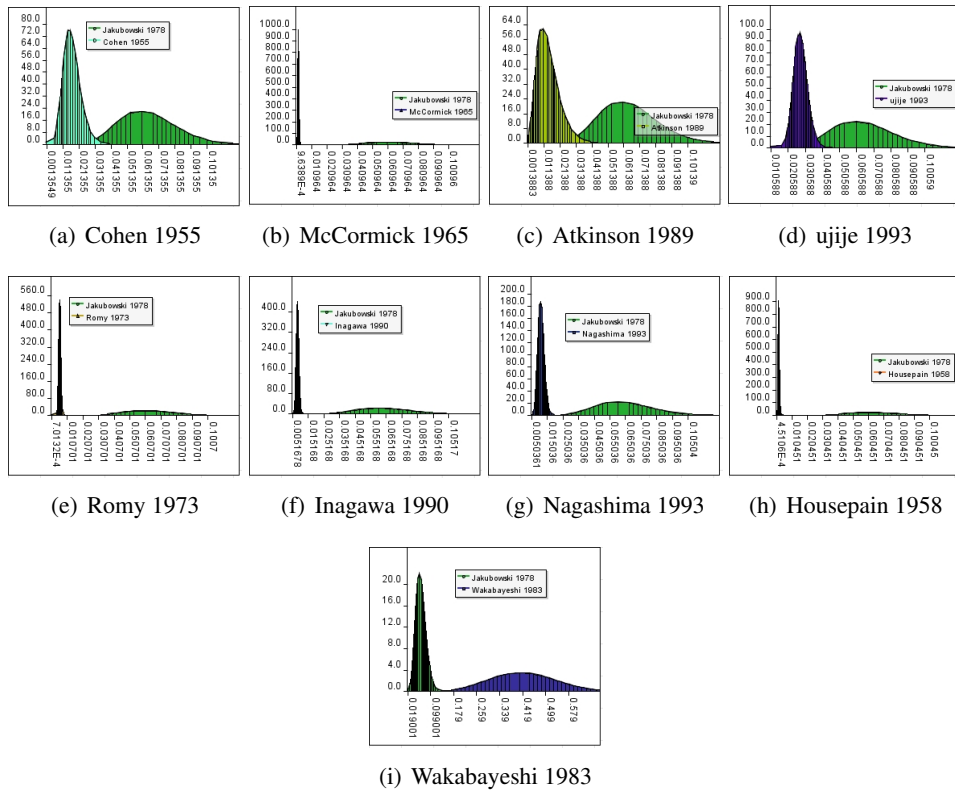


Figure 5.12: Posterior densities from du Jakubowski 1978 and other studies.

tify the degree of inconsistency between any two densities. We propose the use of the KL-distance ($KL(f(x)||g(x))$) to measure the degree of inconsistency between any two studies. We consider $f(x)$ as a partial inference about θ given observation from study i ($\theta|y_i$) and $g(x)$ as another posterior inference about θ given observation from another study j . We then compute the KL-distance between these two partial posterior distributions of θ i.e. $\theta|y_i$ and $\theta|y_j$ as shown in Table 5.11.

Table 5.11: Estimates of the KL-Distance between posterior densities $\theta|y_i$ and $\theta|y_j$

Studies	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
McCormick 1970	1	0	37	10	1	3	51	35	49	84	49	24	14	4	4	1	41	48	0	1	6	21	14
Cohen 1955	2	10	0	1	5	18	3	40	11	84	1	6	1	10	16	8	1	10	8	14	16	28	29
Chason 1958	3	15	15	0	6	46	48	74	44	83	47	19	0	47	39	16	24	46	13	32	36	72	61
Stehbens 1963	4	1	15	2	0	7	47	31	44	83	43	10	3	47	5	3	25	47	1	4	7	18	18
Sugai 1994	5	2	33	10	4	0	47	6	46	80	47	15	13	47	0	6	37	47	2	1	3	2	2
Romy 1973	6	83	47	82	82	83	0	83	21	83	4	62	82	83	82	81	30	34	83	82	81	81	83
Ruggieri 1994	7	1	9	3	2	1	18	0	27	22	15	7	4	29	1	3	12	34	1	1	2	0	0
du Boulay 1965	8	6	1	2	4	9	0	18	0	72	1	4	2	0	8	6	1	0	6	7	8	12	13
Wakabayashi 1983	9	8	15	10	8	5	22	4	30	0	20	11	11	30	6	8	17	35	7	6	6	3	5
Inagawa 1990	10	83	29	81	82	82	4	83	26	82	0	54	76	45	82	77	18	45	82	82	80	81	83
Lublanc 1995	11	0	0	0	0	1	1	2	1	16	0	0	0	1	0	0	0	1	0	0	0	1	1
ujije 1993	12	9	7	0	4	28	38	66	39	82	20	14	0	46	23	12	15	46	8	23	24	54	52
McCormick 1965	13	82	77	82	82	19	83	17	82	42	71	82	0	82	82	81	58	1	82	82	81	81	82
Nakagawa 1994	14	1	15	3	1	0	46	5	43	76	32	11	5	46	0	3	24	45	1	0	2	2	2
Wakai 1979	15	0	4	1	1	0	9	1	16	29	7	3	1	17	0	1	6	17	0	0	0	0	0
Chapman 1992	16	0	1	0	0	3	2	7	35	2	1	0	7	0	1	0	2	7	0	0	0	1	1
Atkinson 1989	17	4	0	1	3	9	0	21	2	79	0	4	1	1	8	12	5	0	4	7	8	15	15
Housepain 1958	18	82	71	81	82	15	83	13	82	32	70	82	0	82	81	81	54	0	82	81	81	81	82
Nagashima 1993	19	0	42	12	2	10	44	52	45	82	45	22	19	45	7	22	11	44	0	7	14	34	36
Jakubowski 1978	20	0	5	1	0	0	13	3	24	56	10	4	1	24	0	1	8	27	0	0	0	1	1
Iwata 1991	21	0	1	0	0	3	1	8	24	2	1	1	7	0	0	0	2	8	0	0	0	0	1
Ronkainen 1995	22	4	37	16	7	1	47	1	67	46	19	18	47	2	2	9	40	46	4	3	5	0	1
Griffiths 1996	23	1	7	2	1	0	15	0	23	12	5	3	24	0	0	2	9	26	1	1	1	0	0

From Table 5.11, it is obvious that both Housepain 1958 and McCormick 1965 seem to be inconsistent with many other studies because they consistently have higher KL-values with other studies. Although discarding a study on the account of heterogeneity is not a good practice in meta-analysis, however, the joint probability approaches zero with an inconsistent observation (y_i). The AgenaRisk software adopts a lower limit of ($\leq 10^{-38}$) for the joint probability and so any observation that results in a joint probability below this threshold is flagged as an inconsistent evidence.

5.7.2 Summary Estimate of the prevalence of Aneurysms: Fixed effect assumption

The summary estimate in this analysis is the prevalence of aneurysm in the population using a hierarchical model that follows.

$$\begin{aligned} y_{Ti} &\sim \text{Bin}(n_{Ti}, \theta) \\ \theta &\sim \text{Beta}(\alpha, \beta) \\ \alpha &\sim U(1, 9999999) \\ \beta &\sim U(1, 9999999) \end{aligned}$$

A BN representation of this model is presented in Figure 5.13 with observation y_{Ti} represented as evidence nodes.

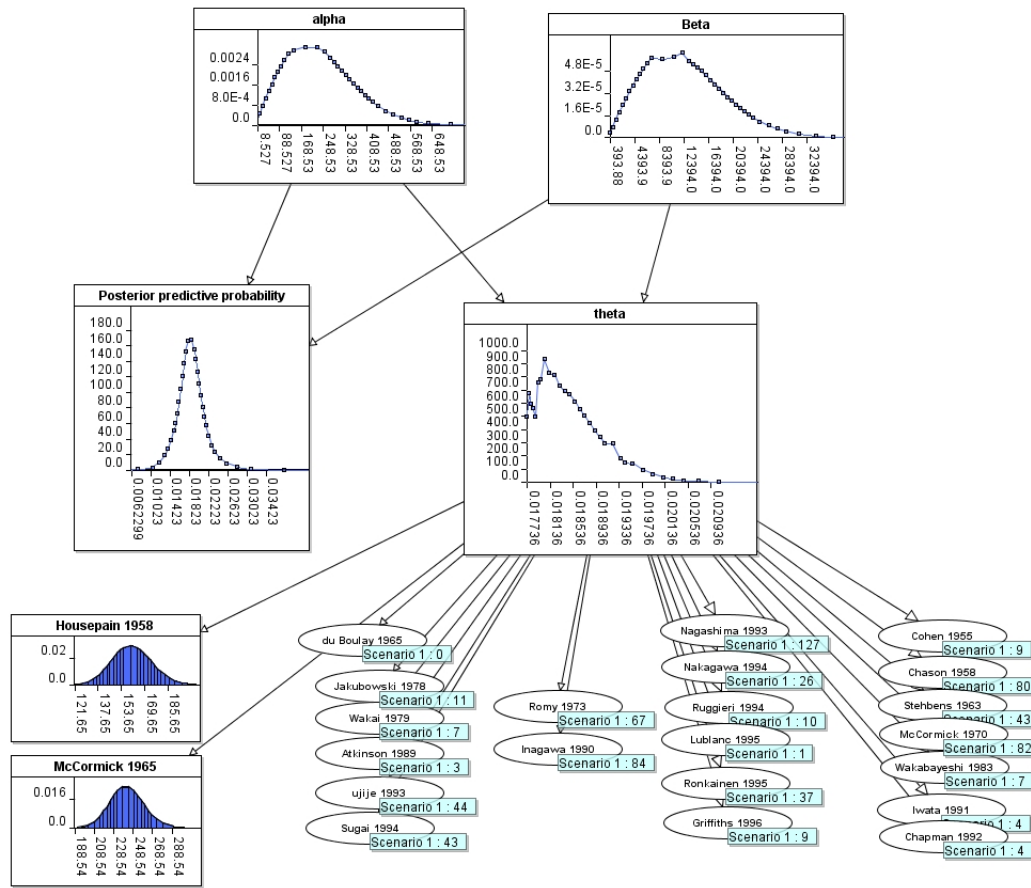


Figure 5.13: Fixed effect model for prevalence of intra-cranial aneurysms

The computation mechanism in AgenaRisk flags the observations from Housepain 1958 and McCormick 1965 as inconsistent observations, therefore, these two observations were not entered as evidence on their corresponding nodes.

Table 5.12 shows the summary statistics for parameters θ and a future study $y_{n+1}|y_1, \dots, y_n$.

Table 5.12: Summary result of prevalence of intracranial aneurysm based on fixed effect assumption (excluding evidence from Housepain 1958 and McCormick 1965)

Parameter	mean	sd	5%	95%
α	127	46	42	192
β	6665	2356	2231	8659
θ	0.019	5.90E-4	0.018	0.020
postpred ($y_{n+1} y_1 \dots y_n, \theta$)	0.018798	0.005485	0.01439	0.02348

Since we cannot discard data from a study on the basis of extreme heterogeneity and the current implementation of the dynamic discretization algorithm in AgenaRisk could not sample a posterior distribution when the joint probability is lower than 10^{-38} , therefore, we repeat the analysis using the MCMC approximation technique. The MCMC

Table 5.13: Summary result of prevalence of intracranial aneurysm based on fixed effect assumption

Parameter	mean	sd	5%	95%
θ	0.0202	7.56E-4	0.019	0.021
postpred	0.0204	0.002864	0.016	0.025

technique is able to accommodate data from all the studies including empirical data from Housepain 1958 and McCormick 1965 that were discarded in the previous analysis. In essence, there is a need for future works on the algorithm so that the dynamic discretization can be robust enough to handle cases when the joint probability of the parameter and the data is less than 10^{-38} . The posterior density of the parameters and their corresponding trace plot after 100,000 iteration with 50,000 burn-in sample are presented in Figure 5.14.

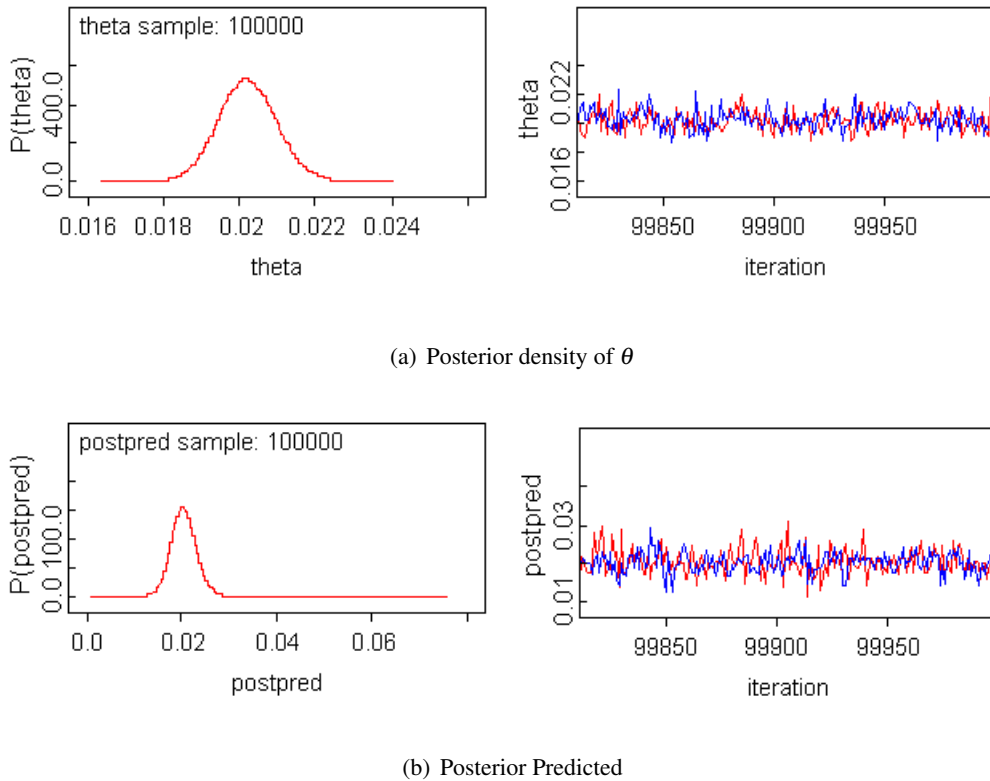


Figure 5.14: Fixed Effect Model parameter from Winbugs

Table 5.13 shows the summary statistics of the *summary effect* based on the fixed effect assumption.

5.7.3 Summary Estimate of the prevalence of Aneurysms: Random effect assumption

The basic sampling model is assumed to have the following form:

$$y_{Ti} \sim \text{Bin}(n_{Ti}, \theta_i)$$

Two different models corresponding to different assumptions about the distribution of θ_i are considered. In the first analysis, we assume that $\theta_i \sim \text{Normal}(\mu, \tau)[0, 1]$. This is a truncated normal distribution whose value is defined in the range $[0, 1]$.

The second model treats θ_i as a random variable from beta distribution with parameters α and β i.e. $\theta_i \sim \text{Beta}(\alpha, \beta)$. In both cases, the posterior distribution of the hyper parameter can be used for prediction. Therefore, we can predict a hypothetical future study y_{n+1} labeled “PostPred”. The posterior distributions of the “Posterior Prediction” based on these two models are overlaid in Figure 5.15.

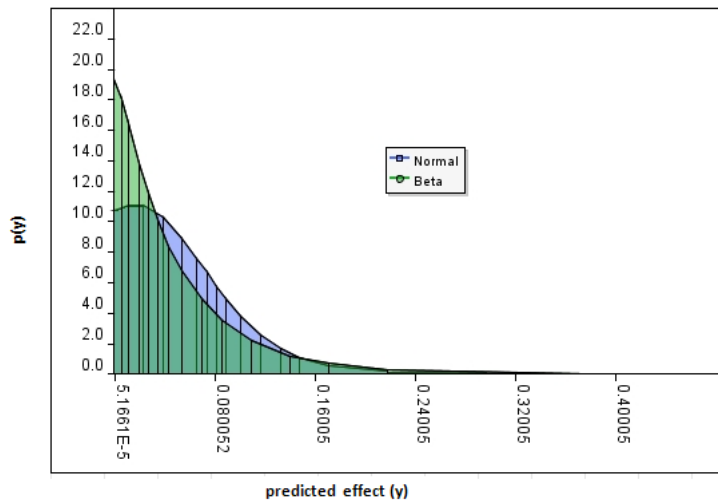


Figure 5.15: Posterior predictive effect from the two models a) θ_i is assumed to be normally distributed and b) θ_i is assumed to have beta distribution

Table 5.14 presents the estimate of the posterior distribution of “PostPred” obtained from these two models.

Table 5.14: Posterior prediction of prevalence

Model	Mean	Median	Variance	5%	95%
Normal	0.055047	0.046552	0.0017626	0.0046765	0.13444
Beta	0.049360	0.033605	0.0026786	0.0025509	0.14775

In this analysis, using two different prior distributions for the hyper-parameters does not drastically change posterior predictive results. But this may not always be the case as

the choice of the prior can affect the posterior result. A good practice is to conduct a sensitivity analysis of the posterior results to the prior assumptions.

In the analysis with the random effect assumption, the estimates of study specific parameters shrink towards the estimate of the “combined effect”. This is because hierarchical BN models allow parameters to borrow strength from each other [219]. Therefore, the random effect analysis is able to incorporate the observations from Housepain 1958 and McCormick 1965 which were rejected in the fixed effect model. Table 5.15 shows the posterior summary statistics of study specific parameters.

Table 5.15: Posterior Summary of study specific prevalence

Studies	Mean	Median	5%	95%
Housepain-1958	0.0017	0.0017	0.0011	0.0025
McCormick-1965	0.0021	0.0020	0.0015	0.0028
Romy-1973	0.0058	0.0058	0.0047	0.0070
Inagawa-1990	0.0083	0.0083	0.0069	0.0098
Cohen-1955	0.0172	0.0167	0.0094	0.0269
Chason-1958	0.0296	0.0295	0.0244	0.0352
Stehbens-1963	0.0430	0.0427	0.0331	0.0539
McCormick-1970	0.0509	0.0507	0.0422	0.0601
du Boulay-1965	0.0074	0.0056	0.0005	0.0207
Jakubowski-1978	0.0604	0.0592	0.0360	0.0894
Wakai-1979	0.0704	0.0683	0.0373	0.1106
Atkinson-1989	0.0131	0.0119	0.0034	0.0268
Ujile-1993	0.0278	0.0276	0.0214	0.0349
Sugai-1994	0.0706	0.0702	0.0546	0.0881
Wakabayeshi-1983	0.1516	0.1459	0.0826	0.2398
Iwata-1991	0.0572	0.0545	0.0244	0.0989
Chapman-1992	0.0478	0.0454	0.0201	0.0835
Nagashima-1993	0.0502	0.0501	0.0432	0.0575
Nakagawa-1994	0.0648	0.0642	0.0465	0.0854
Ruggieri-1994	0.0937	0.0917	0.0557	0.1386
Leblanc-1995	0.0381	0.0338	0.0079	0.0832
Ronkainen-1995	0.0906	0.0900	0.0690	0.1143
Griffiths-1996	0.0823	0.0803	0.0473	0.1239

Let us now visualize these posterior means by plotting the mean and the credible intervals shown in Figure 5.16.

Figure 5.16 shows that the credible interval for the *treatment effects* based on *normal* and *beta* prior distributions cover the credible interval for the study specific *treatment effects*. Some studies have wider credible interval than others because they involve relatively smaller number of patients. In particular, the credible interval for the study, Wakabayeshi-1983 is wider than the rest and is not fully enclosed within the 95% credi-

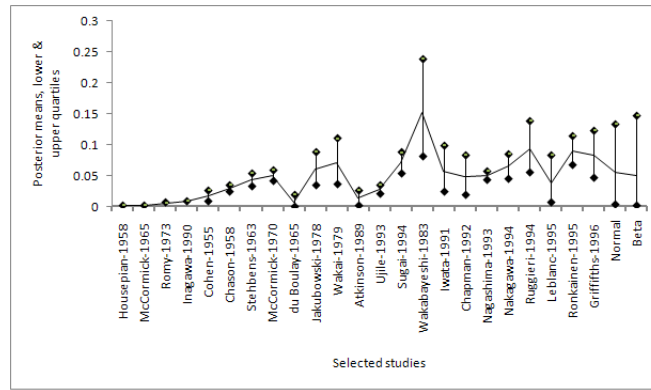


Figure 5.16: Posterior means and lower and upper quartiles.

ble interval of the summary effects.

5.7.4 Fixed parameters for retrospective and prospective studies

Rather than assuming different parameters for all studies, we can assume a fixed parameter for all studies with identical experimental design. In essence, the studies conducted retrospectively have a fixed parameter say θ_1 while those conducted prospectively have a different fixed parameter (θ_2). Then, θ_1 and θ_2 are exchangeable. The study designs are indexed by $i = 1 \dots 2$ retrospective and prospective studies. Let the number of patients x_{ij} with aneurysm in the j^{th} study and i^{th} design follow a binomial distribution with parameter N_i and θ_i . Formally, $x_{ij} \sim Bin(N_{ij}, \theta_i)$ where $i = 1, 2$ and $j = 1 \dots n_i$, n_i being the number of studies in the i^{th} design. The random part of this model treats the parameters θ_i as random parameters from a beta population with parameters α and β such that $\theta_i \sim Beta(\alpha, \beta)$. The hyper parameters α and β are then assigned a non informative prior with uniform distribution $\alpha, \beta \sim U(1, 9999)$. The resulting HRBN model obtained in AgenaRisk with 40 iterations is presented figure 5.17

5.7.5 Fixed parameters for autopsy and angiography

By slightly modifying the model presented earlier, we can now assume new parameters θ'_i and θ''_i representing respectively the prevalence of aneurysm in autopsy and angiography studies. The sampling distribution for observations x_{ij} is also binomial and we can use the uniform prior distribution on the hyper parameter. We again run this model in AgenaRisk with 40 iterations and the resulting BNs is presented figure 5.18.

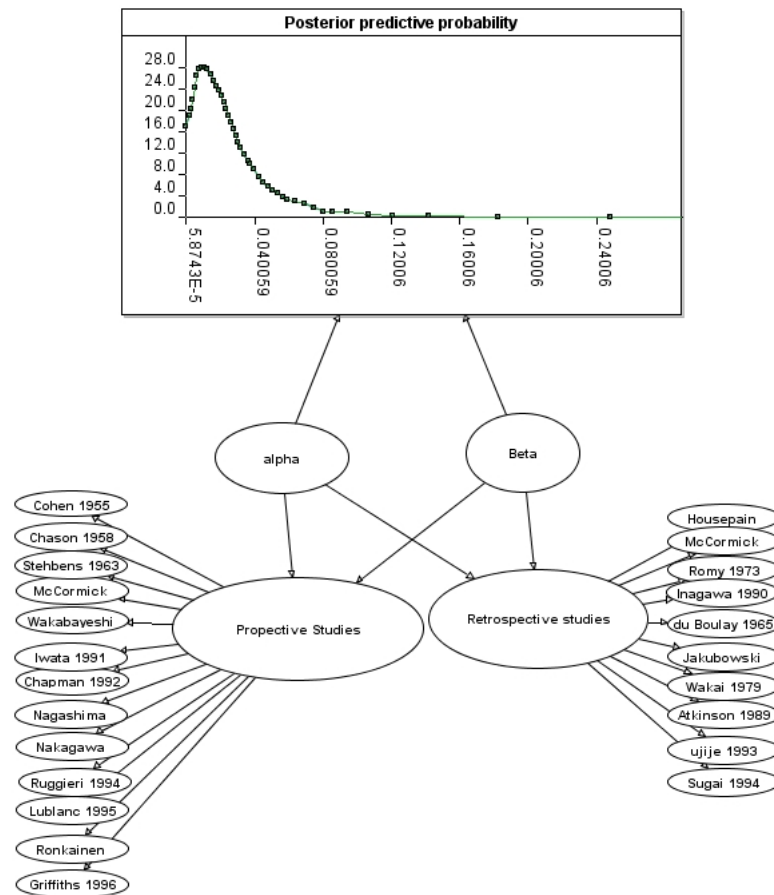


Figure 5.17: Hierarchical BN model for summarizing the prevalence of aneurysms by study designs.

Table 5.16 summarizes the result obtained from the two BN models (Figure 5.18 and 5.18).

Table 5.16: Lower and Upper posterior percentiles of prevalence of aneurysm per 1000 patients in each study design

Parameters	5 th – 95 th	Mean	Variance
θ_1	3 – 6	5	$1.1951E^{-6}$
θ_2	41 – 48	45	$5.3391E^{-6}$
θ'_1	4 – 6	5	$3.2107E^{-7}$
θ'_2	45 – 54	49	$7.9353E^{-6}$

The summary in Table 5.16 shows that studies with retrospective design produced lower prevalence than prospective studies while angiography technique yielded higher estimate than autopsy studies. The posterior predicted prevalence from the combined analysis of retrospective and prospective studies is 3 – 8/1000 while the combined analysis of autopsy and angiography studies is (2 – 8/1000). This result is consistent with

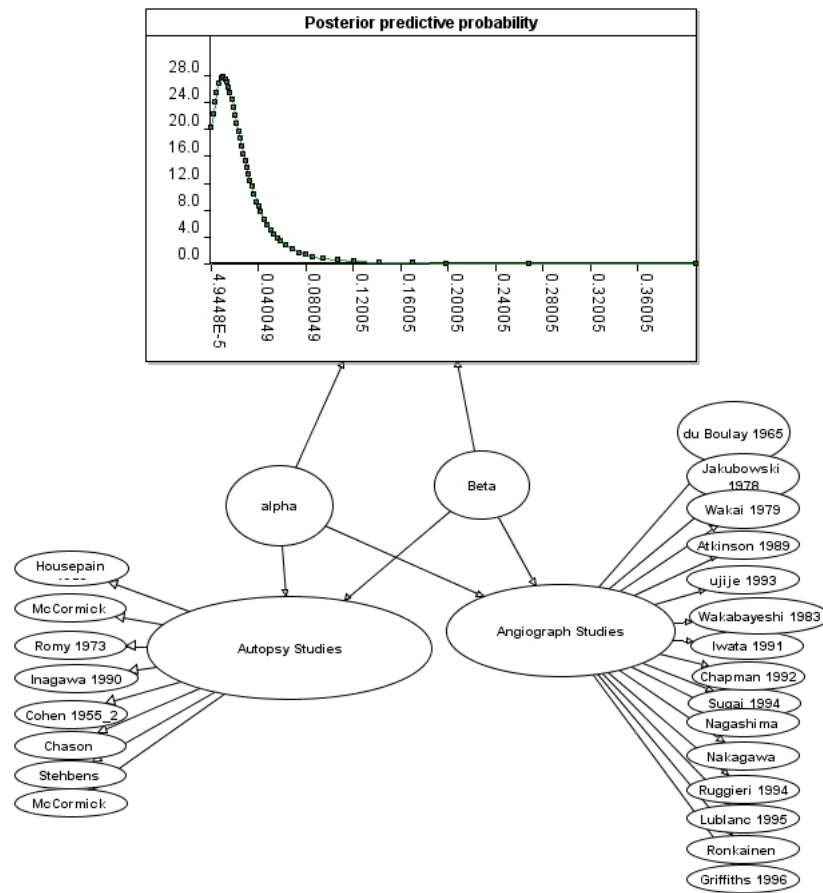


Figure 5.18: Hierarchical BN model for summarizing the prevalence of aneurysms by diagnostic techniques.

the work of Rinkel, *et al's*. [190] on the data who found the estimates for prevalence of aneurysm to vary widely between 2 and 90 per 1000 [190]. They attributed this wide variability to methodological differences between studies, diagnostic tools used, differences in the study populations and discrepancies in the inclusion criteria.

Clearly, the summary estimates of the prevalence based on fixed effect assumptions does not represent all the underlying studies. We found the prevalence of aneurysm to vary between 19 and 21 per 1000 in the fixed effect analysis. The prediction based on this model gave a prevalence of between 16 and 25 intracranial aneurysm per 1000 in the population. Neither of these can be used to represent the empirical information in the underlying studies. In contrast, the estimates from the analysis based on the random effect assumption better represent the underlying data. the result with the analysis with a Normal prior distribution on the hyper parameter gave an estimate of the prevalence to be 4 and 130 per 1000 whereas using a beta distribution to formulate a non informative prior

on the hyper parameter gave a 95% credible interval that correspond to 3 - 148 cases per 1000.

From the modeling point of view, these summary estimates can be used as existing knowledge in a clinical decision model.

5.8 Summary

This chapter described a novel approach for solving meta analysis problems using the Bayesian Network framework augmented by dynamic discretization. Three Meta analysis problems were analyzed in the chapter. The first is the data from a meta-analysis to summarize the treatment effect of Beta-blocker in patient with heart failure condition. The summary effect shows a significant *treatment effect (odd ratio)* indicating a significant reduction in the mortality rate for those who received the treatment compared to those who received a placebo. The second clinical example is the meta-analysis to summarize the effect of magnesium on patients with myocardia infarction. This analysis demonstrated a desirable feature of the BN framework by constraining the posterior distribution of the heterogeneity parameter τ and then performing inference conditional on the constraint. Thus, analysts can make a probabilistic statement about the degree of variability between studies in the analysis. The third model is on the prevalence of intracranial aneurysms. This analysis demonstrated the use of a BN model to summarize data from a non-experimental design. We also used the KL-metric to explore data for inconsistent observations.

In summary this chapter demonstrated the modeling strength of the Hierarchical Bayesian Network and its use in summarizing treatment effects from meta analysis data. Nonetheless, meta-analyses are fundamentally limited by the quality of the underlying studies. The quality of the result depends on the quality of the systematic review that produced the data. Hence, the context within which we have discussed the results from the chapter is to demonstrate the potential use of BN model in solving meta-analysis problems. We envisage a future where both meta-analysis and decision analysis model can be developed within the BN framework.

We can not conclude this chapter without drawing readers attention to the limitations of this approach. As mentioned earlier in the chapter, using the current implementation of

the dynamic discretization algorithm flags an inconsistency error when the posterior joint distribution of the parameter and the data is lower than a threshold. To make this approach suitable for a wider class of meta-analysis problems, we recommend changes in the implementation of the dynamic discretization algorithm in AgenaRisk so that the probability threshold can accommodate empirical data from extremely heterogeneous studies. Also, we like to state that the three clinical meta-analysis problems used in the chapter are arbitrarily chosen from the literature. Other types of clinical effects such as *risk ratio*, *risk difference*, *correlation*, etc have not been summarized with the BN technique. We have described a prototype for solving meta-analysis problems with a BN model but there are still more to be done before we can conclude that the BN approach is generally applicable to solving meta-analysis problems. Hence we recommend more work in summarizing other clinical effects with BN models.

The approach adopted here was pragmatic and focused on choosing those papers that used popular methods of analysis (hence using MCMC as a comparator) where data was available in the research paper itself and also where sufficient detail was provided to enable the model to be reconstructed and then configured and tested. Doing this for the three studies presented here took considerable time and effort, but is not the result of a scientific survey. We do not therefore claim that the analysis neither is free from selection bias nor is comprehensive, however it does show how one can use HBNS and DD for meta-analysis in a flexible way and offers some support to the research hypotheses pursued within this thesis.

Chapter 6

BN Decision Making Models

In the previous chapters we have described the application BN models to parameter learning and meta-analysis problems. This chapter presents two clinical case studies and demonstrates the use of the BN framework to 1) evaluate the impact of a multidisciplinary team meeting on treatment selection for patients with cancer and 2) compare the risk of Magnetic Resonance Imagine and Catheter Angiogram in diagnosing patients with a third nerve palsy.

6.1 Introduction

This chapter applies the techniques described in the preceding chapters to two clinical problems. The main motivation is to use the technique to generate some answers to immediate clinical questions begging for answers. On each of these case studies, our research team has worked in collaboration with clinician with a view of providing answers to some clinical questions.

The first is based on data collected by a multi-disciplinary team (MDT), HepatoPancreaticoBiliary (HPB), based at Barts and the London HPB Centre. The Multidisciplinary Team (MDT) approach has become increasingly important in the decision-making regarding the treatment plan and management of cancer patients [144]. In spite of the increasing popularity of the multidisciplinary approach, there are only few studies on the empirical evaluation of its impact on the treatments and outcomes thereof. Forrest *et al.* [68] exam-

ined the impact of the introduction of a multidisciplinary team on survival of patients with inoperable Non-small-cell lung cancer (NSCLC) and concluded that the introduction of a multidisciplinary team was associated with a change in the treatment of patients with inoperable NSCLC. In particular, they found that more patients now receive chemotherapy and fewer patients received palliative care only.

Whilst multi-disciplinary team meetings (MDTs) are seen as attractive additions to the management process by various stake-holders, there is little evidence or understanding of their utility in rationalising decision-making in complex medical cases. Similarly, MDTs confer a significant resource cost (clinician time, administrative support, governance arrangements) yet their contribution to patient care is poorly understood. The research question worthy of study is to better understand the changes in treatment recommendation for each condition that have arisen as the MDT process has matured (expressed as a function of time since inception). We use the Bayesian Network framework to address this problem.

This particular area is a particularly test of the methodology because it is “messy” in the sense that there are lots of uncertain dependencies, there is a mixture of probabilistic and deterministic reasoning underlying the clinical decision making process and there is a mixture of continuous and discrete data demanding both inference about uncertain causal mechanisms and prediction of effects. Tackling this sort of problem is very difficult for other classical methods because it requires an understanding of the underlying clinical care process and the complexities and constraints therein. This does not mean to say that the analysis performed here is definitive since there are some features of the problem that make MDTs particularly tricky, including the role of counterfactual evidence, confounding variables and, ultimately, data quality. For instance the complex healthcare supply chain means that the causal factors that drive some variables might be outside of the clinician’s control, such as GP referrals. Likewise, treatments are recommended on the basis of what is best for the patient and so any evaluation needs to account for the counterfactual nature of this intervention. Finally continuing changes and improvements to the process, both in terms of operational and treatment efficiency, means the data reflects a moving target. All of these make application of frequentist simulation based modelling near impossible and are challenges for Bayesian analysis, where even if a small contribution can

be made this should be judged a success.

The second case study is based on the work of Fenton *et al.*, [62]. In their work, Fenton *et al.*, [62] used Bayesian arguments to compare the risks of a catheter angiography with the risk of rupture resulting from a false negative result of a magnetic resonance imaging technique. Their objective was to establish if the marginal positive predictive power of the CA technique compensates the attendant risk of complication. This model was presented as experts' witness in the case involving a diabetic woman who suffered a permanent stroke after receiving an invasive diagnostic test.

Section 6.2 applies the multinomial BN formulation to learn some parameters in the model to evaluate the impact of multi-disciplinary team meetings on treatment selection for patients suspected to have cancer. In section 6.3, we apply Bayesian network models to compare the risk of stroke from the CA procedure with the risk of death from undetected aneurysms.

6.2 Evaluating the impact of multidisciplinary team meetings on treatment selection

Multidisciplinary Team (MDT) meetings have become increasingly popular and have become standard decision-making forum for oncology [144]. The idea is to combine expertise from each field in order to generate a comprehensive and coordinated care plan for cancer patients. Previous research has demonstrated that patients, whose care is managed by such meetings have better survival outcomes [15], shorter waiting times [70] and the benefit of more robust treatment decision-making processes [28] compared to those managed without formal multidisciplinary discussions.

Patients are referred by their general practitioners (GPs) to one of the twelve district general hospitals on suspicion of cancer. Further tests are performed at the district hospitals and cases that require further investigation are referred to the Royal London Hospital for MDT discussion and treatment. On the basis of the diagnosis and findings, appropriate treatment pathways are recommended to patients which may include one or more of chemotherapy, palliative care and surgery among others.

6.2.1 Description of Data

The data was collected by the HPB team based in Barts and the London HPB Centre between 2005 and 2009. The team meets weekly to consider possible cancer cases, these include cancer of the liver, pancreas, bile duct and gall bladder. The HPB team consists of surgeons and physicians, oncologists, pathologists, palliative care specialists and radiologists.

In total, 1988 complete records of ‘possible cancerous cases were discussed between 2005 and 2009 by the team. Essential details such as the affected organ, type of a lesion, diagnosis and treatment recommendation were recorded during the period. Figure 6.1 shows a BN fragment of the model presented in Kocher *et al.* [119].

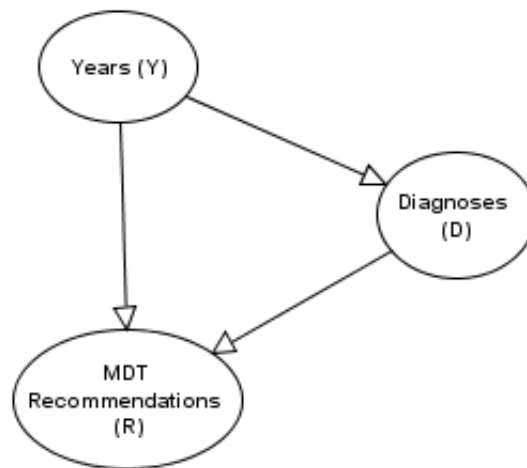


Figure 6.1: BN modeling of treatment recommendations at a multi-disciplinary setting

This fragment consist of three variables, these are: “Year of discussion (Y)”; “Diagnosis (D)” and “MDT treatment recommendations (R)”. We will now take them in turns.

Year of discussion (Year)

This variable is the year a case was discussed by the Multi-Disciplinary Team. This is denoted by a node Y with five possible states $\{2005, 2006, 2007, 2008 \text{ and } 2009\}$. This variable is intended to capture changes that might have occurred over the years. We can think of this variable as a surrogate for (unknown) possible confounders such as the technological advances over the years that might have influenced diagnosis or even treatment recommendations.

Diagnosis

This variable stores all possible diagnoses. Initially, there were 49 possible diagnoses in the data but after due consultations with an expert from the domain, the diagnosis was grouped into eight non overlapping cases as shown in Table 6.1.

Table 6.1: Descriptions of possible diagnosis options

State	Description
BP	For all benign diagnosis of the Pancreas
BL	For all benign diagnosis of the Liver
BGB	For all benign diagnosis of either the Gall bladder or Bile duct (not both)
M/BP	Malignant or borderline diagnosis of the Pancreas
M/BL	Malignant or borderline diagnosis of the Liver
M/BGB	Malignant or borderline diagnosis of Gall bladder or Bile duct (not both)
Unknown	For unknown diagnosis
Multiple	For multiple diagnoses

The “Multiple” diagnosis represents cases with more than one diagnoses. The diagnosis variable is denoted by D . This node has one parent node “Year”. The link from node “Year” to “Diagnosis” was introduced to capture trends in the diagnosis over the five years period.

MDT recommendations

The variable “MDT recommendation” represents treatment options recommended for patients. The options are not mutually exclusive as a patient can have one or more treatment recommendations. It is also possible that a patient is placed on close surveillance or discharged without receiving a treatment. In order to attain mutual exclusivity for this variable, we created a new state labeled “Combination” for cases with more than one treatment option. All cases with more than one treatment strategies are classified as combination treatment. Therefore, the variable “MDT recommendations” has seven states as shown in Table 6.2.

This node has two incoming links from nodes “Year” and “Diagnosis”. Just like the link between “Year” and “Diagnosis”, the link connecting node “Year” to “MDT recommendation” was introduced to capture trends in treatment recommendation over the period. These links do not have causal interpretations. The second link connecting “Diagnosis” to “MDT recommendation” implies that treatment option is determined on the basis of the diagnosis. Indeed some treatment options are forbidden for some diagnoses.

Table 6.2: Descriptions of possible treatment options

State	Description
Chemotherapy	Treated with Chemotherapy only
Combination	Treated with one or more options, e.g surgery followed by chemotherapy
None	No Treatment Recommendation.
Palliative	palliative care recommendation only
Surgery	Surgery only
Intervention Radiology	intervention radiology only
Watchful Waiting	Patients on surveillance

Other factors such as age of the patient and stage of the lesion might have also influenced treatment recommendation but for simplicity, we assume that the information about the diagnosis suffice for treatment recommendation.

For the categories of cancer treated by this multi-disciplinary team, the radiotherapy treatment option is never considered in isolation. It is always performed in addition to other treatments. Therefore the radiotherapy option is not included as a state for the variable.

6.2.2 Summary of the MDT Data

Table 6.3 summarizes the number of cases recorded for each diagnosis. Multiple diagnoses is not very common as only 76 cases were recorded. On the other hand, there are 588 recorded cases of malignant or borderline diagnosis of the Liver.

Table 6.3: Number of cases per Diagnosis

Diagnosis	Number of cases
BP	211
BL	176
BGB	90
M/BP	457
M/BL	588
M/BGB	207
Multiple	76
Unknown	183

Table 6.4 presents the diagnosis data for each year of discussion.

Table 6.5 shows the frequency distribution of treatment recommendations by the multi-disciplinary team. The summary shows that Intervention Radiology is the least recommended option with 94 cases whereas surgery seems to be more frequently recommended than any other treatments.

Table 6.4: Number of cases per Diagnosis by Year of discussion

Diagnosis	Year				
	2005	2006	2007	2008	2009
BP	31	23	34	53	70
BL	32	16	19	39	70
BGB	17	11	15	20	27
M/BP	87	37	95	96	142
M/BL	101	58	136	108	185
M/BGB	36	17	43	49	62
Multiple	14	4	18	38	2
Unknown	44	13	31	56	39

Table 6.5: Treatment recommendation by diagnosis

Treatment	Number of cases
Chemotherapy	219
Combination	194
None	534
Palliative	358
Surgery	403
Intervention Radiology	94
Watchful Waiting	186

Some of these treatments are logically impossible for some diagnoses as a result of health-care policies. For instance, chemotherapy is never administered on benign diagnosis of the pancreas or liver. Table 6.6 shows treatment recommendations by diagnoses. All the recommendation with zero entries in Table 6.6 are logically impossible data.

Table 6.6: Treatment recommendation by diagnosis

Treatment	Diagnosis							
	BP	BL	BGB	M/BP	M/BL	M/BGB	Multiple	Unknown
Chemotherapy	0	0	0	74	111	33	0	1
Combination	0	0	0	71	82	39	0	2
None	101	114	36	56	68	14	16	129
Palliative	0	0	0	124	160	62	2	10
Surgery	27	17	38	112	107	54	37	11
Intervention Radiology	19	10	0	3	55	5	0	2
Watchful Waiting	64	35	16	17	5	0	21	28

We can explore this treatment further by summarizing the frequency of the treatment by diagnoses for each year (Table 6.7).

Table 6.7: Treatment recommendations by diagnoses for the five years period

Year	Treatment	Diagnosis							
		BP	BL	BGB	M/BP	M/BL	M/BGB	Multiple	Unknown
2005	Chemotherapy	0	0	0	8	10	3	0	0
	Combination	0	0	0	35	31	10	0	1
	None	15	21	6	7	12	7	3	30
	Palliative	0	0	0	25	35	13	0	3
	Surgery	4	6	9	9	9	3	7	5
	Int. Radiology	3	2	0	2	4	0	0	0
	Watchful Waiting	9	3	2	1	0	0	4	5
2006	Chemotherapy	0	0	0	0	5	0	0	0
	Combination	0	0	0	6	17	7	0	0
	None	10	12	1	3	7	0	0	10
	Palliative	0	0	0	19	21	7	0	1
	Surgery	4	0	9	8	6	3	3	0
	Int. Radiology	1	1	0	0	2	0	0	0
	Watchful Waiting	8	3	1	1	0	0	1	2
2007	Chemotherapy	0	0	0	8	20	4	0	1
	Combination	0	0	0	18	18	4	0	0
	None	10	9	7	16	19	2	6	21
	Palliative	0	0	0	29	44	21	1	1
	Surgery	5	3	4	20	28	11	8	3
	Int. Radiology	3	2	0	0	6	1	0	0
	Watchful Waiting	16	5	4	4	1	0	3	5
2008	Chemotherapy	0	0	0	7	15	2	0	0
	Combination	0	0	0	11	13	16	0	1
	None	26	22	11	15	18	2	6	40
	Palliative	0	0	0	28	37	14	1	3
	Surgery	4	6	3	32	15	14	19	3
	Int. Radiology	2	4	0	0	8	1	0	1
	Watchful Waiting	21	7	6	3	2	0	12	8
2009	Chemotherapy	0	0	0	51	61	24	0	0
	Combination	0	0	0	1	3	2	0	0
	None	40	50	11	15	12	3	1	28
	Palliative	0	0	0	23	23	7	0	2
	Surgery	10	2	13	43	49	23	0	0
	Int. Radiology	10	1	0	1	35	3	0	1
	Watchful Waiting	10	17	3	8	2	0	1	8

6.2.3 Estimating Conditional probabilities from data

The joint probability distribution for the BN fragment in Figure 6.1 is

$$p(Y, D, R) = p(R|Y, D) \cdot p(D|Y) \cdot p(Y).$$

Let us assume a noninformative prior for the “Year”. Then we need to estimate two conditional probabilities $p(D|Y)$ and $p(R|Y, D)$. For the conditional probability $p(D|Y)$,

we need to estimate eight parameters in Table 6.8 for each year using the data presented in Table 6.4.

Table 6.8: Parameters of the conditional probability $p(D|Y)$ for “Diagnoses”

Parameter	Description
p_1	$p(BP Y)$
p_2	$p(BL Y)$
p_3	$p(BGB Y)$
p_4	$p(M/BP Y)$
p_5	$p(M/BL Y)$
p_6	$p(M/BGB Y)$
p_7	$p(\text{Unknown} Y)$
p_8	$p(\text{Multiple} Y)$

We will use the approach described in section 3.8 to create a multinomial BN model for learning these eight parameters for the conditional Table $p(D|Y)$. Figure 6.2 shows the parameters (p_i) and observation nodes (x_i) in addition to the samples (n_i) and the nodes for summing parameters (v_i).

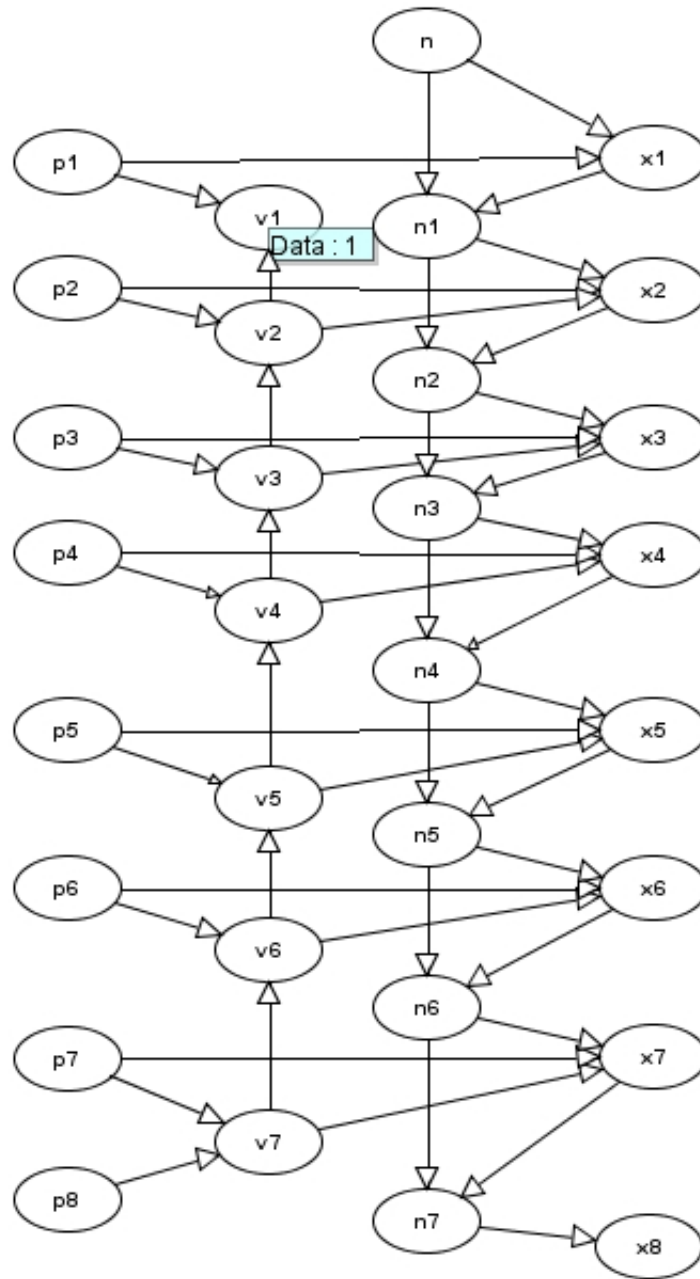


Figure 6.2: Multinomial BN model for learning parameters in Conditional table of “MDT Recommendation”

Suppose we do not have any prior information regarding these parameters then we can define noninformative prior as follows:

$$p_i \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1) \tag{6.1}$$

Also we need to introduce a constraint on the parameters such that $\sum_i^8 p_i = 1$. We do this by entering hard evidence “1” on the node labeled “v1” which represents the sum of

p_i nodes. After entering the data in Table 6.4 as hard evidence on x_i we obtain posterior means $E(p_i|x)$ summarized in Table 6.9.

Table 6.9: Posterior mean for each parameters p_i in conditional Table $p(D|Y)$

p_i 's	2005	2006	2007	2008	2009
p_1	0.0854	0.1266	0.0849	0.1133	0.1151
p_2	0.0881	0.0901	0.0493	0.0844	0.1154
p_3	0.0483	0.0638	0.0392	0.0444	0.0458
p_4	0.2359	0.2020	0.2364	0.2054	0.2337
p_5	0.2733	0.3142	0.3396	0.2329	0.3033
p_6	0.1019	0.0976	0.1140	0.1093	0.1072
p_7	0.0433	0.0287	0.0525	0.0865	0.0063
p_8	0.1238	0.0770	0.0841	0.1239	0.0732

We can now repeat the process to estimate the parameters described in Table 6.10 for the conditional $p(R|Y, D)$.

Table 6.10: Parameters of the conditional probability for ‘‘MDT recommendation’’, $p(R|Y, D)$.

Parameter	Description
p_1	$p(\text{Chemotherapy} Y, D)$
p_2	$p(\text{Combination} Y, D)$
p_3	$p(\text{None} Y, D)$
p_4	$p(\text{Palliative Care} Y, D)$
p_5	$p(\text{Surgery} Y, D)$
p_6	$p(\text{Intervention Radiology} Y, D)$
p_7	$p(\text{Watchful Waiting} Y, D)$

In this analysis the number of parameter to learn varies by diagnoses because some treatment options are not available for some diagnoses. The conventional techniques for learning parameters such as the deal package in R, [20] requires equal number of parameters in the node probability table and therefore cannot distinguish between zero record and logically impossible observations. Recall that the zeros in Table 6.7 does not imply ‘zero’ observations but rather they encode impossible treatment-diagnosis combinations. In other words, the conventional techniques will compute posterior distributions for these parameters even when they are impossible. In this analysis, we can treat the number parameters k_d by diagnoses as unequal. For example, there are four possible treatments available for Benign Pancreas (BP) implying four parameters. Table 6.11 shows the number of parameters to learn for each diagnosis.

Table 6.11: Number of Parameters in table $p(R|Y, D)$ by diagnosis.

Diagnosis	k_d
BP	4
BL	4
BGB	3
M/BP	7
M/BL	7
M/BGB	6
Multiple	4
Unknown	7

The number of parameters in Table 6.11 were used to create multinomial BN models dynamically and we enter observations from Table 6.7 as evidence. The posterior means for these parameters are summarized in Table 6.12.

Table 6.12: Posterior mean for each parameters p_i in conditional Table $p(R|Y, D)$

Year	p's	BP	BL	BGB	M/BP	M/BL	M/BGB	Multiple	Unknown
05	p_1	-	-	-	0.0946	0.09895	0.09435	-	0.0207
	p_2	-	-	-	0.3763	0.2852	0.2591	-	0.0340
	p_3	0.4532	0.6029	0.3469	0.0845	0.1182	0.1894	0.2212	0.5686
	p_4	-	-	-	0.2750	0.3230	0.3315	0.0558	0.0847
	p_5	0.1437	0.1965	0.4974	0.1088	0.0970	0.0997	0.4432	0.1290
	p_6	0.1162	0.0863	-	0.0354	0.0596	0.0260	-	0.0239
	p_7	0.2869	0.1143	0.1557	0.0254	0.0180	-	0.2798	0.1331
06	p_1	-	-	-	0.0225	0.0915	0.0433	-	0.0498
	p_2	-	-	-	0.1578	0.2740	0.3445	-	0.0508
	p_3	0.4033	0.6310	0.1427	0.0905	0.1226	0.0438	0.1241	0.5384
	p_4	-	-	-	-	0.33635	0.3461	0.1247	0.1017
	p_5	0.1861	0.0530	0.7081	0.2052	0.1091	0.1768	0.4966	0.0518
	p_6	0.0754	0.1075	-	0.0246	0.0492	0.0454	-	0.0527
	p_7	0.3352	0.2086	0.1492	0.0496	0.0172	-	0.2546	0.1547
07	p_1	-	-	-	0.0871	0.1451	0.1020	-	0.0517
	p_2	-	-	-	0.1837	0.1317	0.1018	-	0.0266
	p_3	0.2879	0.4302	0.4405	0.1645	0.1390	0.0614	0.3157	0.5702
	p_4	-	-	-	0.2917	0.3125	0.4446	0.0912	0.0536
	p_5	0.1583	0.1748	0.2789	0.2069	0.2024	0.2448	0.4079	0.1078
	p_6	0.1060	0.1324	-	0.0123	0.0522	0.0454	-	0.0277
	p_7	0.4478	0.2626	0.2806	0.0537	0.0171	-	0.1852	0.1624
08	p_1	-	-	-	0.0770	0.1368	0.0547	-	0.0160
	p_2	-	-	-	0.1154	0.1203	0.3068	-	0.0318
	p_3	0.4706	0.5291	0.5177	0.1541	0.1630	0.0548	0.1661	0.6361
	p_4	-	-	-	0.2797	0.3279	0.2724	0.0478	0.0656
	p_5	0.0885	0.1645	0.1752	0.3184	0.1408	0.2711	0.4750	0.0665
	p_6	0.0534	0.1186	-	0.0118	0.0817	0.0403	-	0.0342
	p_7	0.3875	0.1878	0.3071	0.0436	0.0295	-	0.3111	0.1497
09	p_1	-	-	-	0.3390	0.3189	0.3612	-	0.0216
	p_2	-	-	-	0.0134	0.0209	0.0441	-	0.0219
	p_3	0.5458	0.6835	0.3949	0.1064	0.0677	0.0589	0.3303	0.6159
	p_4	-	-	-	0.1617	0.1260	0.1181	0.1662	0.0676
	p_5	0.1510	0.0415	0.4656	0.2939	0.2605	0.3527	0.1667	0.0239
	p_6	0.1519	0.0286	-	0.0187	0.1873	0.0649	-	0.0473
	p_7	0.1512	0.2463	0.1395	0.0670	0.0185	-	0.3369	0.2017

We can now use these parameters in the NPT of the underlying nodes i.e. “Diagnoses” and “MDT recommendation”. Figure 6.3 shows the prior marginal distribution obtained after entering the values in Table 6.9 to populate the NPT for “Diagnoses” and Table 6.12 to populate NPT for “MDT recommendation”.

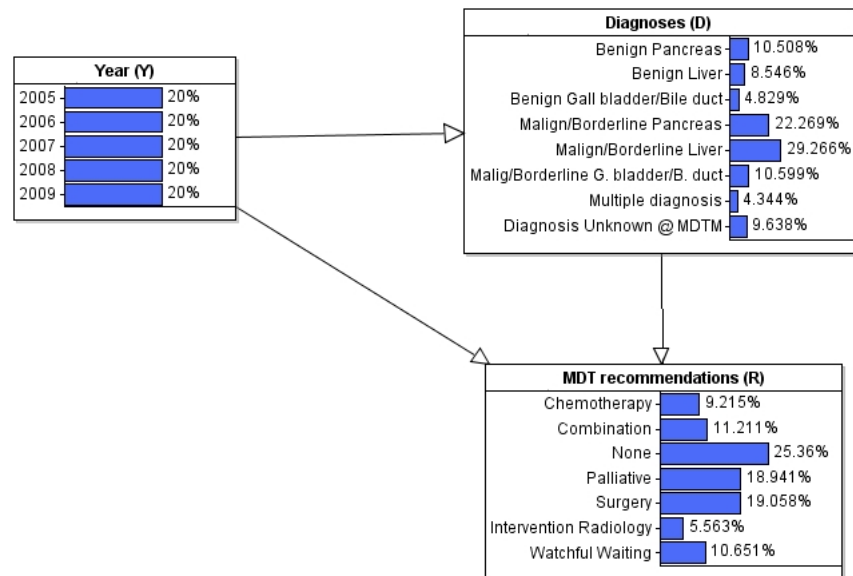


Figure 6.3: BN fragment with two categorical nodes: Parameters were learnt with multinomial BN models

6.2.4 Summary of results

This section uses the BN model in Figure 6.3 to answer some questions about possible trend in treatment recommendations since the commencement of the HPB multidisciplinary team. Figure 6.3 shows the posterior means for each of the treatment recommendations without a hard evidence on nodes “Year” and “Diagnoses”. We refer to this as a baseline result, to which we can compare the posterior means of treatment recommendation after a hard evidence had been entered. From the baseline analysis, about a quarter of patients received no treatment. This is the proportion of cases with no treatment recommendation either because they are non cancerous cases or cases that have reached an advanced stage such that administering a treatment is hopeless. As shown in Figure 6.4, over 20% of cases discussed each year were not given any treatment. Unfortunately, we do not have the information to distinguish non cancerous cases from those in the advanced stages. This information would be useful to access the quality of the referral process to the HPB multidisciplinary team.

From the year 2006 up to 2009, both palliative care option and combination therapy have consistently declined whereas surgery option seems to be gaining popularity within this multi-disciplinary team. It is not clear whether surgery treatment is always preferred

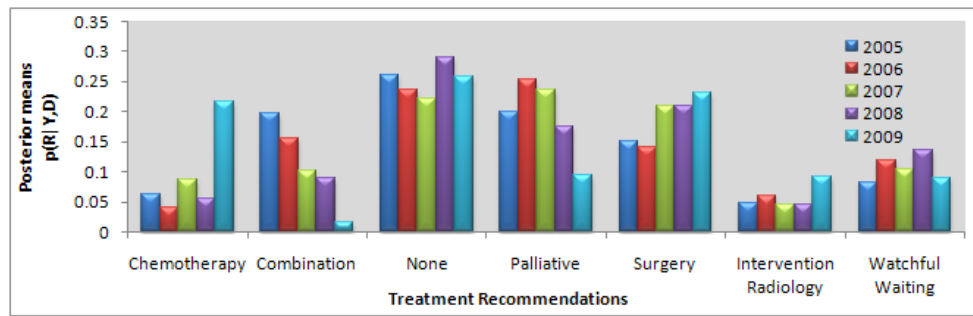
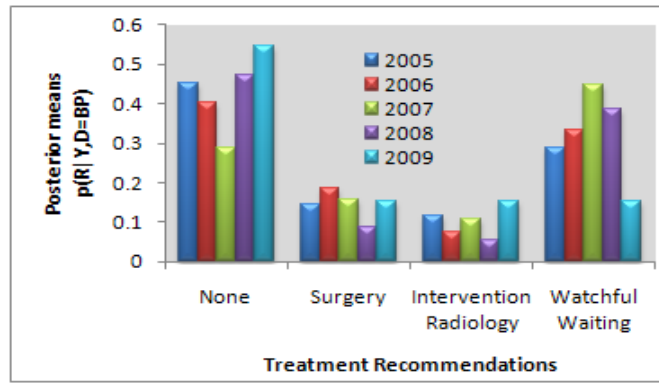
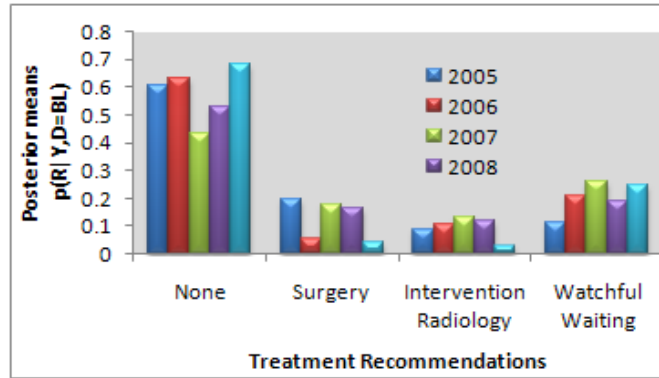


Figure 6.4: Treatment recommendations made by the HPB multi-disciplinary team from 2005 to 2009.

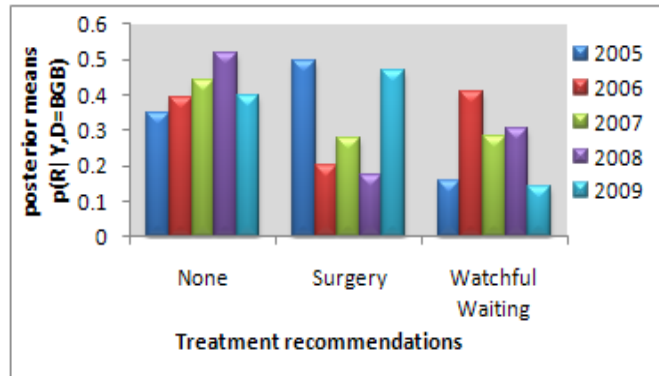
for any diagnosis that could have been offered an alternative treatment recommendation such as chemotherapy or a combination of treatments. One possible reason for this trend could be the technological advances, over this period, that reduced the level of uncertainty confronting the team. For instance, with a high resolution image from a Magnetic resonance Angiography, the team would better understand the location of lesions and make better judgements about their suitability for surgery option much more than before.



(a) Benign diagnosis of pancreas



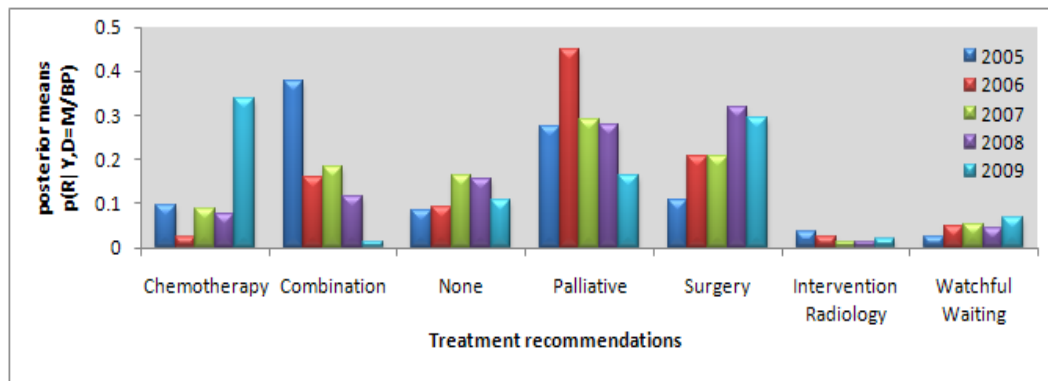
(b) Benign diagnosis of liver



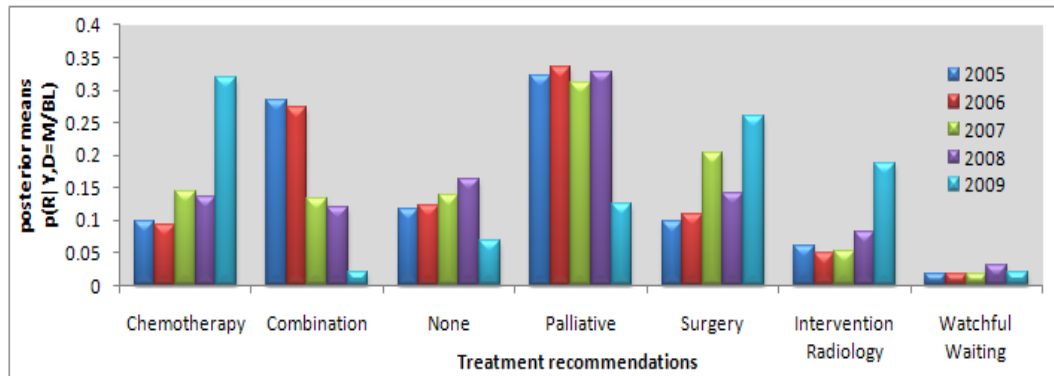
(c) Benign diagnosis of bile duct or gall bladder

Figure 6.5: A graph showing the posterior means of treatment recommendation for benign diagnoses.

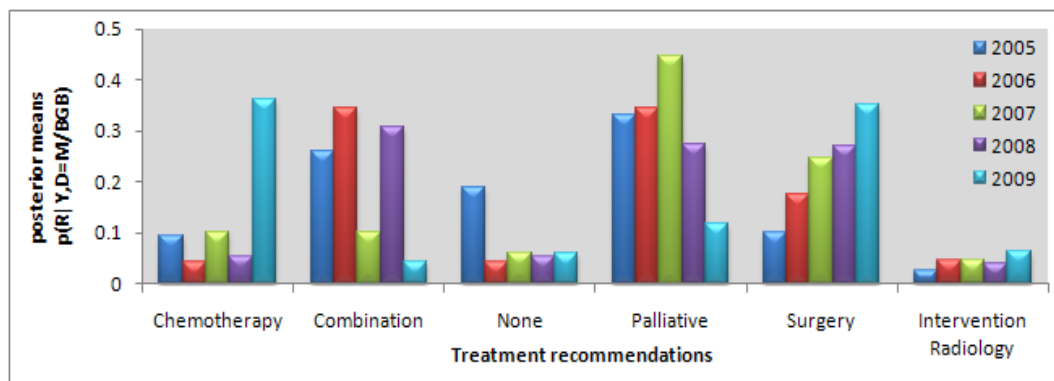
Clearly, the majority of the benign diagnoses are either discharged without a treatment or placed under close surveillance (watchful waiting). From Figure 6.6, there is an apparent dependence between organ and surgery recommendation for benign diagnoses. Surgery is more frequent for benign cases of cancer in the bile duct or gall bladder than those in liver or pancreas. This is really not surprising because the liver and pancreas are physiologically more important than bile duct or gall bladder.



(a) Malignant diagnosis of pancreas



(b) Malignant diagnosis of liver



(c) Malignant diagnosis of bile duct or gall bladder

Figure 6.6: A graph showing the posterior means of treatment recommendation for benign diagnoses.

Also worthy of note in Figure 6.4 and Figure 6.6 is a *surge* in the recommendation of chemotherapy option in 2009. Figure 6.4 shows that more than 20% received chemotherapy treatment in 2009 compared to less than 10% in the remaining years. This pattern is visible for cases with malignant diagnoses of the pancreas, liver, bile duct or gall bladder. There no empirical or clinical justification, in the literature, to support this sudden enthusiasm for chemotherapy option in 2009. Has chemotherapy suddenly become more

effective in 2009 than in 2008 for treating cancer? Therefore, further analysis to establish the cause of this sudden interest in chemotherapy option will be useful.

6.2.5 Direction of future work on the MDT analysis

Some questions arose from our findings presented in the previous section that require further investigations on the MDT. For example, we classified every case into one of the eight diagnoses (see Section 6.2.1) based on the organ and the severity of a condition. Therefore, the model fragment does not include nodes for the affected organ and the level of severity of the condition. Thus, we have assumed that the *Diagnosis node* provides information about organ and the level of severity. However this assumption may not be entirely plausible as Figure 6.6 shows that surgery is more frequent in benign cancer cases affecting bile duct or gallbladder than liver or pancreas. Future work to investigate the dependence of treatment recommendation on the affected organ would be a useful addition to the existing body of knowledge.

We also assume in our analysis that the parameters of treatment recommendation are independent. Again, this may not be entirely correct as some of these recommendations may be correlated. Some cases in the data were discussed more than once either due to lack of information or insufficient information to come up with a recommendation. In such cases, the diagnosis would be set to “unknown” in the first discussion and a new treatment option may be recommended in the subsequent discussion (in the light of new evidence). This may also be the case for those who were initially placed on a close surveillance in the first discussion but later receive treatments. The available data does not capture this possible switching (in treatment recommendation) for cases discussed more than once. As a temporary measure to strengthen our result, cases that were discussed more than once are excluded in the analysis.

It is also interesting to know the proportion of these treatment recommendations that were eventually carried out on the patients. Blazeby *et al* [16] found that following multidisciplinary meetings, about 15.1% of the decisions reached at the MDT were not implemented. The reasons for these changes in treatment plan included patient choice and other clinical information which were not considered during the review process. Therefore, a future research can combine MDT data with post MDT information from *patient*

electronic record in order to establish the proportion of the MDT recommendations that are implemented.

6.3 Comparing the risk of Magnetic Resonance Imagine and Catheter Angiogram

A clinical decision about a diagnostic test requires a careful consideration of the risks and benefits of the test since some diagnostic techniques carry potential risks of complications. Nevertheless a test result normally provide a useful insight into the subsequent clinical decisions. A typical scenario involves a clinician making an observation based on a patient's condition to determine the cause of the symptoms and then deciding on the diagnostic tests best suited to the patient. The tests results would then guide the clinician in deciding on the appropriate treatment option(s).

This section describes a clinical BN model adapted from the work of Fenton *et al.*, [62] to compare the risks of two diagnostic procedures (the Magnetic Resonance Imagine and Catheter Angiogram) for intracranial aneurysms.

6.3.1 Medical Background

Intracranial aneurysms are bulbous expansions of the intracranial vessels. These vessels may rupture leading to a potentially lethal subarachnoid haemorrhage (SAH). We refer to a ruptured intracranial aneurysm as an episode of subarachnoid haemorrhage (SAH) and both terms will be used interchangeably throughout the rest of the chapter. An intracranial aneurysm is mostly asymptomatic and in many cases, an episode of subarachnoid haemorrhage is the first clinical manifestation of an intracranial aneurysm. However, the mass effect of an expanding aneurysm may lead to a third nerve palsy. In such cases, the pupil will be dilated and this is observable by a clinician. In most cases, decisions to diagnose a patient for intracranial aneurysm are taken under a clinical emergency condition [237]. Nevertheless, the risks of complication must be weighed appropriately with the associated benefit in order to come up with an optimal decision for the patient [237].

The two widely used diagnostic procedures are the minimally invasive catheter angiogram and the non-invasive magnetic resonance imaging technique. In catheter angiography (CA), a thin plastic tube (catheter) is inserted into an artery through a small incision in the skin. The catheter is then guided to the area being examined. Once the

catheter has reached the area, a contrast material is injected and images are captured using a small dose of x-rays. Generally, the CA test is used to examine blood vessels in key areas of the body, including the brain. On the other hand, the magnetic resonance angiography (MRA) technique involves the use of a powerful magnetic field, radio waves and a computer to produce the detailed images. Magnetic resonance angiography does not use ionizing radiation (x-rays) and may be performed with or without contrast material. The benefits and risks of both catheter angiogram and magnetic resonance imaging techniques are described in the *radiology information resource for patients*, [176]. The benefit of performing a catheter angiogram includes:

1. Angiography may eliminate the need for surgery and if surgery remains necessary, it can be performed more accurately.
2. The CA technique presents a very detailed, clear and accurate picture of the blood vessels. This may be helpful when a surgical procedure is being considered.
3. The use of a catheter makes it possible to combine diagnosis and treatment in a single procedure, [102]. For example, finding an area of severe arterial narrowing during a CA procedure can be followed by angioplasty¹ and placement of a stent².

However, the catheter angiogram procedure has a number of risks that may result in complications. These are:

1. There is a slight chance of cancer formation as a result of radiation. However, the potential benefit of an accurate diagnosis from the CA far outweighs the risk.
2. The CA technique may not be suitable for patients with histories of allergy to x-ray contrast material.
3. There is a potential risk of skin damage if a large amount of x-ray contrast material leaks out under the skin where the Intravenous cannula (IV) is placed.
4. There is also a rare risk of serious allergic reactions to a contrast material that contains iodine.

¹A surgical procedure to repair a blood vessel by inserting a special catheter (balloon-tipped) to unblock the blood vessel.

²A slender catheter inserted into a blood vessel to provide support for the blood vessel.

5. There is a small risk of blood forming a clot around the tip of the catheter.
6. The technique is not suitable for patients with diabetes.
7. Other risks of CA include contrast nephrotoxicity and the risks associated with the general anaesthetic that is required to perform CA.

The MRA technique has the following benefits:

1. The MRA procedure is a noninvasive imaging technique that does not expose patients to ionizing radiation.
2. There is no risk of damaging an artery because detailed images of a blood vessel and blood can be obtained without using a catheter.
3. The MRA procedure takes a shorter time requires no recovery period.
4. MRA is a good option for patients prone to allergic reactions because it can provide good quality images without using a contrast material.

The MRA examination poses almost no major risk to the average patient when appropriate safety guidelines are followed [176]. Some minor risks of performing an MRA techniques include the following:

1. Although the MRA technique does not expose patients to ionizing radiation, but the strong magnetic field may cause malfunctioning of implanted medical devices that contain metal.
2. There is also a slight risk of allergic reactions if a contrast material is injected.
3. Unlike the CA technique, MRA is not able to capture images of calcium deposits and small vessels, especially those in obscured locations.

Yoshimoto *et al.*, [222] reported that due to the increasing use of the non-invasive magnetic resonance imaging technique, brain screening for aneurysm in Japan has become more popular especially among those with an increased risk of developing aneurysms. Consequently, more un-ruptured intracranial aneurysms are now being detected earlier than before [222].

6.3.2 Background to a clinical negligence case study

Fenton et. al, [62] described a BN model to determine whether the diagnostic procedure offered a patient (Mrs. B) was sub-optimal. This model was used for expert's analysis and witness in a clinical negligence case brought by Mrs. B - a 55-year-old insulin-dependent diabetic woman who was admitted to hospital suffering from headaches and vomiting. She also had a pupil-sparing third Nerve Palsy. She had previously suffered two similar ischaemic³ episodes from which she had fully recovered without treatment. Pupil-sparing third Nerve Palsy caused by transient ischaemic condition is normally temporary and harmless. In a small percentage of cases, the cause of a pupil-sparing third nerve palsy can be due to expanding aneurysms which may rupture and become fatal if undetected and in these circumstances urgent diagnosis and treatment is required.

The clinician in charge of the patient recommended that a passive Magnetic Resonance Angiography (MRA) scan be performed urgently. This is a non-invasive test that is reasonably accurate for detecting expanding aneurysms. Being a Friday evening, the hospital could not offer such a test until the following Monday morning, causing 48 hour delay before treatment. Consequently, the patient was transferred to a specialist hospital that had the equipment to carry out the test immediately. However, the clinicians at the specialist hospital decided to perform a catheter angiogram (CA), an alternative invasive test, contrary to the recommendation of the first clinician. This test is an invasive test known to be more accurate than the MRA for diagnosing aneurysms but carries a known 1% risk of causing a permanent stroke in diabetic patients. The patient suffered a permanent stroke after the CA test and the cause of the third nerve palsy was subsequently found to be ischaemic.

6.3.3 BN model for comparing risk of two screening procedures for intracranial aneurysm

This section describes a BN model to compare these two diagnostic techniques with a view to evaluating the impact of their usage in a patient with a pupil-sparing third nerve palsy. This BN model was reported in Fenton *et al.* The BN model for comparing the risks of CA and MRA screening procedures is presented in Figure 6.7. This model is particularly interesting because intracranial aneurysm is an excellent field to explore in

³An ischaemic condition is caused by a restriction in blood supply due to factors in the blood vessels, with resultant damage or dysfunction of tissue,[225].

terms of uncertainty, competing strategies, and potentially clinically significant events consequent upon poor selection of technique.

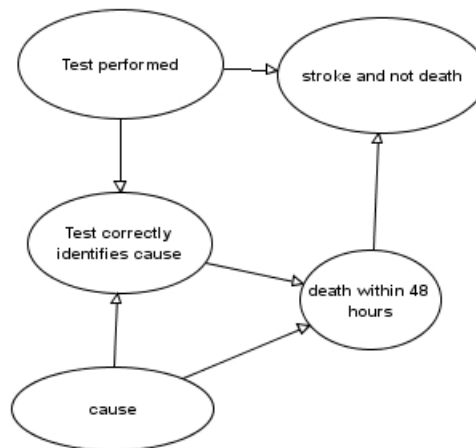


Figure 6.7: A clinical model for accessing risks of two screening procedures for intracranial aneurysms.

This model identifies five variables. These variables and their respective states are listed in Table 6.13.

Table 6.13: Variables in the model

Variable	States
Cause	Ischaemic (Isc) small aneurysm (SA) large aneurysm (LA) cavernous sinus pathology (CSP)
Test performed	MRA CA
Test correctly identifies cause	Yes No
Death within 48 hours	Yes No
Stroke and not death	Yes No

In the following sections some of the variables listed in Table 6.13 are used exactly as they appear in the BN model presented in Figure 6.7. That is, the same conditional independence and conditional probabilities assumptions are made. However, some of the variable are modified by adding additional states as required or by replacing them with a new variable. Section 6.3.4 describes these variables in turn and presents another BN model adapted from the BN model in Figure 6.7.

6.3.4 Screening for intracranial aneurysms

This model is very similar to the model presented in Figure 6.7 in the sense that it compares the risks of the CA and MRA diagnostic procedures. However, this model considers a situation whereby a patient is not screened at all. The similarities and differences between this model and the one presented earlier are emphasized in the descriptions of the variables. Let us examine these variables in turns.

Cause

This node represents causes of a third nerve palsy. A pupil sparing third nerve palsy could be due to an ischaemic condition; an expanding aneurysm or a cavernous sinus pathology (CSP)⁴ [56, 62].

These causes are modeled by a node “cause”. This node has four mutually exclusive causes of third nerve palsy. Hence, the node *cause* can take any of these four possible values “Ischaemic (Isc)”; “small aneurysm (SA)”; “large aneurysm (LA)” and “CSP”. s

Test Performed

Having identified possible causes of third nerve palsy, we can now describe a node that models the two diagnostic procedures for a third nerve palsy. We assume that a clinician can prescribe one of these two procedures at a time. In other words, the two procedures are never performed together on a patient. We model the information by a decision node labeled “Test performed (tp)”. The MRA and CA are two commonly used techniques for the screening procedures for detecting unruptured intracranial aneurysms and other causes of a third nerve palsy. Of fundamental importance, in the decision to perform a CA test, is the likelihood of diabetes mellitus. If diabete mellitus is present, then CA is not advisable because it increases the risk of permanent stroke, [62]. However, if diabete mellitus is absent, the CA procedure offers better accuracy in detecting intracranial aneurysms and CSP compared to the MRA technique.

In addition to these two diagnostic options, we consider a situation whereby neither of these two tests is used. Therefore, we add an additional option “No Test (NT)” to capture

⁴The cavernous sinus syndrome is defined by its resultant signs and symptoms. For example, it can show signs of ophthalmoplegia. This is a paralysis of one or more of the muscles that control the movement of the eye. The condition can be caused by any neurologic disorders.

a decision to place a patient on watchful waiting rather than performing either of these two conventional procedures. Hence, there are three states in this node (“CA”, “MRA”, “NT”).

Test correctly identifies cause

This variable, represented by a node labeled (*Test correctly identifies cause*), summarizes the accuracies of each of the diagnostic tests. The node has three states namely (“Yes”, “No”, “No Test”). The state “Yes”, on one hand, indicates the potential of a treatment option to identify a cause of third nerve palsy. The second state “No”, on the other hand encodes the probability that a test would fail to identify a cause of third nerve palsy. The last state “No Test” option is only relevant when a patient is placed on watchful waiting.

Stroke?

This is a binary node with values “True” or “False” to capture potential complications of each of the diagnostic procedures given clinical circumstances of a patient.

Death within 48 hours

This variable represents the potential risk of death resulting from a rupture of an expanded aneurysm. Fenton *et al* [62] assumed that a ruptured intracranial aneurysm would certainly lead to death within 48 hours. In this thesis, I have assumed that an undetected intracranial aneurysm may rupture within 48 hours but we can treat the event “death” i.e. death from a ruptured intracranial aneurysm as an uncertain event with a probability distribution. Therefore we split the variable into two nodes. These are:

1. Rupture: This variable captures the risk of rupture of aneurysms within 48 hours. This is based on the assumption that an expanded aneurysm (undetected) may rupture within 48 hours. This risk is represented by a binary node “Rupture” with two possible values “True” or “False”. The state “True” is used to capture the probability of rupture within 48 hours and “False” to capture the probability of no rupture within 48 hours;

2. Outcome: This is represented by a node labeled *Outcome*. The node captures potential risk of death resulting from a ruptured intracranial aneurysm. There are four mutually exclusive and collectively exhaustive possibilities (states) namely “Well”, “Stroke”, “Survive SAH” and “Dead”. The “Well” represents the state of wellbeing i.e. no complication whatsoever either due to a complication from a screening procedure or due to rupture of intracranial aneurysms. The second state is “Stroke”, this may be as a result of the initial complication of diagnostic procedures or due to rupture of an undetected aneurysm. Next, we have a state labeled ‘Survive SAH’ representing the state of good recovery following an episode of SAH without any neurological deficiency. Finally, for the mortality associated with aneurysmal rupture (SAH) we introduce a state labeled “Dead”.

A BN model describing the independence assumptions between these variables is presented in Figure 6.8.

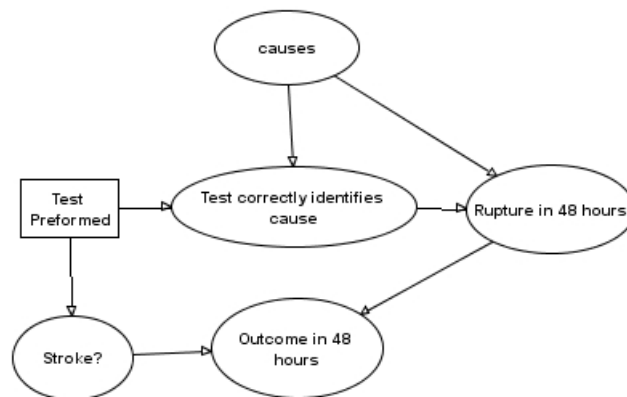


Figure 6.8: BN model to compare risk of the MRA and CA techniques, Adapted from Fenton *et. al's* model, [62].

We have, so far, described the qualitative part of the model, we now need to quantify the strength of each causal link. To do this, we need to describe the node probability table (NPT) for all nodes. We will now describe the NPT in turns.

Cause

The conditional probability table for this node was taken from Fenton *et. al's* model. A third never palsy is largely caused by an ischemic condition (98%) whereas other causes put together only accounted for 2% of cases of third nerve palsy.

$$\begin{aligned}
 p(Isc) &= 0.98 \\
 p(SA) &= 1.0E^{-4} \\
 p(LA) &= 0.0099 \\
 p(CSP) &= 0.01
 \end{aligned}$$

Test Performed

This node has a vacuous probability distribution as shown below:

$$\begin{aligned}
 p(CA) &= 0.3333333 \\
 p(MRA) &= 0.3333333 \\
 p(NT) &= 0.3333333
 \end{aligned}$$

In essence, we are assuming that the three screening options are equally likely. This is a widely used assumption in practice. All decision alternatives are normally taken to be equally likely because a decision maker doesn't know a priori which decision option is better than the other.

Stroke?

We have described the conditional probability table for the screening techniques i.e. "Test Performed". We can now describe our assumption regarding the potential risk of harmful effect resulting from each of these screening options. The risk of complication from a screening option is represented by *stroke?*. The risk of complication for the screening options is as follow:

Table 6.14: The conditional probability distribution $p(\text{Stroke}|tp)$

tp	Stroke?	
	False	True
MRA	0.99	0.01
CA	1.0	0.0
NT	1.0	0.0

The conditional probability presented in Table 6.16 is based on the assumption that a diabetic patient stands 1% risk of complication from a screening with CA test.

Accuracy of the Screening procedures

The CA technique is believed to be more accurate than a MRA option in detecting aneurysms especially small aneurysm and those located in the posterior communication arteries. In some cases a MRA test is followed by a CA in order to confirm or rule out aneurysms. However, we only consider a situation whereby only one or none of these techniques is performed. The node probability table for each configuration of parent nodes is as given in Table 6.15.

Table 6.15: The conditional probability distribution $p(ta|tp, cause)$

tp	cause	Test Accuracy		
		No	Yes	No Test
MRA	ISC	0.0	1.0	0.0
MRA	SA	0.5	0.5	0.0
MRA	LA	0.05	0.95	0.0
MRA	CSP	0.5	0.5	0.0
CA	ISC	0.0	1.0	0.0
CA	SA	0.05	0.95	0.0
CA	LA	0.01	0.99	0.0
CA	CSP	0.1	0.9	0.0
NT	isc	0.0	0.0	1.0
NT	SA	0.0	0.0	1.0
NT	LA	0.0	0.0	1.0
NT	CSP	0.0	0.0	1.0

Where ta and tp stand for nodes *Test correctly identifies cause* and *Test Performed* respectively.

Rupture

The information about rupture is relevant if aneurysm is present. This thesis focuses on cases involving aneurysms rather than considering all causes of third-nerve palsy. The probability values of rupture is set to zero for other causes i.e. $p(Rupture = Yes|ta, cause = ISC) = 0$ and $p(Rupture = Yes|ta, cause = CSP) = 0$. Two large studies before 1998, [111, 190] reported annual rupture rates of 1.4 to 1.9 percent for intracranial aneurysms. Aneurysmal rupture rates are normally higher in a population with aneurysms larger than 10 mm in diameter, symptomatic, or located in the posterior circulation [29]. Francesca *et al.*, [39] reported a rate with lower and narrower interval (0.4% and 1.5%). Therefore, we assume a rupture rate of 0.4% for small aneurysms and 1.9% for large aneurysms.

Outcome within 48 hours

As stated earlier, an episode of SAH may have catastrophic consequence. Full recovery is unlikely in patients with severe SAH. Unruptured Intracranial aneurysm is a fairly common condition that is often asymptomatic until the time of rupture [29]. Subarachnoid hemorrhage (SAH) associated with aneurysmal rupture is a potentially lethal event with more than 50% mortality rate and more than 50% permanent disability among those who survived the initial haemorrhage [96]. Therefore, we can derive probability entries for the node *Outcome within 48 hours (outcome)*. These values are given below:

Table 6.16: The conditional probability distribution $p(\text{Outcome}|\text{Rupture}, \text{Stroke?})$

Rupture	Stroke?	Outcome			
		Well	Stroke	Survive SAH	Dead
False	False	1.0	0.0	0.0	0.0
False	True	0.0	1.0	0.00	0.00
True	False	0.0	0.25	0.25	0.50
True	True	0.0	0.25	0.25	0.50

We now have a fully specified BN model that we can use to answer questions on some hypothetical clinical scenarios. To proceed, let us take a snapshot of prior marginal distributions for all nodes in the model (Figure 6.9). The prior marginal distribution for node *Outcome in 48 hours* shows a very high probability 99.66% for the well state and very small probabilities for other states. The prior marginal distribution for states *stroke*, *Survive SAH* and *Dead* are respectively 0.335%, 0.00167% and 0.00333%. Although these probabilities are very low but if we reason in term of an hypothetical cohort of 1 million patients then the number of patients who would be dead in 48 hours without screening and treatment is roughly 33.

To make inference about outcomes in 48 hours given different screening scenarios, we can consider three different screening scenarios for a patient with a third-nerve palsy. At the point of making the decision about diagnostic option to choose, the cause of the third nerve palsy is unknown to the clinician, so we need not to enter a hard evidence on the node “causes” in our scenarios. We can analyze the impact of each of the three decision options on the outcome in 48 hours. In Figure 6.10 - 6.12, we have posterior marginal distributions for the outcome node for each decision option.

For the first scenario, we enter *MRA* as hard evidence on the decision node *Test per-*

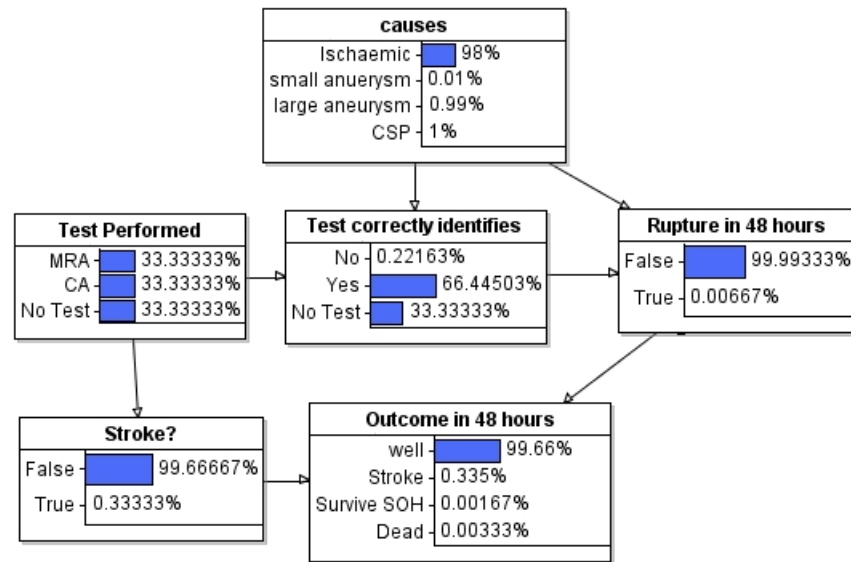


Figure 6.9: The prior Marginal distributions for all nodes

formed. The interest is to assess how the evidence will affect the posterior marginal distribution of the outcome node i.e. *Outcome in 48 hours*. Figure 6.10 shows a posterior marginal distribution given *MRA*. Here we are making inference about what would have happened if a MRA screening option was performed on a patient with a third nerve palsy.

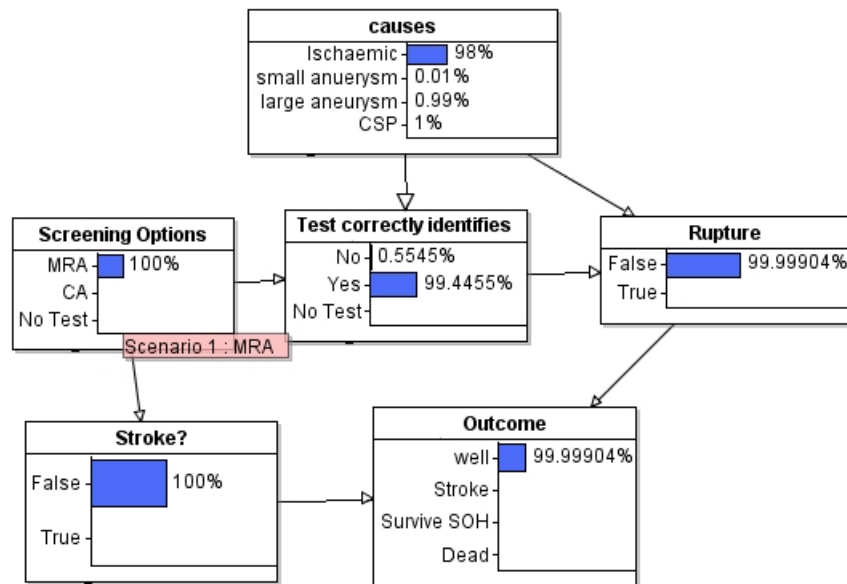


Figure 6.10: The posterior Marginal distributions of BN nodes given MRA test and large aneurysm

Our interest in the second scenario is to assess the impact of preforming a CA test on this patient. That is, we want to access the impact of performing a CA test on a patient with a third nerve palsy. The posterior marginal distribution for this scenario is shown in Figure 6.11.

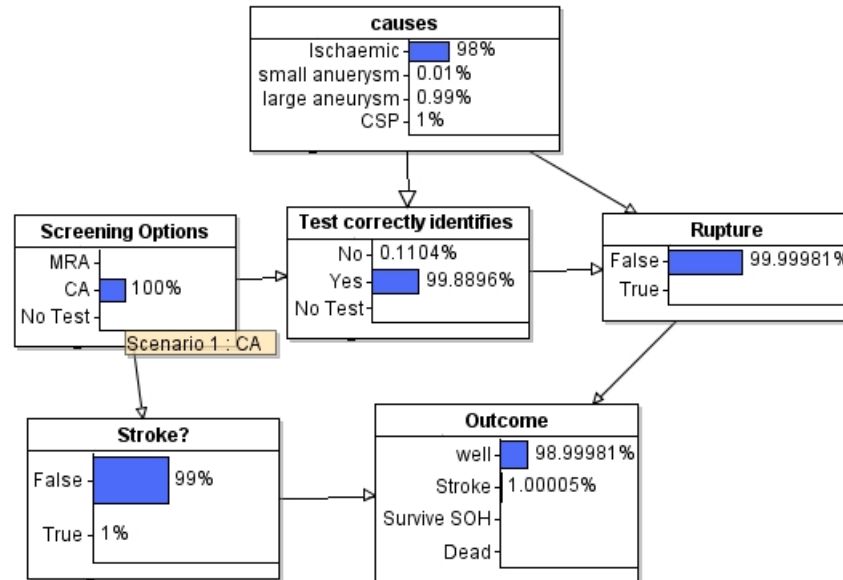


Figure 6.11: The posterior Marginal distributions of BN nodes given CA test and large aneurysm

We consider a third scenario to access what would have happened if neither of the two diagnostic procedures was performed. The posterior marginal distribution for this scenario is presented in Figure 6.12. Although the *No test* scenario would rarely occur in practice. However, we can use the model to simulate what would have happened if a patient with a third nerve palsy is placed on watchful waiting.

The posterior distributions for the outcome node from these scenarios and the prior marginal distribution are presented in Table 6.17.

Table 6.17: Posterior distributions of outcome for different scenarios

Outcomes	Baseline	MRA option	CA option	No treatment
Well	99.66	99.999	98.99981	99.98115
Stroke	0.335	$2.4012E^{-4}$	1.00005	0.00471
Survive SAH	0.00167	$2.4012E^{-4}$	$4.7525E^{-5}$	0.00471
Dead	0.00333	$4.8025E^{-4}$	$9.505E^{-7}$	0.00942

From Table 6.17 the CA test reduces the risk of death but also increases the risk of stroke. This bring us back to the initial question asked at the beginning of the chapter. That is, whether the marginal predictive accuracy of a CA technique compensates the attendant

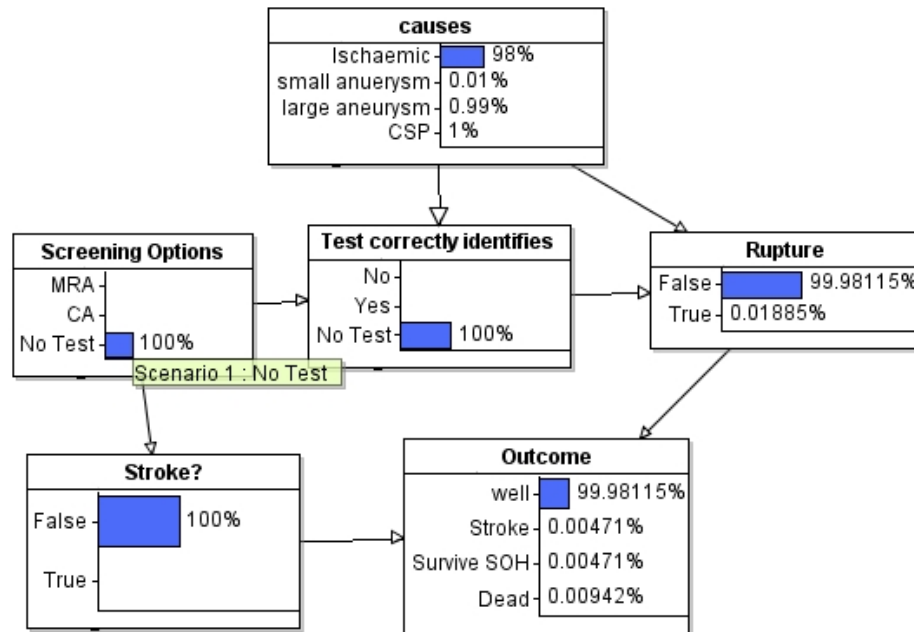


Figure 6.12: The posterior Marginal distributions of BN nodes given (No test) and large aneurysm risk of complication. Supposing the risk of stroke and the risk of death are valued equally, then we can use the posterior marginal distribution in Table 6.17 to access the impact of each of the two screening options. The risk of stroke and death following an episode of SAH from an undetected aneurysm by a MRA test are both $2.4012E^{-4}\%$. Therefore from a hypothetical cohort of 1 million patients with a third nerve palsy, screened with an MRA technique, the expected number of patients that would possibly develop neurological deficiency is 24 and about 48 are expected to die following an episode of SAH. On the other hand, by subjecting this population to a CA screening, only 1 patient would possibly die from an episode of SAH but with about 10,000 patients developing neurological deficiency from screening complication or from an episode of SAH.

The model showed that the risk of the CA test is greater than that of the MRA test. We can run the model with different prior probability assumptions for the screening procedure as well as the complication node “stroke”. If we stick to the current assumptions about the sensitivity and specificity of the two diagnostic techniques, then the risk in the CA pathway is greater, unless the risk of complication from the CA technique is considerably lower than the 1% that was assumed. We must recall that the risk of complication is only available for diabetic patients. Hence we must stress that the hypothetical population of 1 million patients used as an illustration above is a population of insulin dependent diabetic

patients with third nerve palsy.

This model has demonstrated that we can make use of probabilities, from experts and from meta analyses and other sources, to inform the construction of a BN that models key elements of clinical reality and the risks within. It illustrates how a clinician can use a model like this to select the optimum treatment, on average, for a population and as a basis for making decisions for individuals. A key element in the approach is its ability to highlight the key states and probability estimates such that sensitivity to assumptions, in the form of priors, can be tested. Indeed this model can be considered the accumulation of elements central to Buchan, Winn and Bishop's 4th paradigm approach [22].

Even with this success there are a number of weaknesses in the study carried out here and corresponding opportunities for improvement. Firstly, there was some hope that we could conjoin the models for the statistical parameter inference about aneurysms with the diagnostic and predictive modelling offered here. At its simplest this would involve passing probabilities, as parameters from a meta-analysis model to this one. However the overall aim, which was not fully realised, was to consider the full distribution of parameter uncertainty derived from the meta-analysis into account here as a form of sensitivity analysis. This would allow the clinician to test the sensitivity of the risks in this model the uncertainty of the underlying probability estimates based on clinical data. Doing so would complement the work done by Fenton and Neil [62] where they performed sensitivity analysis by investigating the effect on results of changing expert provided priors. The second weakness is that there was insufficient time to develop a means of formally "folding in" the continuous variable distributions into a discrete decision support model, in such a way that the second order uncertainties about every discrete probability are accounted for, rather than treated as definitive point estimates. Indeed this is a continuing and substantial research challenge that should be taken up by others, but remains somewhat unrecognised by the Bayesian research community. Its implications are quite profound since it suggests that the quantitative estimates produced using discrete BNs are over confident.

Finally, dynamic time-based models are clearly the way forward in this sort of analysis where each clinical stage can be modelled as a discrete time slice in the model using Dynamic Bayesian Networks (DBNs). This would enable an object oriented style of

modelling and help account for confounding processes as well as offering a powerful explicit representation of the complexity of the clinical process as a whole.

6.4 Summary

We have so far described different tasks involved in developing clinical decision problems. Two clinical case studies have also been introduced to demonstrate the modeling techniques described in the thesis. These case studies are not by any means exhaustive on the applications of the modeling techniques described in the preceding chapters but as prototypes on how we can use a BN model to learn parameters as well quantifying risks of two competing techniques. We envisage a future where doctors will have access to BN-based decision-support toolboxes that can automatically extract useful information from the literature; use the information to update parameters of clinical models and automatically quantify risks of choosing alternative diagnostic test pathways for any type of condition. Indeed, such decision-support systems will be able to present the results in an intuitive manner such that it is easily understandable to both the patients and doctors. Ultimately, the decision about the suitable clinical pathway to take still rests with the patient and doctors and not with the model, but we believe a *dynamic BN model* would provide a quantitative framework to balance the risks and benefit of different pathways. Such a framework would guide the patient on their perception of risks of different clinical options available to them.

The journey towards the realization of such *dynamic models* has only just begun. There are still more to be done in order to achieve this. Essentially, the way we translate the posterior distributions from meta-analysis into useful priors in decision model requires more work. We have only shown that this is potentially possible with a BN model as we have succeeded in using BN models for both meta-analysis and decision models.

Chapter 7

Summary and Conclusions

7.1 Summary

The BN framework is ideal for modeling clinical problems because it compactly represents probabilistic problems for efficient inference and also allows forward and backward reasoning. This framework has been widely used in the clinical domain to model problems with discrete variables. Many practical applications of BNs either manually discretized continuous variables or restrict their distributions to the conditional linear Gaussian (CLG), [107, 125]. The assumption of the linear Gaussian distribution may be too restrictive for a realistic clinical problem and this has limited the scope of the application of BN models.

Buchan, Winn and Bishop [22] have championed the realignment of Bayesian ideas for medical decision support as the 4th paradigm. However, this 4th paradigm lacks a unifying toolset, methodology or indeed a core algorithm for carrying out the necessary computations needed to transform decision support in medicine. This thesis has gone some way to testing four hypotheses implied by this 4th paradigm approach, specifically centred on the use of a unifying approach to computation, using DD and HBNs, and by illustrating, via case studies, how this approach can be used in practice.

In the following subsections we discuss the contribution of the thesis in helping address each hypothesis.

7.1.1 Modeling existing medical knowledge with plausible statistical distributions (H_1)

One attractive feature of Bayesian analysis is the ability to incorporate subjective knowledge. In a real life application of BN models, domain experts are normally involved in the analysis to provide subjective opinion. Once this has been done, a Bayesian analyst needs to translate the prior information into a mathematical form that can be incorporated into the analysis. For example a clinician might say that the sensitivity of a procedure is 92%. Often analysts use the relative frequency approach to translate existing medical knowledge into a prior distribution in a BN model. This representation implies that the parameter takes a specific value on the parameter space. Using dynamic discretization it is possible to use statistical distribution to model existing medical knowledge. The framework also allows analysts to give a probability that the prior for a medical parameter will be within a range of values in the parameter space. In this thesis, we have demonstrated the flexibility of the Dynamic Discretization Algorithm in modeling existing medical knowledge using appropriate statistical distributions.

A key criticism of standard Bayesian statistical methodology is that the expert is only allowed to express their expertise in the form of conjugate priors (and these mainly for mathematical convenience). This thesis has illustrated that the approach taken using DD is much more flexible, allowing expertise to be specified on sub-ranges of parameters, using non-conjugate priors and using mixtures of prior distributions. This provides great flexibility but there is some cost involved in this. The meaning and effect of a prior on the prediction can be difficult to foresee, so experimentation is needed. However now the choice of priors is wider than before and thus enhances our ability to properly represent a decision maker's prior choices and the effect of these to posterior results and predictions.

7.1.2 Parameter learning with multinomial BN models (H_2)

Another issue related to expertise about the domain concerns situations where deterministic constraints exist but are neglected by conventional methodology. We have identified some research studies where the decisions taken by clinicians are clearly not entirely probabilistic but are treated as such. Particular attention has been paid to Dirichlet style learning for multinomial distributions and we have shown that standard approaches, embedded in software packages, cannot model this situation correctly. This thesis suggests

remedies to this problem.

We described the use of the multinomial BN formulation in learning parameters of categorical variables. Several parameter learning techniques have been proposed to address the challenges of tackling real life problems with Bayesian Network. In this work, we use the dynamic the BN framework and dynamic discretization to learn parameter of a clinical model to evaluate the impact of multi-disciplinary team meetings on the treatment selection for patients with a diagnosis of cancers. In the clinical guideline used by the multi-disciplinary, some treatments are never recommended for some diagnoses. For example, benign diagnoses of the pancreas are never treated with chemotherapy, therefore, it is logically impossible to have any record of a benign diagnosis of pancreas treated with chemotherapy. We encode these logically impossible observations by zero but the conventional technique could not distinguish between logically impossible observations and zero observations. In particular, the Deal package, [20] requires equal number of parameters and treats cases with no record as zero observed values. In other words, the conventional techniques will compute posterior distributions for these parameters even when they are encoded as logically impossible. The approach described in the thesis allows analysts to dynamically create a multinomial model based on the number parameters required.

7.1.3 Causal learning with BN models (H_3)

This thesis also demonstrated the application of the multinomial BN formulation and dynamic discretization to score causal relations between variables. This score allows us to select a more plausible model (given the data) from a pool competing models. This approach requires domain knowledge to formulate a list of hypotheses about a problem. This is not a full-fledged causal learning approach but it has some desirable benefits over the current causal learning techniques. For example, learning causal relations in a problem with logically impossible observations requires learning only the relevant parameters. The Deal package [20] allows analysts to specify a starting BN model into the learning process; it also allows analysts to incorporate existing medical knowledge through a *ban-list*¹ but it cannot handle unequal number of parameters. We extended the use of the multinomial formulation to scoring causal relations. The result our experiments showed

¹A list of impossible arcs between variables

that, using dynamic discretization, we can learn causal relations with BN models. Although this has not been achieved in this thesis but we have been able to show that this is potentially possible.

A key aspect of the 4th paradigm is the mediation of model construction between expert and machine. This thesis takes a hypothesis driven approach by using the expert's views to reduce the a priori search space of possible causal conjectures and then testing these using available data. This stands in contrast to existing machine learning methodology where greedy search is used to determine causal connection. Whilst it cannot be claimed that this will guarantee success, nor is it claimed that the analysis done here is comprehensive, we can claim that it shows potential and is worthy of further investigation.

7.1.4 Solving meta analysis with BN models (H_4)

Another contribution of the thesis is the application of the Hierarchical Bayesian Network (HBN) models (augmented by dynamic discretization technique) to solving meta-analysis problems. Solving meta-analysis with a BN model implies that the full posterior distribution of a summary estimate can be used as prior in clinical decision models. This approach captures the full uncertainty surrounding the summary estimate and also allows analysts to check the sensitivity of a decision to the underlying studies.

Three meta analysis problems were analyzed in the thesis to demonstrate the use of BN in solving meta-analysis problems. Our approach does not cover the system review that produced the data so any inherent deficiencies in the data would affect the reliability of the summary estimates therefore, we have only used these meta-analysis problems to demonstrate how to perform meta-analysis using the BN framework.

The Beta-blocker clinical example showed that the result from the BN approach is comparable with those obtained using the classical technique based on the fixed effect assumption. The second example on the magnesium trial demonstrated a desirable feature of a BN model that allows us to perform inference subject to a constrained parameter. The BN approach allows us to constrain the posterior distribution of the heterogeneity parameter and then perform inference, subject to this constraint. DuMouchel, [53] pointed out that the use of a significance test for homogeneity of the data is questionable, especially when the number of studies in a meta analysis is less than 20. However there are

meta-analysis problems,[53, 223, 232] with less than 10 studies. The Bayesian approach is suitable for meta-analysis with few studies as it does not require a significant test for homogeneity but rather it incorporates the full posterior distribution of heterogeneity parameter in the analysis. Bayesian methodology has been described as a compromise between the two opposing philosophies of meta-analysis [53, 232]; the philosophy of a fixed effect who believe that τ is near 0 and those who believe that τ is large and “borrowing strength” is hopeless in most cases. The BN approach allows us to specify a prior distribution that supports these two philosophies. The third example on the prevalence of intracranial aneurysms demonstrated the use of a BN model to summarize data from a non-experimental design. This example also demonstrated the use of the KL-metric to explore data for inconsistent observations. The conventional approach to meta-analysis of proportions such as the prevalence of intracranial aneurysms uses a normal distribution for the logit transformation of the true proportion, [223]. In this study we assume a beta distribution for the true proportion and binomial likelihood for the observed data. The use of dynamic discretization allows us to model data directly with appropriate distributions.

Ideally when comparing new methods against established techniques one should conduct the comparative analysis in as systematic and objective a way as possible. There are thousands of meta-analyses published every year and unfortunately, due to time and resource constraints it has not been possible to provide an exhaustive review of meta-analysis in the clinical domain as one would have liked. Instead a snap shot has been provided along with caveats and limitations on what might be learned from this alone.

The approach adopted here was pragmatic and focused on choosing those papers that used popular methods of analysis (hence using MCMC as a comparator) where data was available in the research paper itself and also where sufficient detail was provided to enable the model to be reconstructed and then configured and tested. This study does not claim that the analysis neither is free from selection bias nor is comprehensive, however it does show how one can use HBNS and DD for meta-analysis in a flexible way and offers some support to the research hypothesis.

7.2 Conclusions

Four hypotheses were introduced at the beginning of the thesis in chapter 1. Various experiments were conducted to answer these questions. The findings and conclusion based on these experiment are described here, in turns.

Firstly, the thesis is intended to find out whether we can use any plausible statistical distribution to model subjective priors in a BN models. This hypothesis was addressed in chapter 4. Using the Dynamic discretization algorithm, it is possible to use conjugate and non-conjugate priors in a BN model without compromising basic principles of Bayesian inference. Indeed, we can specify subjective prior directly on a parameter or even treat the parameter as a random variable drawn from some distributions. More so, the use of dynamic discretization enables us to specify any plausible distribution for the hyper parameters.

Secondly, the thesis is also intended to establish whether Bayesian Network models can be used for parameter learning. This hypothesis was also addressed in chapter 4. The Multinomial BN formulations was used to learn parameter of a clinical model in chapter 6.

To address the third hypothesis, we extended the use of the multinomial BN formulation to learn causal relations. This hypothesis was address in chapter 4. Parameters learning is an important part of modeling. A great deal of research effort has been devoted into developing parameter learning algorithms and here we have shown that it is potentially possible to use the dynamic discretization algorithm for this purpose.

Finally, chapter 5 addressed the fourth hypothesis i.e. whether it is possible to use BNs models to generate summary estimates from meta-analysis based on fixed and random effect assumptions. The result showed that, using dynamic discretization, the BN framework provides an alternative to the conventional techniques.

7.3 Future work

As we have emphasized at different sections of the thesis, our goal is to develop a unified modeling approach to clinical decision problems. In particular, we want to use the BN models for learning parameters, meta-analysis and for scoring hypotheses about causal relations between variables. We argued that a posterior distribution from a parameter

learning model could be used as an input into a decision making model. While this is conceptually simple and theoretically possible, it has not been implemented in this thesis.

The thesis has also shown that we can use various statistical distributions to specify a subjective prior provided by a medical expert about a parameter. Essentially, different ways to specify subjective prior distributions were described using the reference, enthusiastic and sceptical priors. However, only the summary statistics of the posterior distributions were used as inputs in the decision models. A future study to investigate how to use the full posterior distributions would be an interesting addition to the existing body of knowledge.

Some of the results presented in the thesis were compared with the MCMC (Gibbs Sampling) approach to check the accuracy the dynamic discretization algorithm. While MCMC simulation is by far the most widely used technique for approximate inference, it might also be interesting to compare the results with other approximating techniques such as mixture of truncated exponential, mixture of Gaussian and expectation propagation.

Furthermore, only three meta analysis problems were described in the thesis. Using this technique to summarize other treatment effects would be a valuable contribution to this work. So far, we have demonstrated the use of the technique to summarize, the odds ratio and probability. Other clinical effects such as, risk ratio, risk difference, correlation etc have not been summarized using the BN approach.

Bibliography

- [1] SM Alvarez, BA Poelstra, and RS Burd. Evaluation of a bayesian decision network for diagnosing pyloric stenosis. *J Pediatr Surg*, (41(1)):15561, 2006.
- [2] Sonia M Alvarez, Beverly A Poelstrab, and Randall S Burd. Evaluation of a bayesian decision network for diagnosing pyloric stenosis. *Journal of Pediatric Surgery*, 41:155161, 2006.
- [3] P Antal, G Fannes, D Timmerman, Y Moreau, and De Moor B. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artif Intell Med*, 29(1-2):39–60, 2003 Sep-Oct.
- [4] P Antal, G Fannes, D Timmerman, Y Moreau, and B De Moor. Using literature and data to learn bayesian networks as clinical models of ovarian tumors. artificial intelligence in medicine. *International Journal of Epidemiology*, 30(3):257–81, 2004 Mar.
- [5] D Aronsky, M Fiszman, and WW *et al.* Chapman. Combining decision support methodologies to diagnose pneumonia. *Proc AMIA Symp*, pages 12–6.
- [6] D Ashby and AFM Smith. Evidence-based medicine as bayesian decision-making. *Statistics in Medicine*, 19:32913305, 2000.
- [7] Deborah Ashyby. Bayesian methods. <http://www.wiley.com/Cab001-.pdf>, last accessed on August 8, 2011.
- [8] M Athanasiou and JY Clark. A bayesian network model for the diagnosis of the caring procedure for wheelchair users with spinal injury. In: *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS07)*, page 4338, 2007.
- [9] A Bartolucci. A bayesian meta-analysis approach to address the effectiveness of statins in preventing death after an initial myocardial infarction. *18th World IMACS Congress and MODSIM09 International Congress on Modelling and simulation Bartolucci*, pages 129 –135, 2009.
- [10] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370418.
- [11] I Beinlich, G Suermondt, R Chavez, and G Cooper. The alarm monitoring system: a case study with two probabilistic inference techniques for belief networks. *Second European Conf. Artif Intell Med*, (38):24756, 1989.
- [12] JM Bernardo and AFM Smith. Bayesian theory. *Wiley, New York*, 1994.
- [13] Scott M Berry. Addressing the incremental benefit of histamine dihydrochloride when added to interleukin-2 in treating acute myeloid leukemia: a bayesian meta-analysis. *Cancer investigation*, 29:293, 2011.

- [14] Zweig IBM Birmes, UW. The graphical models toolkit (gmtk). *available from <http://ssli.ee.washington.edu/birmes/gmtk/>*, accessed on the 24th of March 2011.
- [15] M Birchall. Effect of process standards on survival of patients with head and neck cancer in the south and west of england. *Br J Cancer*, (91):1477–81, 2004.
- [16] JM Blazeby, L Wilson, C Metcalfe, J Nicklin, R English, and JL Donovan. Analysis of clinical decision-making in multi-disciplinary cancer teams. *Ann Oncol*, (17(3)):457–60, 2006.
- [17] Michael Borenstein, Larry Hedges, and Hannah Rothstein. Introduction to meta-analysis. *www.Meta-Analysis.com*; 2007 (last accessed on August 8, 2011).
- [18] GEP Box and GC Tiao. Bayesian inference in statistical analysis. *New York: Wiley*, 1973.
- [19] GEP Box and GC Tiao. *Bayesian inference in statistical analysis*. AddisonWesley Reading, 1973.
- [20] Susanne G. Bttcher and Claus Dethlefsen. Learning bayesian networks with r. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing: Vienna, Austria*, pages ISSN 1609–395X, March 2022 2003.
- [21] Susanne Gammelgaard Bttcher. Learning bayesian networks with mixed variables. *Ph.D. Thesis, Department of Mathematical Sciences. AALBORG University*, 2004.
- [22] I Buchan, J Winn, and CA Bishop. Unified modeling approach to data-intensive healthcare in the fourth paradigm: Data-intensive scientific discovery. *Microsoft Research*, 2009.
- [23] ES Burnside. Bayesian networks: Computer-assisted diagnosis support in radiology. *Academic Radiology*, pages 422–430.
- [24] ES Burnside, DL Rubin, JP Fine, RD Shachter, GA Sisney, and WK Leung. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology*, 3(240):66673, 2006.
- [25] BP Carlin and TA Louis. Bayes and empirical bayes methods for data analysis. *Chapman & Hall London*, 1996.
- [26] W Casscells, A Schoenberger, and TB Graboys. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299:999–1001, 1978.
- [27] C Chang, PJ Varghese, J Downey, and S Bloom. Magnesium deficiency and myocardial infarct size in the dog. *J Am Coll Cardiol*, 5:280–9, 1985.
- [28] JH Chang. The impact of a multidisciplinary breast cancer center on recommendations for patient management: the university of pennsylvania experience. *Cancer*, (91):1231–37, 2001.
- [29] Vega Charles, V Kwoon Jeremiah, and D lavine Sean. Intracranial aneurysms: current evidence and clinical practice. *Japanese Ophthalmological Society*, 66(4):601–608, 2008.

- [30] Robert T Clemens. *Making Hard Decisions: Introduction to Decision Analysis*. PWS-Kent Publishing co Boston, 1991.
- [31] Barry R Cobb. Inference and decision making in hybrid probabilistic graphical models. *PhD thesis: University of Northern Iowa*, 1995.
- [32] E Coiera. *Guide to Health Informatics*. Arnold, London,(2nd Edition), 2003.
- [33] Aliferis Constantin, Statnikov Alexander, and Tsamardinos Ioannis. Causal explorer: Causal discovery algorithms. *Discovery Systems Laboratory in the Department of Biomedical Informatics, Vanderbilt University*. available from <http://discover1.mc.vanderbilt.edu/discover/public/>, last accessed on the 24th of March 2011.
- [34] GF Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.
- [35] N Cooper, A Sutton, and K Abrams. Decision analytical economic modelling within a bayesian framework: application to prophylactic antibiotics use for caesarean section. *Stat Methods Med Res*, 6(11):491512, 2002.
- [36] Norsys Software Corp. Netica bayesian network software from norsys. available from <http://www.norsys.com/>, accessed on 24th of March 2011.
- [37] Angela Coulter and Jo Ellins. Patient-focused interventions: A review of the evidence. *The health foundation*, 2006.
- [38] RG Cowell, AP Dawid, TA Hutchinson, and DJ Spiegelhalter. A bayesian expert system for the analysis of an adverse drug reaction. *Artificial Intelligence in Medicine*, (3):25770, 1991.
- [39] Francesca Crawley, Andrew Clifton, and Martin M Brown. Should we screen for familial intracranial aneurysm? *Stroke*, 30:312–316, 1999.
- [40] Ruth J Cronje, James R Freeman, Owen D. Williamson, and Christine J Gutsch. Evidence-based medicine. *American Society for Clinical Pathology*, 35(12), 2004.
- [41] N Cruz-Ramirez, HG Acosta-Mesa, H Carrillo-Calvet, Nava-Fernandez L Alonso, and RE. Barrientos-Martinez. Diagnosis of breast cancer using bayesian networks: a case study. *Comput Biol Med*, (37(11)):155364, 2007.
- [42] P Dagum and M Luby. Approximating probabilistic inference in bayesian belief networks is np-hard. *Artificial Intelligence*, 1(60):141–153, 1993.
- [43] J Wilkinson Darren. Gdagsim: Software for sampling and conditional sampling of gaussian directed acyclic graph. *School of Mathematics & Statistics, University of Newcastle*. Available from <http://www.staff.ncl.ac.uk/~gdagsim/>, accessed on the 24th of March 2011.
- [44] Adnan Darwiche. Sensitivity analysis, modelling, inference and more (samiam). *University of Central Lancashire*, available from <http://reasoning.cs.ucla.edu/samiam/>, accessed on 10th of February 2011.
- [45] F Davidoff, RB Haynes, and DL Sackett. Evidence-based medicine: A new journal to help doctors identify the information they need. *BMJ*, 1995.

- [46] F Davidoff, RB Haynes, DL Sackett, and R Smith. In praise of bayes. *Economist available from <http://www.cim.mcgill.ca/-friggi/bayesfbayesrule/>*, 356:83–84, September 2000.
- [47] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clin Trial*, 7:17788, 1986.
- [48] R DerSimpnian and N Laird. Meta-analysis in clinical trials. *Control Clinical trials*, 7:177–188, 1986.
- [49] JW Dexheimer, LE Brown, J Leegon, and D Aronsky. Comparing decision support methodologies for identifying asthma exacerbations. *Study Health Technol Inform*, (129):8804, 2007.
- [50] L Keefer Donald and E Bodily Samuel. Three-point approximations for continuous random variables. *Sociedad de Estadstica e Investigacion Operativa Test*, 29(5):595–609, May 1983.
- [51] MJ Druzdzel and Gaag LC Van Der. Building probabilistic networks: Where do the numbers come from? *Guest Editors' Introduction. Knowledge and Data Engineering, IEEE Transactions*, 4(12):481–486, 2000.
- [52] DruzdzelMJ. Druzdzel, mj and van, der gaag lc. *Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information*. In: *Besnard & Hanks (eds)*, pages 141–148, 1995.
- [53] WH DuMouchel. Hierarchical bayesian linear models for meta-analysis. *National Institute of Statistical Sciences. Techn, Report*, (27), 1994.
- [54] B Efron. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100:15., 2005.
- [55] PC Elwood, PM Sweetnam, WH Beasley, D Jones, and R France. Magnesium and calcium in the myocardium: cause of death and area differences. *Lancet*, ii:720–2, 1980.
- [56] eMedicine for WebMD. Cavernous sinus syndromes. *available from <http://emedicine.medscape.com>*, last accessed on 7 Mar. 2011.
- [57] Charniak Eugene. Bayesian networks without tears. *American Association for Artificial Intelligence*, pages 50–63, 1991.
- [58] Gagliardi Cozman Fabio. Java bayes: Bayesian networks in java. *Available from <http://www.cs.cmu.edu/javabayes/Home/>*, 1998-2001 (accessed on the 24th of March 2011).
- [59] NE Fenton, P Hearty, M Neil, and L Radlinski. Software project and quality modelling using bayesian networks artificial intelligence. *in Applications for Improved Software Engineering Development: New Prospects. (Eds Meziane, F. and Vadera, S. Hershey), New York, USA, IGI Global: Chapter 1,1-25*.
- [60] NE Fenton, M Neil, P Hearty, W Marsh, D Marquez, P Krause, and R Mishra. Predicting software defects in varying development lifecycles using bayesian nets. *Information & Software Technology*, 2007.

- [61] NE Fenton, M Neil, and D Marquez. Using bayesian networks to predict software defects and reliability. *5th International Mathematical Methods in Reliability Conference (MMR 07) Glasgow*, 1-4 July 2007.
- [62] Norman Fenton and Martin Neil. Comparing risks of alternative medical diagnosis using bayesian arguments. *Journal of Biomedical Informatics*, 2010.
- [63] Norman Fenton, Lukasz Radlinski, and Martin Neil. Improved bayesian networks for software project risk assessment using dynamic discretisation. *IFIP Conference Software Engineering Techniques (SET 2006), Warsaw, Poland, 17-20 Oct 2006, in Software Engineering Techniques: Design for Quality*, Springer Boston, ISBN 978-0-387-39387-2:139–148.
- [64] Stephen E. Fienberg. When did bayesian inference become “bayesian”? *Bayesian Analysis*, 1:140., 2006.
- [65] RA Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd Edinburgh, <http://psychclassics.yorku.ca/Fisher/Methods.>, 1925.
- [66] RA Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd Edinburgh, 1956.
- [67] Fisheries and Aquaculture Department. Bayes-sa - bayesian stock assessment methods in fisheries - user’s manual. *FAO dorporate document repository; available at <http://www.fao.org/DOCREP/005/Y1958E/y1958e06.htm>*, last accessed on Sept 2, 2011.
- [68] LM Forrest, DC McMillan, CS McArdle, and DJ Dunlop. An evaluation of the impact of a multidisciplinary team, in a single centre, on treatment and survival in patients with inoperable non-small-cell lung cancer. *British Journal of Cancer*, (93):977–978, 2005.
- [69] Schwarz G. Estimating the dimensions of a model. *Annals of Statistics*, 6:461–464, 1978.
- [70] M Gabel. Multidisciplinary breast cancer clinics. do they work? *Cancer*, (79):2380–84, 1997.
- [71] D Geiger and D Heckerman. Learning gaussian networks. *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, USA.,* page 235243.
- [72] A Gelman, FY Bois, and J Jiang. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91:14001412, 1996.
- [73] Andrew Gelman, B Carlin John, S Stern Hal, and B Rubin Donald. *Bayesian Data Analysis*. Chapman & Hall/CRC,(2nd Edition), 2003.
- [74] Andrew Gelman and Donald B Rubin. Markov chain monte carlo methods in biostatistics. *Statistical Methods in Medical Research*, 5:339–55, 1996.
- [75] O Gevaert. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Hum Reprod*, 7(21):182431, 2006.

- [76] MF Ghani and M Rabah. Effect of magnesium chloride on electrical stability of the heart. *Am Heart J*, 94:600–2, 1977.
- [77] WR Gilks, S Richardson, and DJ Spiegelhalter. *Markov chain Monte Carlo methods in practice*. New York: Chapman and Hall, 1996.
- [78] IJ Good. *The Estimation of Probabilities*. Cambridge, MA: MIT Press., 1965.
- [79] Peter J Green. Grappa: R functions for probability propagation. *University of Bristol, UK*. Available from <http://www.stats.bris.ac.uk/~peter/Grappa/>, November 20, 2005 (accessed on the 24th of March 2011).
- [80] S Greenland. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev*, 9:1–30, 1987.
- [81] Evidence-Based Working Group. Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*, pages 2420–2425, 1992.
- [82] ISIS-4 Collaborative Group. Isis-4 - a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58050 patients with suspected acute myocardial infarction. *Lancet*, 345:669–85, 1995.
- [83] JH Gurwitz, NF Col, and J Avorn. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA*, 1992.
- [84] G Guyatt and R Drummond. Users' guides to the medical literature. a manual for evidence-based clinical practice. *American Medical Association*, 2002.
- [85] Akaike H. A new look at statistical model identification. *IEEE Trans. Automatic Control*, 19:716–723, 1974.
- [86] P Haddawy, Jr CE Kahn, and M Butarbutar. A bayesian network model for radiological diagnosis and procedure selection: work-up of suspected gallbladder disease. *Med Phys*, 21:1185–92.
- [87] PW Hamilton, N Anderson, PH Bartels, and Thompson D. Expert system support using bayesian belief networks in the diagnosis of fine needle aspiration biopsy specimens of the breast. *J Clin Pathol*, 1994.
- [88] Alexander J. Hartemink. Banjo: Bayesian network inference with java objects. *Department of Computer Science, Duke University*. Available from <http://www.cs.duke.edu/~amink/software/banjo/>, accessed on 24th of March 2011.
- [89] J Hartigan. Invariant prior distributions. *Annals of mathematical statistics*, 35:836–845, 1964.
- [90] P Hearty, NE Fenton, M Neil, and P Cates. Automated population of causal models for improved software risk assessment. *20th IEEE/ACM International Conference on Automated Software Engineering, Long Beach, California, USA, November 7-11*, ISBN: 1-59593-993-4:433–435, 2005.
- [91] D Heckerman. A tutorial on learning with bayesian networks. *Technical Report: Microsoft Research.*, March 1995 (Revised November 1996).
- [92] DE Heckerman, EJ Horvitz, and BN Nathwani. Towards normative expert systems: part i - the pathfinder project. *Methods Inf Med*, (31):90105, 1992.

- [93] Julian Higgins and Anne Whitehead. Borrowing strength from external trials in meta-analysis. *Statistics in Medicine*, 15:2733–2749, 1996.
- [94] Julian PT Higgins and David J Spiegelhalter. Being sceptical about meta-analyses: a bayesian perspective on magnesium trials in myocardial infarction. *International Journal of Epidemiology*, 31:96–104, 2002.
- [95] N Hoot and D Aronsky. Using bayesian networks to predict survival of liver transplant patients. *AMIA Annu Symp Proc*, page 3459, 2005.
- [96] JW Hop, GJ Rinkel, A Algra, and J van Gijn. Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review. *JStroke*, 28:660–4, 1997.
- [97] Cecil Huang and Adnan Darwiche. Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning*, 11, 1994.
- [98] Hugin. Hugin experts. Available from <http://www.hugin.com/>, last access on July 17, 2011.
- [99] R Ihaka and R Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, (5):299314, 1996.
- [100] S Aleem Ilyas, H Schemitsh Emil, and P Hanson Beate. What is a clinical decision analysis study? *J Can Chiropr Assoc*, (49), 2005.
- [101] SAS Institute Inc. Preliminary capabilities for bayesian analysis in sas/stat[®] software. *SAS working paper, SAS Institute Inc.*, 2006.
- [102] Angels Health International. Benefits and risks of coronary angiogram procedure? Available from <http://www.angeleshealth.com/angiogram-benefits.aspx>, accessed March 6, 2011.
- [103] ET Jaynes. Papper on probability, statistics, and statistical physics,. ed. *RD Rosenkrantz. Dordrecht, Nederlandrs: Reidel*, 1983.
- [104] H Jeffreys. Theory of probability. *Third edition, Oxford University press*, 1961.
- [105] M Jenicek. *Foundations of Evidence-Based Medicine. London. Parthenon Publishing Group*, 2003.
- [106] FV Jensen. Bayesian networks and decision graphs. *United States., Springer-Verlag. New York*, 2001.
- [107] FV Jensen, SL Lauritzen, and KG Olesen. Bayesian updating in causal probabilistic networks by local computations. *Comp. Stat. Quart.*, 4(269), 1990.
- [108] Lin JH and Haug PJ. Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *J Biomed Inform*, 1(41):114, 2008.
- [109] Wu Jian, Zhang Guangming, Cao Yanyan, and Cui Zhiming. Research on cerebral aneurysm image recognition method using bayesian classification. *Proceedings of the 2009 International Symposium on Information Processing (ISIP09)*, pages 058–062, August 21–23, 2009.

- [110] Cheng Jie. Belief net power constructor. *Department of Computing Science, University of Alberta*. Available from <http://webdocs.cs.ualberta.ca/jcheng/bnpc.htm>, accessed on the 24th of March 2011.
- [111] S Juvela, M Porras, and O Heiskanen. Natural history of unruptured intracranial aneurysms: a long-term follow-up study. *J Neurosurg*, 79:174–82, 1993.
- [112] CE Kahn Jr, JJ Laur, and GF Carrera. A bayesian network for diagnosis of primary bone tumors. *J Digit Imaging*, 2 Suppl. 1(14):567, 2001.
- [113] CE Kahn Jr, LM Roberts, KA Shaffer, and P. Haddawy. Construction of a bayesian network for mammographic diagnosis of breast cancer. *Comput Biol Med*, 1(27):1929, 1997.
- [114] JI Kazi, PN Furness, and M Nicholson. Diagnosis of early acute renal allograft rejection by evaluation of multiple histological features using a bayesian belief network. *J Clin Pathol*, 2(51):10813, 1998.
- [115] CE Khan, L Roberts, K Wang, D Jenks, and P Haddawy. Preliminary investigation of a bayesian network for mamographic diagnosis of breast cancer. In *R.M. Gardner, editor, Proceedings of the American Medical Informatics Association*, pages 208–212, 1995.
- [116] Vahid Khodakarami. Applying bayesian networks to model uncertainty in project scheduling. *PhD Thesis, Queen Mary, University of London*, 2009.
- [117] JH Kim and J Pearl. A computational model for causal and diagnostic reasoning in inference engines. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, International Joint Conference on Artificial Intelligence*, pages 190–193, 1983.
- [118] JA Kline, AJ Novobilski, C Kabrhel, PB Richman, and DM Courtney. Derivation and validation of a bayesian network to predict pretest probability of venous thromboembolism. *Ann Emerg Med*, 45:282– 90.
- [119] HM Kocher, OV Ogunsanya, and Martin Neil. An evaluation of the impact of a multidisciplinary team meeting on treatment selection and the outcome of treatment for patients with cancer. *working papper*, 2010.
- [120] Alexander V Kozlov. Efficient inference in bayesian networks. *PhD Thesis, Stanford University*, 1998.
- [121] Alexander V Kozlov and Daphne Koller. Nonuniform dynamic discretization in hybrid networks. In *the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 314–325, 1997.
- [122] Paul J Krause. Learning probabilistic networks. *Working paper Philips Research Laboratories.*, 2010.
- [123] S Kullback and RA Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
- [124] Kathryn Blackmond Laskey and Suzanne M Mahoney. Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, (12).

- [125] SL Lauritzen and DJ Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224, 1988.
- [126] Steffen Lauritzen. graphical models in r. *available from Available from <http://www.ci.tuwien.ac.at/gR/>*, 2002 accessed on the 24th of March 2011.
- [127] WP Leary, AJ Reyes, CJ Lockett, DD Arbuckle, and K Van der Ryl. Magnesium and deaths ascribed to ischemic heart disease in south africa. a preliminary report. *S Afr Med J*, 64:775–6, 1983.
- [128] J Leegon, I Jones, K Lanaghan, and D Aronsky. Predicting hospital admission for emergency department patients using a bayesian network. *AMIA Annu Symp Proc*, page 1022, 2005.
- [129] H Lehmann and E Thomas Shortliffe. Building bayesian statistical expert systems to aid in clinical decision making. *Comput Methods Programs Biomed*, 4(35):25160, 1991.
- [130] Z Li and B D’Ambrosio. Efficient inference in bayes networks as a combinatorial optimization problem. *International Journal of Approximate Reasoning*, 11:55–81, 1994.
- [131] Agena Limited. Agenarisk software. *Available from <http://www.agenarisk.com/>*, 2008, accessed on March 14, 2011.
- [132] Agena Limited. Agenarisk 5.0 user manual. *Available from <http://www.agenarisk.com/>*, last access on July 17, 2011.
- [133] Bayesware Limited. Bayesware discoverer: Knowledge discovery using bayesian networks. *available from <http://www.bayesware.com/>*, accessed on 24th of March 2011.
- [134] DeriveIt LLC. Deriveit’s bayesian network modeler. *available from <http://www.deriveit.com/>*, accessed on the 24th of March 2011.
- [135] Oteniya Lloyd. Bayesian belief networks for dementia diagnosis and other applications: a comparison of hand-crafting and construction using a novel data driven technique. *PhD thesis: University of Stirling, Stirling, FK9 4LA, Scotland*, 2008.
- [136] Bayesia Ltd. Bayesialab. *available from <http://www.bayesia.com/>*, accessed on 24th of March 2011.
- [137] PJ Lucas, der van, LC Gaag, and A Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 3(30):20114, 2004.
- [138] D Luciani, S Cavuto, L Antiga, M Miniati, S Monti, M Pistolesi, and G Bertolini. Bayes pulmonary embolism assisted diagnosis: a new expert system for clinical use. *Emerg Med J*, 3(24):15764, 2007.
- [139] Alan Mackworth and David Poole. Cspace: Tools for learning artificial intelligence. *Laboratory for Computational Intelligence at the University of British Columbia. Available from <http://www.aispace.org/index.shtml>*, accessed on the 24th of March 2011.

- [140] D Madigan and A Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *J. Am. Statist. Association*, 89:1535–1546, 1994.
- [141] S Mani, M Valtorta, and S McDermott. Building bayesian network models in medicine: the mentor experience. *Appl Intell*, 2(22):93108, 2005.
- [142] N Mantel and W Haenszel. Statistical aspects of the analysis of data from retrospective study of disease. *J Natl cancer Inst*, 22:719–748, 1959.
- [143] C Marcelo, D Flather Shibata, and Wang Duolao. Systematic review of the impact of beta blockers on mortality and hospital admissions in heart failure. *European Journal of Heart Failure*, 48:351–357, 2001.
- [144] A Mark and MG Sidhom. Multidisciplinary care in oncology: medicolegal implications of group decisions. *Lancet Oncology*, 7(11):951–4, Nov 2006.
- [145] D Marquez, M Neil, and N Fenton. Solving dynamic fault trees using a new hybrid bayesian network inference algorithm. in *16th Mediterranean Conference on Control and Automation (MOD 08) Ajaccio, Corsica, France.*, 2008.
- [146] D Marquez, M Neil, and N Fenton. Improved dynamic fault tree modelling using bayesian networks. *The 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2007*, Edinburgh 2007.
- [147] D Marquez, M Neil, and Fenton N. Reliability modelling using hybrid bayesian networks. in *ISBIS-2008 International Symposium on Business and Industrial Statistics. Prague, Czech Republic.*, 2008.
- [148] AH Marshall, SI McClean, CM Shapcott, IR Hastie, and PH. Millard. Developing a bayesian belief network for the management of geriatric hospital care. *Health Care Manag Sci*, 1(4):2530, 2001.
- [149] Sara Martino and Havard Rue. Approximate bayesian inference using integrated nested laplace approximation. *Department of Mathematical Sciences, NTNU, Norway*. Available from <http://www.math.ntnu.no/hrue/GMRFsim/>, accessed on the 24th of March 2011.
- [150] Egger Matthias. Meta-analysis: Potentials and promise. *BMJ*, 315, 197.
- [151] IJ McKendrick, G Gettinby, Y Gu, SW Reid, and CW Revie. Using a bayesian belief network to aid differential diagnosis of tropical bovine diseases. *Prev Vet Med*, 3(47):14156, 2000.
- [152] MesdicineNet.com. Seroconversion: Medical definitions of popular medical terms easily defined on medterms. <http://www.medterms.com/art.asp?articlekey=9388>, accessed on March 10 2011.
- [153] M Meyer and J Booker. Eliciting and analyzing expert judgement. practical guide. *Cambridge, Academic Press.*, 1991.
- [154] Microsoft. Winmine. Available from <http://research.microsoft.com>, accessed on 10th of February 2011.
- [155] Microsoft. Microsoft bayesian network editor. Available from <http://research.microsoft.com/msbnx/>, accessed on the 24th of March 2011.

- [156] R Montironi, WF Whimster, Y Collan, PW Hamilton, D Thompson, and PH Bartels. How to develop and use a bayesian belief network. *J Clin Pathol*, 3(49):194201, 1996.
- [157] Wong Moore. Optimal reinsertion. Available from <http://www.autonlab.org/autonweb/10530>, accessed on the 24th of March 2011.
- [158] K Murphy. An introduction to graphical models. *Technical report, University of British Columbia*, 2001.
- [159] K Murphy. Bayes net toolbox for matlab (bnt). available from <http://code.google.com/p/bnt/>, accessed on the 17th of July 2011.
- [160] K Murphy. Software packages for graphical models / bayesian networks. Available from <http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html>., October 2005. (Accessed on July 18, 2011).
- [161] P Myllymki, T Silander, H Tirri, and P Uronen. B-course: A web-based tool for bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools*. Available from <http://b-course.cs.helsinki.fi/obc/>, 11(3):369–387, 2002, accessed on the 24th of March 2011.
- [162] M Negrn and F Vzquez-Polo. Bayesian cost-effectiveness analysis with two measures of effectiveness: the cost-effectiveness acceptability plane. *Health Econ*, 4(15):36372, 2006.
- [163] M Neil, X Chen, and NE Fenton. Optimizing the calculation of conditional probability tables in hybrid bayesian networks using binary factorization. *IEEE Transactions on Knowledge and Data Engineering*, 2011.
- [164] M Neil, N Fenton, and M Tailor. Using bayesian networks to model expected and unexpected operational losses. *Risk Analysis: An International Journal*, 2005.
- [165] M. Neil, D Marquez, and N Fenton. Using bayesian networks to model the operational risk to information technology infrastructure in financial institutions. *Journal of Financial Transformation*, 2008.
- [166] M Neil, D Marquez, and NE Fenton. Improved reliability modeling using bayesian networks and dynamic discretization. *Reliability Engineering & System Safety*, 4(95):412–425, 2010.
- [167] M Neil, M Tailor, D Marquez, NE Fenton, and P Hearty. Modelling dependable systems using hybrid bayesian networks. *Reliability Engineering and System Safety*, 93(7):954–955, 2008.
- [168] Martin Neil, Manesh Tailor, and David Marquez. Inference in hybrid bayesian networks using dynamic discretisation. *Stat Comput*, pages 219–233, 2007.
- [169] G Nikiforidis and G Sakellaropoulos. Expert system support using bayesian belief networks in the prognosis of head-injured patients of the icu. *Med Inform (Lond)*, 1(23):118, 1998.
- [170] D Nikovski. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Trans Knowledge Data Eng*, 4(12):50918, 2000.

- [171] JP Nordmann and G Berdeaux. Use of a bayesian network to predict the nighttime intraocular pressure peak from daytime measurements. *Therapeutics*, pages 1751–1760.
- [172] Breslow Norman. Odds ratio estimators when the data are sparse. *Biometrika*, 68:73–84, 1981.
- [173] Markku Nurminen. To use or not to use the odds ratio in epidemiologic analyses? *European Journal of Epidemiology*, 365371.
- [174] EA Oakley and PL Barnett. Is acid base determination an accurate predictor of pyloric stenosis? *J Paediatr Child Health*, 36:587–9, 2000.
- [175] University of Helsinki. Bassist: A tool for mcmc simulation of statistical models. *Department of Computer Science, University of Helsinki*. Available from <http://www.cs.helsinki.fi/research/fdk/bassist/>, accessed on 24th of March 2011.
- [176] Radiology Society of North America and American College of Radiology. The radiology information resource for patients. available from <http://www.radiologyinfo.org/>, accessed on 27 Jan 2011.
- [177] J Pearl. A constraint-propagation approach to probabilistic reasoning. In: L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence*, Amsterdam, North Holland, pages 3718–382, 1986.
- [178] J Pearl. Fusion, propagation and structuring in belief networks. *UCLA Computer Science Department Technical Report 850022 (R-42); Artificial Intelligence*, 29:241–288, 1986.
- [179] J Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. Morgan Kaufmann Publishers: San Francisco, California, 1988.
- [180] J Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
- [181] K Pearson. Report on certain enteric fever inoculation statistics. *BMJ*, 3:12436, 1904.
- [182] Roberts Penny. Dbnbox: Matlab dynamic bayes nets toolbox. available from <http://www.robots.ox.ac.uk/~parg/software.html>, accessed on the 24th of March 2011.
- [183] JF Lucas Peter, Boot Henk, and Taal Babs. A decision-theoretic network approach to treatment management and prognosis. *Knowledge-Based Systems*, 11:321330, 1998.
- [184] JF Lucas Peter, de Bruijn Nicolette, Schurink Karin, and Hoepelman Andy. A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. *Artificial Intelligence in Medicine*, 19(3):251–279, 2000.
- [185] Diana B Petitti. *Meta-Analysis, Decision Analysis AND Cost-Effective Analysis*. Oxford university press, second edition edition, 1997.
- [186] U. Pittsburgh. Structural modeling, inference, and learning engine. available from <http://genie.sis.pitt.edu/>, accessed on the 24th of March 2011.

- [187] RL Prentice and DB Thomas. On the epidemiology of the oral contraceptive and disease. *Adv Cancer Res*, 49:285–401, 1986.
- [188] Rumí Rafael and Salmerón Antonio. Estimating mixtures of truncated exponentials in hybrid bayesian networks. *Sociedad de Estadística e Investigación Operativa Test*, 15(2):397–421, 2006.
- [189] S Renooij. Probability elicitation for belief networks: Issues to consider. *The Knowledge Engineering Review*, 3(16):255–269, 2000.
- [190] Gabriel JE Rinkel, Mamuka Djibuti, Ale Algra, and Gijn J van. Prevalence and risk of rupture of intracranial aneurysms: A systematic review. *American heart Association*, 1997.
- [191] J Rissanen. Stochastic complexity (with discussion). *J. Roy. Statist. Soc. B*, 49:223–239, 1987.
- [192] AS Robbins, SY Chao, and VP Fonseca. What’s the relative risk? a method to directly estimate risk ratios in cohort studies of common outcomes. *Annals of Epidemiology*, October 2002.
- [193] Dodier Robert. Riso: distributed belief networks. *available from <http://sourceforge.net/projects/riso/>*, accessed on 10th of February 2011.
- [194] DB Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.
- [195] DB Rubin. Bayesian inference for causal effects: The role of randomisation. *Annals of Statistics*, 6:34–58, 1978.
- [196] DB Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47:1213–1234, 1991.
- [197] L Rychetnik, P Hawe, and E Waters. A glossary of evidence based public health. *J Epidemiol Community Health.*, 2004.
- [198] D Sackett, S Straus, and W Richardson. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, 2000.
- [199] DL Sackett, RB Haynes, and GH Guyatt. *Clinical Epidemiology. A Basic Science for Clinical Medicine*. Boston: Little, Brown and Company, 1991.
- [200] DL Sackett and D Rennie. The science of the art of the clinical examination. *JAMA*, 267:2650–2652, 1992.
- [201] DL Sackett, WS Richardson, W Rosenberg, and RB Haynes. *EvidenceBased Medicine: How to Practise and Teach EBM*. New York:Churchill Livingstone, 1997.
- [202] DL Sackett, WC Rosenberg, and JAM Gray. Evidence-based medicine: What it is and what it isn’t. *BMJ*, 312:71–72, 1996.
- [203] GC Sakellaropoulos and GC Nikiforidis. Development of a bayesian network for the prognosis of head injuries using graphical model selection techniques. *Methods Inf Med*, 1(38):742, 1999.

- [204] D Sanders and D Aronsky. Prospective evaluation of a bayesian network for detecting asthma exacerbations in a pediatric emergency department. *AMIA Annu Symp Proc*, (1085), 2006.
- [205] DL Sanders and D Aronsky. Detecting asthma exacerbations in a pediatric emergency department using a bayesian network. *AMIA Annu Symp Proc*, (1085):6848, 2006.
- [206] Kumar Santosh. Evidence-based medicine: The new paradigm of clinical decision making. *Indian Journal of Urology*, pages 73–78, 2000.
- [207] Richard Scheines, Peter Spirtes, Clark Glymour, and Christopher Meek. The tetrad project: Causal models and statistical data. available from <http://www.phil.cmu.edu/projects/tetrad/tetrad2.html>, 1994 (accessed on 10th of February 2011).
- [208] R Shachter. Intelligent probabilistic inference. In *J. F. Lemmer and L. N. Kanal, editors, Uncertainty in Artificial Intelligence, North Holland, Amsterdam*, pages 371–382, 1986.
- [209] R Shachter. Evidence absorption and propagation through evidence reversals. In *M. Henrion, R. D. Shachter, J. F. Lemmer, & L. N. Kanal (Eds.), Uncertainty in Artificial Intelligence*, 5:173–190, 1990.
- [210] RD Shachter, B D’Ambrosio, and BD Del. Favero. symbolic probabilistic inference in belief networks. *Proc. 8th National Conference on Artificial Intelligence, MIT Press, Boston*, page 126131, 1990.
- [211] Straus Sharon, Richardson Scott, Glasziou Paul, and Haynes Brian. *Evidence-Based Medicine: How to practice and teach EBM*. Elsevier Churchill Livingstone, 2005.
- [212] P Shenoy and G Shafer. Axioms for probability and belief-function propagation. in shachter, r. et al. *Uncertainty in AI*, 4:169–198, 1990.
- [213] GA Smith, L Mihalov, and BJ Shields. Diagnostic aids in the differentiation of pyloric stenosis from severe gastroesophageal reflux during early infancy: the utility of serum bicarbonate and serum chloride. *Am J Emerg Med*, 17:28–31, 1999.
- [214] TC Smith, DJ Spiegelhalter, and A Thomas. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine*, 14:26852699, 1995.
- [215] SourceForge. Unbbayes. Available from <http://sourceforge.net/projects/unbbayes/>, accessed on 10th of February 2011.
- [216] DJ Spiegelhalter, JP Myles, DR Jones, and KR Abrams. Bayesian methods in health technology assessment: a review. *Health Technol Assess*, 38(4), 2000.
- [217] DJ Spiegelhalter, A Thomas, and NG Best. Winbugs version 1.3 user manual mrc biostatistics unit, cambridge 2000. Available from <http://www.mrc-bsu.cam.ac.uk/bugs>, last access on July 17, 2011.
- [218] JAC Sterne, MJ Bradburn, and M Egger. *Meta-analysis in STATA*. In: *Egger M, Davey Smith G, Altman DG (eds). Systematic Reviews in Health Care: Meta-analysis in Context*. London: BMJ Books pp.34769., 2001.

- [219] AJ Sutton, KR Abrams, DR Jones, TA Sheldon, and F Song. Systematic reviews of trials and other studies. *HTA*, 2:19, 1998.
- [220] Lumina Decision Systems. Analytica: Risk analysis, monte carlos simulation and optimization software. Available from <http://www.lumina.com/>, accessed on 24th of March 2011.
- [221] Noetic systems. Ergo: Bayesian network editor by noetic systems. available from <http://www.noeticsystems.com/>, accessed on the 24th of March 2011.
- [222] Yoshimoto Takashi and Mizoi Kazuo. Importance of management of unruptured cerebral aneurysms. *4th international Workshop on Cerebrovascular surgery*, June 11-14 1995.
- [223] H Hamza Taye, C van Houwelingen Hans, and Stijnen Theo. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology*, 61:41–51, 2008.
- [224] KK Teo, S Yusuf, R Collins, PH Held, and R Peto. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ*, 303:1499503, 1991.
- [225] Wikipedia the free encyclopedia. Ischemia. <http://emedicine.medscape.com>, last accessed on 7 Mar. 2011.
- [226] SG Thompson, TC Smith, and SJ Sharp. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine*, 16:27412758, 1997.
- [227] Friedman (Hebrew U). Libb for windows/linux programs. Available from <http://www.cs.huji.ac.il/labs/compbio/LibB/>, (accessed on the 24th of March 2011).
- [228] Elisabeth Umkehrer, Hong Xu, and Alessandro Saffiotti. Pulcinella: Propagating uncertainty through local computations. *Universit Libre de Bruxelles available from <http://iridia.ulb.ac.be/pulcinella/Welcome.html>*, accessed on 10th of February 2011.
- [229] Warnes U.Wash. Hydra. Available from http://www.warnes.net/index_html?Tab=MCMC, laccessed on the 24th of March 2011.
- [230] Gaag LC Van Der and S Renooij. How to elicit many probabilities. *15th Conference of Uncertainty in Artificial Intelligence*, 1999.
- [231] YT van der Schouw, MT van der Velden, and *et al.* Hitge-Boetes C. Diagnosis of hypertrophic pyloric stenosis: value of sonography when used in conjunction with clinical findings and laboratory data. *AJR Am J Roentgenol*, 163:905–9, 1994.
- [232] I. Verdinelli, K. Andrews, K. Detre, and P. Peduzzi. The bayesian approach to meta-analysis: a case-study. *Working papper: Department of Statistics, Carnegie Mellon University, Pittsburgh*, Last access on July 17, 2011.
- [233] M Verduijn, N Peek, PMJ Rosseel, E de Jonge, and BA de Mol. Prognostic bayesian networks: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*,, pages 609–618.
- [234] M Verduijn, PM Rosseel, N Peek, E Jonge, and BA Mol. Prognostic bayesian networks ii: an application in the domain of cardiac surgery. *Journal of Biomedical Information*, 6(40):61930, 2007.

- [235] AJ Viera. Odds ratios and risk ratios: what's the difference and why does it matter? *South. Med.*, July 2008.
- [236] XH Wang, B Zheng, and WF *et al.* Good. Computer-assisted diagnosis of breast cancer using a data-driven bayesian belief network. *Int J Med Inf*, 54:115–26.
- [237] JM Wardlaw and PM White. The detection and management of unruptured intracranial aneurysms. *Brain: Oxford university press*, 123:205–221, 2000.
- [238] HR Warner, AF Toronto, and LG Veasy. Experience with bayes theorem for computer diagnosis of congenital heart disease. *Ann NY Acad Sci*, (115):55867, 1964.
- [239] R Winkler. The assessment of prior distributions in bayesian analysis. *American Statistical Association Journal.*, 62:776–800, 1967.
- [240] Winn and Bishop. Vibes: Variational inference for bayesian networks. Available from <http://www.inference.phy.cam.ac.uk>, accessed on 10th of February 2011.
- [241] FM Wolf. *Meta-Analysis: Quantitative methods for research synthesis*. Sage publication, 1986.
- [242] Viechtbauer Wolfgang. Meta-analysis package for r. <http://www.wvbauer.com/>, accessed on March 10 2011.
- [243] KL Woods. Possible pharmacological actions of magnesium in acute myocardial infarction. *Br J Clin Pharmacol*, 32:3–10, 1991.
- [244] KL Woods, S Fletcher, C Roffe, and Y Haider. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second leicester intravenous magnesium intervention trial (limit-2). *Lancet*, 339:155358, 1992.
- [245] S Wright. Correlation and causation. *Journal of Agricultural Research*, pages 557–585.
- [246] Xiang. Web weaver. *Department of Computer Science, University of Regina*. Available from <http://www.inference.phy.cam.ac.uk/jmw39/>, accessed on 10th of February 2011.
- [247] I Liu Yueyi, Kamaya Aya, S. Desser Terry, and L. Rubin Daniel. A bayesian classifier for differentiating benign versus malignant thyroid nodules using sonographic features. *AMIA 2008 Symposium Proceedings*, pages 419–423, 2008.
- [248] S Yusuf, R Peto, J Lewis, R Collins, and P Sleight. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *prog Cardiovasc dis*, 27:335–371, 1985.
- [249] J Zhang and KF Yu. What's the relative risk? a method of correcting the odds ratio in cohort studies of common outcomes. *Journal of America Medical Association*, 4(49), November 1998.
- [250] NL Zhang and D Poole. A simple approach to bayesian network computations. *In Proc. of the Tenth Canadian Conference on Artificial Intelligence*, pages 171–178, 1994.