Identifying Causes of Neonatal Mortality from Observational Data: A Bayesian Network Approach

K. A. Wilson[1,3], D. D. Wallace[1], S. S. Goudar[2], D. Theriaque[1], E. M. McClure[1]
[1]RTI International, Biostatistics and Epidemiology Division, Durham, NC, USA
[2]KLE University's Jawaharlal Nehru Medical College, Belgaum, Karnataka, India
[3]University of Liverpool, School of Medicine, Liverpool, UK

Abstract - *Despite improvements in access to birth facilities, neonatal mortality remains a critical health issue in many developing countries and causes are not fully understood. The Global Network Maternal Newborn Health Registry provides a rich source of data of neonatal mortality risk factors and outcomes to identify direct causes and higher-level determinants, however performing causal inference using observational data is difficult and remains an open problem in epidemiology. In this paper we sought to determine whether Bayesian networks can be used to identify the complex causal pathways leading to neonatal mortality outcomes and to quantify the effect of each cause on mortality. Our analysis identified a complex network of causes that contribute to neonatal mortality, including maternal death, pre-term birth, movement and breathing at birth. For variables identified as direct causes we estimated the average causal effect using logistic regression models that controlled for known confounders.*

Keywords: Causal inference, Bayesian network, neonatal mortality.

1 Introduction

While there has been a significant reduction in neonatal deaths from 5.6 million per year in 1990 to 4.0 million per year in 2000, neonatal mortality remains a major global public health issue [1]. Of the 130 million children born annually, approximately 4 million will die in the first month of life, 75% within the first 7 days and 25% in the first 24 hours [2]. According to the UN, the 3.6 million neonatal deaths that occurred in 2008 comprised 41% of all deaths under age 5. This underscores the importance of reducing neonatal mortality, and has been formalized as the fourth UN Millennium Development Goal [3]. To meet this goal of a two-thirds reduction in mortality of children under age 5 by 2015, the current rate of improvement must be increased 6-fold.

In India, while the overall neonatal mortality is 31 deaths per 1000 live births, the rate varies widely by region and birth facility [4]-[5]. And, despite improvements in access to birth facilities, neonatal mortality remains high, suggesting that the causes of neonatal death may be more complex than previously thought [6]. The Global Network Maternal and Child Health Registry provides a unique source of data relating to maternal and neonatal risk factors that can potentially explain these complex causal relationships. More than 70 variables were collected from pregnant women enrolled at 20 geographic clusters in Belgaum, India, including demographics, antenatal care, maternal and neonatal health conditions, delivery characteristics, and medical treatments. Although this data set is comprehensive, identifying the complex causal pathways between risk factors and outcomes is challenging due to it being observational in nature [5].

Observational studies are particularly susceptible to selection bias and confounding, which can result in biased estimates of effect [7]. Under certain assumptions, Bayesian networks (BNs) have shown promise in performing causal inference using observational data. So-called causal BNs can be used to model relationships between random variables, where the direction of the edges in the graph signifies a direct causal relationship [8]. Algorithms exist to identify the graph structure directly from data in the presence of confounding and selection bias [9]-[10]. Once the causal structure has been identified, the BN can be used to estimate the effect of manipulating key variables on a specified outcome variable, such as neonatal mortality [11]. Thus the BN approach promises to be a useful technique for identifying the causes of neonatal mortality given a rich observational data set.

The goal of this work is to extend and enhance existing Bayesian network methods to perform causal inference and to estimate causal effects of neonatal mortality using observational data from the Global Network Maternal and Child Health Registry. The remainder of this article is organized as follows. In section 2 we discuss the challenges of causal inference and approaches to overcome some of these challenges to obtain valid inferences based on analyses of observational data. In section 3 we describe our Bayesian network-based methods of identifying causal factors and estimating effects. Our results are presented and discussed in sections 4 and 5. In section 6, we present our conclusions and ideas for future work.

2 Background and Related Work

2.1 Causal Inference

The fundamental problem of causal inference is that it is not possible to measure the difference in outcome for an

individual for different levels of a variable of interest [12]. As a result, estimating a causal effect can only be accomplished by comparing groups of similar individuals at different levels of a given variable. To ensure that the true causal effect is estimated, this comparison requires both the *manipulation* of a variable and measurement of the change in the outcome variable while *accounting for clinical and environmental variables that could confound the conclusions about the variable of interest*. Valid causal inference is often achieved in randomized controlled trials through the use of an intervention, with the randomized assignment to this intervention, which theoretically balances known and unknown confounders across treatment groups.

In comparison, observational studies are problematic because of non-random group assignment and the absence of manipulation [13]. As a result, measures of effect can easily biased due to confounding, and thus estimating the average causal effect of changes in one variable on an outcome of interest requires controlling for potential confounders, some of which may be unobserved [14]. Most analytical methods used with observational data focus on ensuring that comparison groups are as similar as possible with respect to measured and unmeasured confounders [13]-[16]. However, the absence of a true manipulation or intervention, at best, results in unbiased estimates of association and not causal inference. Bayesian networks, and in particular, *causal Bayesian networks* can potentially address this weakness. The reader is referred to the seminal work by Pearl for a more complete discussion of causal inference algorithms [8].

2.2 Bayesian Networks

A Bayesian network (BN) is a probabilistic graphical model, in which the nodes in a directed acyclic graph represent random variables and the edges represent probabilistic associations between the variables. A BN models the joint distribution over all the variables in the graph, factored into a series of conditional probability distributions, resulting in a compact and efficient representation [17], [11].

Spirtes' *PC-algorithm* can be used to learn a causal BN [14]. The algorithm performs a series of conditional independence tests to determine directed relationships between the variables [18]. Kalisch and Bühlmann achieve a true positive rate of over 80% and false positive rate of less than 1 percent [18]. Nguefack-Tsangue and Zucchini argue that in the absence of unmeasured confounders, causal BNs are able to identify all causal relationships up to sampling error [19]. Shrier and Platt confirmed that this approach does not introduce additional conditional associations or bias [20]. Li, Shi and Satz used the *PC-algorithm* to successfully estimate the causal relationship between risk factors and disease using case-control data [7]. Kalisch et al. provide an efficient implementation that supports both categorical and continuous variables in *R* [21].

2.3 Estimating Causal Effects

Estimating causal effects from observational data can be achieved by simulating an intervention on a variable, a process known as manipulation. Pearl provides a theoretical background for estimating causal effects through the *do()* operator, which performs a manipulation on the variable of interest while accounting for clinical and environmental variables not on the causal pathway that could confound conclusions about the variable of interest. [12]. With this approach, parent nodes of the manipulated variable are included as covariates, a process known as adjusting for the direct causes, which captures the prior state of the probability distribution. Applying a manipulation to a variable removes the influence of any other variables and sets the value of that variable for all members of the sample. Maathuis, Kalisch and Bülmann show that the average causal effect can be estimated using a linear regression model, and that this approach is equivalent to Pearl's *do()* operator [22]. This method is implemented as the *ida* algorithm by Kalisch et al. in the *pcalg R* package [21]. One limitation of this implementation is its use of linear regression to estimate causal effects, which prevents it from being used to estimate the casual effect on a dichotomous variable.

2.4 Bayesian Network Assumptions

In standard BNs the directions of the edges do not imply any specific causal direction and probabilistic inference is agnostic to the directions of the edges. For a BN to be causal, additional assumptions are required, including the Causal Markov Assumption and the Causal Faithfulness Assumption. The Causal Markov Assumption attributes a direct causal relationship when two variables are connected by a directed edge, and states that each variable is independent of its non-effects given its causes [23]. The Causal Faithfulness Assumption states that the graph structure and the independence relationships in the data are isomorphic [10]. Additional assumptions include the absence of hidden common causes, causal feedback loops, and selection bias [24]. While methods exist to accommodate the existence of unmeasured confounders, causal effect estimates are undefined in these latent confounder models. As a consequence, most methods assume that all potential confounders are included in the graph. In this case, un-confounded estimates of causal effects can be determined [22].

3 Methods

Our methods consist of three steps: data processing, learning the optimal causal Bayesian network, and estimating the causal effects for direct causes of neonatal mortality.

3.1 Data Processing

Data were collected on all mothers and nenonates at three time points. At enrollment, basic demographic information was collected for all eligible and consented women. Maternal and neonatal outcomes were collected at the time of delivery and subjects were followed up at 42 days after birth to collect the 28-day neonatal mortality outcome.

Data from these time points were combined into a single analysis dataset using SAS 9.3, with one observation for each birth outcome. Data that were missing due to skip

patterns in the data collection forms were coded as "not collected." All variables were categorical except for hemoglobin level and BMI. These continuous variables were discretized using standard categories

Missing data analysis was performed for all variables included in the model. We assumed data were missing at random. Where the amount of missing data was significant and could potentially introduce bias, multiple imputation was performed prior to the estimation of causal effects. Imputation was performed in R using the MICE package, using polytomous regression with 20 imputed datasets [25]. To test our missing at random assumption and to ensure that the imputation did not introduce bias into the causal effect estimates, we built models based on the original un-imputed data and performed a sensitivity analysis.

The final analysis dataset contained 70 variables and 60,985 observations.

3.2 Learning the Causal Bayesian Network

The causal Bayesian network was learned using the *PC-Algorithm*, which was initially developed by Spirtes et al. and implemented in the *R* package, *pcalg*, by Maathuis, Kalisch and Bülmann [26], [22]. Stacked output from multiple imputation was used as the training dataset.

The PC-Algorithm is a constraint-based algorithm that estimates the conditional probability distribution over all variables using a series of conditional independence tests. One problem with this approach is the use of multiple comparisons, which can result in false positives. In the context of a causal Bayesian network, a false positive implies that two nodes are not independent given a set of conditioning nodes. Residual dependence after conditioning results in a graph that less sparse, where spurious causal relationships are uncovered. To address this issue, we treated the P-value used in the conditional independence tests as a tuning parameter and built a series of models using different P-values. We optimized the P-value using the Bayesian Information Criterion (BIC), and selected the model with the lowest BIC. A P-value of 0.0005 was used to learn the final model.

3.3 Estimating Causal Effects

One limitation of the PC-Algorithm, and constraint-based methods in general, is an inability to learn a unique Bayesian network. This problem arises from a failure to uniquely identify a network's structure using only conditional independence tests, as multiple graphs can encode the same conditional independencies. We addressed this issue through the development of an enhanced method for estimating causal effects, which we have named *ida+*. Our method is an extension of the *ida* method developed by Maathuis, Kalisch and Bülmann [22].

The *ida* algorithm uses the Markov blanket of a specific variable to build a multiple linear regression model and estimate the causal effect of predictor variable on an outcome using the direct parents of the predictor as covariates in the model. Because the PC-Algorithm is often unable to identify a single causal Bayesian network, the *ida* algorithm returns a multi-set of possible causal effects. An additional limitation of this method is that the estimation of causal effect may not be valid when the outcome is dichotomous or multinomial. Thus, our *ida+* algorithm incorporates several key enhancements:

- Logistic regression is used to estimate causal effects for dichotomous outcomes;
- Polytomous regression is used to estimate causal effects for categorical variables with more than two outcomes;
- The Cox goodness of fit test for non-nested models is used to determine which of the multiset of possible causal effect estimates is most likely correct;
- Confidence intervals, standard errors, and p-values are returned to quantify the precision of the estimates.

Logistic and polytomous regression models enable the estimation of odds ratios for categorical outcomes. The Cox goodness of fit test for non-nested models determines the best set of covariates for a model on the principle that if a given model contains the correct covariates then fitting a second model to these covariates should add no explanatory value [27]. Because calculated odds ratios are estimates subject to sampling error, quantifying their precision is essential.

The *ida+* algorithm is shown in Fig. 1.

```
Input: Set of Causal BNs (G),   Predictor (x),
  Outcome (y), Outcome type {linear | logistic
                    | polytomous}
Output: Causal Effect of Predictor on Outcome with
95% confidence intervals and p-value

for each graph in G {
  if y in parents(x)
     model ← null
  else
  {
      if length(parents(x)) > 0
      {
         model <- glm(y ~
           x + parents(x))
      }
      else
      {
         model ← glm(y ~ x)
      }
   }
   model_array ← model
}

lowest_p_val ←1
correct_model ← null

for each model in model_array {
  if p_value(model) < lowest_p_val
  {
     lowest_p_val <-
             p_value(model)
     correct_model ← model
   }
  return correct_model
```

Fig. 1. The *ida+* algorithm.

4 Results

Fig. 2 provides a simplified view of the Bayesian network generated by the PC-Algorithm with P=0.0005 and neonatal (28-day) mortality as the outcome. The summarized view is presented for clarity and includes only the outcome variable, its direct causes, and the parents of the direct causes. The algorithm identified 9 direct causal factors of neonatal mortality: maternal mortality, gender, pre-term birth, multiple birth, whether the baby moved upon birth, whether the baby was breathing when born, the presence of one or more neonatal conditions, whether transport was available if a hospital referral was needed, and whether the neonate was seen at a facility. For each of these causal factors, direct upstream causes were also identified and the relationship between all variables in the model can be seen.

For each of the direct causes, the *ida+* algorithm estimated the average causal effect using a logistic regression model with neonatal mortality as the dependent variable, the direct causes as the primary independent variable, and the parents of the direct causes as covariates in the model. The overall causal effect of the 9 direct causes is displayed in Table 1 along with 95% confidence intervals and P-values. The effect estimate is an odds ratio calculated as the exponent of the beta coefficient of the primary dependent variable in each model. Included covariates for each model are summarized in Table 2. The addition of some covariates introduced multicollinearity into the models. Multicollinearity generally occurred as a result of the structure of the questionnaires that generated the dataset. For example, the Bayesian network model shows Maternal Cause of Death as a direct cause of Maternal Mortality. Multicollinearity was likely introduced because cause of death was not collected for mothers that did not die. A similar phenomenon occurred with hospital referral and admission variables. As a result of multicollinearity, the estimates produced were not deemed reliable, and these covariates were dropped from the model.

Maternal mortality (OR: 7.972), gender (OR: 1.264), pre-term birth (OR: 1.247), neonatal conditions (OR: 21.704), breathing (OR: 9.974) and movement of the baby (OR: 30.139) all exhibit a substantial and significant effect on neonatal mortality, with baby movement and neonatal conditions having the largest effects. Multiple births appears to have a protective effect with an odds ratio of 0.774. Nenonate Seen at Facility is uninformative, likely due to the multicollinearity issues described above.

5 Discussion

The main finding of our research is that constraint-based methods of learning Bayesian networks can be used to identify direct and in-direct causes of neonatal mortality from an observational data source, and that the effects of these causes can be estimated using logistic regression models that control for appropriate confounders. The constraint-based PC-Algorithm identified 9 direct causes of neonatal mortality: maternal mortality, gender, pre-term birth, movement at birth, breathing at birth, presence of neonatal conditions, transport to facility, and neonate seen at facility. The use of the Cox goodness of fit test for non-nested models employed by our *ida+* algorithm was able to disambiguate multiple possible Bayesian networks, identify the single most likely graph, and estimate the causal effect of the 9 component causes on the mortality outcome. In contrast to standard associational approaches, such as linear or logistic regression modeling, the Bayesian network was able to identify more complex relationships between variables.

The causal effects shown in the results tables represent the odds of a neonate dying within 28 days of birth when the given variable is manipulated with an intervention and all other variables are held constant. In contrast to standard observational approaches, these estimates give greater insight into the impact of these direct causes on the mortality outcome in a situation where direct, real intervention with a randomized controlled trial is infeasible, primarily for ethical reasons. The causes identified by the algorithm can be considered component causes that contribute to the overall cause of morality. The largest causal factors are maternal death (8 times increase in odds), movement at birth (24 times increase in odds), breathing at birth (10 times increase in odds), and the presence of one of a number of neonatal health conditions (baby stopped feeding, high fever, hypothermia, difficulty breathing, bleeding from umbilicus – 22 times increase in odds). While the correctness of the graph cannot be determined formally, in general, the algorithm was able to identify several major causes of neonatal mortality. Developing public health interventions aimed at prevention or treatment of these causes should result in reduced mortality.

Fig. 2. Simplified Bayesian network for identified causes of 28-day mortality.

Table 1. Causal estimates for direct causes of 28-day mortality.

Variable (Reference Value)	*Causal Effect*	*Lower 95% CI*	*Upper 95% CI*	*P-Value*
Maternal Mortality (No)				
Yes	7.972	3.736	17.010	0.000
Gender (Female)				
Male	1.264	1.129	1.414	0.000
Pre-term Birth (Term (≥ 37 wks))				
Preterm (< 37 wks)	1.247	0.951	1.636	0.110
Multiple Birth (No)				
Yes	0.774	0.585	1.025	0.074
Movement at Birth (Yes)				
No	30.139	24.285	37.404	0.000
Breathing at Birth (Yes)				
No	9.974	7.995	12.443	0.000
Neonatal Conditions Present (No)	21.704			
Yes		17.879	26.347	0.000
Transport to Facility (Yes)				
No	0.984	0.259	3.738	0.981
Neonate Seen at Facility (Baby dead at arrival)				
Did not reach facility	0.000	0.000	Inf	0.972
No	0.000	0.000	Inf	0.970
Yes	0.000	0.000	Inf	0.966

The validity of the estimates of causal effects relies on the ability of the PC-Algorithm to correctly identify the causal relationships in the data and to reflect these relationships in the structure of the Bayesian network. In the absence of test data, it is impossible to formally validate the correctness of the resultant network, although Maathuis, Kalisch and Bülmann argue that the PC-Algorithm is guaranteed to uncover the correct causal graph up to sampling error [22]. It is difficult to determine whether this statement is true and the degree to which it is necessary to adhere to the underlying assumptions of the model. Nevertheless, the fact that the model identified pre-term birth and neonatal health conditions is consistent with the literature, particularly Bassani et al. who argue that pre-term birth, low-birth weight and neonatal health conditions account for 78% of all neonatal deaths in India. Although low birth weight is not identified in the Bayesian networks as a direct cause, it clearly defines several causal pathways that lead to neonatal mortality: it is shown to cause neonatal health conditions, which in turn causes neonatal mortality, and it also appears to be strongly associated with facility referral and pre-term birth, although the directionality of the pathways in these cases is questionable [28]. Additional factors, such as antenatal care and the administration of cost-effective interventions discussed by Bhaumik are reflected in the Bayesian network as higher-level determinants [29]. In fact, lack of antenatal care is on the causal pathway for neonatal mortality and has a direct effect on the presence of neonatal health conditions, which in turn affects mortality. There are also a number of spurious causal relationships, such as the association of antenatal care with hemoglobin level. Although this relationship is present in the data from a probabilistic perspective, hemoglobin is likely only collected during antenatal visits, and as a result, this pathway introduces bias into the model.

There are a number of limitations to our research. While the direct causes of mortality identified are consistent with the literature, some of the indirect causes appear to be problematic. For example, Transport to Facility is identified as a cause of Pre-term Birth. While there is clearly an association between these two variables, it is more likely that Pre-term birth is a cause of Transport to Facility. Thus, the model was unable to correctly orient the edge between these variables. Another example is the identification of Bag and Mask Resuscitation as a cause of Neonatal Conditions, which is also likely to be reversed.

These errors in identification could be attributed to a number of factors, including lower sample sizes of these higher-level causes resulting in a lack of power to detect the true relationships, an absence of temporal information (e.g., the fact that Neonatal Conditions must occur before Bag and Mask Resuscitation is used), and the inability of current methods to extract this information from the conditional probability distribution. In addition, the lack of formal evaluation of the Bayesian network or the causal effect estimates is a weakness. The best solution to this problem would be the use of an independent validation dataset; this approach would also assess the generalizability of the model. A more viable approach, however, would to use cross-validation techniques to assess the fit of the model to held-out data using an objective metric, such as Bayesian Information Criterion. One additional limitation is the lack of validation of causal estimates, through traditional approaches, such as cross

validation. However, a comparison of causal effects and associations estimated using standard regression models provides some useful insight. For example, for neonatal mortality, an intervention on maternal mortality has an estimated effect of 7.972 (95% CI: 3.3736 – 17.010), which is substantially larger than the association odds ratio of 2.299 (95% CI: 0.554 – 9.538). Therefore, an intervention on maternal mortality results in an 8-times increased risk of neonatal mortality, whereas the association when controlling for other factors, results in only a 2 times increase in risk. Similar differences in effect (including for protective factors) are noted for the other direct causes. One additional limitation of our methodology is the relatively strong assumption of no unmeasured confounders. In practice, unmeasured confounders very likely exist, and our inability to account for these may limit our ability to identify uncounfounded causes and higher level determinants of neonatal mortality.

Table 2. Covariates included in each logistic regression model.

Cause	*Covariates*
Maternal mortality	None
Gender	Parity
	Antenatal Location
	Birth Location
	Fetal Heartrate
Pre-term Birth	Cluster Resident
	Birth Weight
	Transport to Facility
Multiple Birth	Age of Mother
	Antenatal Location
	Maternal Conditions
	Maternal Mortality
	Birth Attendant
	Birth Location
	Birth Weight
	Pre-term Birth
Baby Move	Born in Cluster
Baby Breathe	Baby Move
	Baby has Heartbeat
Neonatal Conditions	Age of Mother
	Antenatal Location
	Maternal Conditions
	Birth Weight
	Pre-term Birth
	Multiple Birth
	Baby Breathe
	Bag and Mask Resuscitation
Transport to Facility	Birth Weight
	Multiple Birth
	Neonatal Conditions
	Oxygen Treatment
Neonate Seen at Facility	Prenatal Vitamins
	Multiple Birth

These limitations point to a number of paths for future research. Improved methods of learning and evaluating causal Bayesian networks are needed, along with methods to evaluate the accuracy of causal effect estimates. One possible approach is to compare results of these methods with results from a RCT where an actual intervention was performed. Theoretically, odds ratios obtained from an RCT should be equivalent to those generated by these methods. Further work with this dataset should include validation of the graph and estimates with cross-validation approaches.

6 Conclusions

Although formal validation of results is needed, this research has demonstrated the promise of Bayesian networks as a method for identifying causal factors from observational data. The methods described, and the specific application to neonatal mortality, are of strong public health relevance for several reasons. First, the ability to perform causal inference from observational data is a critical issue in epidemiology where a major emphasis in any study is the identification and control of confounding factors, particularly when conduct of a randomized controlled trial is not possible. Second, the inductive nature of the Bayesian network learning algorithms provides an opportunity to uncover previously unknown causal factors and pathways. Although these methods are imperfect, their use in exploratory data analysis can augment traditional research hypothesis generation. Application of these tools can thus inform future research studies, increasing our ability to the identify causes of, and develop effective interventions for, critical public health issues.

7 Acknowledgements

Data were originally collected by the Global Network for Women's and Children's Health funded by grants U01HD042372 and U01HD040636 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the US National Institutes of Health. All participants signed informed consent prior to study participation. This secondary analysis was conducted under the auspices of the University of Liverpool and RTI International. Institutional Review Board approval was obtained from both organizations prior to gaining access to the data. The authors acknowledge the support of Belgaum site of the Global Network in conducting this research.

8 References

[1] Lawn, J. E., Kerber, K., Enweronu-Laryea, C. & Cousens, S. (2010). '3.6 million neonatal deaths — what is progressing and what is not', Semin Perinatol, 34, pp.371-386, [Online]. Available from: http://www.healthynewbornnetwork.org/sites/default/files/resources/Epidemiology_Lawn.pdf (Accessed: 26 November 2013).

[2] Jehan, I. et al. (2009). 'Neonatal mortality, risk factors and causes: a prospective population-based cohort study in urban Pakistan', Bulletin of the World Health Organization, 87, pp.130-138, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636189/ (Accessed: 30 November 2013).

[3] United Nations (2013). 'Goal 4: Reduce Child Mortality', Millennium Development Goals, [Online]. Available from: http://www.un.org/millenniumgoals/childhealth.shtml (Accessed: 7 September 2013).

[4] World Bank (2013). 'Data: Mortality Rate, Neonatal (per 1,000 live births)', [Online]. Available from: http://data.worldbank.org/indicator/SH.DYN.NMRT (Accessed: 9 November 2013).

[5] Goudar, S.S. et al. (2012a). 'The maternal and newborn health registry study of the Global Network for Women's and Children's Health research', International Journal of Gynecology & Obstetrics, 118 (3), pp.190-193, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22738806 (Accessed: 30 November 2013).

[6] World Health Organization (2013). 'Newborn health epidemiology', Maternal, newborn, child and adolescent health, [Online]. Available from: http://www.who.int/maternal_child_adolescent/epidemiology/newborn/en/index.html (Accessed: 7 September 2013).

[7] Li, J., Shi, J. & Satz, D. (2008). 'Modeling and analysis of disease and risk factors through learning bayesian networks from observational data', Qual Reliab Engng Int, 24, pp.291-302, [Online]. Available from: http://141.213.232.243/bitstream/handle/2027.42/58076/893_ftp.pdf?sequence=1 (Accessed: 30 November 2013).

[8] Pearl, J. (2009). Causality. Cambridge University Press.

[9] Kleinberg, S. & Hripcsak, G. (2011). "A review of causal inference for biomedical informatics', J Biomed Inform, 44 (6), pp.1102-1112, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21782035 (Accessed: 30 November 2013).

[10] Cooper, G. F. (1999). An overview of the representation and discovery of causal relationships using Bayesian networks. In: Glymour, C and Cooper, G. F. Computation, Causation, and Discovery. AAAI Press.

[11] Darwiche, A. (2010). 'Bayesian Networks', Communications of the ACM, 53 (12), pp.80-90, [Online]. Available from: http://cacm.acm.org/magazines/2010/12/102122-bayesian-networks/abstract (Accessed: 30 November 2013).

[12] Höfler, M. (2005). 'Causal inference based on counterfactuals', BMC Medical Research Methodology, 5 (28), [Online]. Available from: http://www.biomedcentral.com/1471-2288/5/28 (Accessed: 3 January 2014).

[13] Trojano, M. et al. (2009). 'Observational studies: propensity score analysis of non-randomized data', The International MS Journal, 16, pp.90-97, [Online]. Available from: http://www.msforum.net/journal/download/20091690.pdf (Accessed: 3 January 2014).

[14] Spirtes, P. (2010). 'Introduction to causal inference', Journal of Machine Learning Research, 11, pp.1643-1662, [Online]. Available from: http://jmlr.org/papers/volume11/spirtes10a/spirtes10a.pdf (Accessed: 3 January 2013).

[15] Hernán, M. A. & Robins, J. M. (2006). 'Estimating causal effects from epidemiological data', J Epidemiol Community Health, 60 (7), pp.578-586, [Online]. Available from: http://jech.bmj.com/content/60/7/578.abstract (Accessed: 3 January 2014).

[16] Winship, C. & Morgan, S. L. (1999). 'The estimation of causal effects from observational data', Annual Review of Sociology, 25, pp.659-706, [Online]. Available from: http://dash.harvard.edu/bitstream/handle/1/3200609/Winship_EstimationCausal.pdf?sequence=1 (Accessed: 18 October 2013).

[17] Ben-Gal, I. (2007). 'Bayesian networks', In: Ruggeri, F., Failtin, F. & Kennet, R. Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007), [Online]. Available from: http://www.eng.tau.ac.il/~bengal/BN.pdf (Accessed: 11 January 2014).

[18] Kalisch, M. & Bühlmann, P. (2007). 'Estimating high-dimensional directed acyclic graphs with the PC-algorithm', Journal of Machine Learning Research, 8, pp.613-636, [Online]. Available from: http://jmlr.org/papers/volume8/kalisch07a/kalisch07a.pdf (Accessed: 11 January 2013).

[19] Nguefack-Tsague, G. & Zucchini, W. (2011). 'Modeling hierarchical relationships in epidemiologic studies: a Bayesian networks approach', Epidemiol Health, 33, [Online]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3132659/ (Accessed: 11 January 2013).

[20] Shrier, I. & Platt, R. W. (2008). 'Reducing bias through directed acyclic graphs', BMC Medical Research Methodology, 8 (70), [Online]. Available from: http://www.biomedcentral.com/1471-2288/8/70 (Accessed: 11 January 2014).

[21] Kalisch, M. et al. (2012). 'Causal inference using graphical models with the r package pcalg', Journal of Statistical Software, 47 (11), [Online]. Available from: http://www.jstatsoft.org/v47/i11 (Accessed: 11 January 2014).

[22] Maathuis, M. H., Kalisch, M. & Bühlmann, P. (2009). 'Estimating high-dimensional intervention effects from observational data', The Annals of Statistics, 37 (6A), [Online]. Available from: http://arxiv.org/pdf/0810.4214.pdf (Accessed: 11 January 2014).

[23] Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). 'Using Bayesian networks to analyze expression data', Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, [Online]. Available from: http://www.cs.huji.ac.il/~nirf/Papers/FLNP1Full.pdf (Accessed: 11 January 2014).

[24] Neapolitan, R. E. (2009). Probabilistic Methods for Bioinformatics with an Introduction to Bayesian Networks. Elesvier.

[25] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/.

[26] Spirtes, P., Glymour, C. & Scheines, R. (2000). Causation, prediction, and search. MIT Press, Cambridge, MA.

[27] R. Davidson & J. MacKinnon (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. Econometrica, 49, 781-793.

[28] Bassani, D. et al. (2010). 'Causes of neonatal and child mortality in India: a nationally representative mortality survey', The Lancet, 376 (9755), pp.1853-1860, [Online]. Available from: http://search.ebscohost.com.ezproxy.liv.ac.uk/login.aspx?direct=true&db=cmedm&AN=21075444&site=eds-live&scope=site (Accessed: 7 September 2013).

[29] Bhaumik, S. (2013b), 'India tops world table for number of babies who die on day of birth', BMJ, 346, p.f3123, [Online]. Available from: http://www.bmj.com/content/346/bmj.f3123 (Accessed: 7 September 2013).