A Risk Calculator for the Pulmonary Arterial Hypertension Based on a Bayesian Network

Jidapa Kraisangka & Marek J. Druzdzel *
Decision System Laboratory,
School of Information Sciences,
University of Pittsburgh,
Pittsburgh, PA

Raymond L. Benza
Advanced Heart Failure, Transplant,
MCS and Pulmonary Hypertension
Allegheny Health Network
Allegheny General Hospital
Pittsburgh, PA

Abstract

Pulmonary arterial hypertension (PAH) is a severe and often deadly disease, originating from an increase in pulmonary vascular resistance. Its prevention and treatment are of vital importance to public health. A group of medical researchers proposed a calculator for estimating the risk of dying from PAH, available for a variety of computing platforms and widely used by health-care professionals. The PAH Risk Calculator is based on the Cox's Proportional Hazard (CPH) Model, a popular statistical technique used in risk estimation and survival analysis, based on data from a thoroughly collected and maintained Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension Disease Management (REVEAL Registry). In this paper, we propose an alternative approach to calculating the risk of PAH that is based on a Bayesian network (BN) model. Our first step has been to create a BN model that mimics the CPH model at the foundation of the current PAH Risk Calculator. The BN-based calculator reproduces the results of the current PAH Risk Calculator exactly. Because Bayesian networks do not require the somewhat restrictive assumptions of the CPH model and can readily combine data with expert knowledge, we expect that our approach will lead to an improvement over the current calculator. We plan to (1) learn the parameters of the BN model from the data captured in the REVEAL Registry, and (2) enhance the resulting BN model with medical expert knowledge. We have been collaborating closely on both tasks with the authors of the original PAH Risk Calculator.

1 Introduction

Pulmonary arterial hypertension (PAH) is a fatal, chronic, and life-changing disease originating from an increase in pulmonary vascular resistance, and leading to high blood pressure in the lung (Benza et al., 2010; Subias et al., 2010). Patients with PAH suffer from shortness of breath, chest pain, dizziness, fatigue, and possibly other symptoms depending on the progression of disease (Hayes, 2013). Currently, there is no cure for PAH and treatment is often determined based on the symptoms. With an early diagnosis and proper treatment, patients' lives can be extended by five or more years.

With the long-term goal to characterize the clinical course, treatment, and predictors of outcomes in patients with PAH in the United States, a group of medical researchers established a Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension Disease Management (REVEAL Registry) (Benza et al., 2010). The REVEAL registry is quite likely the most comprehensive collection of data of patients suffering from PAH and it has led to interesting insights improving the diagnosis, prediction, and treatment of PAH. One of the prominent applications of the REVEAL Registry is the PAH Risk Calculator (Benza et al., 2012), a statistical model learned from the REVEAL Registry data and predicting the survival of patients at risk for PAH. A computer implementation of the PAH Risk Calculator is available for a variety of computing platforms and widely used by health-care professionals (see http://www.pah-app.com/ for more information).

The PAH Risk Calculator is based on the Cox's Proportional Hazard (CPH) model (Cox, 1972), a popular statistical technique used in risk estimation and survival analysis. One weakness of this approach is that the underlying model can be only learned from data and is not readily amenable to refinement based on expert knowledge. Another possible weakness is that the CPH model rests on several assumptions simplifying the interactions between the risk factors and the disease. While these assumptions are reasonable and the CPH model has been successfully used for decades,

* Also Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

it is interesting to question them with a possible benefit in terms of model accuracy.

In this paper, we propose an alternative approach to calculating the risk of PAH that is based on a Bayesian network (BN) (Pearl, 1988) model. BNs are acyclic directed graphs in which vertices represent random variables and directed edges between pairs of vertices capture direct influences between the variables represented by the vertices. A BN captures the joint probability distribution among a set of variables both intuitively and efficiently, modeling explicitly independences among them. A representation of the joint probability distribution allows for calculation of probability distributions that are conditional on a subset of variables. This typically amounts to calculating the probability distributions over variables of interest given observations of other variables (e.g., probability of one-year survival given a set of observed risk factors). There is a well developed theory expressing the relationship between causality and probability and often the structure of a BN is given a causal interpretation. This is utmost convenient in terms of user interfaces, notably knowledge acquisition and explanation of results. The first step in our work has been to create a BN model that mimics the CPH model at the foundation of the current PAH Risk Calculator. In this, we use the BN interpretation of the CPH model proposed by Kraisangka and Druzdzel (2014). Our BN-based calculator reproduces the results of the current PAH Risk Calculator exactly.

Because Bayesian networks do not require the assumptions of the CPH model and can readily combine data with expert knowledge, we expect that our approach will eventually lead to an improvement over the current PAH Risk Calculator. Our mid- to long terms plans include (1) learning the parameters of the BN model directly from the data captured in the REVEAL Registry, and (2) enhancing the resulting BN model with medical expert knowledge. We are collaborating on both tasks with the team maintaining the REVEAL Registry and the authors of the original PAH Risk Calculator.

The remainder of this paper is structured as follows. Section 2 describes the problem of PAH, the CPH model, and the PAH Risk Calculator. Sections 3 and 4 describe application of Bayesian networks to risk estimation and the proposed BN-based PAH Risk Calculator. Finally, Section 5 describes our conclusions and future work.

2 Pulmonary Arterial Hypertension

This section introduces some facts related to the pulmonary arterial hypertension (PAH), notably its risk factors, the Cox's Proportional Hazard (CPH) model, and the PAH Risk Calculator based on the CPH model.

PAH Risk Factors

Risk can be defined as the rate of an occurrence of a particular disease or adverse event (Irvine, 2004). Although PAH can occur at any age, in any races, and any ethnic background (Hayes, 2013), there are risk factors that make some people more susceptible. For example, females are at least two and a half times more susceptible than men to idiopathic PAH. Recently, medical care professionals treating PAH have relied on existing patient registries to understand PAH better. Several risk factors have been identified and used to develop prognostic models for guiding their therapeutic decision making. For example, a study based on the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL) (Benza et al., 2010) extracted several demographic, functional, laboratory, and hemodynamic parameters associated with patient survival in PAH (Benza et al., 2012) by means of a multivariate Cox's proportional hazard model (CPH) (discussed in more detail in the following section). By developing a prognosis model, physician can access a short-term and long-term patient survival in the context of current treatment and clinical variables (Benza et al., 2012). Although prognostic tools for patient survival have improved the quality of predictions, the models are still imperfect and more research is needed on improving them.

Cox's Proportional Hazard Model

Hazard is a measure of *risk* at a small time interval t, which can be considered as a rate (Allison, 2010). In survival analysis, the hazard function can be represented by probability distributions (e.g., exponential distribution) or can be modeled by regression techniques. The Cox's proportional hazard model (CPH) (Cox, 1972) is a set of regression methods used in the assessment of survival based on its risk factors or explanatory variables. The probability of an individual surviving beyond time t can be estimated with respect to a hazard function (Allison, 2010). As defined originally by Cox (1972), the hazard regression model is expressed as

$$\lambda\left(t\right) = \lambda_0\left(t\right) \exp^{\beta' \cdot \mathbf{X}} \ . \tag{1}$$

This hazard model is composed of two main parts: the baseline hazard function, $\lambda_0\left(t\right)$, and the set of effect parameters, $\beta' \cdot \mathbf{X} = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. The baseline hazard function determines the risks at an underlying level of explanatory variables, i.e., when all explanatory variables are absent. The βs are the coefficients corresponding to the risk factors, $\mathbf{X}$. According to Cox (1972), this $\lambda_0\left(t\right)$ can be unspecified or can follow any distribution and be estimated from data.

The application of the CPH model relies on the assumption that the hazard ratio of two observations is constant over time (Cox, 1972). For example, a hazard ratio of a group of

PAH patients having renal insufficiency to a group of PAH without renal insufficiency (control/baseline group) is estimated as 1.90. This assumption means that patients with renal insufficiency always have a 90% higher risk for dying from PAH than patients without renal insufficiency by Cox's assumptions. The ratio of two hazards is defined as γ:

$$\gamma = \frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\exp(\beta' X_2)}{\exp(\beta' X_1)} \ . \quad (2)$$

If the risk factors X are binary, their value could be expressed as *presence* $(X = 1)$ or as *absence* or *baseline* $(X = 0)$ of the risk factor. Once, we know the hazard ratio of one group toward another group, we can estimate the survival probability (Casea et al., 2002) by

$$S(t) = S_0(t)^{\gamma} \ . \quad (3)$$

$S_0(t)$ is the baseline survival probability estimated from the data, i.e., when all risk factor are absent or at their baseline value $(X = 0)$ at any time t, while γ is hazard ratio of an interested group to the baseline group. In other words, the survival probability of any patients relative to the baseline group can be estimated from

$$S(t) = S_0(t)^{\exp^{\beta' \cdot \mathbf{x}}} \ . \quad (4)$$

An example of CPH model used as a prognosis model for PAH patients is from the REVEAL Registry Risk Score Calculator (Benza et al., 2012). The model, including 19 risk factors, was developed to predict a one-year survival probability. The main survivor function is

$$S(t=1) = S_0(1)^{\exp^{\beta' \cdot \mathbf{x}_\gamma}} \ , \quad (5)$$

where $S_0(1)$ is the baseline survivor function of 1 year (0.9698) and γ in this equation is the shrinkage coefficient after model calibration (0.939) (Benza et al., 2010). The risk factors X (listed in Table 1) included PAH associated with portal hypertension (APAH-PoPH), PAH associated with connective tissue disease (APAH-CTD), family history of PAH (FPAH), modified New York Heart Association (NYHA)/World Health Organization(WHO)functional class I, III, and IV, men aged > 60, renal insufficiency, systolic blood pressure(SBP) < 110 mm Hg, heart rate > 92 beats per min, mean right atrial pressure (mRAP) > 20 mm Hg, 6-minute walking distance(6MWD), brain natriuretic peptide (BNP)> 180 pg/ml, 165 m, brain natriuretic peptide (BNP), 180 pg/mL, pulmonary vascular resistance(PVR)> 32 Wood units, percentage predicted diffusing capacity of lung for carbon monoxide (Dlco) $\leq 32\%$, and presence of pericardial effusion on echocardiogram. Most of the risk factors were associated with increasing mortality rate (indicated by positive sign in β in Table 1), while only four factors were associated with increased one-year survival (indicated by negative sign in β in Table 1).

Risk factors X_i	β	$exp(\beta)$
APAH-CTD	0.7737	1.59
FPAH	1.2801	3.60
APAH-PoPH	0.4624	2.17
Male >60 years age	0.7779	2.18
Renal insufficiency	0.6422	1.90
NYHA Class I	-0.8740	0.42
NYHA Class III	0.3454	1.41
NYHA Class IV	1.1402	3.13
SBP <110 mmHg	0.5128	1.67
Heart Rate >92bmp	0.3322	1.39
6MWD ≥440 m	-0.5455	0.58
6MWD <165 m	0.5210	1.68
BNP <50 pg/ML	-0.6922	0.50
BNP >180 pg/ML	0.6791	1.97
Pericardial effusion	0.3014	1.35
% Dlco ≥80%	-0.5317	0.59
% Dlco ≤32%	0.3756	1.46
mRAP > 20 mmHg	0.5816	1.79
PVR >32 Wood units%	1.4062	4.08

Table 1: A list of 19 binary risk factors, their corresponding coefficients β, and hazard ratio $exp(\beta)$ reported for the PAH REVEAL system (Benza et al., 2010).

To be able to summarize from the model, patients were stratified into five risk groups according to their range of survival probability (Benza et al., 2010) including the low risk group where the predicted 1-year survival probability $> 95\%$, average risk with 90% to 95% survivals, moderately high risk with 85% to 90% survivals, high risk with 70% to 85% survival, and very high risk group with survival probability $< 70\%$.

PAH Calculator

Based on the CPH model, the further application of the CPH model is in the form of a risk calculator. This simplified calculator are useful in everyday clinical practice by helping physicians to decide patient therapies based on level of risk (Benza et al., 2012). The calculator was designed from assigning score to variables according to their hazard ratio. For the risk factors associated with increasing mortality (positive β coefficients), score of two points were assigned for the risk factors which has their hazard ratio ($exp(\beta)$) at least two or more folds, i.e., those with $exp(\beta) \geq 2$, and one point were assigned for other risk factors. Risk factors associated with decreasing mortality (negative β coefficients) were assigned a negative score. Figure 1 shows all risk factors and the interpretation of their hazard ratio rate.

Figure 2 shows the user interface of the PAH Risk Calculator. Each risk factor from the CPH model is listed and mapped with the score. The calculator allows for adding and subtracting the score based on the data entered for an individual patient case. To avoid a negative total score, the base score of 6 is set as a starting score. The total score is interpreted in the same way as the survival probability

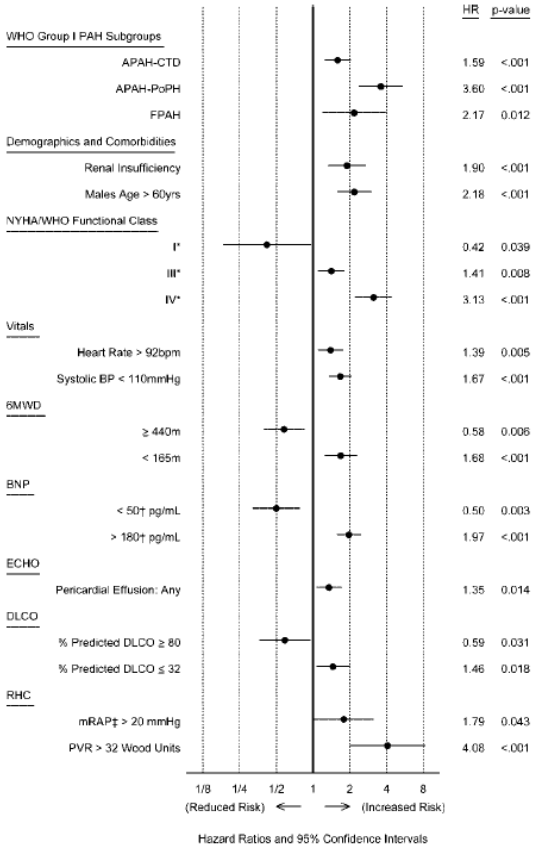

Figure 1: Cox's proportional-hazards of 1-year PAH patients survival variables (Benza et al., 2010) indicating increasing/decreasing mortality rate for each risk factor

given by the CPH model, i.e., it includes the low risk group with the score ≤ 7, average risk with score $= 8$, moderately high risk score $= 9$, high risk with score between 10 and 11, and very high risk group with score ≥ 12. The score, defined as above, makes it simpler for health care providers to use than probabilities.

3 Application of Bayesian Networks to Risk Calculation

An alternative approach to the traditional survival analysis is the use of Bayesian networks (Pearl, 1988) to estimate risks. Compared to the CPH model and several other Artificial Intelligence and Machine Learning techniques, a Bayesian network can model explicitly the structure of the relationships among explanatory variables with their probability (Hanna and Lucas, 2001). A Bayesian network can be built from expert knowledge, available data, or combination of both. If there exists a probabilistic interpretation of existing modeling tool, like in case of the CPH model, a BN model can also be an interpretation of the existing model. The structure of a Bayesian network can depict a complex structure of a problem and provide a way to infer posterior

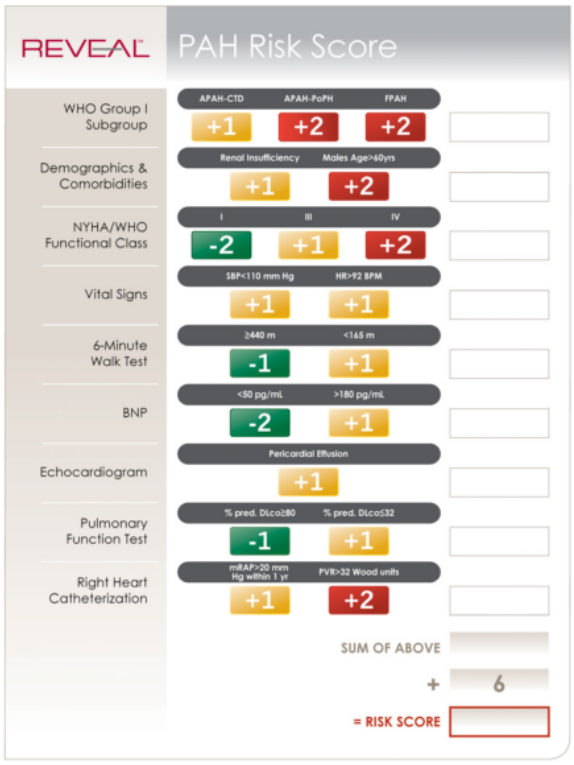

Figure 2: PAH risk score calculator (Benza et al., 2012) (electronic version developed by the United Therapeutics Europe Limited)

conditional probability distributions, useful for prognosis and diagnosis, including medical decision support systems (Husmeier et al., 2005).

To estimate risks using Bayesian network, the prognosis can be created as a static model, i.e., it can predict the survival at a future point in time. For example, the work of Loghmanpour et al. (2015) focuses on risk assessment models for patients with the left ventricular assist devices (LVADs). Bayesian network have been shown to estimate the risk at various points in time (including 30 days, 90 days, 6 months, 1 year, and 2 years) with accuracy higher than traditional score-based methods (Loghmanpour et al., 2015). An alternative, more complex approach could use dynamic Bayesian networks (DBN), which are an extension of Bayesian networks modeling time explicitly. van Gerven et al. (2007) implemented a DBN for prognosis of patients that suffer from low-grade midgut carcinoid tumor. Instead of treating risk factors independently at each time point, the DBN model considered how the state of patient changed under the influence of choices made by physicians. This model was shown suitable to temporal nature of medical problems throughout the course of care and provide detailed prognostic predictions. However, DBNs requires additional effort during model construction, for example expertise to structure of temporal interaction, large amounts

of (complete) data, which translates to time-consuming efforts (van Gerven et al., 2007).

4 Bayesian Network PAH Risk Calculator

BN Cox model

With no access to the REVEAL Registry data, we created a Bayesian network model that is a formal interpretation of the CPH model, for which the parameters are reported in the literature (Benza et al., 2010). To this effect, we used the method proposed by Kraisangka and Druzdzel (2014). We first created a Bayesian network structure by using all risk factors of the PAH CPH models. We converted all binary risk factors to random variables, which were the parents of the *survival* node. In our case, we have omitted the *time* variable, as the purpose of the PAH Risk Calculator is to capture the risk at one point in time (in this case, it is one year). Figure 3 shows the structure of the BNCox model for the BN-based calculator.

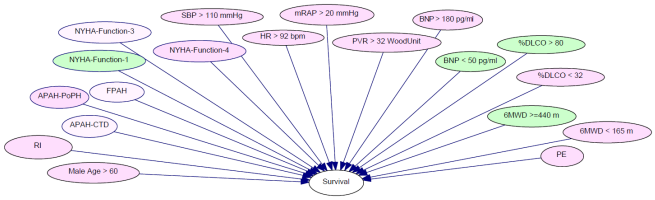

Figure 3: A Bayesian network representing the interaction among variables for the PAH CPH model. All random variables are from the original PAH CPH model and the *Survival* node was added to capture the survival probabilities from the CPH model.

In the next step, we created the conditional probability table for the survival node. The survival probabilities from a CPH model can be encoded into the conditional probabilities as

$$Pr(s \mid X_i, T = t) = S_0\,(t)^{e^{(\beta' x_i)}}\ , \qquad (6)$$

where s means the state of *survived* in the survival node, X_i are all risk factors, T is the time point which is 1 in this case.

We configured all risk factors cases (all binary risk factors generated 2^{19} cases) and obtained all survival probabilities filled in the conditional probability table of a *survival* node. This allowed us to reproduce fully the PAH CPH model by means of a Bayesian network.

BN Interpretation of the PAH Calculator

The original PAH Risk Calculator uses the hazard ratios in the CPH model to derive the risk score for the calculator (Benza et al., 2012). We apply the same approach in our model. Equation 6 captures the survival probabilities s given the states of risk factor. We can extract a hazard ratio of each variable by configuring states of other risk factors to be absent. For example, the hazard ratio of a risk factor x_j can be estimated from

$$\gamma = \frac{\log(Pr(s \mid \bar{x}_1, \ldots, \bar{x}_{j-1}, \mathbf{x_j}, \bar{x}_{j+1}, \ldots, \bar{x}_n))}{\log(Pr(s \mid \bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{\mathbf{x}}_\mathbf{j}, \bar{x}_{j+1}, \ldots, \bar{x}_n))}\ . \qquad (7)$$

The term $\log(Pr(s \mid \bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{\mathbf{x}}_\mathbf{j}, \bar{x}_{j+1}, \ldots, \bar{x}_n))$ is similar to the baseline survival probability in the CPH model ($S_0(1) = 0.9698$). Hence, with this equation, we can track back all hazard ratios.

We use the same criteria as the original PAH Risk Calculator to convert the hazard rate to the score, i.e., score of 2 indicates at least two-fold increase in risk of mortality compared to the baseline risk.

Figure 4 shows a screen shot of our prototype of the Bayesian network risk calculator. The left-hand pane allows for entering risk factors for a given patient case. The right-hand pane shows the calculated score and survival probabilities. Currently, our calculator is a Windows app running on a local server. The numerical risks that produced by the BN calculator are identical to those of the original CPH-based PAH Risk Calculator (Benza et al., 2012).

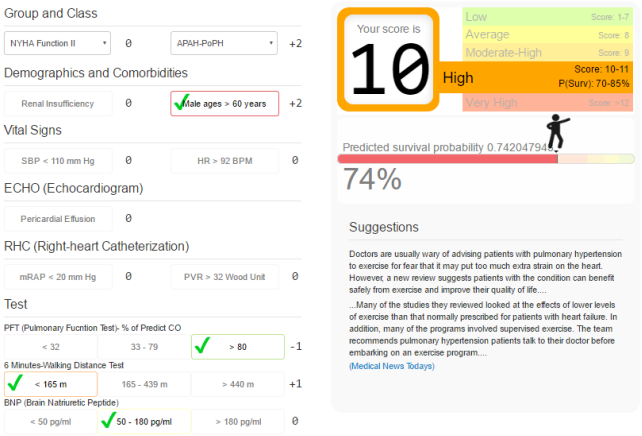

Figure 4: A prototype for Bayesian network risk score calculator for a 1-year PAH prognosis model. The left-hand pane allows for entering risk factors for a given patient case. The right-hand pane shows the calculated score and survival probabilities.

5 Conclusions and Future Work

In this paper, we propose an alternative the the existing Pulmonary Arterial Hypertension (PAH) Risk Calculator that replaces the original Cox Proportional Hazard (CPH) model with a Bayesian network. Because we did not have access to the REVEAL Registry data, we created a Bayesian network model that uses the CPH parameters

learned from the REVEAL Registry data and available in the literature. To this effect, we used a Bayesian network interpretation of the CPH model (Kraisangka and Druzdzel, 2014).

Our calculator reproduces the results of the current PAH Risk Calculator exactly. From this point of view, we have not yet offered a superior calculator. However, we plan to refine the calculator by (1) learning the parameters of the BN model from the data captured in the REVEAL Registry, and (2) enhancing the resulting BN model with medical expert knowledge. The extended model will relax the assumption of the multiplicative character of interactions between the risk factors and the survival variable. It will also relax the assumption that the risk ratio is constant over time. Another direction of our work is allowing risk variables that are not binary. Instead of having 19 binary risk factors, we will be able to group those risk factors that are mutually exclusive, e.g., WHO Group or NYHA/WHO Functional Class. As a result, we can control the number of risk factors and reduce complexities of the model. Yet another direction is allowing dependencies between the risk factors, something that is not straightforward in the CPH model. We should be able to refine the Bayesian network model by using expert knowledge or by training its elements from available data. The current calculator produces a patient-specific score based on hazard ratio. Because the new Bayesian network model will no longer use the multiplicative CPH model, we plan to create new risk score criteria based on the probability of survival rather than the hazard ratio. We have little doubt that with some further modeling effort we should be able to obtain a superior calculator in the sense of producing higher accuracy of the risk estimate than the original CPH-based risk calculator.

Acknowledgements

We acknowledge the support the National Institute of Health under grant number U01HL101066-01 and the Faculty of Information and Communication Technology, Mahidol University, Thailand. Implementation of this work is based on GeNIe and SMILE, a Bayesian inference engine available free of charge for academic teaching and research use at http://www.bayesfusion.com/. While we take full responsibility for any remaining errors and shortcomings of this paper, we would like to thank anonymous reviewers for their valuable suggestions.

References

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide, Second Edition*. SAS Institute Inc., Cary, NA.

Benza, R. L., Gomberg-Maitland, M., Miller, D. P., Frost, A., Frantz, R. P., Foreman, A. J., Badesch, D. B., and McGoon, M. D. (2012). The REVEAL registry risk score calculator in patients newly diagnosed with pulmonary arterial hypertension. *Chest*, 141(2):354–362.

Benza, R. L., Miller, D. P., Gomberg-Maitland, M., Frantz, R. P., Foreman, A. J., Coffey, C. S., Frost, A., Barst, R. J., Badesch, D. B., Elliott, C. G., Liou, T. G., and McGoon, M. D. (2010). Predicting survival in pulmonary arterial hypertension: Insights from the Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension disease management (REVEAL). *Circulation*, 122(2):164–172.

Casea, L. D., Kimmickb, G., Pasketta, E. D., Lohmana, K., and Tucker, R. (2002). Interpreting measures of treatment effect in cancer clinical trials. *The Oncologist*, 7(3):181–187.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Hanna, A. A. and Lucas, P. J. (2001). Prognostic models in medicine- AI and statistical approaches. *Method Inform Med*, 40:1–5.

Hayes, G. B. (2013). *Pulmonary Hypertension: A Patient's Survival Guide - Fifth Edition, 2013 Revision*. Pulmonary Hypertension Association.

Husmeier, D., Dybowski, R., and Roberts, S. (2005). *Probabilistic modeling in bioinformatics and medical informatics*. Springer.

Irvine, E. J. (2004). Measurement and expression of risk: optimizing decision strategies. *The American Journal of Medicine Supplements*, 117(5):2–7.

Kraisangka, J. and Druzdzel, M. J. (2014). Discrete Bayesian network interpretation of the Coxs proportional hazards model. In van der Gaag, L. C. and Feelders, A. J., editors, *Probabilistic Graphical Models*, volume 8754 of *Lecture Notes in Computer Science*, pages 238–253. Springer International Publishing.

Loghmanpour, N. A., Kanwar, M. K., Druzdzel, M. J., Benza, R. L., Murali, S., and Antaki, J. F. (2015). A new bayesian network-based risk stratification model for prediction of short-term and long-term lvad mortality. *ASAIO Journal*, 61(3):313–323.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Subias, P. E., Mir, J. A. B., and Suberviola, V. (2010). Current diagnostic and prognostic assessment of pulmonary hypertension. *Revista Española de Cardiología (English Edition)*, 63(5):583–596.

van Gerven, M. A., Taal, B. G., and Lucas, P. J. (2007). Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics*, 41(4):515–529.