Hamstring Surgery Outcome Prediction with Linear Continuous Bayesian Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Cerebral palsy is a neuromotor condition that leads to a range of walking disorders. Symptoms can be alleviated by musculoskeletal surgery, such as hamstring lengthening. As surgical outcomes vary across patients, there is a need to predict which patients will benefit from surgery and the magnitude of the expected improvement. We built a linear continuous Bayesian network to predict a quantitative surgery outcome variable following hamstring lengthening, trained on a group of patients who received the intervention. We trained another Bayesian network on patients who did not receive a hamstring lengthening. We designed a classifier that compares the predictions from the two Bayesian networks to decide whether a given patient is a favorable candidate for hamstring lengthening. Our classifier advised in favor of 79.4% of the observed hamstring lengthening interventions. 78.8% of those recommended surgeries yielded patient state improvement, while only 70.5% of interventions performed by the clinical team did. Thus, the classifier could be used to assess clinical surgery decisions, potentially increasing the likelihood of success for the subjects selected for the intervention.

1 Introduction

Cerebral palsy is a neurological condition that affects body movement and muscle coordination. Hamstring lengthening is a surgical intervention commonly performed in patients with cerebral palsy who walk with excessive knee flexion during gait, with the goal of achieving a more upright walking posture. Surgical outcomes vary across patients: in our dataset, about 30% of hamstring operations performed did not yield an improvement in patient state, and 50% did not yield a meaningful improvement. Moreover, if an intervention is expected to yield a minor improvement in patient knee flexion, the clinical team could decide to prescribe a less invasive treatment, such as physical therapy bracing. Therefore, there is a need to predict which patients would benefit from surgery, and the magnitude of the expected improvement.

In addition to predicting surgical benefit, clinicians seek to understand the factors that influence patient knee flexion during gait. While past studies (e.g., [1, 2]) have identified the factors that influence surgical outcomes, few have simultaneously tried to predict outcomes. Here, we aim to train a model that can forecast patient knee flexion following different types of surgery, in addition to highlighting the most predictive factors. Hicks and colleagues [7] developed a linear regression model with this goal using input variables selected by domain knowledge experts. In contrast, we adopt a knowledge-agnostic approach leveraging Bayesian networks. Bayesian networks [9] bridge the gap between descriptive studies and forecasting systems. Networks are built manually from expert knowledge [17, 14] or automatically using structure learning [6, 14, 13]. Bayesian networks are interpretable, which allows clinicians to check model coherence. In addition, when the structure is

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

learned from data, it may reveal unexpected correlations between variables that may not be expected by domain experts [13, 8, 14].

In healthcare applications, Bayesian networks mostly use discrete variables to solve classification tasks, since continuous variables pose challenges such as high sample complexity, long training runtime and the lack of closed form solutions [4, 15, 16, 18]. Data describing patients with cerebral palsy consists primarily of continuous variables (e.g., measures of joint range of motion or joint angles throughout a gait cycle). There is growing interest in integrating continuous variables in Bayesian networks [11]. However the most common approach to incorporating them is discretization [17, 13, 8] – most frequently achieved by binning all variables into quantiles. Discretizing the features results in an increase of the number of parameters which need to be learned. Further, the parameter space increases exponentially with the number of parents that a BN variable can have, resulting in unlearnable conditional probability tables. This means that, in order to keep the problem tractable, the number of parents must be limited. In contrast, continuous Bayesian networks do not suffer from this constraint. Under the assumption that simple probability distributions are chosen, continuous Bayesian networks can perform fine-grained prediction following training on limited datasets, while keeping a highly interpretable network structure.

In this paper, we describe a continuous Bayesian network to predict post-surgical patient state as a continuous variable by training the model on a set of retrospective patient data from a large clinical center. We employed structure learning, a process by which the correlations and dependencies between the variables, which give the form of the network graph itself, are determined from the training data, as opposed to having them hard coded. Our model outperforms random forest regression, a baseline which is extensively used in the domain, by 6.5% in mean absolute error (MAE). We created a decision support system that uses the network predictions to identify favorable surgery candidates with 8.3% higher precision than the clinical team on observed data. Finally, we analyzed the structures of the Bayesian networks to ensure that the models concur with the experience of clinicians and to identify key variables related to the evolution of patient health status.

2 Methods

Data. We analyzed anonymized, retrospective data for patients with cerebral palsy collected as part of standard care at Gillette Children's Specialty Healthcare in Minneapolis, MN. Our dataset consists of 6039 clinical visits from 2904 patients. The number of visits per patient ranges from 2 to 10, and the typical gap between consecutive visits is approximately one year. A training example consists of a pair of consecutive visits with surgery performed in between. We excluded visits separated by more than two years. Data collected at each visit includes patient characteristics such as age, clinical measurements such as strength assessments, kinematic data in the form of 11 joint angle time series obtained during a gait cycle, and hand-engineered features obtained from kinematic data, for a total of 1162 variables. Since hamstring surgery can be performed on either side of the patient, or on both sides, we considered the left and right sides of a patient as separate training examples.

Task definition and preprocessing. We designed a regression system that predicts the post-surgical patient's average knee flexion during the stance phase of the gait cycle, which we will refer to as the post-surgical "KneeScore". We built a classifier that uses the predictions made by the model to recommend patients for surgery or not. We defined a hamstring lengthening to be a success if the patient's KneeScore decreases following intervention, which translates to a more upright walking posture.

For a given patient, we predicted post-surgical KneeScore with a linear continuous Bayesian network trained on a population of 285 patients who received a hamstring lengthening. Hamstring lengthening was performed as part of a single event multilevel surgery (SEMLS), a common treatment consisting of multiple surgical operations including at least one major and one minor orthopedic operation. We removed from the dataset visits that are followed by a SEMLS that includes surgeries with a large potential effect on knee flexion[1]. To isolate the impact of hamstring lengthening from the rest of the SEMLS, we examined the same prediction under a control scenario, computed by a Bayesian network trained on 1149 patients who received a SEMLS without hamstring lengthening. Then, a classification rule identified the subject as a hamstring surgery candidate if the predicted post-surgical KneeScore in the hamstring lengthening scenario exceeds a decision margin below the predicted

[1] Distal femoral extension osteotomy and patellar tendon advancement interventions

value in the control scenario. Therefore, clinical application is straightforward with our decision support system, and is considered an extension of the work by Hicks and colleagues [7]. We tuned the decision margin through cross-validation and the value that best balanced precision and recall was $3°$. As both patient groups were selected for a SEMLS, we make the assumption that the difference in distributions between the two populations is limited.

We removed variables whose correlations to the target variable were below a threshold determined by a grid search with a step size of 2.5%. When the correlation of a variable pair was above another threshold, set by a similar grid search, we removed the variable with the smaller correlation to post-surgical KneeScore. Variable number after processing was approximately between 20 and 50.

Structure learning with continuous Bayesian networks. We chose a knowledge discovery approach as we learned Bayesian network structure from the data. We assumed conditional probability laws were Gaussian. Linear relationships are easier to learn, which helps as our hamstring lengthening dataset contains only 285 samples. Structure learning algorithms fall into two categories – constraint-based algorithms, which better fit problems with a small number of variables, and score-based algorithms, which are more popular and versatile [9]. From the latter, we chose the often-used K2 algorithm [3] because the continuous setup is implemented in the Bayesian Network Toolbox in Matlab [12]. The K2 algorithm is a greedy algorithm which takes as input an initial variable ordering, tests all parent-child variable relationships compatible with that order and assigns a score to each resulting structure. We used the Bayesian Information Criterion score to regularize the network. To make the K2 search faster, we bounded the number of children per node to [6,11,16,21], with 6 yielding the best results. Initial ordering has a strong influence on learned structure, so we tried multiple initializations [10, 5]. Random orderings can be used, but the number of variables made this option intractable. We experimented with the common Maximum Weighted Spanning Tree as the initial ordering [5]. The best performing initialization followed the order of correlation to the post-surgical KneeScore, which is intuitive to use as highly correlated variables are close to the post-surgical KneeScore at initialization. Once structure was learned, parameters were derived by Maximum Likelihood Estimation.

We compared the Bayesian networks to benchmark models: 1 and 2 layer neural network, linear regression, k-nearest neighbor regression, and random forests (RF). We used 5 fold cross-validation for the Bayesian network, and averaged 5-fold cross-validation errors on 4 random cross-validation splits for the other benchmarks.

3 Results

KneeScore prediction results. In Table 1, we present the performances of the Bayesian networks and the best benchmark, random forests.

The approximate mean absolute error (MAE) is 7 degrees. First and third KneeScore quantiles in our dataset are $14.4°$ and $29.1°$ respectively. The regression problem is challenging for several reasons. The amount of data is limited. We also do not account for variability in surgical technique. Pre and post-surgical visits are separated by variable time gaps often longer than a year. However, the predictions computed by the Bayesian networks were accurate enough to be used as inputs to a classifier for surgery candidates.

Table 1: Performances of regression algorithms

	RF		BN	
Surgery group	MAE	MSE	MAE	MSE
Hamstring	7.56	96.8	7.45	90.70
No hamstring	7.04	81.2	6.53	69.52

Analysis of Bayesian network variables. We analyzed the resulting Bayesian network to (1) ensure that the model concurs with the experience of clinicians and (2) identify key variables in the evolution of patient health status. We differentiate two types of variables, depending on their impact on the KneeScore. First, there are direct predicting variables which are the parents of the KneeScore variable in the Bayesian network. The KneeScore is conditionally independent of all the other variables in the network given the direct predicting variables. Second, there are intermediary variables which are related to KneeScore through the direct predicting variables or their ancestors in the graph. The network learned on patients who received SEMLS with hamstring surgery is represented in Figure 1, with less consequential intermediary variables omitted.

Structure learning selected a small number of direct predicting variables: 2 for the 'Hamstring' model and 5.4 for the 'No hamstring' model, on average using cross-validation. Inferring the value of post-surgical KneeScore only requires those few direct predicting variables, so the model is intuitive to understand. Both models selected a 'Popliteal Angle' variable and the 'Knee Flexion at maximum extension in the gait cycle'. It hints that those are key in tracking patient state evolution, which was judged coherent by domain knowledge experts as they relate to hamstring function and tightness. Analysis of intermediary variables supports the clinical consistency of the model. The model confirmed that the health statuses of both sides of the patients were correlated as several variables were related to their opposite side equivalent in the graph. Then, the central position of pre-surgical KneeScore in the network suggests this variable is a good fit to quantify patient health status.

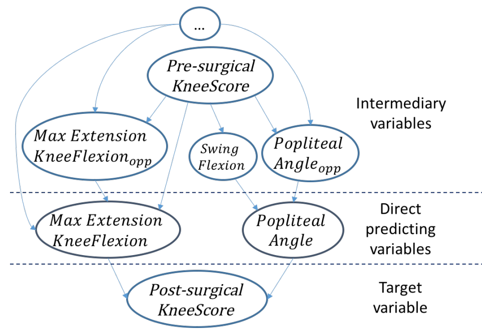

Figure 1: Partial network structure learned on patients who received hamstring lengthening. Only two variables are directly related to the target variable. The opposite side equivalent of a variable is denoted by the suffix 'opp'.

Performance of the classifier for surgery candidates. Evaluation of the classifier built on the Bayesian networks' predictions showed it could be used to confirm the surgical decisions of the clinical team. We assigned the ground truth label 'True' to hamstring lengthening patients if their KneeScore decreased (improved) after surgery, and 'False' otherwise. We have no labels for the patients who received a SEMLS without hamstring, since we cannot verify whether hamstring lengthening should have been included. Therefore, we cannot compute recall of the clinical team's decisions. We compare the decisions of our classifier to those of the clinical team on observed examples. Out of the 285 examples of hamstring surgery, 201 resulted in improved KneeScore, which translates to a clinical team precision of 70.5%. To compute the recall of our classifier, we assumed all the patients who needed a hamstring lengthening received it. We present the results in Table 2.

Table 2: Performance of our classifier evaluated on patients who received hamstring lengthening. Precision is the key metric to examine as our goal is to verify surgery decisions of the clinical team.

Model	Accuracy	Precision	Recall
Clinical team	70.5	70.5	N/A
Classifier	70.5	78.8	79.6

Our classifier exceeds the clinical team's precision by 8.3%. Recall is less than 100%, as the system did not select some of the subjects selected by the clinical team, and only the latter were used for evaluation. Hamstring lengthening is not a vital intervention, it could be performed at a later time if necessary. So precision is the key evaluation metric and surgery should be performed only when confidence in post-surgical patient improvement is high. It follows that the envisioned use case of our system is to improve confidence in the decision of including hamstring lengthening to SEMLS by confirming the clinical team's consensus with classifier prediction. If the classifier predicts a patient is not a favorable candidate for surgery, the clinical team can decide to further investigate the patient's case. They can also refer to the predictions of post-surgical KneeScores which led to the classifier's result.

4 Conclusion

We designed a classifier that predicts whether patients with cerebral palsy are favorable candidates for hamstring lengthening. The classifier is built on the quantitative predictions yielded by two linear continuous Bayesian networks: one trained on patients who received the surgery, one trained on patients who did not. The Bayesian networks compare favorably in performance to random forests, while making the predictive model interpretable, which allowed us to validate the clinical relevance of the model and to identify key variables regarding patient state evolution. Directions for future work include leveraging the learned Bayesian network structure to introduce latent variables representing patient state over time or subpopulation, investigating a setup with multiple surgery outcome variables, and testing the model on other datasets that require surgical decision support.

References

[1] Allison S. Arnold, May Q. Liu, Michael H. Schwartz, Sylvia Õunpuu, and Scott L. Delp. The role of estimating muscle-tendon lengths and velocities of the hamstrings in the evaluation and treatment of crouch gait. *Gait & Posture*, 23(3):273 – 281, 2006.

[2] Allison S. Arnold, May Q. Liu, Michael H. Schwartz, Sylvia Õunpuu, Luciano S. Dias, and Scott L. Delp. Do the hamstrings operate at increased muscle–tendon lengths and velocities after surgical lengthening? *Journal of Biomechanics*, 39(8):1498 – 1506, 2006.

[3] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[4] Cláudia Camila Dias, Fernando Magro, and Pedro Pereira Rodrigues. Preliminary study for a bayesian network prognostic model for crohn's disease. In *Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on*, pages 141–144. IEEE, 2015.

[5] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[6] Eren Gultepe, Jeffrey P Green, Hien Nguyen, Jason Adams, Timothy Albertson, and Ilias Tagkopoulos. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association*, pages 315–325, 2014.

[7] Jennifer L. Hicks, Scott L. Delp, and Michael H. Schwartz. Can biomechanical variables predict improvement in crouch gait? *Gait & Posture*, 34(2):197 – 201, 2011.

[8] Nathan Hoot and Dominik Aronsky. Using bayesian networks to predict survival of liver transplant patients. In *AMIA annual symposium proceedings*, volume 2005, page 345. American Medical Informatics Association, 2005.

[9] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[10] Boaz Lerner and Roy Malka*. Investigation of the k2 algorithm in learning bayesian network classifiers. *Applied Artificial Intelligence*, 25(1):74–96, 2011.

[11] Michael J McGeachie, Hsun-Hsien Chang, and Scott T Weiss. Cgbayesnets: conditional gaussian bayesian network learning and inference with mixed discrete and continuous data. *PLoS computational biology*, 10(6):e1003676, 2014.

[12] Kevin Murphy et al. The bayes net toolbox for matlab. *Computing science and statistics*, 33(2):1024–1034, 2001.

[13] Jung Hun Oh, Jeffrey Craft, Rawan Al Lozi, Manushka Vaidya, Yifan Meng, Joseph O Deasy, Jeffrey D Bradley, and Issam El Naqa. A bayesian network approach for modeling local failure in lung cancer. *Physics in medicine and biology*, 56(6):1635, 2011.

[14] M Berkan Sesen, Timor Kadir, Rene-Banares Alcantara, John Fox, and Michael Brady. Survival prediction and treatment recommendation with bayesian techniques in lung cancer. In *AMIA Annual Symposium Proceedings*, volume 2012, page 838. American Medical Informatics Association, 2012.

[15] Siriwan Suebnukarn, Nattaya Rungcharoenporn, and Sujinda Sangsuratham. A bayesian decision support model for assessment of endodontic treatment outcome. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 106(3):e48–e58, 2008.

[16] Marcel AJ Van Gerven, Babs G Taal, and Peter JF Lucas. Dynamic bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics*, 41(4):515–529, 2008.

[17] Emily W Watt and Alex AT Bui. Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. In *AMIA annual symposium proceedings*, volume 2008, page 788. American Medical Informatics Association, 2008.

[18] Barbaros Yet, Zane Perkins, William Marsh, and Norman Fenton. Towards a method of building causal bayesian networks for prognostic decision support. *Probabilistic Problem Solving in BioMedicine*, 107, 2011.