Concept Bottleneck Model for Adolescent Idiopathic Scoliosis Patient Reported Outcomes Prediction

Micky C. Nnamdi[†], Wenqi Shi[†], J. Ben Tamo[†], Henry J. Iwinski [§], J. Michael Wattenbarger[§], and May D. Wang[†]
[†] Georgia Institute of Technology, Atlanta, Georgia 30332–0250
email: mnnamdi3@gatech.edu; wqshi@gatech.edu; maywang@gatech.edu
[§] Shriners Children's, Greenville, SC, USA
email: hiwinski@shrinenet.org; mwattenbarger@shrinenet.org

***Abstract*—Post-surgical patient-reported outcomes (PROs) serve as a crucial subjective measure of surgical success for adolescent idiopathic scoliosis (AIS) patients. Leveraging pre-operative patient information to predict post-operative PROs is instrumental in improving pediatric patient care and providing invaluable insights for clinical decision-making. Recently, deep learning techniques have demonstrated encouraging results in developing predictive models for clinical decision support. However, the inherent black-box nature makes them non-interactive and challenging to troubleshoot during the training phase. To mitigate this issue, our study introduces an interactive concept bottleneck model to predict subjective rehabilitation outcomes for AIS patients. We assess three learning schemas - independent, sequential, and joint - to first comprehend the concepts, which are a set of post-operative radiographic data available during the training phase. Subsequently, these acquired concepts are employed to predict post-operative patient rehabilitation outcomes across five domains: pain, function, general satisfaction, self-image, and mental health. Our results demonstrated improvement compared to the existing baseline, with the joint learning schema yielding the highest F1 score in the function and pain domains, while sequential learning recorded the highest F1 score in the mental health and self-image domains. This proposed framework harbors the immense potential to aid pre-operative surgical planning and further enhance the transparency of AI models, thereby supporting real-world clinical decision-making.**

***Index Terms*—concept bottleneck model, explainable artificial intelligence, adolescent idiopathic scoliosis, pediatric healthcare**

I. Introduction

Adolescent idiopathic scoliosis (AIS), an abnormal lateral curvature of the spline, is one of the most common types of spinal deformity for pediatric patients [1]. The National Scoliosis Foundation estimates a total number of approximately 6 to 9 million cases in the United States. Although the spinal condition is manageable, it can impact the quality of life of most people by limiting their respiratory function, activities, and self-esteem, or even increasing the pain experienced [2] Specifically, posterior spinal fusion (PSF) surgery aims to stabilize the spine and provide relief to patients afflicted with severe scoliosis. When it comes to this rare severity of scoliosis, it is estimated that merely 38,000 such surgeries are performed across the nation annually [3].

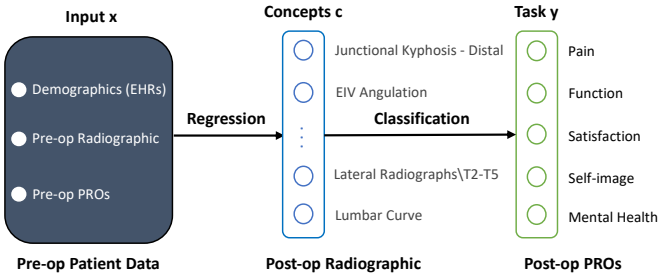

Fig. 1: Concept bottleneck model leverages the input of pre-operative patient demographics, radiographic measurements, and survey outcomes. It utilizes a set of post-operative radiographic as concepts to predict the post-operative PROs.

The Scoliosis Research Society 22R instrument (SRS-22R) questionnaire[1], initially designed as a disease-specific instrument, is widely adopted to measure health-related quality of life in patients with adolescent idiopathic scoliosis [2]. This instrument offers a nuanced understanding of patient outcomes and their well-being following medical interventions, facilitating continued enhancements in scoliosis care. It measures the quality of life with 22 questions across five main domains, including pain, self-image, mental health, satisfaction, and function [4]. PROs, along with objective surgical correction metrics determined by radiographic patient imaging, serve as an important subjective measurement of treatment success. However, pre-operative radiographic parameters currently used during surgical planning have not been shown to correlate with PROs in AIS [1].

The potential to accurately prognosticate surgical outcomes, inclusive of life quality metrics, presents significant benefits for AIS patients [5]. This predictive ability could serve as a crucial tool for patients and their families, allowing them to set up their expectations effectively. Moreover, it could enhance the efficacy of pre-operative deliberations and counseling sessions, thus ensuring a more comprehensive understanding and preparation for the forthcoming procedure.

This research has been supported by Shriners Children's and Georgia Institute of Technology in Greenville-Lexington Shriner Multi-site AI-enabled Rehabilitation Technology for Scoliosis Patients Care (GL-SMART) project. This research has also been supported by a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research to Professor May D. Wang.

[1]https://www.srs.org/professionals/online-education-and-resources/patient-outcome-questionnaires

In recent years, advanced deep learning techniques have been adopted in clinical settings to facilitate clinical decision support [6]. However, these models typically necessitate end-to-end training, which is often non-interactive, making it challenging to intercept errors that might transpire during training. Additionally, the black-box nature of deep learning models prevents their broader adoption in practical clinical settings [6]. To address this concern, we develop an explainable machine learning model, concept bottleneck neural network, to enable the post-operative PROs predictions for surgical decision-making of AIS patients (Fig. 1). Specifically, the proposed interactive predictive model can learn the patient radiographic measurements as important concepts and use these concepts to predict the self-reported answers to the post-operative SRS-22R questionnaire. The key contributions of this work include:

- We develop a concept bottleneck model for comprehensively predicting post-operative PROs with intermediate concepts to improve model transparency for wide adaptation in clinical settings.
- We utilize the implicit association between radiographic measurements and PROs to enhance the efficacy of the model; this association has not been explored or leveraged in prior research.
- The proposed clinical decision support system could serve as a shared decision-making support tool for AIS patients and families to set expectations about surgical outcomes in pre-operative clinical visits.

II. Related Works

Predicting the quality of life a patient experiences after a PSF surgical operation could prove invaluable for pre-operative counseling and patient rehabilitation [5], [7]. With recent advances in deep learning, researchers have adopted AI-enabled tools in clinical settings to predict subjective surgical outcomes and facilitate clinical decision support. In scoliosis studies, Ames et al. [8] proposed a binary classification approach for the post-operative SRS-22R responses prediction. This method categorizes the lower three responses (scores 1-3) as 'poor' and the upper two responses (scores 4-5) as 'good' across five domains. Besides, several studies [9], [10] sought to refine predictions further by focusing on specific domains, such as function and satisfaction. However, previous studies usually focused on a narrow domain or few specific questions and failed to provide comprehensive predictions to describe post-operative patient results.

End-to-end machine learning models are usually non-interactive, which complicates the process of error detection during training [6], [11], [12]. In the computer vision task, Koh et al. [13] introduced the idea of utilizing human-understandable concepts to interpret model behavior. They highlighted the potential of this training method to detect and rectify errors that may surface during the prediction of these concepts, and subsequently to adjust the final prediction. Similar innovative training methodology has undergone refinement and has found application across diverse domains [14]. However, to the best of our knowledge, none of the existing studies leverage an explainable or interactive training process for patient outcomes predictions to guide surgical decision-making.

III. Data Collection

The inclusion criteria includes pediatric patients from 10 to 18 years diagnosed with AIS who underwent PSF surgery. A cohort comprising 455 pediatric patients, with an average age of 11.98 years, from Shriners Children's hospitals satisfied the inclusion criteria. Complete pre-operative and post-operative follow-up data were available for 428 patients. All patients were enrolled in a protocol that had obtained Institutional Review Board (IRB) approval at each site. Pre-operative demographic information such as age, gender, and race, along with smoking history, comorbidities, and results of neurological assessments, were extracted from standardized EHRs during patient visits. Clinical experts manually measured and curated radiographic parameters from full-length, free-standing posterior/anterior and lateral spine radiographs.

IV. Methodology

In this study, we develop an interactive concept bottleneck model of post-operative PROs prediction for AIS patients, with a special focus on SRS-22R questionnaires.

A. Concept Bottleneck Model Learning Schema

Our proposed concept bottleneck model incorporates two distinct models: given the input pre-operative patient data x, one model focuses on learning the intermediate set of concept c (i.e., post-operative radiographic data), and another dedicates to classifying post-operative PROs y. In addition, we explore three distinct learning schemas, including independent learning, sequential learning, and joint learning:

- *Independent bottleneck* learns the ground truth labels $\hat{y} = \hat{f}(c)$ and concepts $\hat{c} = \hat{g}(x)$ independently. It uses the true concepts c and learned concepts $\hat{c}$ to predict $\hat{y}$ at the training and testing, respectively.
- *Sequential bottleneck* first learns $\hat{c} = \hat{g}(x)$, then using this learned concepts $\hat{c}$ it learns the labels $\hat{y} = \hat{f}(\hat{c})$.
- *Joint bottleneck* learns both concepts $\hat{c}$ and labels $\hat{y}$ by minimizing the weighted sum $\hat{y}, \hat{c} = \hat{f}(\hat{g}(x))$.

Although the independent and sequential bottleneck learns concepts similarly, they differ in terms of access to labels. The sequential bottleneck uses predicted concepts to learn the labels, while the independent bottleneck directly learns from the true concepts and uses the predicted concepts at testing time. On the other hand, the joint bottleneck uses a hyperparameter λ to control the trade-off between the concepts and label loss. Following previous studies [13], we set $\lambda = 0.01$ for all joint learning experiments with the best validation results.

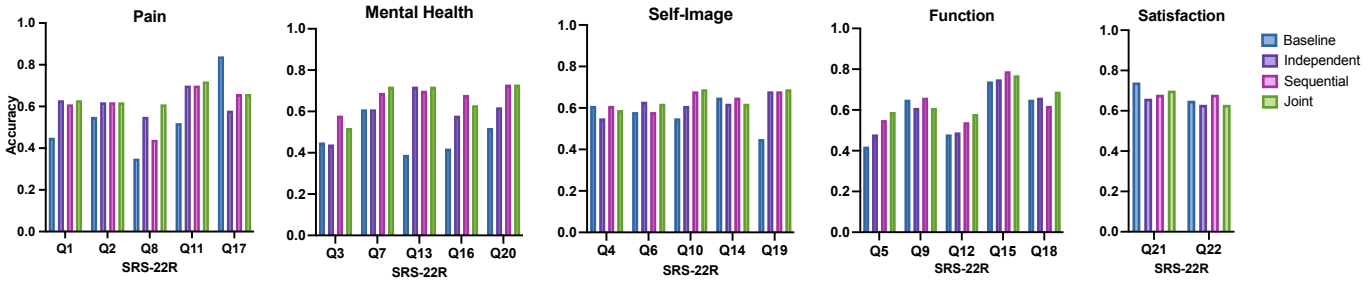

Fig. 2: Accuracy on predicting each individual outcome to the SRS-22R questionnaire. Questions are categorized into five different subjective evaluation domains, including function, mental health, pain, satisfaction, and self-image.

TABLE I: The average precision, recall, and F1 score across each domain for the different learning schema and baseline model.

Domain	Precision				Recall				F1 Score			
	Baseline	Ind.	Sequential	Joint	Baseline	Ind.	Seq	Joint	Baseline	Ind.	Sequential	Joint
Function	0.618	0.574	0.620	**0.646**	0.588	0.598	0.632	**0.648**	0.590	0.566	0.614	**0.624**
Pain	0.592	0.550	0.558	**0.614**	0.542	0.616	0.606	**0.636**	0.554	0.570	0.564	**0.604**
Mental Health	0.460	0.544	**0.640**	0.620	0.476	0.614	**0.676**	0.664	0.446	0.566	**0.642**	0.606
Self-Image	0.568	0.536	**0.596**	0.590	0.568	0.618	0.640	**0.642**	0.556	0.558	**0.608**	0.602
Satisfaction	**0.700**	0.565	0.630	0.630	**0.695**	0.645	0.680	0.665	**0.690**	0.580	0.625	0.630

B. Inputs, Concepts, and Outputs

The input for our model comprises patient demographic information, preoperative radiographic measurements, and preoperative PROs (results from the SRS-22R questionnaire). Subsequently, we utilize the 1-year post-operative radiographic measurements, which serve as follow-up data on surgical operation outcomes, as the concepts. These concepts encompass 76 features such as ThL-Lumbar Apical Translation, Lumbar Curve, and others. The final outputs or labels are the 1-year post-operative PROs in response to the SRS-22R questionnaire, which includes patient responses to each of the 22 questions. We split the entire patient data into 60%, 20%, and 20% for training, validation, and testing, respectively.

C. Post-operative PROs Prediction

Our proposed model includes using a deep neural network to discern the underlying concepts. This neural network comprises three Dense layers for the learning of non-linear mappings between features, augmented with two Batch Normalization layers for stability and speed. To ensure stability and speed, a Batch Normalization layer is connected to the outputs of the first two Dense layers. The activation function Rectified Linear Unit (ReLU) is integrated into the first two Dense layers, aiding the network in learning complex patterns and relationships. To prevent overfitting, we implement an early stopping monitor to keep track of the loss during training. Overall, this deep neural network model is designed to learn the underlying concepts efficiently and facilitate accurate predictions. The concept neural network was trained through 200 epochs, utilizing the Adam optimizer and Mean Absolute Squared Error (MAE) as the loss function. These parameters ensured that the models were well-optimized and able to accurately predict the desired concepts. For the final classification, we employ an additional Multi-Layer Perceptron (MLP) with Adam optimizer for multi-class classification.

V. Results and Discussions

A. Main Results

We evaluate the effectiveness of the proposed concept bottleneck model on the 1-year post-operative PROs prediction using multiple evaluation metrics, including accuracy, F1 score, precision, and recall. Fig. 2 presents the accuracy of each training schema on individual post-operative question outcomes across all domains. Additionally, we present the precision, recall, and F1 score in Table I to provide a more comprehensive evaluation of the performance of each training schema. Specifically, the F1 score provides a comprehensive evaluation of multi-class classification by computing the harmonic mean of precision and recall across each domain.

Upon comparing the individual performance of each training approach against the others and the baseline neural network without concept learning, it was observed that concept bottleneck models generally outperformed the baseline models across most questions. Specifically, the sequential and joint training methods manifested marginally better results in most instances. This trend is discernible in several other domains, excluding the general satisfaction domain, where the baseline surpassed the proposed training approaches in Q21, and the sequential approach outperformed other methods in Q22.

B. Learning Schema Comparison

We then evaluate the effectiveness of the three proposed training schemas, independent, sequential, and joint bottleneck. Specifically, we use MAE and RMSE, which measure the difference between the predicted and actual values to evaluate the performance of the concept prediction model. From Table II, we observed that the independent and sequential training achieved the highest MAE and RMSE, indicating an accurate estimation of the concept. The joint training method, which prioritizes the prediction of the labels, achieved relatively worse regression performance. Besides, the correlation between the predicted and true concepts is shown in Fig. 3.

TABLE II: The regression model performance of the concepts prediction with different learning schemas

Schema	MAE	RMSE
Independent	**2.07**	**3.46**
Sequential	**2.07**	**3.46**
Joint	2.26	3.86

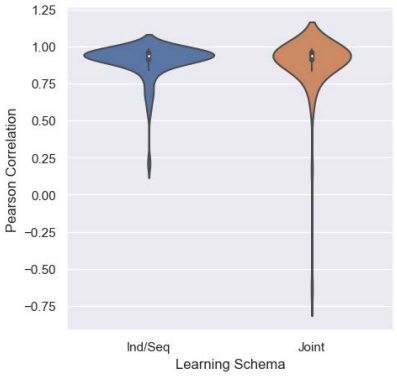

Fig. 3: The Pearson Correlation between predicted and true concepts within different learning schema.

It revealed that the various learning schemas are capable of predicting individual concepts accurately.

Comparing three learning schemas, we observed a similar performance on concept predictions between the sequential and independent learning schemas. The potential reason could be the training is contingent upon the concepts learned at both the training and inference stages. On the other hand, in the independent training approach, the training of 1-year post-operative PROs relies on the true concepts, with the learned concepts introduced to the model at the testing stage, as shown in Fig. 4. If errors arising during concept prediction are detected at the testing stage, the independent training approach is likely to achieve better performance, given that it learns based on the true concepts.

VI. Conclusion

We developed a clinical decision support system to facilitate shared decision-making between orthopedic surgeons and pediatric patients with idiopathic scoliosis. Considering the rare condition with PSF surgery, our study leveraged a relatively large pediatric AIS patient cohort from multiple clinical sites. We developed a concept bottleneck model for comprehensively predicting quality-of-life-related questions to improve model transparency for wide adoptions in real-world clinical scenarios. Specifically, we utilized the implicit relationship between radiographic measurements and PROs to enhance the efficacy of the model, which could be generalized into future studies. These efforts provide a scientific rigorous foundation for guiding shared clinical decision-making utilizing precision medicine to enhance pediatric scoliosis patient care. Our long-term objective is to evaluate the comparative efficacy of AIS treatments and to utilize patient quality-of-life instruments to personalize treatment decisions for pediatric patients.

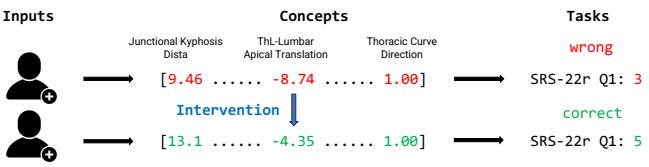

Fig. 4: Example of intervention within the independent training schema during test stage, leading to a better prediction.

References

[1] L. P. D'Andrea, R. R. Betz, L. G. Lenke, D. H. Clements, T. G. Lowe, A. Merola, T. Haher, J. Harms, G. K. Huss, K. Blanke *et al.*, "Do radiographic parameters correlate with clinical outcomes in adolescent idiopathic scoliosis?" *Spine*, vol. 25, no. 14, pp. 1795–1802, 2000.

[2] K. Chen, X. Zhai, K. Sun, H. Wang, C. Yang, and M. Li, "A narrative review of machine learning as promising revolution in clinical practice of scoliosis," *Annals of Translational Medicine*, vol. 9, no. 1, 2021.

[3] V. Chidambaran, L. Ding, D. Moore, K. Spruance, E. Cudilo, V. Pilipenko, M. Hossain, P. Sturm, S. Kashikar-Zuck, L. Martin *et al.*, "Predicting the pain continuum after adolescent idiopathic scoliosis surgery: a prospective cohort study," *European journal of pain*, vol. 21, no. 7, pp. 1252–1265, 2017.

[4] M. Dufvenberg, E. Diarbakerli, A. Charalampidis, B. Öberg, H. Tropp, A. Aspberg Ahl, H. Möller, P. Gerdhem, and A. Abbott, "Six-month results on treatment adherence, physical activity, spinal appearance, spinal deformity, and quality of life in an ongoing randomised trial on conservative treatment for adolescent idiopathic scoliosis (contrais)," *Journal of Clinical Medicine*, vol. 10, no. 21, p. 4967, 2021.

[5] P. Trobisch, O. Suess, and F. Schwab, "Idiopathic scoliosis," *Deutsches Ärzteblatt International*, vol. 107, no. 49, p. 875, 2010.

[6] F. Giuste, W. Shi, Y. Zhu, T. Naren, M. Isgut, Y. Sha, L. Tong, M. Gupte, and M. D. Wang, "Explainable artificial intelligence methods in combating pandemics: A systematic review," *IEEE Reviews in Biomedical Engineering*, 2022.

[7] W. Shi, F. O. Giuste, Y. Zhu, A. M. Carpenter, H. J. Iwinski, C. Hilton, J. M. Wattenbarger, and M. D. Wang, "A fhir-compliant application for multi-site and multi-modality pediatric scoliosis patient rehabilitation," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 1524–1527.

[8] C. P. Ames, J. S. Smith, F. Pellisé, M. Kelly, J. L. Gum, A. Alanay, E. Acaroğlu, F. J. S. Pérez-Grueso, F. S. Kleinstück, I. Obeid *et al.*, "Development of predictive models for all individual questions of srs-22r after adult spinal deformity surgery: a step toward individualized medicine," *European Spine Journal*, vol. 28, pp. 1998–2011, 2019.

[9] K. Hayashi, L. Boissière, D. Larrieu, A. Bourghli, O. Gille, J.-M. Vital, F. Guevara-Villazón, F. Pellisé, F. J. S. Pérez-Grueso, F. Kleinstück *et al.*, "Prediction of satisfaction after correction surgery for adult spinal deformity: differences between younger and older patients," *European Spine Journal*, vol. 29, pp. 3051–3062, 2020.

[10] K. Hayashi, L. Boissière, F. Guevara-Villazón, D. Larrieu, A. Bourghli, O. Gille, J.-M. Vital, F. Pellisé, F. J. S. Pérez-Grueso, F. Kleinstück *et al.*, "Mental health status and sagittal spinopelvic alignment correlate with self-image in patients with adult spinal deformity before and after corrective surgery," *European Spine Journal*, vol. 29, pp. 63–72, 2020.

[11] W. Shi, L. Tong, Y. Zhu, and M. D. Wang, "Covid-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2376–2387, 2021.

[12] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ser. ACM-BCB '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 233–240. [Online]. Available: https://doi.org/10.1145/3107411.3107445

[13] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5338–5348.

[14] C. Wu, S. Parbhoo, M. Havasi, and F. Doshi-Velez, "Learning optimal summaries of clinical time-series with concept bottleneck models," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 648–672.