Web-based Long-term Spine Treatment Outcome Forecasting

Hangting Ye*
School of Artificial Intelligence,
Jilin University
Changchun, China
yeht2118@mails.jlu.edu.cn

Zhining Liu*
University of Illinois
Urbana-Champaign
Champaign, United States
zhining.liu@outlook.com

Wei Cao
Microsoft Research
Beijing, China
weicao@microsoft.com

Amir M. Amiri
Microsoft Research
Redmond, United States
aamiri@microsoft.com

Jiang Bian†
Microsoft Research
Beijing, China
jiang.bian@microsoft.com

Yi Chang†
School of Artificial Intelligence,
Jilin University
Changchun, China
yichang@jlu.edu.cn

Jon D. Lurie
The Dartmouth Institute
Hanover, United States
jon.d.lurie@dartmouth.edu

Jim Weinstein
Microsoft Research
Redmond, United States
jim.weinstein@microsoft.com

Tie-Yan Liu
Microsoft Research
Beijing, China
tyliu@microsoft.com

ABSTRACT

The aging of global population is witnessing increasing prevalence of spinal disorders. According to latest statistics, nearly five percent of the global population is suffering from spinal disorders. To relieve the pain, many spine patients tend to choose surgeries. However, recent evidences reveal that some spine patients can self-heal over time with nonoperative treatment and even surgeries may not ease the pain for some others, which raises a critical question regarding the appropriateness of such surgeries [31, 32]. Furthermore, the complex and time-consuming diagnostic process places a great burden on both clinicians and patients. Due to the development of web technology, it is possible for spine patients to obtain decision making suggestions on the Internet. The uniqueness of web technology, including its popularity, convenience, and immediacy, makes intelligent healthcare techniques, especially Treatment Outcome Forecasting (TOF), able to support clinical decision-making for doctors and healthcare providers. Despite a few machine-learning-based methods have been proposed for TOF, their performance and feasibility are mostly unsatisfactory due to the neglect of a few practical challenges (caused by applying on the Internet), including *biased data selection*, *noisy supervision*, and *patient noncompliance*. In light of this, we propose `DeepTOF`, a novel end-to-end deep learning model to cope with the unique challenges in web-based long-term continuous spine TOF. In particular, we combine different patient groups and train a unified predictive model to eliminate the data selection bias. Towards robust learning, we further take advantage of indirect but fine-grained supervision signals to mutually calibrate with the noisy training labels. Additionally, a feature selector was co-trained with `DeepTOF` to select the most important features (i.e., answers/indicators that need to be collected) for inference, thus easing the use of `DeepTOF` during web-based real-world application. The proposed `DeepTOF` could bring great benefits to the rehabilitation of spine patients. Comprehensive experiments and analysis show that `DeepTOF` outperforms conventional solutions by a large margin[1].

CCS CONCEPTS

• **Applied computing** → **Forecasting**; **Health informatics**; • **Computing methodologies** → *Semi-supervised learning settings*; *Machine learning*; *Feature selection*.

KEYWORDS

Intelligent healthcare, Medical forecasting, Health informatics

ACM Reference Format:
Hangting Ye, Zhining Liu, Wei Cao, Amir M. Amiri, Jiang Bian, Yi Chang, Jon D. Lurie, Jim Weinstein, and Tie-Yan Liu. 2023. Web-based Long-term Spine Treatment Outcome Forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599545

1 INTRODUCTION

Spine disease is one of the most significant causes of disability in the world [32, 36]. Associated with a variety of clinical symptoms, including lower extremity pain and low back pain of varying levels of severity, spinal disorders usually lead to a significant reduction in the quality of life for patients [44] (we primarily focus on back pain-related spinal diseases in this paper). In the 2010 Global

*Both authors contributed equally to this research. This work was conducted when the first author was an intern at Microsoft Research Asia.
†Corresponding authors.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599545

[1]The source code are available at https://github.com/HangtingYe/DeepTOF.

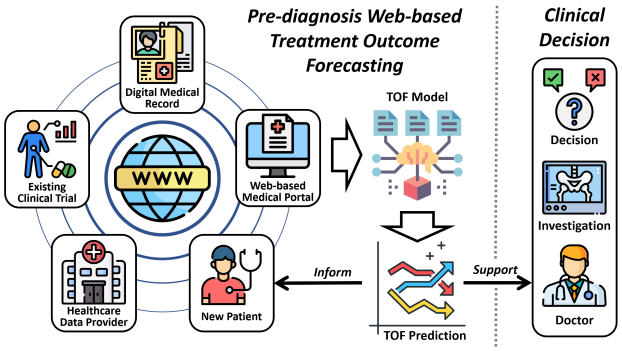


Figure 1: Web-based Treatment Outcome Forecasting.

Burden of Disease Study, among a total of 291 studied conditions, low back pain was ranked top in terms of years lost to disability, as it gave rise to 83 million disability adjusted life years lost in 2010 [19]. To ease the pain, spine patients are mainly treated with either surgical or nonoperative measures. Although it has long been believed that surgical treatment would be more effective, recent studies have found that some patients could self-heal over time and, more importantly, there was little difference between patients who underwent surgery and those who did not in the long run (2-8 years) [1, 31, 32]. This means that surgery is not suitable for every spine patient. It raises an important question: how to determine the most appropriate treatment for a specific patient?

Appropriate decision making during the spine disease diagnostic-treatment cycle is a complex activity [20], which requires extensive collaboration between patients and clinicians [23]. Clinicians have long envisioned the day when computers could make long-term medical predictions to assist patients in making decisions in complex clinical situations. Due to the development of web technology, it is possible for spine patients to obtain decision making suggestions through web and understand their condition in advance with machine learning techniques. Specifically, web-based healthcare tools have the following uniqueness: (i) Popularity. Web technology allows healthcare technology to serve a large population of spine patients. With the help of web technology, patients are no longer constrained by scarce medical resources, anyone can use web technology to know their condition in advance at any time and location. (ii) Convenience. Contrary to the time-consuming clinical diagnostic process necessary in a hospital (e.g. myelography and computed tomography (CT)), patients could provide information on the Internet by completing a short questionnaire. (iii) Immediacy. Patients can obtain treatment suggestion at the early diagnosis during disease diagnostic-treatment cycle. An early understanding of the spine condition will aid in patients' recovery during the subsequent treatment.

One important way to achieve informed decision making is Web-based Treatment Outcome Forecasting (WTOF). We first describe the Treatment Outcome Forecasting (TOF) problem. It allows for personalized prediction of the expected treatment outcome for each patient, and thus supports clinicians to anticipate the course of treatment and make better-informed decisions [34, 48, 62]. Furthermore, with the development of web technology, TOF systems can be widely deployed anywhere (healthcare organizations or patients' home) to support decision making. In this paper, we focus on the Web-based Long-term Continuous Spine TOF task: predicting the health conditions of a spine patient on WTOF system, with or without taking surgical treatment, in each of the next few years [63]. This specific task is supported by a web-based system to assist clinicians and patients at the early diagnostic part of the disease diagnostic-treatment cycle. It is important to note that the predicted health conditions are not intended to be equivalent to recommending surgery or not, and therefore, they do not substitute the expertise of clinicians. Inspired by the rapid development of digitization in the medical industry, an increasing number of efforts have attempted to leverage machine learning approaches to build up data-driven TOF models [26, 57]. However, very few of the existing approaches could achieve practically satisfactory performance due to their weakness in learning effective models from the biased, inconsistent, and noisy real-world data. Specifically, to construct an accurate WTOF model, one faces following key challenges:

- **Data selection bias.** WTOF system serves a large population of spine patients. Although the system could obtain enormous amounts of training data, there still exists data bias. The training data is naturally grouped as a patient can only undergo one kind of treatment (e.g., surgical or nonoperative treatment). To predict the outcome of different treatments, a straightforward solution is separate modeling, i.e., build multiple predictive models, each trained with data from patients who took a specific treatment. However, individual patient's treatment choice is usually highly dependent on a few critical factors (e.g., age and education level), mainly due to pre-existing surgical knowledge. This inevitably causes significantly variant data distribution across different subgroups of patients by treatment. As a result, sub models will carry over the data selection bias from their training data into the prediction of new patients. Such bias prevents the model from answering the question of "what if": what if a patient takes another treatment? Therefore, the first challenge is *how to build an unbiased TOF model with the presence of data selection bias across different treatment groups.*
- **Noisy supervision.** Although web technology has many benefits for spine TOF, it has a serious problem that the information submitted to the WTOF system is not obtained through clinical test. The target of the model prediction (which is also the supervision signal in model training) is a subjective health status score that is self-reported by the patients. However, many existing studies have shown that patient self-reporting is not always reliable [14, 25, 52]. This may disrupt model training and lead to inaccurate predictions. Other sources of information, such as health surveys, can easily be done online and provide more precise health status estimations, albeit this is not applicable to all patients. Hence, the second challenge is *how to exploit the indirect patient information to calibrate the training labels and thus getting a more robust prediction.*
- **Patient noncompliance.** Patients voluntarily provide all their personal information to the spine WTOF system. There are no rules that can compel patients to provide information due to the privacy and freedom attributes of the Internet. Therefore the final critical factor that affects the accuracy and feasibility of WTOF

models is the problem of patient noncompliance [16, 45, 50]. For example, subjects assigned to surgery may delay or subsequently refuse surgery, whereas nonsurgical subjects may ultimately seek and receive surgery. Such information might not be updated timely to WTOF system. A patient may also be absent from a particular survey or refuse to provide certain information for privacy or other concerns. This will result in missing data or even labels in the training set, and the model can only perform inference based on limited input information at test-time. Thus, the third challenge is *how to cope with patient noncompliance*, which includes *robust learning from inconsistent and incomplete data* as well as *robust inference given limited input information.*

In light of the above challenges and the great potential of deep learning in healthcare artificial intelligence applications [18, 34, 37], we propose a novel end-to-end deep learning framework DeepTOF for accurate and trustworthy web-based long-term continuous spine TOF. Different from prevailing methods that treat predictions for different treatments as separate tasks, DeepTOF is a unified learning framework that can handle multi-treatment outcome forecasting. Specifically, all the massive patients data are used to train DeepTOF. To eliminate the data selection bias with respect to the assigned treatment, we introduce counter-factual training [6] to enforce the model to learn treatment-agnostic representations from the data. DeepTOF also makes additional predictions about the patient's physical condition during training. These extra predictions are used to mutually calibrate with the noisy training labels, thus exploiting the finer-grained supervision to achieve robust learning. In response to the missing training data and labels brought about by patient noncompliance and data unification, we adopt semi-supervised learning to better mine information from incomplete supervision. To sum up, this paper makes following 3-fold contributions:

- We present the first systematic discussion of the difficulties and challenges of building learning-based models at early diagnosis of the diagnostic-treatment cycle for Web-based Long-term Continuous Spine Treatment Outcome Forecasting, including data selection bias, noisy supervision, and patient noncompliance.
- To handle the unique challenges of this complex task, we develop the DeepTOF, which demonstrates superior performance in real-world spine WTOF task. To our best knowledge, it is the first unified end-to-end deep learning framework for spine WTOF. And it has a great effect on the rehabilitation of spine patients.
- We also present a detailed quantitative analysis and discussion of real-world clinical spine patients TOF data with respect to commonly adjusted demographic variables: gender, race, and education level. Our findings can shed some light on developing better WTOF algorithms for similar tasks.

2 RELATED WORK

In this section, we systematically review the existing works related to machine-learning-based treatment outcome forecasting and deep learning applications in medical Artificial Intelligence (AI).

2.1 Medical Treatment Outcome Forecasting

The forecasting problem is a big topic in medical artificial intelligence (AI). It encompasses many different kinds of tasks, such as forecasting of health risk [9, 21, 58], health-care costs [5], treatment compliance [4], and even pandemic trends of COVID-19 [13, 33]. In this paper, we focus on Treatment Outcome Forecasting (TOF), which is one of the most common and challenging medical forecasting problems for its importance in improving patient-centered healthcare outcomes [57]. Canonical regression analysis is the most common method for TOF, but recent studies show that utilizing machine learning algorithms provides better results in many scenarios [49, 65]. As a result, a few machine-learning-based TOF methods have been proposed in recent years. For example, *Rahman et al.* [42] used Support Vector Machine (SVM) to predict pain volatility, *Wang et al.* [60] adopted naïve Bayes for oral health prediction, and *d'Hollosy et al.* [35] predicted low back pain with decision tree. Recently, *Verma et al.* [57] provided a systematic review of works that apply machine learning for predicting treatment outcomes.

Table 1: An overview of existing TOF settings.

Work	Task Type	Length of Prediction	Target of Prediction	Output Space
[35]	CLF	immediate	referral advice	$\{0, 1, 2\}$
[2]		10 weeks	level of depression	$\{0, 1\}$
[42, 43]		6 months	low/high pain volatility	$\{0, 1\}$
[15]		1 year	presence of complication	$\{0, 1\}$
[12]		2 years	significant outcome	$\{0, 1\}$
[8]	REG	12 weeks	quality of life	$\mathbb{R}$
[47]		2 years	level of depression	$\mathbb{R}$
This work	**LC-REG**	**1-8 years**	**bodily pain & physical function**	$\mathbb{R}^{2\times2\times T}$

* Classification (CLF), Regression (REG), Long-term Continuous Regression (LC-REG).

However, many existing TOF methods are simply based on linear learning algorithms, e.g., logistic regression [8, 47], linear kernel SVM [12, 42, 43], decision tree [35], and quadratic discriminant analysis [15]. Many of them also tested ensemble learning frameworks, such as random forest [35, 42, 43] and gradient boosting machine [60]. Table 1 summarizes the different TOF settings of these related works. It is worth noting that most of them, though applying machine learning methods, did not focus on making targeted improvements to the models for the TOF problem. More importantly, the vast majority of existing research works simplified the TOF into binary classification [2, 12, 15, 35, 43], while continuous-value prediction [57] can create more practical value since it can facilitate finer-grained treatment evaluation. We also must note that most existing works follow the typical TOF setting, which only requires the model to provide the "taking treatment" prediction for a single time-point in the future (e.g., 6-months later). In contrast, we consider the most challenging long-term spine TOF setting: the model needs to give separate "taking treatment" and "not taking treatment" predictions in each of the next 8 years for a given patient.

2.2 Deep Learning for Medical AI

For nearly half a century, logic-based methods like expert systems and graphical models are dominant in medical artificial intelligence. However, we have witnessed the recent rapid uptake of Deep Learning (DL) in various fields of intelligent healthcare for its demonstrable strengths in intricate pattern recognition and predictive model building [18, 34]. Thus far, DL has been primarily employed

in image-intensive domains like radiology, radiotherapy, pathology, ophthalmology, dermatology, and image-guided surgery [34].

In the field of medical forecasting, existing research efforts have mainly focused on employing DL for health risk prediction. Specifically, they apply deep learning models to predict a healthy individual's risk of developing a particular type or class of disease(s), e.g., chronic diseases [10], Parkinson's disease [64], Alzheimer's disease [29], and cancers [59, 66]. We can see that the health risk prediction is oriented to healthy people and does not involve medical treatments and corresponding outcome assessment. So it is a task distinct from treatment outcome forecasting which focuses on assessing the effectiveness of medical treatments for patients [57].

We notice that there are a few recent studies that try to apply artificial neural networks for TOF [39, 41, 47], but are limited to direct application of fully connected Multi-Layer Perceptron (MLP), without any adaptive modification to cope with the unique challenges in TOF tasks [57]. To the best of our knowledge, our work is the first to discuss the difficulties and challenges of building data-driven predictive models for spine TOF. The proposed DeepTOF is also the first unified deep learning framework designed for long-term continuous spine treatment outcome forecasting.

3 METHODOLOGY

In this section, we will first cover some preliminaries, including a description of the background task of this paper, i.e., the application scenarios, the *Spine Patient Outcomes Research Trial* [2] (**SPORT**), the definitions of basic notations, and the formulation of web-based long-term continuous spine TOF problem. After that, we will elaborate the details of DeepTOF framework.

3.1 Preliminaries

Application Scenarios. The proposed DeepTOF is a spine TOF system deployed on the Internet. Our goal is to assist decision making at early diagnosis during spine disease diagnostic-treatment cycle. For instance, before a patient goes to the hospital for treatment, some personal information (which will be kept confidential) can be provided to the WTOF system and then each treatment's outcomes in the future will be returned. This can be done on the Internet, eliminating a lot of complex diagnostic processes. We first use a local real-world dataset to train DeepTOF and then release the trained model on the Internet for use. While with the patient's consent and adhering to the principle of protecting patient privacy, we use the data gathered by the WTOF system to further fine-tune the model. For more detailed information of how the DeepTOF utilizes web system, please refer to the Appx. A.

The SPORT dataset. SPORT [1, 63] is a clinical trial comparing surgical and nonoperative treatment for several different back conditions. It was conducted at 13 US medical centers with multidisciplinary spine practices across 11 states, with a total of 2505 participants enrolled. The SPORT dataset provides more than 130 attributes about the patient that could be used for forecasting. The primary outcome measures are changes from baseline for the *bodily pain* (**BP**) and *physical function* (**PF**) scales, which are also the targets of model predictions. For available patients, SPORT also collected their self-reported scores on the *Short Form-36* (**SF-36**), which is a widely-used generic health survey with 36 questions, scored from 0 (low) to 100 (high) [61]. All the outcome scores (BP, PF, and SF-36 if available) were measured in follow-up visits at 6 weeks, 3 months, 6 months, and yearly out to **8 years** after diagnosis. SPORT is a questionnaire form of data that could easily be applied to the web service. Specifically, the SPORT data defines a challenging long-term continuous spine TOF problem: *a model takes the patient's information as input, and simultaneously predicts BP and PF for each of the next 8 years given he/she takes the surgical or nonoperative treatment.*

[2]funded by the NIH, supported by grant (U01-AR45444-01A1).

Table 2: Definitions of basic notations.

Notation	Definition
d	Number of input dimensions (features).
T	Maximum time step.
N	Number of patients.
$\mathcal{X} : \mathbb{R}^d$	Input feature space.
$\mathcal{Y} : \mathbb{R}^{2\times2\times T}$	Output label space. (2 treatments × 2 outcomes × T timesteps)
$\mathcal{S} : \mathbb{R}^{2\times36\times T}$	SF-36 score space. (2 treatments × 36 questions × T timesteps)
$\mathcal{T} : \{\tau_n, \tau_s\}$	Treatment space, where τ_n/τ_s denotes nonoperative/surgical treatment.
$p^i : (x^i, y^i, s^i, \tau^i)$	Data of patient p^i, where $x^i \in \mathcal{X}, y^i \in \mathcal{Y}, s^i \in \mathcal{S}, \tau^i \in \mathcal{T}$.
$D : \{p^i\}_{i=1,2,...,N}$	Dataset D of N patients.

Notations and Problem Formulation. With the previous description, we can now formally define the spine WTOF problem considered in this paper. Formally, let d denote the number of input features, T as the maximum time steps of prediction, N as the number of patients, and $\mathcal{T} : \{\tau_n, \tau_s\}$ as the treatment space, where τ_n and τ_s represent the nonoperative and surgical treatments, respectively. We can define the input space $\mathcal{X} : \mathbb{R}^d$ and the output space $\mathcal{Y} : \mathbb{R}^{2\times2\times T}$, since the model needs to predict 2 outcomes (BP and PF) of 2 very different treatments (τ_n and τ_s) at each of the T time steps. Similarly, the SF-36 score space is $\mathcal{S} : \mathbb{R}^{2\times36\times T}$. Then the data of a patient p^i can be defined as $p^i : (x^i, y^i, s^i, \tau^i)$, where $x^i \in \mathcal{X}, y^i \in \mathcal{Y}, s^i \in \mathcal{S}$, and $\tau^i \in \mathcal{T}$. A dataset D including N patients is thus $D : \{p^i\}_{i=1,2,...,N}$. Table 2 summarizes the above notation definitions. Please note that for simplicity, we use **X**, **Y**, **S**, **T** to represent the inputs, outputs, sf-36 scores, and treatments of the dataset D unless otherwise stated. Specifically, $x^i : [x^i_1, x^i_2, ..., x^i_d]$ is the feature vector of p^i, where x^i_k represents the k-th feature. For label y^i, the first dimension corresponds to the choice of treatment, the second to different outcomes, and the third to the time steps. For clarity, we use $BP^i_{t,\tau}/PF^i_{t,\tau}$ to represent the BP/PF score of patient p^i at the t-th time step taking the treatment τ, where $\tau \in \mathcal{T}$. Likewise, $SF^{i,j}_{t,\tau}$ denotes the j-th score of SF-36 under the same condition. Let's denote the predictive model as $F(\cdot)$, then the objective of long-term continuous spine TOF is to *learn an accurate* $F : \mathcal{X} \to \mathcal{Y}$ *from the given dataset* $D : \{p^i\}_{i=1,2,...,N}$.

3.2 The Proposed Framework

We now introduce our DeepTOF framework. To start with, let θ be the learnable parameters of predictive model $F(\cdot;\theta)$ and $\mathcal{L}$ be the

loss function. Then in order to learn a spine WTOF model F from dataset D, we aim to solve the following optimization problem:

$$\underset{\theta}{\operatorname{argmin}} \mathcal{L}(F(\mathbf{X};\theta),\mathbf{Y}). \tag{1}$$

However, as discussed in Section 1, directly optimizing this objective will lead to sub-optimal performance due to the unique challenges in spine WTOF: *data selection bias*, *noisy supervision*, and *patient noncompliance*. Thus, the primary motivation of our DeepTOF design is to overcome these issues. Specifically, within the whole DeepTOF framework, we take advantage of *counter-factual training* to obtain unbiased data representation, *multi-task calibration* to conquer noisy supervision, and *semi-supervised learning* to address patient noncompliance. Furthermore, we propose using *feature selection* to make the DeepTOF more clinically feasible. Please refer to Figure 2 for an overview of the proposed DeepTOF framework. We will describe these components in detail in the rest of this section.

3.2.1 Counter-factual training. As mentioned before, the data distribution varies across different treatment subgroups, thus introducing *data selection bias* to the model predictions. Unlike traditional clinical studies, we are unable to do the randomized control trials for web-based services. Thus, the treatment on any individual patient depends on his/her input features x^i. Simply building a model F_{τ_n} (F_{τ_s}) based on D_{τ_n} (D_{τ_s}) and directly using it to predict the outcome of a new patient actually assumes that this new patient was implicitly assigned to the treatment τ_n (τ_s). To counter such a bias in regards to the assigned treatment, we propose the counter-factual training to drive DeepTOF to learn treatment-agnostic representations from the complete ungrouped dataset D.

Specifically, we aim to learn an unbiased parameterized feature extractor $G_f : \mathcal{X} \to \mathcal{R}$, where $\mathcal{R}$ is the representation space. A treatment classifier $G_\tau : \mathcal{R} \to \mathcal{T}$ is used to detect treatment information in $G_f(\mathbf{X})$: better classification accuracy indicates more discriminative information w.r.t. the assigned treatment that retained in $G_f(\mathbf{X})$. Formally, let $G_f(\cdot;\theta_f)$ be the feature extractor of F. In addition to the label predictor $G_y(\cdot;\theta_y) : \mathcal{R} \to \mathcal{Y}$, we add the treatment classifier $G_\tau(\cdot;\theta_\tau)$, and re-write the original objective described in Eq. (1) as:

$$\underset{\theta_f,\theta_y}{\operatorname{argmin}} \mathcal{L}(G_y(G_f(\mathbf{X};\theta_f);\theta_y),\mathbf{Y}). \tag{2}$$

To achieve counter-factual training, we additionally optimize:

$$\underset{\theta_f}{\operatorname{argmax}} \ \underset{\theta_\tau}{\operatorname{argmin}} \ \mathcal{L}(G_\tau(G_f(\mathbf{X};\theta_f);\theta_\tau),\mathbf{T}) \tag{3}$$

This objective encompasses two coupled aspects. On one hand, G_τ is optimized to perform accurate treatment classification, so as to find all possible discriminative information that remains in the extracted representations $G_f(\mathbf{X})$; on the other hand, the G_f is optimized adversarially to learn treatment-agnostic representations to disturb G_τ. This allows DeepTOF to remove treatment-related information to the greatest extent, thereby eliminating *data selection bias*. The adversarial optimization of θ_f is achieved by adding a gradient reversal layer R_λ, which will multiply the backpropagate gradient by a negative number $-\lambda$, i.e., $\theta_f = \theta_f - \mu(-\lambda\frac{\partial\mathcal{L}}{\partial\theta_f})$, where μ is the learning rate. With R_λ, we can reformulate the objective in Eq. (3) as:

$$\underset{\theta_f,\theta_\tau}{\operatorname{argmin}} \ \mathcal{L}(G_\tau(R_\lambda(G_f(\mathbf{X};\theta_f));\theta_\tau),\mathbf{T}) \tag{4}$$

3.2.2 Multi-task calibration. As we know, the training labels $\mathbf{Y}$ in spine WTOF are treatment outcomes that are self-reported by the patients. However, in medical research, many factors associated with patient characteristics, such as the education level, can affect the reliability and accuracy of self-reporting [14, 25]. Therefore, the training labels $\mathbf{Y}$ may be incorrect, leading to *noisy supervision* in spine WTOF training. In response to this, we introduce multi-task calibration to simultaneously (1) exploit extra fine-grained supervision and (2) calibrate the noisy training labels, thus achieving more robust learning from *noisy supervision*.

Firstly, we add a SF-36 score predictor $G_s(\cdot;\theta_s) : \mathcal{R} \to \mathcal{S}$. Compared to the BP and PF, SF-36 [61] provides a more detailed evaluation of patients' health condition. It also comes with detailed instructions to help patients understand the questionnaire and give a more accurate self-assessment[3], which makes it a reliable source of external supervision signals. Formally, we optimize:

$$\underset{\theta_f,\theta_s}{\operatorname{argmin}} \mathcal{L}(G_s(G_f(\mathbf{X};\theta_f);\theta_s),\mathbf{S}). \tag{5}$$

On the other hand, we incorporate medical domain knowledge to calibrate the noisy training labels (i.e., BP and PF) with the SF-36 score predictions. Note that for clarity, we use the $\widehat{*}$ to denote the prediction value of $*$, e.g., $[\widehat{BP^i},\widehat{PF^i}] := G_y(G_f(x^i))$ and $\widehat{SF^i} := G_s(G_f(x^i))$. Then according to [27, 61], BP and PF score could also be derived from the predicted SF-36 scores by:

$$\widetilde{BP} = \frac{1}{2}\sum_{i=21}^{22}(\widehat{SF}_i);\ \widetilde{PF} = \frac{1}{10}\sum_{i=3}^{12}(\widehat{SF}_i), \tag{6}$$

where the subscript i denotes the i-th score of SF-36. The $\widehat{BP},\widehat{PF}$ and $\widetilde{BP},\widetilde{PF}$ are then used for mutual calibration, thereby achieving more robust learning for both $G_y(\cdot;\theta_y)$ and $G_s(\cdot;\theta_s)$:

$$\underset{\theta_f,\theta_y,\theta_s}{\operatorname{argmin}} \mathcal{L}([\widehat{BP},\widehat{PF}],[\widetilde{BP},\widetilde{PF}]) \ \Leftrightarrow \ \underset{\theta_f,\theta_y,\theta_s}{\operatorname{argmin}} \mathcal{L}(\widehat{\mathbf{Y}},\widetilde{\mathbf{Y}}). \tag{7}$$

3.2.3 Semi-supervised learning. We note that a considerable portion of the data is missing due to *patient noncompliance*. For example, patients' absence from follow-up visit(s) or death will cause missing outcome reports in $\mathbf{Y}$. Moreover, one may notice that with our uniform modeling strategy, half of the training labels are counterfactual that do not exist in the original data, e.g., for a patient p_i whose $\tau_i = \tau_n$, the outcome labels of taking another treatment τ_s (such as BP_{τ_s}, PF_{τ_s}) are absent in $\mathbf{Y}$. All these factors lead to a significant number of missing values in the training set, which motivates us to introduce semi-supervised learning into DeepTOF.

Inspired by [54], we propose a simple yet effective solution to this problem by maintaining a temporal moving average teacher network $F_{teacher}(\cdot;\theta_{teacher})$ for the student model $F(\cdot;\theta)$. Note that F is defined as the combination of feature extractor G_f and label predictor G_y, i.e., $F(x;\theta) := G_y(G_f(x;\theta_f);\theta_y)$. Specifically, during the learning process of the student network F, the weights (aka

[3] An online example of SF-36: SF-36 Score (Short Form Health Survey).

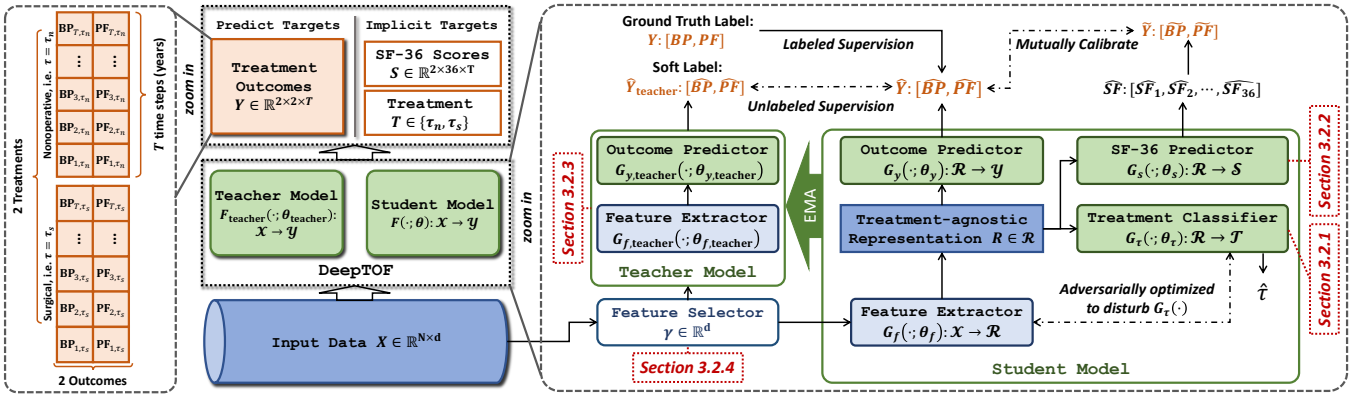


Figure 2: The DeepTOF framework.

parameters) $\theta_{teacher}$ of teacher network are simply the Exponential Moving Average (EMA) of student model weights θ. Formally,

$$\theta_{teacher,t} = \alpha \cdot \theta_{teacher,t-1} + (1-\alpha) \cdot \theta_t, \tag{8}$$

at the training step t, with a smoothing coefficient hyper-parameter α. The student model F is encouraged to make consistent predictions with $F_{teacher}$ as the EMA teacher is generally more accurate than the student [40]. For labeled data (x, y), F learns from both the ground truth label y and the output of $F_{teacher}$. For unlabeled data $(x, -)$, F learns solely from the teacher's prediction $F_{teacher}(x)$, i.e., $F_{teacher}(x)$ is used as a soft target for x. Then learning from the teacher model corresponds to the following objective:

$$\underset{\theta,\theta_{teacher}}{\text{argmin}} \ \mathcal{L}(F(\mathbf{X};\theta), \mathbf{Y}) + \mathcal{L}(F(\mathbf{X};\theta), F_{teacher}(\mathbf{X};\theta_{teacher})). \tag{9}$$

3.2.4 Feature extractor. As mentioned above, the feature extractor G_f is crucial to the process of mining information. The output label $y^i \in \mathbb{R}^{2\times2\times T}$ of an individual can be viewed as a sequence, we consider that the knowledge gained from from previous timesteps may be useful for the model to forecast future outcomes. To better capture the dependencies in physical condition during the follow-up visit(s) under a certain treatment, we adopt attention mechanism which is achieved through the Transformer Encoder [56]. The Transformer Encoder could relate different positions of a sequence to compute a representation. Specifically, we first map the $\mathbf{X} \in \mathbb{R}^{N\times d}$ to a sequence representation $\mathbf{H} \in \mathbb{R}^{N\times T\times h}$, where T is the maximum timesteps of prediction and h is hidden dimension. In addition, positional embeddings of different timesteps are added to $\mathbf{H}$, which are then fed to the Transformer Encoder. While extracting information for a certain timestep, Transformer Encoder enables the feature extractor to adaptively learn dependencies from previous timesteps. As a result, the extracted representation $G_f(\mathbf{X})$ contains both feature and temporal dependencies information.

3.2.5 Feature selection. Finally, proper feature selection is also critical to make the spine WTOF model clinically feasible. In practice, patients are less likely to be equally cooperative due to issues like privacy and cost. To this end, we co-train an adaptive feature selector with DeepTOF to find the most important questions/indicators, reducing the information needed for inference, and thus largely increase the inference efficiency of DeepTOF. By modeling the feature

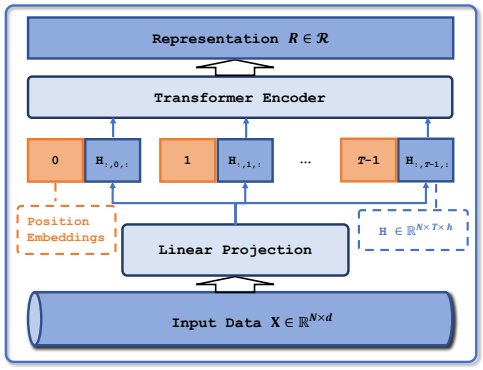


Figure 3: The Feature Extractor of DeepTOF, $G_f : \mathcal{X} \to \mathcal{R}$

selector as a input mask layer γ, we optimize:

$$\begin{aligned} \underset{\theta,\ \gamma}{\text{argmin}} \quad & \mathcal{L}(F(\gamma \circ \mathbf{X};\theta), \mathbf{Y}) \\ \text{subject to} \quad & \gamma \in [0,1]^d, \\ & \frac{\|\gamma\|_1}{\|\gamma\|_2^2} + \max(0, |M - \|\gamma\|_1|) = 1, \end{aligned} \tag{10}$$

where M is the number of features to be retained, and "$\circ$" denotes the element-wise product operation. In the second constraint, the first term $\frac{\|\gamma\|_1}{\|\gamma\|_2^2}$ forces the mask vector $\gamma : [\gamma_1, \gamma_2, ..., \gamma_d]$ to be sparse, i.e., all components of γ to be close to 0 or 1, while the second term motivates $\|\gamma\|_1 \to M$, i.e., only M components close to 1. Thereby, we can get a binary feature selection mask code γ with length d, where $\gamma_i = 0/1$ represents to drop/keep the i-th feature.

3.2.6 Objective function. Altogether, DeepTOF aims to minimize the following objective function w.r.t. $\theta_f, \theta_y, \theta_s, \theta_\tau$ and $\theta_{teacher}$:

$$\begin{aligned} \mathcal{O} = & \mathcal{L}(G_y(G_f(X;\theta_f);\theta_y), \mathbf{Y}) + \\ & \mathcal{L}(G_\tau(R_\lambda(G_f(X;\theta_f));\theta_\tau), \mathbf{T}) + && \triangleright \text{Section 3.2.1} \\ & \mathcal{L}(G_s(G_f(X;\theta_f);\theta_s), \mathbf{S}) + \mathcal{L}(\widehat{\mathbf{Y}}, \widetilde{\mathbf{Y}}) + && \triangleright \text{Section 3.2.2} \\ & \mathcal{L}(F(\mathbf{X};\theta), F_{teacher}(\mathbf{X};\theta_{teacher})). && \triangleright \text{Section 3.2.3} \end{aligned} \tag{11}$$

Table 3: Comparison of all models' predictive performance in terms of Bodily Pain (BP) and Physical Function (PF) with a different number of features. Feature selector was co-trained with DeepTOF to select a different number of features (i.e. M is set to 20, 30, 40, 50, 60 and 70, here, full represents 131 features). Using M features as input, the results show the comparison of all models' performance in terms of correlation. For each method, correlation scores averaged over surgical and nonoperative treatment are reported (mean±std, higher is better). The reported performance is averaged over 10 independent runs. The best results are highlighted in bold.

Correlation of Bodily Pain (BP) predictions with different number of input features M							
Method \ Performance	M=20	M=30	M=40	M=50	M=60	M=70	Full
Lasso [55]	0.4204±0.021	0.4201±0.030	0.4280±0.023	0.4148±0.014	0.4365±0.022	0.4230±0.012	0.4188±0.012
SVR [3]	0.4085±0.013	0.4160±0.009	0.4056±0.007	0.4015±0.003	0.3956±0.001	0.4193±0.008	0.4145±0.007
K-NN [38]	0.4019±0.006	0.4127±0.003	0.4302±0.010	0.4067±0.012	0.4275±0.011	0.4133±0.009	0.4187±0.014
RandomForest [7]	0.4099±0.018	0.4167±0.005	0.4129±0.020	0.4031±0.008	0.4242±0.003	0.4047±0.012	0.4023±0.012
LightGBM [24]	0.4086±0.006	0.4093±0.004	0.4141±0.010	0.4080±0.004	0.4221±0.0120	0.4105±0.012	0.4059±0.020
ResNet [17]	0.4286±0.007	0.4376±0.002	0.4358±0.024	0.4320±0.002	0.4274±0.008	0.4233±0.013	0.4288±0.006
DeepTOF	**0.4596**±0.011	**0.4459**±0.021	**0.4402**±0.030	**0.4595**±0.004	**0.4406**±0.021	**0.4469**±0.024	**0.4583**±0.036

Correlation of Physical Function (PF) predictions with different number of input features M							
Method \ Performance	M=20	M=30	M=40	M=50	M=60	M=70	Full
Lasso [55]	0.4174±0.026	0.4283±0.016	0.4176±0.033	0.4132±0.017	0.4125±0.021	0.4213±0.015	0.4138±0.013
SVR [3]	0.4209±0.001	0.4216±0.013	0.4303±0.011	0.4228±0.017	0.4202±0.015	0.4142±0.020	0.4116±0.005
K-NN [38]	0.4067±0.015	0.4190±0.010	0.4213±0.007	0.4228±0.001	0.4308±0.026	0.4332±0.013	0.4290±0.013
RandomForest [7]	0.4118±0.013	0.4137±0.004	0.3981±0.018	0.4174±0.004	0.4059±0.022	0.4210±0.009	0.4124±0.005
LightGBM [24]	0.4063±0.014	0.4009±0.005	0.4033±0.004	0.4070±0.024	0.4179±0.010	0.3980±0.017	0.4076±0.008
ResNet [17]	0.4434±0.019	0.4543±0.015	0.4572±0.018	0.4650±0.005	0.4589±0.009	0.4503±0.005	0.4513±0.001
DeepTOF	**0.4890**±0.022	**0.4944**±0.004	**0.4836**±0.008	**0.4921**±0.003	**0.4861**±0.011	**0.4869**±0.016	**0.4992**±0.013

4 EXPERIMENT & ANALYSIS

In this section, we conduct experiments to evaluate the performance of DeepTOF in real-world spine WTOF task. We also discuss the importance of the selected features and the impact of patient characteristics on the predictive model. Finally, the ablation study validates the effectiveness of the proposed techniques (i.e., counter-factual training, multi-task calibration, and semi-supervised learning).

4.1 Experimental Setup

4.1.1 Data settings. The format of the data collected by WTOF system is consistent with the *Spine Patient Outcomes Research Trial* (**SPORT**, *funded by the NIH, supported by grant (U01-AR45444-01A1)*) described in Section 3.1, where the inputs are all feature-based. And thus it is easy for patients to fill out the input information online. We first train DeepTOF by local real-world dataset SPORT and then release DeepTOF on the Internet for use. We intend to use the data collected by the spine WTOF system (we will protect patient privacy) to fine-tune DeepTOF. Due to the lack of data collected on the Internet, we use the real-world SPORT dataset in our experiments to validate the effectiveness of DeepTOF. The SPORT dataset contains 2430 valid patient samples (of the total 2505 patients, 2430 had at least one follow-up through eight years), each of which can be defined as $p^i : (x^i, y^i, s^i, \tau^i)$. The input x^i has 131 features, which can be divided into 5 groups. Each group has the following number of features: i) Basic information: 13; ii) Self evaluation: 8; iii) Symptom: 34; iv) Previous treatment: 46; v) Health condition: 30. Since in SPORT, the outcomes are measured yearly out to 8 years after diagnosis, DeepTOF makes predictions annually with a total time step T=8. Therefore, the output label y^i and SF-36 score s^i can be rewritten as $y^i \in \mathbb{R}^{2\times2\times8}$ and $s^i \in \mathbb{R}^{2\times36\times8}$. The statistic of missing values in dataset is as follows: i) Input features: 10.0%; ii) Output label: 64.4%; iii) SF-36 score: 65.1% [4]. We utilize three-fold cross-validation to obtain reliable and stable evaluations. To reduce the effect of randomness, the reported performance is averaged over 10 independent runs.

4.1.2 Experimental details. For the Transformer Encoder component of feature extractor, the number of layers is set to 6, hidden dimension is set to 128, the activation function in each layer is Leaky ReLU and the number of heads in the multiheadattention models is set to 8. The target of the model prediction is a subjective health status score that is self-reported by the patients. Many studies have shown that self-reporting is not always reliable [14, 25, 52]. It may be challenging to measure model performance by directly evaluating prediction accuracy. However, an individual's subjective sensations may alter as his body's actual state of health does, which shows that the trend of the outcome from time to time is significant. As a consequence, we evaluate the prediction performance on the correlation between predictions and output labels. Here, we adopt the Pearson Correlation Coefficient as our evaluation metric. A higher correlation represents a better prediction. Collecting a large number of features indeed requires many human efforts that may cause a big burden on both doctors and patients. To ease the burden, we introduce a feature selector (described in Section 3.2.5). Here the number of features to be retained (M) is set to 20, 30, 40, 50, 60 and 70, i.e. feature selector is co-trained with DeepTOF to select the M most important features. Then these selected features are used as input to re-train DeepTOF.

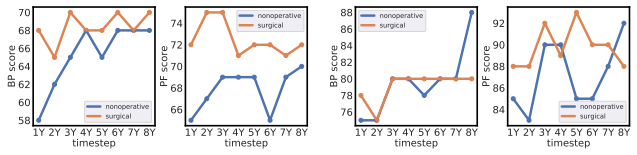
[4]Note that, if there exists missing SF-36 score, Eq. 6 will be adjusted to compute the mean of all available scores.

4.2 Experimental Results & Analysis

4.2.1 Baselines Compared. A few research efforts [35, 42, 43] have been tried to adopt machine-learning methods for spine disease TOF (Section 2.1). Here we adopt five popular conventional machine-learning algorithms as our baseline models. To our best knowledge, we are the first to design a unified end-to-end deep learning framework for spine WTOF. In order to fully illustrate the contribution of DeepTOF, we also compare it against the ResNet [17]. The ResNet is a deep learning architecture, which has been widely used in a variety of tasks for its capability of learning discriminative representations from complex data. The baseline models are as follows: **Lasso** [55], **SVR** [3], **K-NN** [38], **RandomForest** [7], **LightGBM** [24] and **ResNet** [17] (please refer to the Appx. A).

Note that the conventional machine-learning algorithms (i.e. Lasso, SVR, K-NN, RandomForest and LightGBM) are not naturally designed to handle multi-target prediction, thus we adopt separate modelling to implement them for multi-treatment continuous spine TOF. Specifically, for each of them, we train a model for each of the 32 dimensions of the output label space $\mathbf{Y} : \mathbb{R}^{2\times2\times8}$, each of which represents the value of BP or PF in a certain year under a certain treatment. Meanwhile, we adopt a 6-layer neural network ResNet, and hidden dimension is set to 128. All hyper-parameters of baselines are carefully tuned to achieve their optimal performance.

4.2.2 Main Results. We compare the performance of DeepTOF and baseline methods with a different number of features, which are reported in Table 3. We can observe that DeepTOF outperforms all baseline methods in all settings, which reveals the benefit of using DeepTOF. One may note that correlation of PF is higher than BP in most cases. We argue that Bodily Pain (BP) is difficult to describe during a questionnaire survey, and it may be difficult for the patients to accurately give the value of BP. To show the individual treatment outcome forecasting, among 2430 samples, we choose two representative patients and predict their outcomes after taking different treatments, which are shown in Figure 4. In terms of prediction value, surgical treatment effect is better than nonoperative treatment for patient 1. However, for patient 2, nonoperative treatment effect is better than surgical treatment effect in long run (8 years). This also corresponds to the studies of [31, 32] that some spine patients can self-heal over time with nonoperative treatment. Furthermore, we also provide additional results on simulated data in our repository[5].

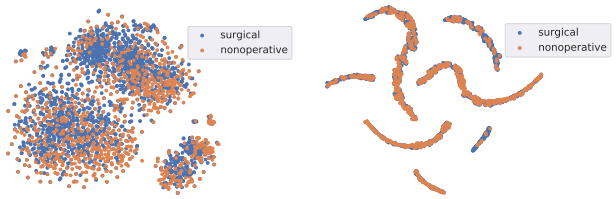


(a) Forecasting for patient 1 (b) Forecasting for patient 2

Figure 4: Individual treatment outcome forecasting for two patients in the next 8 years (higher score is better).

[5]https://github.com/HangtingYe/DeepTOF.

4.2.3 Validating Counter-factual Training. Furthermore, to verify whether DeepTOF can obtaine unbiased data representations, we show the T-SNE embeddings of original patient data and representations learned by DeepTOF in Figure 5. As shown in Figure 5(a), the data distribution varies across different treatment subgroups, while Figure 5(b) shows that with counter-factual training, DeepTOF removes treatment-related information in the extracted representations, i.e., capable of learning treatment-agnostic representations.



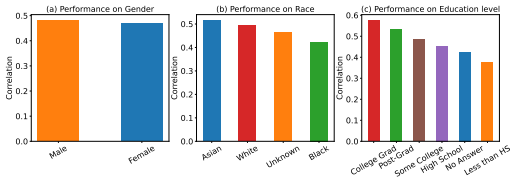
(a) Original patient representations (b) Treatment-agnostic representations

Figure 5: T-SNE embeddings of original patient data and treatment-agnostic representations. Notice that each patient is naturally assigned one treatment (surgical or nonoperative treatment) in original dataset.

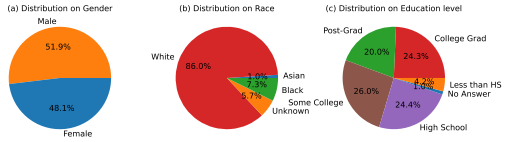
Table 4: Explanation of selected features.

Group	Feature	Explanation
Basic information	dx	type of spine disease
	age	
	gender	
	racenew_4grp	race
	educ	education level
	bmi	body mass index
	smoke	smoking habit
	income2	income
	work4grp2	working time
	insurance	insurance type
Self evaluation	WorkLift	importance of ability to lift heavy objects
	GuessSSU	expectation for surgical treatment
	GuessSNO	expectation for nonoperative treatment
	Episode fixd	duration of symptoms
Symptom	SLR	straight leg raise test
	HernLoc2	herniation location
	HernTyp	herniation type
Health condition	pf (enrollment)	physical functioning score
	hp	general health score
	satis	satisfaction with symptoms

4.2.4 Feature Analysis. We conduct feature analysis to verify the impact of feature selector in DeepTOF. First, to demonstrate the influence of feature selection on prediction performance, we test DeepTOF and baseline models with a different number of features, which is reported in Table 3. We find that DeepTOF's performance has not been compromised while the workload has been drastically reduced, which indicates that the original feature set contains significant information redundancy, and our feature selection strategy is extremely effective. In order to further analyze the selected features, we use the retained 20 features for analysis (i.e. M is set to 20). According to the taxonomy in Section 4.1, the selected 20



(a) Performance (higher is better) on different features.



(b) Data distribution on different features.

Figure 6: Data analysis and statistic

features are from 4 groups: i) Basic information; ii) Self evaluation; iii) Symptom; iv) Health condition. Please see Table 4 for more detailed information.

The majority of the 20 features are consistent with human intuition and should be retained. The most basic features (e.g. *type of spine disease, age* and *gender*) and health condition information (e.g. *general health score (hp)*) are selected because they are important indicators of patient's condition and health status. *WorkLift* (see group Self evaluation) is related to the health of spine. [11] concluded that the erector spine muscles get fatigued with repeated lifting, and then affect the lumbar spine. Additionally, the study of [28] showed that *straight leg raise test (SLR)* (see group Symptom) can be used as a screen of lumbar spine related symptoms, [30] indicated that failure to diagnose and precisely localize herniations can affect surgical treatment, thus it is essential to collect features in group Symptom (i.e. *SLR, HernLoc2* and *HernTyp*) for describing patients' condition.

Furthermore, a few selected features (e.g. *smoke, insurance*) are of interest. They are not necessarily intuitive as related to the spine, but they are worth considering. Prior studies [22] demonstrated that smokers are more likely to experience pseudarthrosis and postoperative infection after surgery in lumbar spines. Therefore, smoking information can be used to better forecasting post-surgery recovery. *Insurance* is a feature that is sometimes ignored but is quite important. Lack of insurance may create perverse incentives. [53] indicated that insurance status is associated with the accessibility and quality of health care.

4.2.5 Performance Analysis. To further show the performance of our model on different aspects, we report the predicted results (correlation averaged over BP and PF) on gender, races and education level in Figure 6(a). The distribution of dataset on the corresponding features is shown in Figure 6(b). As can be seen from the Figure 6(a), there is a small difference of correlation between female and male, which shows the fairness of our model. We can find that the prediction varies between different races. This may be due to the fact that the dataset contains different proportions of different races (shown in Figure 6(b)). Meanwhile, the results have prompted us to consider whether there are differences in clinical manifestations between races. If possible, researchers could conduct research to study the effect of race on outcomes in spine patients undergoing different treatments. Additionally, the performance also varies on different education levels. The results show that patients with higher education level may obtain more accurate predictive results. Previous medical research [51] indicated that a high education level can influence positively on alleviation of symptoms. Patients with a good education may have better psychological mechanisms to cope with treatment and postoperative symptoms positively.

Table 5: Ablation results. Here, C = counter-factual training, M = multi-task calibration, S = semi-supervised learning. The results show the relative change in terms of correlation of the ablated DeepTOF compared to the full DeepTOF.

Variants	BP	PF
DeepTOF	0.4596	0.4890
DeepTOF w/o C	-0.0215	-0.0246
DeepTOF w/o M	-0.0294	-0.0118
DeepTOF w/o S	-0.0188	-0.0147
DeepTOF w/o M,S	-0.0290	-0.0294
DeepTOF w/o C,S	-0.0222	-0.0041
DeepTOF w/o C,M	-0.0225	-0.0281
DeepTOF w/o C,M,S	-0.0280	-0.0365

4.2.6 Ablation Study. Since DeepTOF model adopts several essential techniques (i.e. counter-factual training, multi-task calibration, semi-supervised learning), we conduct ablation study to analyze the contributions of different techniques by considering 7 variants. Let **C, M, S** denote counter-factual training, multi-task calibration and semi-supervised learning respectively. The results are reported in Table 5. From this table, we can find that the removal of any of the components degrades the performance of DeepTOF, which demonstrated the effectiveness of our framework.

5 CONCLUSION

In this paper, we presented the unique challenges (i.e. data selection bias, noisy supervision and patient noncompliance) in web-based spine treatment outcome forecasting task, which need to be solved to support clinical decision-making. To handle these challenges, we proposed DeepTOF, a novel end-to-end deep learning model. Several novel techniques (i.e. counter-factual training, multi-task calibration and semi-supervised learning) for this specific task were adopted in DeepTOF. Additionally, an adaptive feature selector effectively selected the most important features to reduce the workload for patients and clinicians. The empirical results on *Spine Patient Outcomes Research Trial* (SPORT) dataset demonstrated the effectiveness of using DeepTOF framework in spine WTOF task. The proposed DeepTOF could bring great benefits to the rehabilitation of spine patients. Our work can shed some light on developing better algorithms for similar tasks.

6 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.61976102, No.U19A2065). The *Spine Patient Outcomes Research Trial* (**SPORT**) is funded by the NIH and supported by grant (U01-AR45444-01A1).

REFERENCES

[1] William A Abdu, Olivia A Sacks, Anna NA Tosteson, Wenyan Zhao, Tor D Tosteson, Tamara S Morgan, Adam Pearson, James N Weinstein, and Jon D Lurie. 2018. Long-term results of surgery compared with nonoperative treatment for lumbar degenerative spondylolisthesis in the Spine Patient Outcomes Research Trial (SPORT). *Spine* 43, 23 (2018), 1619.

[2] JA Andrews, RF Harrison, LJE Brown, LM MacLean, F Hwang, T Smith, Elizabeth A Williams, C Timon, T Adlam, H Khadra, et al. 2017. Using the NANA toolkit at home to predict older adults' future depression. *Journal of affective disorders* 213 (2017), 187–190.

[3] Mariette Awad and Rahul Khanna. 2015. Support vector regression. In *Efficient learning machines*. Springer, 67–80.

[4] Marshall H Becker, Lois A Maiman, John P Kirscht, Don P Haefner, and Robert H Drachman. 1977. The health belief model and prediction of dietary compliance: A field experiment. *Journal of Health and Social behavior* (1977), 348–366.

[5] Dimitris Bertsimas, Margrét V Bjarnadóttir, Michael A Kane, J Christian Kryder, Rudra Pandey, Santosh Vempala, and Grant Wang. 2008. Algorithmic prediction of health-care costs. *Operations Research* 56, 6 (2008), 1382–1392.

[6] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083* (2020).

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] Adam M Chekroud, Ralitza Gueorguieva, Harlan M Krumholz, Madhukar H Trivedi, John H Krystal, and Gregory McCarthy. 2017. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry* 74, 4 (2017), 370–378.

[9] Chacha Chen, Junjie Liang, Fenglong Ma, Lucas Glass, Jimeng Sun, and Cao Xiao. 2021. UNITE: Uncertainty-based Health Risk Prediction Leveraging Multi-sourced Data. In *Proceedings of the Web Conference 2021*. 217–226.

[10] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 432–440.

[11] P Dolan and MA Adams. 1998. Repetitive lifting tasks fatigue the back muscles and increase the bending moment acting on the lumbar spine. *Journal of biomechanics* 31, 8 (1998), 713–721.

[12] Mark Alan Fontana, Stephen Lyman, Gourab K Sarker, Douglas E Padgett, and Catherine H MacLean. 2019. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical orthopaedics and related research* 477, 6 (2019), 1267.

[13] Junyi Gao, Rakshith Sharma, Cheng Qian, Lucas M Glass, Jeffrey Spaeder, Justin Romberg, Jimeng Sun, and Cao Xiao. 2021. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association* 28, 4 (2021), 733–743.

[14] Mariëlle Gerritsen, Nadine Berndt, Lilian Lechner, Hein de Vries, Aart Mudde, and Catherine Bolman. 2015. Self-reporting of smoking cessation in cardiac patients: how reliable is it and is reliability associated with patient characteristics? *Journal of addiction medicine* 9, 4 (2015), 308–316.

[15] Alex HS Harris, Alfred C Kuo, Yingjie Weng, Amber W Trickey, Thomas Bowe, and Nicholas J Giori. 2019. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research* 477, 2 (2019), 452.

[16] R Brian Haynes, D Wayne Taylor, David L Sackett, Edward S Gibson, Charles D Bernholz, and Jayanti Mukherjee. 1980. Can simple clinical measurements detect patient noncompliance? *Hypertension* 2, 6 (1980), 757–764.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Geoffrey Hinton. 2018. Deep learning—a technology with the potential to transform health care. *Jama* 320, 11 (2018), 1101–1102.

[19] Damian Hoy, Lyn March, Peter Brooks, Fiona Blyth, Anthony Woolf, Christopher Bain, Gail Williams, Emma Smith, Theo Vos, Jan Barendregt, et al. 2014. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Annals of the rheumatic diseases* 73, 6 (2014), 968–974.

[20] Sajjad Hussain and Syed Sibte Raza Abidi. 2009. Integrating healthcare knowledge artifacts for clinical decision support: Towards semantic web based healthcare knowledge morphing. In *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 171–175.

[21] Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar, R Sujatha, Jyotir Moy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai, and Ohyun Jo. 2020. COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in public health* 8 (2020), 357.

[22] Keith L Jackson and John G Devine. 2016. The effects of smoking and smoking cessation on spine surgery: a systematic review of the literature. *Global spine journal* 6, 7 (2016), 695–701.

[23] Andrei F Joaquim and Alpesh A Patel. 2013. Thoracolumbar spine trauma: Evaluation and surgical decision-making. *Journal of Craniovertebral Junction and Spine* 4, 1 (2013), 3.

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017), 3146–3154.

[25] Bhairavi V Kharod, Paul B Johnson, Heather A Nesti, and Douglas J Rhee. 2006. Effect of written instructions on accuracy of self-reporting medication regimen in glaucoma patients. *Journal of glaucoma* 15, 3 (2006), 244–247.

[26] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4, 1 (2021), 1–6.

[27] Nicholas C Laucis, Ron D Hays, and Timothy Bhattacharyya. 2015. Scoring the SF-36 in orthopaedics: a brief guide. *The Journal of bone and joint surgery. American volume* 97, 19 (2015), 1628.

[28] Craig Liebenson, Amy M Karpowicz, Stephen HM Brown, Samuel J Howarth, and Stuart M McGill. 2009. The active straight leg raise test and lumbar spine stability. *PM&R* 1, 6 (2009), 530–535.

[29] Branimir Ljubic, Shoumik Roychoudhury, Xi Hang Cao, Martin Pavlovski, Stefan Obradovic, Richard Nair, Lucas Glass, and Zoran Obradovic. 2020. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. *Computer methods and programs in biomedicine* 197 (2020), 105765.

[30] Joseph C Maroon, Thomas A Kopitnik, Larry A Schulhof, Adnan Abla, and James E Wilberger. 1990. Diagnosis and microsurgical approach to far-lateral disc herniation in the lumbar spine. *Journal of neurosurgery* 72, 3 (1990), 378–382.

[31] Morio Matsumoto, Kazuhiro Chiba, Masayuki Ishikawa, Hirofumi Maruiwa, Yoshikazu Fujimura, and Yoshiaki Toyama. 2001. Relationships between outcomes of conservative treatment and magnetic resonance imaging findings in patients with mild cervical myelopathy caused by soft disc herniations. *Spine* 26, 14 (2001), 1592–1598.

[32] MD Ming Zhong and Jin Tao Liu. 2017. Incidence of spontaneous resorption of lumbar disc herniation: a meta-analysis. *Pain physician* 20 (2017), E45–E52.

[33] Stephen S Morse, Jonna AK Mazet, Mark Woolhouse, Colin R Parrish, Dennis Carroll, William B Karesh, Carlos Zambrana-Torrelio, W Ian Lipkin, and Peter Daszak. 2012. Prediction and prevention of the next pandemic zoonosis. *The Lancet* 380, 9857 (2012), 1956–1965.

[34] C David Naylor. 2018. On the prospects for a (deep) learning health care system. *Jama* 320, 11 (2018), 1099–1100.

[35] Wendy Oude Nijeweme-d'Hollosy, Lex van Velsen, Mannes Poel, Catharina GM Groothuis-Oudshoorn, Remko Soer, and Hermie Hermens. 2018. Evaluation of three machine learning models for self-referral decision support on low back pain in primary care. *International journal of medical informatics* 110 (2018), 31–41.

[36] World Health Organization and International Spinal Cord Society. 2013. *International perspectives on spinal cord injury*. World Health Organization.

[37] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. 2009. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine* 46, 1 (2009), 5–17.

[38] Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia* 4, 2 (2009), 1883.

[39] Evan M Polce, Kyle N Kunze, Michael C Fu, Grant E Garrigues, Brian Forsythe, Gregory P Nicholson, Brian J Cole, and Nikhil N Verma. 2021. Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery* 30, 6 (2021), e290–e299.

[40] Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* 30, 4 (1992), 838–855.

[41] Yong-Hao Pua, Hakmook Kang, Julian Thumboo, Ross Allan Clark, Eleanor Shu-Xian Chew, Cheryl Lian-Li Poon, Hwei-Chi Chong, and Seng-Jin Yeo. 2020. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. *Knee Surgery, Sports Traumatology, Arthroscopy* 28, 10 (2020), 3207–3216.

[42] Quazi Abidur Rahman, Tahir Janmohamed, Hance Clarke, Paul Ritvo, Jane Heffernan, and Joel Katz. 2019. Interpretability and class imbalance in prediction models for pain volatility in manage my pain app users: analysis using feature selection and majority voting methods. *JMIR medical informatics* 7, 4 (2019), e15601.

[43] Quazi Abidur Rahman, Tahir Janmohamed, Meysam Pirbaglou, Hance Clarke, Paul Ritvo, Jane M Heffernan, and Joel Katz. 2018. Defining and predicting pain volatility in users of the Manage My Pain app: Analysis using data mining and machine learning methods. *Journal of medical Internet research* 20, 11 (2018), e12001.

[44] Vijay M Ravindra, Steven S Senglaub, Abbas Rattani, Michael C Dewan, Roger Härtl, Erica Bisson, Kee B Park, and Mark G Shrime. 2018. Degenerative lumbar spine disease: estimating global incidence and worldwide volume. *Global spine journal* 8, 8 (2018), 784–794.

[45] Fred Rosner. 2006. Patient noncompliance: causes and solutions. *The Mount Sinai Journal of Medicine, New York* 73, 2 (2006), 553–559.

[46] Robert E Schapire. 1999. A brief introduction to boosting. In *Ijcai*, Vol. 99. Citeseer, 1401–1406.

[47] Hon-Yi Shi, Jinn-Tsong Tsai, Yao-Mei Chen, Richard Culbertson, Hong-Tai Chang, and Ming-Feng Hou. 2012. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. *Breast cancer research and treatment* 135, 1 (2012), 221–229.
[48] Edward H Shortliffe and Martin J Sepúlveda. 2018. Clinical decision support in the era of artificial intelligence. *Jama* 320, 21 (2018), 2199–2200.
[49] Amit G Singal, Ashin Mukherjee, B Joseph Elmunzer, Peter DR Higgins, Anna S Lok, Ji Zhu, Jorge A Marrero, and Akbar K Waljee. 2013. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *The American journal of gastroenterology* 108, 11 (2013), 1723.
[50] Colleen M Sitlani, Patrick J Heagerty, Emily A Blood, and Tor D Tosteson. 2012. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Statistics in medicine* 31, 16 (2012), 1738–1760.
[51] Javier Cobo Soriano, Marcos Sendino Revuelta, Martín Fabregate Fuente, Ignacio Cimarra Díaz, Paloma Martínez Ureña, and Roberto Deglané Meneses. 2010. Predictors of outcome after decompressive lumbar surgery and instrumented posterolateral fusion. *European spine journal* 19, 11 (2010), 1841–1848.
[52] Robert J Straka, Jeffrey T Fish, Steven R Benson, and J Thomas Suh. 1997. Patient self-reporting of compliance does not correspond with electronic monitoring: an evaluation using isosorbide dinitrate as a model drug. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 17, 1 (1997), 126–132.
[53] Joseph E Tanenbaum, Vincent J Alentado, Jacob A Miller, Daniel Lubelski, Edward C Benzel, and Thomas E Mroz. 2017. Association between insurance status and patient safety in the lumbar spine fusion population. *The Spine Journal* 17, 3 (2017), 338–345.
[54] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems* 30 (2017).
[55] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.
[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[57] Deepika Verma, Kerstin Bach, and Paul Jarle Mork. 2021. Application of Machine Learning Methods on Patient Reported Outcome Measurements for Predicting Outcomes: A Literature Review. In *Informatics*, Vol. 8. Multidisciplinary Digital Publishing Institute, 56.
[58] Andrew J Walley, Alexandra IF Blakemore, and Philippe Froguel. 2006. Genetics of obesity and the prediction of risk for health. *Human molecular genetics* 15, suppl_2 (2006), R124–R130.
[59] Hsiao-Han Wang, Yu-Hsiang Wang, Chia-Wei Liang, and Yu-Chuan Li. 2019. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer. *JAMA dermatology* 155, 11 (2019), 1277–1283.
[60] Y Wang, RD Hays, M Marcus, CA Maida, J Shen, D Xiong, ID Coulter, SY Lee, VW Spolsky, JJ Crall, et al. 2020. Developing Children's Oral Health Assessment Toolkits Using Machine Learning Algorithm. *JDR Clinical & Translational Research* 5, 3 (2020), 233–243.
[61] John E Ware Jr and Cathy Donald Sherbourne. 1992. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical care* (1992), 473–483.
[62] William B Weeks and James N Weinstein. 2015. Patient-reported data can help people make better health care choices. *Harvard Business Review.[cited 2016 Feb 24]. Available from: https://hbr. org/2015/09/patient-reported-data-can-help-people-make-better-health-care-choices* (2015).
[63] James N Weinstein, Tor D Tosteson, Jon D Lurie, Anna NA Tosteson, Brett Hanscom, Jonathan S Skinner, William A Abdu, Alan S Hilibrand, Scott D Boden, and Richard A Deyo. 2006. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. *Jama* 296, 20 (2006), 2441–2450.
[64] James Wingate, Ilianna Kollia, Luc Bidaut, and Stefanos Kollias. 2020. Unified deep learning approach for prediction of Parkinson's disease. *IET Image Processing* 14, 10 (2020), 1980–1989.
[65] DJN Wong, CM Oliver, and SR Moonesinghe. 2017. Predicting postoperative morbidity in adult elective surgical patients using the Surgical Outcome Risk Tool (SORT). *BJA: British Journal of Anaesthesia* 119, 1 (2017), 95–105.
[66] Wan Zhu, Longxiang Xie, Jianye Han, and Xiangqian Guo. 2020. The application of deep learning in cancer prognosis prediction. *Cancers* 12, 3 (2020), 603.

A APPENDIX

A.1 Baseline models

The baseline models are as follows:

- **Lasso** [55]: The Lasso is a linear model that estimates sparse coefficients. It tends to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features.
- **SVR** [3]: SVR is a regressor version of Support Vector Classification. Samples whose predictions are close to their targets will be ignored by cost function during training.
- **K-NN** [38]: K-NN is a regressor based on K-nearest neighbors. The prediction of a sample is computed based on the mean of the labels of its nearest K neighbors in the training set.
- **RandomForest** [7]: Random forest consists of a number of decision trees which are fitted on various subsets and sub-spaces of the dataset. It uses average results of the decision trees to improve predictive accuracy.
- **LightGBM** [24]: LightGBM is a gradient boosting [46] framework that uses tree based learning algorithms.
- **ResNet** [17]: A residual neural network (ResNet) is an artificial neural network (ANN). Typical ResNet models are implemented with double- or triple- layer skips that contain nonlinearities (ReLU) and batch normalization in between.

A.2 Regarding the internet portal

We would like to provide a more concise and clearer description of our data sources and future goals:

(i) *Data source for our current setting*:

- We collected self-reported scores on the Short Form-36 (SF-36) for each patient, which is a widely-used generic health survey with 36 questions. It also comes with detailed instructions to help patients understand the questionnaire and give a more accurate self-assessment. The format of the 36 questions can be found at https://www.mdapp.co/sf-36-score-short-form-health-survey-calculator-521/. As an example, one question from the SF-36 questionnaire is "Did you feel tired?", with answer options ranging from: 1) all of the time; 2) most of the time; 3) a good bit of the time; 4) some of the time; 5) a little of the time; 6) none of the time. Since SF-36 is a structured, self-reported questionnaire that takes around 10 minutes to complete, SF-36 scores could be collected from the internet portal.
- For each question, we mapped each answer to a score ranging from 0-100, making it compatible with our multi-task calibration model.

(ii) *For future goals*:

In the future, our aim is to collect other features that can be easily self-reported online through questionnaires, in addition to SF-36 scores. The questionnaire format will be similar to that of the SF-36. Take the selected 20 features by feature selector as an example (shown in Table 4 in original paper). Among them, features which can be answered without clinical experience (e.g. dx, age, gender, smoke, etc.) will be collected through online questionnaires. Our goal is to utilize web technology to assist physicians in making decisions rather than providing direct decisions. With the collected data from the website, we could (1) further fine-tune our model, and then (2) utilize it to assist in early diagnosis decision making (please refer to original paper in Section Preliminaries.Application Scenarios).