Journal of the American Medical Informatics Association, 29(11), 2022, 1931–1940
https://doi.org/10.1093/jamia/ocac130
Advance Access Publication Date: 29 August 2022
Research and Applications

Research and Applications

Dynamic prediction of work status for workers with occupational injuries: assessing the value of longitudinal observations

Erkin Ötleş [1,2], Jon Seymour[3], Haozhu Wang [4], and Brian T. Denton[1]

[1]Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, Michigan, USA, [2]Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor, Michigan, USA, [3]Peers Health, Chicago, Illinois, USA, and [4]Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, Michigan, USA

Corresponding Author: Erkin Ötleş, MS, Department of Industrial & Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48105, USA; eotles@umich.edu

Received 27 August 2021; Revised 22 June 2022; Editorial Decision 17 July 2022; Accepted 20 July 2022

ABSTRACT

Objective: Occupational injuries (OIs) cause an immense burden on the US population. Prediction models help focus resources on those at greatest risk of a delayed return to work (RTW). RTW depends on factors that develop over time; however, existing methods only utilize information collected at the time of injury. We investigate the performance benefits of dynamically estimating RTW, using longitudinal observations of diagnoses and treatments collected beyond the time of initial injury.

Materials and Methods: We characterize the difference in predictive performance between an approach that uses information collected at the time of initial injury (baseline model) and a proposed approach that uses longitudinal information collected over the course of the patient's recovery period (proposed model). To control the comparison, both models use the same deep learning architecture and differ only in the information used. We utilize a large longitudinal observation dataset of OI claims and compare the performance of the two approaches in terms of daily prediction of future work state (working vs not working). The performance of these two approaches was assessed in terms of the area under the receiver operator characteristic curve (AUROC) and expected calibration error (ECE).

Results: After subsampling and applying inclusion criteria, our final dataset covered 294 103 OIs, which were split evenly between train, development, and test datasets (1/3, 1/3, 1/3). In terms of discriminative performance on the test dataset, the proposed model had an AUROC of 0.728 (90% confidence interval: 0.723, 0.734) versus the baseline's 0.591 (0.585, 0.598). The proposed model had an ECE of 0.004 (0.003, 0.005) versus the baseline's 0.016 (0.009, 0.018).

Conclusion: The longitudinal approach outperforms current practice and shows potential for leveraging observational data to dynamically update predictions of RTW in the setting of OI. This approach may enable physicians and workers' compensation programs to manage large populations of injured workers more effectively.

Key words: machine learning, deep learning, prediction model, occupational injuries, workers' compensation

OBJECTIVE

Occupational injuries (OIs) cause an immense burden on the US population and economy. Millions of workers are injured annually, leading to pain, emotional suffering, and economic. In addition to resulting in time away from work, OIs increase medical expenditures and shorten lifespans; furthermore, they disproportionately affect minorities.[1–8] OIs have far-reaching economic consequences due to decreases in corporate productivity and major costs to government organizations.[1–3] As in other facets of medicine, timely and clinically appropriate intervention is critical to promote injured worker healing and recovery.[9–11] In occupational medicine, the primary clinical outcome is return to work (RTW).

The RTW process, like most medical episodes, is complex.[12] It requires individual medical management by highly trained physicians; additionally, injuries are often reviewed for treatment utilization by reviewers, or *recovery managers*, who oversee thousands of simultaneous cases on behalf of workers' compensation programs.[13] The current state of the art for injury recovery prediction are models and guidelines that are used at the onset of the injury.[14–16] These models are often used by payers to estimate a worker's RTW date. Predicted RTW duration is both a clinical and administrative tool that has been ingrained into the occupational health framework.[17] The most prevalent modeling techniques used for this approach are Cox proportional hazards models, which are used to estimate the likelihood that workers will RTW in a given time period.[18–21] These models estimate RTW based on information at the time of a worker's injury, thus, providing guidance on the expected resources needed for a worker's recovery, and enabling stratification of the currently injured worker population. While these models assist initial triage of resources for injured workers, their utility decreases over time as they fail to account for diagnoses and treatments workers experience over the course of their recovery. To the best of our knowledge, longitudinal data available in the form of insurance claims streams are not currently used to generate or update RTW predictions.

To support decision-making over the course of worker recovery, we investigate the predictive performance benefits of using longitudinal observations collected over the course of a workers compensation case. The use of longitudinal observations has been shown to improve performance in the prediction of cardiovascular events.[22] However, to the best of our knowledge, this has not been characterized for the prediction of RTW. In this work, we measure the difference in predictive performance between the current approach to RTW prediction (baseline model), which only uses information collected near the time of injury, to an approach that uses longitudinal observations (proposed model) collected over the course of a worker's recovery.

To do this, we present a new framework to dynamically predict the RTW of injured workers. The proposed model reframes the prediction of RTW into a dynamic prediction task. For injured workers, it seeks to learn the relationship between observations collected daily and the worker's future *work status*, that is, whether the worker has returned to work or not. To evaluate whether longitudinal observation data collected beyond the first week of injury can help predict work status, we estimated a deep learning model capable of ignoring missing longitudinal observation data. We trained this proposed model with the entire history of longitudinal observations available in the training dataset. Given daily longitudinal observations, the model will return future work status predictions. Although the predictions are dynamic, the underlying model parameters are static.

We compare the performance of this proposed model against the baseline model, which is representative of current RTW prediction approaches used in practice as it is limited to data collected around the time of injury.[16] The baseline model only utilizes information collected around the time of a patient's initial injury (the first week). To assess the benefit of the proposed approach, we use a large claims dataset from the state of Ohio's workers' compensation program to develop the models. Both models are implemented as recurrent neural networks, a type of deep learning model, to learn this relationship.[23–26] We evaluate the predictive performance difference between these two approaches using a held-aside portion of the claims dataset, based on the daily predictions they each produce of future work status. The main contributions from this work are as follows:

1. An evaluation of the predictive performance impact of the use of longitudinal observations on the RTW prediction task
2. Introduction of the RTW prediction problem as an important area of research that should be addressed by the informatics community
3. A reformulation of the OI recovery problem as a dynamic work status prediction problem
4. Recurrent neural network implementation and training procedure used for both the proposed and baseline models
5. A Python framework to automatically build dynamic health-status prediction models from longitudinal datasets from electronic health records or claims-like systems.

BACKGROUND AND SIGNIFICANCE

The prediction of RTW for an OI is fundamental to decision-making by employers, occupational health physicians, and recovery managers—all of whom share the common goal of minimizing the employee's absence. Disability management is a human resources process conducted by many employers who recognize it as a key component of overall workplace productivity.[27–29] On an individual basis, if the predicted RTW duration is short, then minimal personnel shifting need occur. On the other hand, with a longer predicted RTW duration, employers face more operational decisions, including whether to hire temporary workers and/or to offer the injured employee modified duty during the recovery.[30] On an aggregate basis, actual RTW durations are compared aginst predictions forming an important benchmark for many businesses.[31] Questions that an employer may seek to use a RTW model for are: Will the employer need to replace the worker on a temporary or permanent basis? Is modified duty a worthwhile option for this worker? Is the organization measuring up to RTW benchmarks?

RTW predictions are related to the expert prognoses generated by occupational health physicians, who are often asked or required to supply absence notes for injured patients.[17,32] Importantly, RTW patients are often seen by generalist primary care physicians or non-occupational health specialists.[33] As such, RTW predictions are used as a part of treatment guidelines for nonspecialist physicians to benchmark OIs.[34] A question that a physician may seek to answer using a RTW model is: How long is this patient's absence from work expected to be?

Recovery managers, typically working on behalf of insurance organizations, are often assigned to cases based on the RTW prediction. Cases with longer predicted RTW durations are usually classified as severe or difficult cases. These cases are often directed to experienced recovery managers. In any scenario, the RTW prediction is used to manage expectations and to dictate operational processes across clinical and corporate stakeholders. A recovery manager may use an RTW model to answer: How should this case be triaged? Should I alert other stakeholders that the RTW duration has exceeded the prediction?

Due to the close relationship between RTW prediction and treatment, predictive models are bundled with guidelines for treatment and resource management.[35,36] Despite this relationship we focus our work on the task of RTW prediction and leave additional guidance to future work. From the existing RTW literature, it is important to note that state of the art in OI modeling has several potential avenues for further exploration. The first is that models are generally based on a static time-to-event prediction of RTW, designed for usage only at the time of injury, and incapable of handling newly observed information.[18–21] The second is that models are traditionally made for specific diseases with custom collected data.[18,20,21,37–40] The existing work presents a gap to be explored. Specifically, what is the value of producing RTW predictions using longitudinal observations?

To address this question, we reformulate OI modeling as a dynamic prediction task, where the prediction of a worker's RTW is made sequentially over the time horizon of their injury. These repeated predictions would be based on observational data commonly available to decision-makers, like physicians and workers' compensation programs. For example, each day, new claims observations may be fed to a model, which returns the likelihood that the injured worker will be back to work in a week. This is a type of *sequence-to-sequence learning task*, where a model captures the mapping between a given sequence of observations and a sequence of predictions. Markov chain-based models have successfully been used for sequence-to-sequence learning tasks.[41–44] However, we would like the model to learn to use the longitudinal observations directly (eg, no grouping or curation of diagnoses or treatments) and we would like the proposed model to build a representation of the accumulated observations (or *history*). *Recurrent neural networks* (RNNs), a type of deep neural network, are naturally well suited for this task. This is due to their ability to handle sequences with long-range time dependencies[23–26] and capability to learn representations for high-cardinality categories (*eg*, diagnoses and treatment codes) with minimal modification.[45–49] Thus, we use RNNs for this study to establish the predictive difference between the two approaches.

MATERIALS AND METHODS

We assess the value of utilizing longitudinal observations by reframing RTW prediction as a dynamic task and comparing this to a baseline model that only uses information collected around the time of injury. Our proposed model reframes the RTW prediction problem to produce future work status predictions using observations of diagnoses and treatments collected over time. In the following subsections, we describe the dataset used to train this model, formalize the sequence-to-sequence prediction task, and then discuss the experimental setup.

Dataset

We utilized the Peers Health Ohio Workers' Compensation Dataset for this work. This dataset contains longitudinal workers' compensation claims information for over 1.2 million workplace injuries collected in the state of Ohio from January 2001 to October 2010. For each injury record, there is demographic information describing the age, sex, and job type of the worker at the time of their injury. This demographic information is accompanied by time-stamped longitudinal information that describes the diagnoses and treatments (treatments are procedures or activities rendered by healthcare providers to improve the health of a patient, like physical rehabilitation) that the worker experienced throughout their injury recovery. Finally, for each injury record, the dates of a worker's departure from and return to work are recorded (an injury record may contain multiple depart from and return to work dates). This work was conducted with approval from the University of Michigan Institutional Review Board. The data underlying this study were provided by Peers Health by permission. Data will be shared on request to the corresponding author with permission of Peers Health.

Based on preliminary experiments, we sample 300 000 injuries to achieve a suitable trade-off between model training time and predictive performance. We then exclude all injuries with case durations of less than 7 days, as a predictive model would provide marginal utility for these cases. Finally, we split the dataset evenly between training, development, and test datasets (1/3, 1/3, 1/3, respectively). The development dataset was used for hyperparameter search. After the model hyperparameters were found, final model training was conducted using a dataset that consisted of the combined training and development datasets. The held-aside test dataset was used to evaluate the performance of our proposed model against the baseline model.

Problem statement

We seek to learn a model, $f(\cdot)$, that when given a sequence of diagnoses and treatments observations, $\boldsymbol{x}_{i,t}$, over time, t, for a given worker injury, i, produces an estimate of the likelihood of return to work within a defined period, $\Pr(y_{i,t} = 1)$.

Approach

In the following 2 sub-subsections, we provide an overview of the variable definitions and then follow with a discussion of the mathematical model implemented using RNNs.

The set of worker injuries are denoted by I which is indexed by $i \in I$. Time was discretized using a fixed time-step duration set to 1 day for this study and $t = 1, 2, \ldots, 365$. We limited case durations to the typical cut-off for maximal medical improvement (365 days).[50] This discretization and transformation are further described in the Supplementary Material S2 and S3. Moreover, Figure 1 depicts an example of an injury transformation.

Each injury, i, had two types of data collected. The first type is *characteristics*, which includes all time-invariant demographic data (*eg*, biological sex and job classification). The second type is *observations*, which includes time-stamped longitudinal information that was collected over time (*eg*, procedure information). For every injury, i, we create a characteristic vector $\boldsymbol{c}_i$ of equivalent size (d_c), to represent time-invariant information that was collected before or at the time of injury. We create an observation vector $\boldsymbol{o}_{i,t}$ of size d_o for every injury, i, at every time-step, t; these vectors represent information collected each day of an injured worker's recovery. Characteristic and observation vectors both contain information encoded as either real numbers and or as integers (for categorical data). Missing observations were denoted with a special missing value (see Supplementary Material S2 for more detail). We let $o^W_{i,t}$ denote the work

	Day	Jan 5	Jan 6	Jan 7	Jan 8	Jan 9	Jan 10
	Diagnoses	Ankle Sprain				Tendon Rupture	
	Procedures		Ankle X-ray				
Age 55 Bio. Sex Male Job Postal Worker	Work-status	Not Working	Not Working	Not Working	Working	Working	Not Working

$\mathbf{c}_1^R = (55)$
$\mathbf{c}_1^C = (1,78)$

$t =$	1	2	3	4	5	6
$\mathbf{o}_{1,t}^R =$	(1,0)	(0,1)	(0,0)	(0,0)	(1,0)	(0,0)
$\mathbf{o}_{1,t}^C =$	(10,0)	(0,15)	(0,0)	(0,0)	(169,0)	(0,0)
$y_{1,t} =$	0	0	1	1	0	0

Figure 1. Example worker timeline and corresponding characteristic and observation data. In this example, worker injury 1, is a 55-year-old male postal worker is injured on January 5. This information is encoded in the real characteristic vector, $c_1^R = (55)$, which contains the age information, and the categorical characteristic vector, $c_1^C = (1,78)$, which encodes biological sex (male = 1) and job code (postal worker = 78). His injury case runs until the last observed date, January 10. Throughout the course of injury, diagnoses and procedures are observed. This information is encoded in daily observation vectors. On the first day of the worker's injury, January 5, the real observation vector, $o_{1,1}^R = (1,\ 0)$, contains information regarding the number of diagnoses and procedures observed for the injured worker at $t = 1$ (1 diagnosis and 0 procedures). The categorical observation vector $o_{1,1}^C = (10, 0)$, encodes diagnosis (ankle sprain = 10) and the no procedures observed token (0). The input vector at $t = 1$, $x_{1,1}$ is the concatenation of the observation vectors at that time and the characteristic vectors $(55,\ 1,\ 78,\ 1,\ 0,\ 10,\ 0)$. The model will then map the input vector to the output, $y_{1,1}$ which is the work status 1 day in the future ($\phi = 1$ for this example).

status of an injured worker over time where $o_{i,t}^W = 1$ denotes "working" status and $o_{i,t}^W = 0$ denotes "not working" status.

Characteristic and observation vectors are used to generate the input features and output labels of the model. Model input features, $\boldsymbol{x}_{i,t}$, denote the vector of injured worker's characteristics and observations over all time-steps, $\boldsymbol{x}_{i,t} = (\boldsymbol{c}_i, \boldsymbol{o}_{i,t})\ \forall\ i\ \in\ I,\ t\ \in T$. An example calculation of $\boldsymbol{x}_{i,t}$ is depicted and explained in Figure 1. The model output label, the future work status, denoted as $y_{i,t}$ is also indexed in terms of injuries and time-steps. Each $y_{i,t}$ is related to the observed work status. We define $y_{i,t} = o_{i,t+\phi}^W$, where ϕ is termed the *offset*, a positive integer value for the number of time-steps in the future we would like to predict work status.

Model definition

At every time-step, the model maps all observed input features about an injury i up until time t to estimate the probability of being in the future work status of working, $\Pr(y_{i,t} = 1)$. We denote the overall model as $f(\cdot)$ and the model's parameters as θ, formally $f : (x_{i,1},\ \ldots,\ x_{i,t}) \rightarrow \Pr(y_{i,t} = 1)$. The model is composed of 3 functions, the input encoder $f_{\text{in}}(\cdot)$, the history encoder $f_{\text{mid}}(\cdot)$, and the output estimator $f_{\text{out}}(\cdot)$. Each function is described in more detail below. The parameters of the overall model, $f(\cdot)$, θ is the combination of the parameters of these functions $\theta_{e_{\text{in}}}$, $\theta_{e_{\text{mid}}}$, and θ_{out}.

The model does not directly use $\boldsymbol{x}_{i,t}$ to predict $y_{i,t}$; instead, it uses two intermediary lower-dimensional approximations: the encoded observation vector $\tilde{\boldsymbol{x}}_{i,t}$ and the encoded history vector $\tilde{\boldsymbol{h}}_{i,t}$. The encoded feature vector $\tilde{\boldsymbol{x}}_{i,t}$ is a transformation of $\boldsymbol{x}_{i,t}$ that replaces the categorical integer values with real-valued embeddings.[46,48] We compute $\tilde{\boldsymbol{x}}_{i,t}$ using $f_{\text{in}}(\boldsymbol{x}_{i,t})$ which transforms $\boldsymbol{x}_{i,t}$ using $\theta_{e_{\text{in}}}$ parameters into $\tilde{\boldsymbol{x}}_{i,t}$, $f_{\text{in}} :\ \boldsymbol{x}_{i,t} \rightarrow \tilde{\boldsymbol{x}}_{i,t}$. Thus, $\tilde{\boldsymbol{x}}_{i,t}$ is a real valued vector with dimension $d_{\tilde{\boldsymbol{x}}}$.

Similarly, the encoded history vector $\tilde{\boldsymbol{h}}_{i,t}$ approximates the full history of the injury's observations, $(x_{i,1},\ \ldots,\ x_{i,t})$. The encoded history vector $\tilde{\boldsymbol{h}}_{i,t}$ is a real-valued vector of size $d_{\tilde{\boldsymbol{h}}}$ that is updated by the middle function, $f_{\text{mid}}(\cdot)$, a recursive function that takes the current timestep's encoded input $(\tilde{\boldsymbol{x}}_{i,t})$ along with the encoded history from the previous time-step $(\tilde{\boldsymbol{h}}_{i,t-1})$ and returns an updated encoded history for the current time-step $(\tilde{\boldsymbol{h}}_{i,t})$. It uses $\theta_{e_{\text{mid}}}$ parameters and is formally denoted as $f_{\text{mid}} :\ \left(\tilde{\boldsymbol{x}}_{i,t}, \tilde{\boldsymbol{h}}_{i,t-1}\right) \rightarrow \tilde{\boldsymbol{h}}_{i,t}$.

Since the encoded history vector $\tilde{\boldsymbol{h}}_{i,t}$ is a representation of the injury's entire history up to and including the current time-step t it can be used to estimate the output label $y_{i,t}$. This mapping is controlled by the out function $f_{\text{out}}(\cdot)$ which takes $\tilde{\boldsymbol{h}}_{i,t}$ and returns a probability estimate of the output label (work status) being equal to 1. The out function $f_{\text{out}}(\cdot)$ is parameterized by θ_{out} and formally, $f_{\text{out}} :\ \tilde{\boldsymbol{h}}_{i,t} \rightarrow [0, 1]$. This probability estimate can then be used to estimate the outcome based on a threshold, τ, as follows:

$$\hat{y}_{i,t} \ = \ \begin{cases} 1 \text{ if } \Pr(y_{i,t} = 1) \geq \tau \\ 0 \text{ otherwise} \end{cases}$$

In summary, the sequence of functions transforming the inputs, $\boldsymbol{x}_{i,t}$, to the probability estimate, $\Pr(y_{i,t} = 1)$, is:

$$\tilde{\boldsymbol{x}}_{i,t} = f_{\text{in}}(\boldsymbol{x}_{i,t})$$

$$\tilde{\boldsymbol{h}}_{i,t} = f_{\text{mid}}\left(\tilde{\boldsymbol{x}}_{i,t}, \tilde{\boldsymbol{h}}_{i,t-1}\right)$$

$$\Pr(y_{i,t} = 1) = f_{\text{out}}\left(\tilde{\boldsymbol{h}}_{i,t}\right)$$

This recurrent approach yields a model that maps to clinical decision-making. Note, although the model updates the history encoding in response to observations collected across time, the underlying parameters of the model remain static across time-steps.

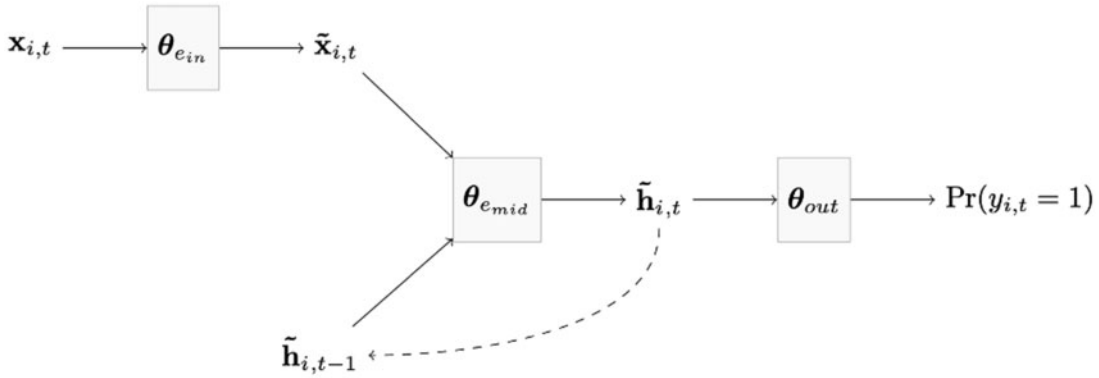

Figure 2. Model diagram. Inputs of observations, $x_{i,t}$, and the prior history encoding, $\tilde{h}_{i,t-1}$, are used as inputs at every time-step, t. Observations are encoded with $\theta_{e_{in}}$ in order to learn representations of high-cardinality categories ($\tilde{x}_{i,t}$). Along with $\tilde{h}_{i,t-1}$ these representations are then encoded into the current history, $\tilde{h}_{i,t}$, using $\theta_{e_{mid}}$. Finally, θ_{out} converts the current history to a prediction of the injured worker's likelihood of being at work in the future, $\Pr(y_{i,t}=1)$.

A block diagram of the model is depicted in Figure 2, with all future work statuses being a function of the current observation and all historical observations.

Experimental setup

The overall model processes information through a series of submodels that each consist of one or more neural network layers, allowing the entire model to be trained via *back-propagation*.[24] Training was conducted using stochastic gradient descent, with a batch size of 64 injuries, and using the Adam optimizer.

The size of the submodels and the activation functions for all their layers (except the last one of the out-submodel) are model hyperparameters. Thus, they were selected as a part of the hyperparameter search process. We used *hyperband search*, training on the training dataset and selecting hyperparameters which yielded the best performance in terms of area under the receiver-operating characteristic curve (ROC) on the development dataset.[51] Additional hyperparameters included the width and depth of each of the submodels, drop-out rate applied to the inputs and between layers, learning rate, and layer activation functions. For full details on the possible values for each hyperparameter, please see the "Hyperparameters" of the Supplementary Material S2.

Baseline model

Industry standards for predicting RTW are based on regression models that predict the case durations given static patient covariates, including age, gender, job class, and comorbidities, together with diagnosis information available at the beginning of the case. ODG and MDGuidelines have deployed these models in their web-based subscription service for treatment guidance and resource management.[35,36] The model used by ODG is developed with the same dataset we use for this study. However, these models are proprietary and not available without purchase.[34] As such, we sought to create a baseline regression model that is analogous to the industry standard proprietary models.[16] We did this by limiting the data presented to the model while retaining the daily prediction capability of the proposed approach. The baseline model uses patient covariates and observations collected during the first 7 days of the initial patient injury. After the seventh day, the baseline model was then fed with "missing observations" for the observation component of its inputs. Specifically, $o_{i,t}$ = "missing observations" $\forall\ i \in \mathrm{I},\ t \in \mathrm{T} > 7$. These missing observations allow both models to function over time-steps where there might not have been longitudinal observation data collected. Utilizing these missing observations provides valid inputs for all time-steps, allowing the model to return predictions across the entire case duration. For more information, see "High-Cardinality Category Embeddings" in Supplementary Material S2. Thus, simulating the information available to the proprietary models. Note, this data observation limitation was applied to the baseline model at both training and testing time.

To have a controlled comparison between the baseline model (representing existing approaches that do not use longitudinal observations) and our proposed model (representing the usage of longitudinal observations), we sought to ensure that they had the same overall capacity and used the same training procedure. As such, we used the same framework and searched over the same hyperparameters; and only limited the baseline so that it only used information typically used for the proprietary models we sought to replicate. This setup replicates the data used to create the Cox proportional hazards models traditionally used for this task with an added benefit; the baseline model can learn from the initial observation data and the observation timing. By using the same architecture, searching over the same hyperparameter space, and by only using the first 7 days of observations we seek to create a capable baseline that represents the best possible performance of existing proprietary models. The restriction to the first 7 days is an optimistic interpretation of existing proprietary models that only use information that was collected at the time of injury. By comparing our proposed method against the baseline model, we can estimate the potential improvement of using longitudinal observations over the current industry approach of using information collected around the time of injury.

Evaluation

To evaluate the performance of both models, we generated daily predictions on the held-out test dataset of OIs. All of the daily predictions were used to calculate performance measures, we use this window-level approach (also known as time-horizon approach)[52] as users of the model have the ability to intervene on patients every day. Performance was measured in terms of discriminative performance, using the ROC and the area under it (AUROC). Calibration performance was also assessed with calibration curves[53] and expected calibration error[54] (ECE). For each injury, all daily predictions, $\Pr(y_{i,t}=1)$, were compared against the true label ($y_{i,t}$). To assess the variation in performance, we computed 90% confidence intervals for all curves and measures. Confidence intervals were generated using bootstrap sampling; in this procedure, the population

of injuries in the test set was resampled-with-replacement 100 times to estimate model performance under varying distributions of injured workers.

We implemented the entire data transformation, model training, and evaluation pipeline using python 3.6.9, using the TensorFlow docker container (tag: latest-gpu-jupyter accessed on June 9, 2021) running on an Ubuntu 18.04 workstation with an Intel Xeon 6146 CPU, 256 gigabytes of RAM, and a NVIDIA Titan V graphics card.[55] The proposed and baseline models were implemented using TensorFlow and Keras.[56–58] Additionally, we utilized the SQLite, SKLearn, NumPy, pandas, and tableone python packages.[59–64] We have released our data transformation code and model training python framework on GitHub (https://github.com/eotles/TemporalTransformer). The methods and approaches described above are covered by a US utility patent application.

In addition to evaluating the proposed deep learning-based model and the baseline model, we also evaluated several models using simpler machine learning architectures. We used L2-regularized logistic regression and random forest regression as these architectures are more directly interpretable than the proposed deep learning approach. These additional evaluations can be found in Supplementary Material S5.

RESULTS

For our study, we set the offset to 1 week (or 7 days, thus $\phi = 7$) so that we predict work status for 1 week in the future. Note, when $t + \phi$ surpasses the last observed $o^W_{i,t}$ value, the last observed $o^W_{i,t}$ value is filled forward. The choice to forward fill the future work status may not be appropriate for all use cases; however, it is appropriate for this RTW task. Injury cases are only considered to have reached completion once the injured worker has reached their maximal recovery or has transitioned to long-term disability at the 365-day cutoff we employ above.[50] Thus, the work status observed on their last day is likely to be their lasting work status. After applying our minimum case duration of 7 days exclusion criteria to the 300 000 randomly sampled OIs, we had 294 103 OI cases.

The median age of injured workers at the time of injury was 35 years old, with an interquartile range (IQR) between 26, 45 years old. Most of the workers were male, with only 31.9% having a biological sex of female. In total, these workers represented 595 different occupation classifications, with the 5 most common occupations being: city employees, restaurant workers, school district employees, nursing home workers, and automobile service workers (Supplementary Table S1). The median number of diagnoses observed per injured worker was 1 (IQR: 1, 2), the number of procedures was 5 (3, 10). When limited to observing the first week of the worker's recovery, as in the case of the baseline model, the number of diagnoses observed was 0 (0, 0), and the number of procedures was 4 (2, 6), see Supplementary Material S2 for discussion of this. The most commonly observed diagnoses and procedures are categorized in Supplementary Tables S2 and S3. Since the RTW observations were not limited to the first week, both the baseline and our proposed model observed the 1.1 (SD: 0.5) return-to-work events per injured worker. These numbers are also depicted in Table 1.

When evaluated on daily predictions generated over the test dataset, our proposed model had an AUROC of 0.728 (90% confidence interval: 0.723, 0.734), compared to the baseline model's AUROC of 0.591 (0.585, 0.598). In terms of calibration, our proposed model had an ECE of 0.004 (0.003, 0.005) versus 0.016 (0.009, 0.018) for the baseline model. The values along with ROC curves and calibration curves are displayed in Figure 3. Despite underestimation of RTW likelihood in low-likelihood cases, the proposed model displays better overall calibration (smaller ECE) than the baseline model. The baseline model shows underestimation of RTW likelihood in both low- and high-likelihood cases but also shows overestimation in midlikelihood cases. When examining the performance of our proposed model and the baseline model in subpopulations of injuries occurring in workers of different ages or sexes, their performance varies slightly. However, our proposed model generally outperforms the baseline model in each of these subpopulations, Supplementary Figures S1 and S2.

Table 1. Population characteristics

Population characteristics	*n*: 294 103	
Demographic characteristics	Entire population	
Age, median (IQR)	35 (26, 45)	
Biological sex		
n missing: 3876		
F, *n* (%)	92 674 (31.9)	
M, *n* (%)	197 553 (68.1)	
Case duration (days), mean (SD)	88.9 (111.7)	
Observation characteristics per worker	Baseline	Model
Number of diagnoses, median (IQR)	0 (0, 0)	1 (1, 2)
Number of procedures, median (IQR)	4 (2, 6)	5 (3, 10)

Note: Demographic information (age and biological sex) is equally used between the baseline model (baseline) and the proposed model (model). Observations such as diagnoses and procedures are not equally used by both models, as the baseline model is limited to observations that occur within the first week of injury. As such, these observation characteristics are counted per worker for the baseline model and proposed model. The case duration is measured in days.

The additional experiments utilizing simpler model architectures discussed in Supplementary Material S5 reinforce the findings of the main experiments. These simpler model architectures generally had worse discriminative performance than the proposed deep learning-based model. However, simpler architectures using longitudinal observations outperformed baselines without longitudinal observations. For example, the logistic regression model using longitudinal observations had an AUROC of 0.607 (0.606, 0.607) compared an AUROC of 0.581 (0.580, 0.581) for the logistic regression model without longitudinal observations (see Supplementary Figure S5 for full details).

Additionally, when we examined the importance of longitudinal observations, we saw that the longitudinal observation data played a large role in the prediction of future work status. We observed that 9 out of the top 25 features of the logistic regression model corresponded to longitudinal observations. These features and their coefficients are displayed in Supplementary Table S9. Using permutation importance, we also observed the importance of longitudinal observations. Procedure Codes, a type of longitudinal observation, constituted the second largest group of features that impacted the discriminative performance of the logistic regression model. This is depicted in Supplementary Figure S7.

DISCUSSION

We found that utilizing longitudinal observations improves the performance of RTW prediction compared to approaches that only use

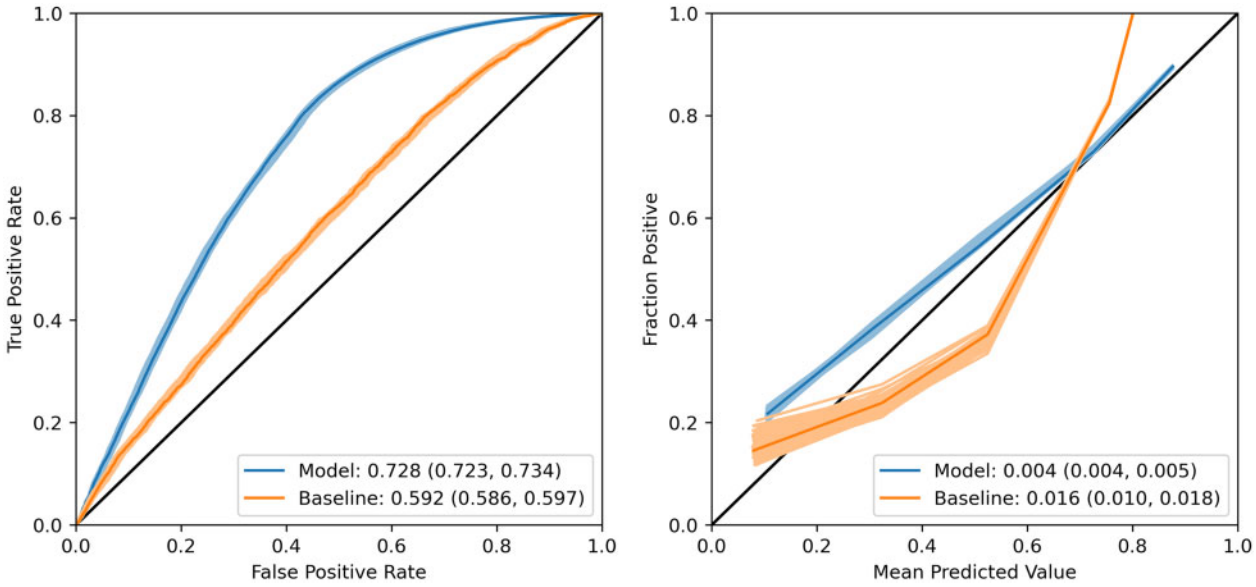

Figure 3. Predictive performance of the proposed model compared to baseline model. In the left subfigure, discriminative performance in terms of the receiver-operating characteristic (ROC) curve of the proposed model (blue) and the baseline model (orange) are plotted. The proposed model has a significantly better discriminative performance by dominating the ROC curve of the baseline model and having a larger area under the ROC curve, which is depicted in the legend. In the right subfigure, quintiled calibration curves for the proposed model (blue) and baseline model (orange) are displayed. Despite underestimation of RTW likelihood in low-likelihood cases, the proposed model displays better overall calibration (smaller expected calibration error) than the baseline model. The baseline model shows underestimation of RTW likelihood in both low- and high-likelihood cases but also shows overestimation in midlikelihood cases.

information at the time of injury. Despite both models using the same RNN-based deep learning architecture, the proposed model outperformed the baseline model in both discrimination and calibration. The baseline model is analogous to the current state of the art in RTW prediction, as it uses information collected at the time of injury to generate predictions. In contrast, the proposed model uses treatment information that is collected daily to update RTW predictions. The performance differences we observed between our proposed model and the baseline model show the potential practical benefit for reframing the RTW task as a dynamic prediction task.

Our proposed approach uses standard longitudinal data that is routinely collected by workers' compensation programs and exploits the capabilities of deep learning to build a dynamic model that outperforms approaches that only use information collected around the time of injury. Our python framework transforms readily available injury claims data into sequences of daily observations. These observations encode time-variant information, like diagnoses and treatment codes, and are combined with time-invariant data (eg, worker demographics). The framework then trains an RNN-based deep learning model to map these daily observations to the future work status of an injured worker. Thus, the learned model could be used to repeatedly generate RTW predictions given a sequence of longitudinally observed diagnoses and treatments.

The updating of RTW in response to observed diagnostic and treatment information could be valuable for employers, physicians, and OI recovery managers. Existing RTW prediction models coupled with treatment guidelines software have already been implemented into EHR systems.[10,11] Our proposed approach may provide additional value as the dynamic assessment of the worker's future work status relates to how physicians and other clinicians assess injuries over time. Like the proposed model, physicians update their understanding of an injured worker's recovery and future recovery prognoses based on information they collect over time. Additionally, this formulation helps to monitor populations effectively. As near real-time observations are collected for individual injured workers, the proposed model can generate RTW estimates. RTW estimates can then be used by OI recovery managers to allocate treatment resources to injured workers. These estimates can also be used by people with managerial responsibility for workforce coverage in industry organizations. Furthermore, this model may eventually be used to help answer "what-if questions"; using the model to assess the impact of potential treatment choices on work status could help support clinical decision-making. Altogether, the dynamic prediction of work status and may assist in the management of OIs, ultimately positively impacting injured workers and organizations that support them (*eg*, workplaces and governmental organizations).

To be useful, this dynamic model needs to be implemented within feasible workflows. We will briefly sketch a potential implementation method that would enable predictions to be used by recovery managers. This implementation would utilize insurance claims data. A hosted model fed claims data, in an automated or manual manner, could provide predictions for recovery managers and employers. Another potential implementation mirrors a project implemented at Kaiser Permanente[10,11] and is described in Supplementary Material S4. Both potential implementations raise many questions, ranging from privacy concerns to data infrastructure issues.[65] Of note, evaluation of OIs in terms of RTW is dependent on desired use-case and implementation. For this initial development study, we chose to use a daily evaluation as it is the most plausible evaluation frequency. We present potential implementations, not as finalized solutions, but as ideas to inform future study in this space.

A key set of issues that arise as we consider the translation of this model from "bench to bedside" are the issues of algorithmic bias and fairness. These must be very carefully considered and studied before, during, and after any implementation of this work.[66] As noted in the Supplementary Figures S1 and S2 our results show that the proposed model outperforms the baseline model for all age and

sex subpopulations. This is an example of some of the analysis necessary, but not sufficient, to identify sources of algorithmic bias. Although this assessment was not the primary focus of this work, we present a brief discussion of some potential issues that may arise in terms of bias and fairness of this proposed approach.

One potential issue is the nonrepresentativeness of the underlying claims data employed to develop the models. For example, undocumented workers may be underrepresented in this dataset. Generally, these workers are less likely to have their OIs be properly documented, treated, and assigned to workers compensation resources by employers.[67] Other socio-economic factors obscured from claims data may also exert pressure on RTW decision-making, for example RTW duration has been shown to correlate with the size of an injured worker's family.[68] Blindly developing and implementing models may reinforce negative structures in society that harm vulnerable groups of people. As such, it would be problematic to blindly implement the proposed model. Instead, we emphasize that these challenges are areas for careful future study, which should combine additional analytical work with further data collection and study.

Although this work presents new challenges and opportunities, it comes with several limitations. Several of these limitations pertain to the dataset we used to create and validate our model. We utilized a large dataset from the state of Ohio's workers' compensation program, containing OIs and subsequent observations observed between 2001 and 2010. Using data from a single state limits the potential generalizability of the model to other regions, as some of the data collected is specially tailored to Ohio (eg, procedure codes specific to the state of Ohio's workers' compensation program). Additionally, other US states or regions outside of the United States may have a different composition of occupations, injuries, and treatments. Moreover, diagnoses and treatments may have changed since the end of the data collection. For example, the recent shift away from opioid-based analgesics in the treatment of pain is likely not captured by this dataset.[13,69] Despite validating the model on a single region, our work provides a valuable foundation for which to replicate our study for other regions.

Another set of limitations pertain to the inaccessible baseline models and the deep learning architecture used for this study. To assess the improvement that the proposed approach yields, we must compare it against a representative baseline. We trained a baseline analogous to proprietary models by limiting the data to the first week after injury.[16] We tried to ensure parity in terms of capacity between our proposed model and the baseline model by using the same framework and the same hyperparameter space. We believe this yielded a generous baseline, representing the predictive performance of using information collected around the time of injury. We note that this is not an attempt to measure the performance of existing proprietary models. In addition, we employ deep learning approaches, which are powerful, but problematic in terms of complexity, power usage, and interpretability.[70–72] Work described in Supplementary Material S5 examines the impact of longitudinal observations on prediction of future work status using other machine learning architectures. These results suggest that longitudinal observation data plays an important role in predicting future work status. Of note, these models demonstrate worse discriminative performance than the proposed model implemented with deep learning. Further study is needed to fully explore architecture tradeoffs. Although we observed performance degradation when using simpler model architectures, some of the benefits of longitudinal data are still realized under the simpler architectures. It is possible that there may be modeling approaches that provide similar performance benefits to deep learning with less complexity and more interpretability.[73–75] This could be a fruitful direction for future research.

Given the scope of this work, we focused entirely on utilizing retrospectively collected data. To fully assess the utility of using longitudinal observations in real-world usage, the proposed model would need to be studied with a prospective implementation. Finally, the usage of claims-based workers' compensation data provides a limited view into the recovery process, especially when viewed from the lens of algorithmic bias. Although our claims-dataset contains time-stamped information regarding diagnoses and treatments, this is an incomplete depiction of recovery from OIs. For example, job type is a very limited representation of the occupation of the worker and a great deal of recovery depends on psychosocial factors that are not explicitly captured through claims.[12,76] With additional psychosocial information, the proposed framework would likely be able to create models with greater predictive performance that account for these factors.

Our study is, to the best of our knowledge, the first to evaluate the potential of dynamically predicting RTW for injured workers using longitudinal observations. Future work using other large claims or electronic health records datasets may address some of the limitations described above.

CONCLUSION

In this article, we establish the value of using longitudinal observations for the RTW prediction task by comparing approaches that use information collected in the first week of an OI to longitudinal information collected over the course of recovery. For this comparison, we proposed a new formulation for OI prediction as a dynamic work status prediction task. We utilized an approach that transforms longitudinal claims data into a sequence of observations. These longitudinal observations are fed to a recurrent neural network-based model to generate predictions about an injured worker's future work status, and can be used to update estimates over time. Thus, the longitudinal observation approach could help physicians and payers efficiently manage large populations and enable industrial organizations to better plan for their workforce needs. If our initial findings are borne out through subsequent modeling and validation studies, the dynamic prediction of RTW may provide crucial support in clinical decision-making, providing aid for a problem that plagues many insurers, governments, and workers.

FUNDING

This research was sponsored in part by Peers Health through a research agreement with the University of Michigan. Sponsors were given the opportunity to review and comment on all proposed publications; however, the corresponding authors had final right and responsibility for the decision to submit the findings for publication.

AUTHOR CONTRIBUTIONS

All authors contributed substantially to the conception, design, drafting, and revising of this study. Individual author contributions are highlighted here. EÖ: study design, data analysis, data interpretation, manuscript drafting, manuscript revisions. JS: study design, data acquisition, data interpretation, manuscript drafting, manuscript revisions. HW: study design, data interpretation, manuscript drafting, manuscript revisions. BD: study supervision, study

design, data interpretation, manuscript revisions. Additionally, all authors approved the final manuscript and agree to be accountable for all aspects of this work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

EÖ would like to thank Dakota Lambert for his assistance in the configuration and maintenance of the high-performance workstation used for this work. All the authors would like to thank Suyanpeng Zhang and Dr Jenna Wiens for their help in bridging the gap between OIs and machine learning. Additionally, the authors would like to highlight their gratitude to the anonymous reviewers and associated editors at JAMIA. Their feedback has greatly strengthened this work.

CONFLICT OF INTEREST STATEMENT

The University of Michigan has filed a patent application for intellectual property (IP) resulting from this research with EÖ and BTD named as inventors. They and the University may one day benefit financially should the IP be licensed and result in royalties.

DATA AVAILABILITY

The data underlying this study were provided by Peers Health by permission. Data will be shared on request to the corresponding author with permission of Peers Health.

REFERENCES

1. Leigh JP. Economic burden of occupational injury and illness in the United States. *Milbank Q* 2011; 89 (4): 728–72.
2. National Safety Council. Work safety introduction. https://injuryfacts.nsc.org/work/work-overview/work-safety-introduction/. Accessed March 1, 2019.
3. U.S. Bureau of Labor Statistics. *Injuries, Illnesses, and Fatalities*. 2019. https://www.bls.gov/iif/. Accessed 2019.
4. Boden LI, O'Leary PK, Applebaum KM, Tripodis Y. The impact of nonfatal workplace injuries and illnesses on mortality. *Am J Ind Med* 2016; 59 (12): 1061–9.
5. Boden LI, Galizzi M. Economic consequences of workplace injuries and illnesses: lost earnings and benefit adequacy. *Am J Ind Med* 1999; 36 (5): 487–503.
6. Okechukwu CA, Bacic J, Velasquez E, Hammer LB. Marginal structural modelling of associations of occupational injuries with voluntary and involuntary job loss among nursing home workers. *Occup Environ Med* 2016; 73 (3): 175–82.
7. Dong XS, Wang X, Largay JA, Sokas R. Economic consequences of workplace injuries in the United States: findings from the National Longitudinal Survey of Youth (NLSY79). *Am J Ind Med* 2016; 59 (2): 106–18.
8. Seabury SA, Terp S, Boden LI. Racial and ethnic differences in the frequency of workplace injuries and prevalence of work-related disability. *Health Aff (Millwood)* 2017; 36 (2): 266–73.
9. Ben-Shalom Y, Bruns S, Contreary K, Stapleton D. *Stay-at-Work/Return-to-Work: Key Facts, Critical Information Gaps, and Current Practices and Proposals*. Washington, DC: Mathematica Policy Research; 2017.
10. Abstracts for the International Forum on Disability Management (IFDM), London, England, September 10–12, 2012. THE ELECTRONIC ACTIVITY PRESCRIPTION TOOL (ARX): CHANGING THE WORK DISABILITY PARADIGM. *Int J Disabil Manag* 2012; 7: 40–61.
11. Wiesner S, Guerriero J, Garcia M. From patient to productivity: effectiveness of evidence-based guidelines in the clinical environment. In: IBI Annual Forum. Oakland, CA: Integrated Benefits Institute; 2016; San Francisco.
12. Bible JE, Spengler DM, Mir HR. A primer for workers' compensation. *Spine J* 2014; 14 (7): 1325–31.
13. Schieber LZ, Guy GP, Seth P, Losby JL. Variation in adult outpatient opioid prescription dispensing by age and sex—United States, 2008–2018. *MMWR Morb Mortal Wkly Rep* 2020; 69 (11): 298–302.
14. Return-to-Work Guidelines/Modeling. ODG. https://www.mcg.com/odg/odg-solutions/return-work-guidelines-modeling/. Accessed June 2021.
15. MDGuidelines. ReedGroup. https://www.mdguidelines.com. Accessed June 2021.
16. Gaspar F. Duration views methodology. 2017. https://www.reedgroup.com/wp-content/uploads/2017/03/Duration-Views-Methodology.pdf. Accessed June 2021.
17. Mueller K, Konicki D, Larson P, Hudson TW, Yarborough C; ACOEM Expert Panel on Functional Outcomes. Advancing value-based medicine: why integrating functional outcomes with clinical measures is critical to our health care future. *J Occup Environ Med* 2017; 59 (4): e57–62.
18. Hou WH, Tsauo JY, Lin CH, Liang HW, Du CL. Worker's compensation and return-to-work following orthopaedic injury to extremities. *J Rehabil Med* 2008; 40 (6): 440–5.
19. Haldorsen EM. The right treatment to the right patient at the right time. *Occup Environ Med* 2003; 60 (4): 235–6.
20. Hogg-Johnson S, Cole DC. Early prognostic factors for duration on temporary total benefits in the first year among workers with compensated occupational soft tissue injuries. *Occup Environ Med* 2003; 60 (4): 244–53.
21. Steenstra IA, Busse JW, Tolusso D, *et al*. Predicting time on prolonged benefits for injured workers with acute back pain. *J Occup Rehabil* 2015; 25 (2): 267–78.
22. Zhao J, Feng Q, Wu P, *et al*. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep* 2019; 9 (1): 717.
23. Zachary C, Lipton JB, Elkan C. A critical review of recurrent neural networks for sequence learning. *ArXiv* 2015. https://doi.org/10.48550/arXiv.1506.00019.
24. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
25. Graves A. Supervised sequence labelling with recurrent neural networks. *Stud Comput Intell* 2012; 385: 1–141.
26. Graves A, Wayne G, Danihelka I. Neural turing machines. *arXiv* 2014. https://doi.org/10.48550/arXiv.1410.5401.
27. Tompa E, De Oliveira C, Dolinschi R, Irvin E. A systematic review of disability management interventions with economic evaluations. *J Occup Rehabil* 2008; 18 (1): 16–26.
28. Bunn WB, Baver RS, Ehni T, *et al*. Impact of a musculoskeletal disability management program on medical costs and productivity in a large manufacturing company. *Am J Manag Care* 2006; 12: S27–32.
29. Burton WN, Conti DJ. Disability management: corporate medical department management of employee health and productivity. *J Occup Environ Med* 2000; 42 (10): 1006–12.
30. Krause N, Lund T. *Returning to Work after Occupational Injury*. Washington, DC: American Psychological Association; 2004.
31. American College of Occupational and Environmental Medicine. Integrated Health & Safety Index: Guide to a Healthy & Safe Workplace. 2017. https://acoem.org/Guidance-and-Position-Statements/Reference-Materials-Related-OEM-Documents/Integrated-Health-and-Safety-(IHS)-Index. Accessed June 2021.
32. Guideline A. Preventing needless work disability by helping people stay employed. *J Occup Environ Med* 2006; 48 (9): 972–87.
33. Michas MG, Iacono CU. Overview of occupational medicine training among US family medicine residency programs. *Fam Med* 2008; 40 (2): 102.
34. Nuckols TK, Harber P, Lim Y-W, *et al*. *Evaluating Medical Treatment Guideline Sets for Injured Workers in California*. Vol. 400. Santa Monica, CA: RAND; 2005.

35. ODG by MCG. MCG Health. https://www.mcg.com/odg/. Accessed June 2021.
36. MDGuidelines. MDGuidelines Home Page—MDGuidelines. https://www.mdguidelines.com/. Accessed June 2021.
37. Clay FJ, Newstead SV, McClure RJ. A systematic review of early prognostic factors for return to work following acute orthopaedic trauma. *Injury* 2010; 41 (8): 787–803.
38. Papic M, Brdar S, Papic V, Loncar-Turukalo T. Return to work after lumbar microdiscectomy—personalizing approach through predictive modeling. *Stud Health Technol Inform* 2016; 224: 181–3.
39. Gragnano A, Negrini A, Miglioretti M, Corbiere M. Common psychosocial factors predicting return to work after common mental disorders, cardiovascular diseases, and cancers: a review of reviews supporting a cross-disease approach. *J Occup Rehabil* 2018; 28 (2): 215–31.
40. Ervasti J, Joensuu M, Pentti J, *et al*. Prognostic factors for return to work after depression-related work disability: a systematic review and meta-analysis. *J Psychiatr Res* 2017; 95: 28–36.
41. Ghahramani Z. An introduction to hidden Markov models and Bayesian networks. *Int J Patt Recogn Artif Intell* 2001; 15 (01): 9–42.
42. Bolano D, Berchtold A, Ritschard G. A discussion on hidden Markov models for life course data. In: Sequence Analysis and Related Methods (LaCOSA II); June 8–10, 2016: 241; Lausanne.
43. Dymarski P. *Hidden Markov Models: Theory and Applications*. Rijeka, Croatia: InTech; 2011.
44. Bartolucci F, Farcomeni A, Pennoni F. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman and Hall/CRC; 2012.
45. Bai T, Chanda AK, Egleston, BL, Vucetic S. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. *BMC Med Inform Decis* 2018; 18: 123.
46. Beam AL, Kompa B, Fried I, *et al*. Clinical concept embeddings learned from massive sources of medical data. *arXiv preprint arXiv:180401486* 2018. https://doi.org/10.1186/s12911-018-0672-0.
47. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:14090473* 2014. https://doi.org/10.48550/arXiv.1409.0473.
48. Google. Word Embeddings. https://www.tensorflow.org/tutorials/text/word_embeddings. Accessed January 2021.
49. Google. Embeddings. https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture. Accessed January 2021.
50. Szymendera SA, Library of Congress. Congressional Research Service IB. Workers' compensation: overview and issues. Congressional Research Service; 2018. Report/Congressional Research Service; R44580. https://purl.fdlp.gov/GPO/gpo112439. Accessed March 24, 2021.
51. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: a novel bandit-based approach to hyperparameter optimization. 2018-06-18T23:01:43 2018.
52. Wong A, Ötleş E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021; 181 (8): 1065–70.
53. Bröcker J, Smith LA. Increasing the reliability of reliability diagrams. *Weather Forecast* 2007; 22 (3): 651–61.
54. Seedat N, Kanan C. Towards calibrated and scalable uncertainty representations for neural networks. 2019-12-04T02:19:35; 2019.
55. Rossum G. *Python Reference Manual*. CWI; 1995. https://ir.cwi.nl/pub/5008.
56. Abadi M, Agarwal A, Barham P, *et al*. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:160304467* 2016. https://doi.org/10.48550/arXiv.1603.04467.
57. Géron A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: O'Reilly Media, Inc.; 2017.
58. Chollet F. *Keras*. 2015. https://keras.io.
59. Owens M, Allen G. *The Definitive Guide to SQLite*. 2nd ed. New York, NY: Springer; 2010.
60. Oliphant TE. *A Guide to NumPy*. Vol. 1. USA: Trelgol Publishing; 2006. http://numpy.scipy.org.
61. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* 2011; 13 (2): 22–30.
62. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference; 2010: 51–56; Austin, TX.
63. Pollard TJ, Johnson AE, Raffa JD, Mark RG. tableone: an open source Python package for producing summary statistics for research papers. *JAMIA Open* 2018; 1 (1): 26–31.
64. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12 (Oct): 2825–30.
65. Ötleş E, Oh J, Li B, *et al*. Mind the performance gap: examining dataset shift during prospective validation. 2021-07-23T14:30:59; 2021. https://doi.org/10.48550/arXiv.2107.13964.
66. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019; 322 (24): 2377–8.
67. Stuesse A. When they're done with you: legal violence and structural vulnerability among injured immigrant poultry workers. *Anthropol Work Rev* 2018; 39 (2): 79–93.
68. He Y, Hu J, Yu ITS, Gu W, Liang Y. Determinants of return to work after occupational injury. *J Occup Rehabil* 2010; 20 (3): 378–86.
69. Ayres I, Jalal A. The impact of prescription drug monitoring programs on U.S. opioid prescriptions. *J Law Med Ethics* 2018; 46 (2): 387–403.
70. Chakraborty S, Tomsett R, Raghavendra R, *et al*. *Interpretability of Deep Learning Models: A Survey of Results*. San Fransisco, CA: IEEE; 2017.
71. Thompson NC, Greenewald K, Lee K, Manso GF. The computational limits of deep learning. 2020-07-10T18:26:17; 2020. https://doi.org/10.48550/arXiv.2007.05558.
72. Marcus G. Deep learning: a critical appraisal. 2018-01-02T12:49:35; 2018. https://doi.org/10.48550/arXiv.1801.00631.
73. Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N. Survival analysis with electronic health record data: experiments with chronic kidney disease. *Stat Anal Data Min* 2014; 7 (5): 385–403.
74. Daumé H III. Frustratingly easy domain adaptation. 2009-07-10T13:25:48; 2009.
75. Oh J, Makar M, Fusco C, *et al*. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018; 39 (4): 425–33.
76. Rajamani S, Chen ES, Lindemann E, Aldekhyyel R, Wang Y, Melton GB. Representation of occupational information across resources and validation of the occupational data for health model. *J Am Med Inform Assoc* 2018; 25 (2): 197–205.