Simultaneous Measurement Imputation and Outcome Prediction for Achilles Tendon Rupture Rehabilitation

Charles Hamesse* CHARLES.HAMESSE@MIL.BE
KTH Royal Institute of Technology *Royal Military Academy*
Stockholm, Sweden *Brussels, Belgium*

Ruibo Tu* RUIBO@KTH.SE
KTH Royal Institute of Technology
Stockholm, Sweden

Paul Ackermann PAUL.ACKERMANN@SLL.SE
Karolinska University Hospital
Stockholm, Sweden

Hedvig Kjellström HEDVIG@KTH.SE
KTH Royal Institute of Technology
Stockholm, Sweden

Cheng Zhang CHENG.ZHANG@MICROSOFT.COM
Microsoft Research
Cambridge, UK

Abstract

Achilles Tendon Rupture (ATR) is one of the typical soft tissue injuries. Rehabilitation after such a musculoskeletal injury remains a prolonged process with a very variable outcome. Accurately predicting rehabilitation outcome is crucial for treatment decision support. However, it is challenging to train an automatic method for predicting the ATR rehabilitation outcome from treatment data, due to a massive amount of missing entries in the data recorded from ATR patients, as well as complex nonlinear relations between measurements and outcomes. In this work, we design an end-to-end probabilistic framework to impute missing data entries and predict rehabilitation outcomes simultaneously. We evaluate our model on a real-life ATR clinical cohort, comparing with various baselines. The proposed method demonstrates its clear superiority over traditional methods which typically perform imputation and prediction in two separate stages.

1. Introduction

Soft tissue injuries, such as Achilles Tendon Rupture (ATR), are increasing in recent decades (Huttunen et al., 2014). Such injuries require lengthy healing processes with abundant complications, which can cause severe incapacity in individuals. Influences of various factors such as patient demographics and different treatment methods are not clear for the rehabilitation outcome due to large variations in symptoms and the long healing process. Additionally, many medical examinations are not carried out for a large portion of patients

* Both authors contributed equally to this manuscript.

since they can be costly and/or painful. Thus, accurately predicting the ATR rehabilitation outcome at different stages using existing measurements is highly interesting, and can be used for decision support for practitioners. Moreover, ATR is one example of a wider class of medical conditions. In these situations, patients first need acute treatments, then go through a long-term and uncertain rehabilitation process (Horstmann et al., 2012). Decision support tools for practitioners are in general of great need, and outcome prediction plays an important role in the medical decision-making.

Predicting ATR rehabilitation outcomes is extremely challenging for both medical experts and machines. This is mainly due to large number of noisy or absent measurements from various medical instruments. Medical tests and outcome scores for ATR involve a large variety of metrics. The total number of those metrics is on the magnitude of hundreds. However, only a subset of all possible medical measurements are used for a patient. Thus, the observations are very sparse. In this work, we use an ATR cohort which is collected from multiple hospitals in the past five years. The sparsity of this cohort is the consequence of several phenomena: firstly, they are aggregated from different studies realized by different clinicians who have different procedures; secondly, some measurements can be painful, costly or time-consuming such that not all patients are willing to take them. Such phenomena are common in many medical cohorts. Moreover, among those measurements, many are noisy and establish highly non-linear relationship to the rehabilitation outcome, which makes the outcome prediction task difficult.

Leveraging data-driven approaches, we design machine learning models to predict potential ATR rehabilitation outcomes with sparse and noisy data for patients, and provide decision support for practitioners. In particular, we develop a probabilistic model to address two problems at once: imputing the missing values of the costly medical measurements for patients, and predicting patients' final rehabilitation outcome. We focus on predicting the ATR rehabilitation outcome, while our framework can be further applied to a wider domain of conditions beyond ATR.

Technical significance. We propose a novel probabilistic framework where probabilistic matrix factorization is combined with a Bayesian Neural Network (BNN) for rehabilitation outcome prediction with the noisy and sparse dataset. Our method shows clear improvement for this task comparing to traditional methods. For outcome prediction with such a cohort, traditional methods commonly need two stages: Firstly, missing values are imputed using methods such as mean-imputing or zero-imputing; secondly, a linear model is used to predict the outcome with the imputed data (Arverud et al., 2016; Bostick et al., 2010). These methods commonly lead to a low prediction quality because of the low imputation quality and the linear relationship assumption between measurements and outcomes.

Our framework simultaneously imputes the missing values and predicts the rehabilitation outcomes. We first adopt a probabilistic latent variable model to predict the missing entries in the hospital stay measurements. Prediction based methods, such as using latent variable models, in general, demonstrate superior performance for data imputation compared to traditional methods such as mean imputation (Scheffer, 2002; Buuren and Groothuis-Oudshoorn, 2010; Keshavan et al., 2010; Ma et al., 2018). These latent variables summarize patients' underlying health situation in a low-dimensional space and impute the missing entries based on the patient status. We then combine the latent variable imputation

model with BNN to predict ATR rehabilitation outcomes as one integrated probabilistic model. BNN is highly flexible, thus can handle non-linear relationships between rehabilitation outcomes and measurements. Moreover, our model is fully probabilistic, thus provides uncertainty estimation of the prediction results. In an end-to-end manner, our framework provides significant improvement in the clinical standard.

Clinical relevance. ATR rehabilitation is a prolonged process with unpredictable variation in the individual long-term outcome. The optimal and individualized rehabilitation protocol is unknown and therefore inappropriate treatments may often be provided leading to worse outcome for the patient and increased cost for society. In this case, prediction of ATR rehabilitation outcome can shorten the healing process by helping clinicians to choose effective treatments based on varying patient characteristics.

There is a large number of ATR treatments and rehabilitation protocols and also assessments made on the patients. Choosing suitable treatments for patients is still challenging. Moreover, different measurements may vary in price and time to perform. Given imputed values for measurements and the predicted values for outcomes with calibrated uncertainty using our method, the clinicians can make decisions on the patient treatment and monitoring more easily. For example, if the model predicts an unobserved measurement value with high confidence and the predicted value is in a clinical normal range, the clinician does not need to measure this value anymore. Thus, time and cost are saved by not performing unnecessary medical assessments. Otherwise, if the prediction indicates any abnormal situation or high uncertainty for an important measurement, it is worthwhile to apply this medical instrument and obtain the measurement value for this patient. The predicted outcome also helps the clinician to better estimate the patient status in general and aids the treatment decisions.

The paper is structured as follows: We discuss related work (Section 2) and describe the ATR cohort (Section 3). We then introduce the proposed model (Section 4). Finally we evaluate our proposed method against multiple baselines. The experimental results demonstrate clear improvement for ATR rehabilitation outcome prediction using our proposed model (Section 5).

2. Related Work

Our work focuses on utilizing machine learning methods for ATR rehabilitation outcome prediction. We use a latent variable model based on probabilistic matrix factorization to address the missing entry problem, and then use the estimated patient state to predict the rehabilitation outcome through BNN. There is very limited work on using machine learning to address the ATR outcome prediction. We revisit the related work in the following three aspects: ATR analysis, AI in a generic health-care setting and missing value imputation, which is a key component for this type of applications.

Achilles tendon rupture analysis. Numerous studies have been carried out on understanding the treatment and rehabilitation of ATR due to its importance in health-care. However, most studies are performed with a clinical approach, and use traditional statistic analysis, typically linear regression. Machine learning based approaches have not been widely adopted in the field of ATR research. As such, tools for rehabilitation outcome pre-

diction using machine learning are of great interest. Here, we briefly review some related work on ATR.

Olsson et al. (2014) employ linear regression to predict the rehabilitation outcome using variables such as age, sex, body mass index (BMI) or physical activity. The result shows that using traditional statistical models such as linear regression yields a limited prediction ability despite having a wide range of clinically relevant variables. A more recent study shows that assessing clinical markers of tendon callus production (procollagen type I N-terminal propeptide (PINP) and type III N-terminal propeptide (PIIINP)) shortly after operation can help improve the prediction of long-term patient-reported outcomes applying multiple linear regression on Achilles Tendon Total Rupture Score (ATRS) one year post-injury (Alim et al., 2016). Additionally, microcirculation in the tendon was also shown to be a strong predictor of the patient outcome after ATR (Praxitelous et al., 2017). Although insightful, this research also shows that some accurate measurements, such as microcirculation, can be expensive or difficult to obtain. Therefore, utilizing a large range of cost-efficient data to predict the rehabilitation outcome is desirable.

AI in health-care. There is a broad spectrum of machine learning methods used for generic medical applications. When dealing with large amounts of data, deep learning algorithms show promising results. For example, long short-term memory networks (LSTM) and convolutional neural networks (CNN) has been applied to various clinical tasks such as mortality prediction in ICU setting where the data are often gathered from sensor readings (Suresh et al., 2017; Chalapathy et al., 2016; Jo et al., 2017; Purushotham et al., 2017).

However, health-care datasets often have a limited number of patients with large numbers of variables from different instruments. This often leads to datasets with many missing entries. At the same time, being able to encode existing medical research results in new models and providing interpretable results are desirable features in many health-care related applications. In this case, probabilistic models are needed. Depending on the medical context, different types of models are used. For example, Lasko (2014) employs Gaussian processes to predict irregular and discrete medical events. Schulam and Saria (2015) design a hierarchical latent variable model to predict the trajectory of an individual's disease. These models are developed for different medical contexts and are not directly applicable to our application setting. In this work, we use ATR as an example and design a model to predict patients' rehabilitation outcomes after acute treatments.

Missing value imputation. Most real-life medical cohorts have a large amount of missing values. Traditional methods such as zero imputation or mean imputation ease the analysis but may lead to low imputation accuracy. For the datasets with missing values, matrix factorization based methods are shown to be effective for many missing value imputation applications (Shi et al., 2016; Troyanskaya et al., 2001), and frequently used for other applications of the matrix completion problem, i.e., collaborative filtering (Ocepek et al., 2015). Many efficient algorithms have been proposed, such as Singular Value Thresholding (SVT) (Cai et al., 2010), Fixed Point Continuation (FPC) (Ma et al., 2011), and Inexact Augmented Lagrange Multiplier (IALM) (Lin et al., 2010). Typically, these methods construct a matrix factorization objective and optimize it using traditional convex optimization techniques. Extensions, such as Singular Value Projection (SVP) (Jain et al., 2010) and OptSpace (Keshavan and Oh, 2009) consider observation noise in the objective. How-

ever, the sparse dataset can damage the performance of matrix factorization based methods (Mnih and Salakhutdinov, 2008). In this case, probabilistic matrix factorization (Mnih and Salakhutdinov, 2008) is an alternative solution for sparse and imbalanced datasets.

In this work, we combine a probabilistic matrix factorization approach similar to that of Matchbox (Stern et al., 2009), with a supervised learning approach using models such as Bayesian neural networks (Neal, 2012). Therefore, we can impute missing values based on latent patient traits with the sparse ATR dataset and predict the rehabilitation outcome in an end-to-end probabilistic framework.

3. Cohort

In our work, the cohort is a real-life dataset collected from multiple previous studies by an orthopedic group (Valkering et al., 2017; Domeij-Arverud et al., 2016). A snapshot of the dataset is shown in Figure 1. There are 442 patients in the dataset ($N = 442$). The number of measurements is $M = 297$, and the number of the outcome scores is $S = 63$. We denote the first $N \times M$ part of the dataset as the *predictors*, $\mathbf{P}$, and the second $N \times S$ part as the *scores*, $\mathbf{S}$. The percentage of missing values is 69.5% in the predictors and 64.2% in the scores.

	Length	Weight	...	DVT_2	...	ATRS_12_stiff
1	190	79.8	...	×	...	8
2	×	76.5	...	0	...	×
3	×	×	...	1	...	10
4	178	96.7	...	0	...	×

Figure 1: A snapshot of the Achilles Tendon Rupture (ATR) cohort. Each row represents a patient's medical record and each column represents a measurement. In this example, DVT_2 refers to the presence of deep venous thrombosis after two weeks and the ATRS_12_stiff measurements refer to the Achilles Tendon Rupture Score (ATRS) metrics of stiffness after 12 months. × indicates the entry is missing.

Problem setting. We review a typical case of patient journey first and then introduce the problems. ATR patients typically go to hospital to get a treatment immediately after an injury. There, their demographic data are registered. As part of the treatment process, they go through a number of tests from various medical instruments. Due to the complexity of these tests (e.g. in terms of time, cost, pain, invasiveness, accuracy), not all patients go through the same procedure. This leads to a lot of missing data and a lot of variation in which measurements are missing. After the treatment, patients are discharged from the hospital to heal. To monitor the healing process, they are asked to return to the hospital for rehabilitation examination after 3, 6 and 12 months. Not all tests are applied for all patients in the study, since not all patients return on time for rehabilitation examination. Thus, the rehabilitation outcome scores also have a large amount of missing entries.

Based on the patient journey, we split these variables into two categories. The first one contains patient demographics and measurements realized during their stay at a hospital.

These measurements include features such as age, BMI, blood tests of various chemicals related to tendon callus production, whether there was surgical intervention, or information on post-operative treatment. Variables in this category are referred to as *predictors* in the following text. The second category is the *scores*, and includes all metrics of rehabilitation outcomes such as ATRS or Foot and Ankle Outcome Score (FAOS). An example snapshot of the dataset is depicted in Figure 1. In this work, we will impute the *predictors* and predict the *scores*.

4. Methods

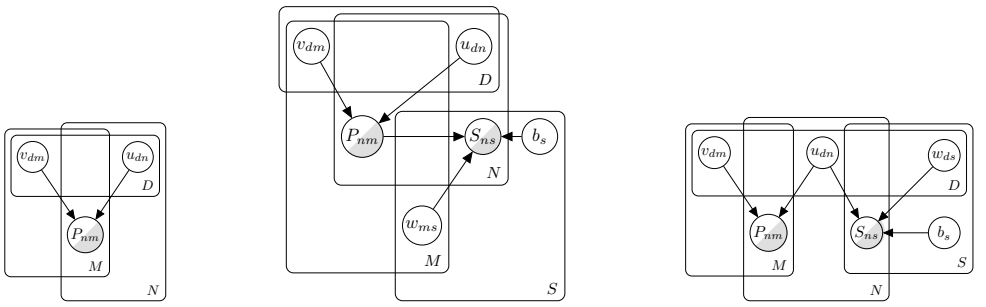
We design an end-to-end probabilistic model to simultaneously impute the missing entries in the predictors and predict the rehabilitation outcomes. The data imputation part is a latent variable model which can be used separately or be part of the end-to-end model. For the prediction part, we provide multiple alternatives of modeling choices, including Bayesian linear regression and Bayesian neural networks, using either the learned latent variables or the imputed predictors as inputs. In this section, we first introduce the basic data imputation unit and then introduce our end-to-end model for simultaneous data imputation and rehabilitation outcome prediction.

4.1. Measurement imputation

We first present the component of the model which aims to recover the missing measurements in the predictors part of the matrix, $\mathbf{P} \in \mathbb{R}^{N\times M}$. We formulate the missing data imputation problem into a collaborative filtering problem. Typically, matrix factorization models are used in collaborative filtering for recommender systems. They work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices to predict unseen ratings. Thus, this technique can be used for data imputation. We adopt a probabilistic matrix factorization based method (Stern et al., 2009), where the latent traits are used to model the personal preference of users and the ratings for all items are predicted, but in this work we model the patient state and predict the missing measurements. The result is a latent variable model with Gaussian distributions, and its graphical representation is shown in Figure 2(*a*).

For N patients, M measurements, S scores and a latent space of size D, the model assumes that the patient measurement affinity matrix $\mathbf{A} \in \mathbb{R}^{N\times M}$ is generated from the patient traits $\mathbf{U} \in \mathbb{R}^{N\times D}$, which reflect the health status of the patient, and predictor traits $\mathbf{V} \in \mathbb{R}^{M\times D}$, which map different health status to measurements from various medical instruments. We use Gaussian distributions to model these entries. Thus, $p(\mathbf{U}|\sigma_{\mathbf{U}}^2) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{u}_i|\mu_{\mathbf{U}}, \sigma_{\mathbf{U}}^2\mathbf{1})$, $p(\mathbf{V}|\sigma_{\mathbf{V}}^2) = \prod_{j=1}^{M} \mathcal{N}(\mathbf{v}_j|\mu_{\mathbf{V}}, \sigma_{\mathbf{V}}^2\mathbf{1})$. Not all measurements are observed. The measurement imputation model is

$$p(\mathbf{P}|\mathbf{U}, \mathbf{V}, \sigma_{\mathbf{P}}^2) = \prod_{n=1}^{N}\prod_{m=1}^{M} \left[\mathcal{N}(P_{nm}|\mathbf{u}_n^T\mathbf{v}_m, \sigma_{\mathbf{P}}^2)\right]^{\mathbf{I}(n,m)}, \tag{1}$$



(*a*) Matrix factorization (*b*) Using dense predictors $\hat{\mathbf{P}}$ (*c*) Using patient traits $\hat{\mathbf{U}}$

Figure 2: Graphical representation of the proposed models with probabilistic matrix factorization and Bayesian linear regression (or BNN). Half-shaded nodes describe partially observed variables. Panel (a) is the graphical representation of the probabilistic matrix factorization model for data imputation only. Panel (b) shows the model which uses the imputed measurements to predict the rehabilitation outcome. Panel (c) shows the model which uses the patient traits, a latent representation of the patient state, to predict the rehabilitation outcome.

where $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2. $\mathbf{I}$ is an observation indication matrix, defined as

$$\mathbf{I}(n, m) = \begin{cases} 1 & \text{if } P_{n,m} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Thus, we can use the observed measurements to train the model, and use the generated measurements affinity $\mathbf{A}$ to impute the missing data. Eventually, we want the observed entries in $\mathbf{P}$ to be as close as possible to the corresponding entries in $\mathbf{A}$.

4.2. Simultaneous data imputation and outcome prediction

We present our proposed models which impute the missing entries and predict the rehabilitation outcome. Based on the model presented before, we add the second component to predict the scores matrix $\mathbf{S} \in \mathbb{R}^{N \times S}$ using the patient information. The patient information can be the imputed measurement matrix as shown in Figure 2(*b*), or the patient trait vector which is a low-dimensional summary of the patient state as shown in Figure 2(*c*).

Bayesian linear regression. We consider a Bayesian linear regression model first. The score is modeled as

$$p(\mathbf{S} \mid \mathbf{W}, \mathbf{b}, \mathbf{X}) = \prod_{n=1}^{N} \prod_{s=1}^{S} \left[\mathcal{N}(S_{ns} \mid \mathbf{x}_n \mathbf{w}_s + b_s, \sigma^2_{\mathbf{S}}) \right]^{\mathbf{I}'_{ns}}, \tag{3}$$
$$p(\mathbf{W}) = \mathcal{N}(\mathbf{W} \mid \mathbf{0}, \sigma_w^2 \mathbf{1}),$$
$$p(\mathbf{b}) = \mathcal{N}(\mathbf{b} \mid \mathbf{0}, \sigma^2_{\mathbf{b}}),$$

where the input $\mathbf{X}$ is either the predictors or the patient traits, $\mathbf{W} \in \mathbb{R}^{M \times S}$ and $\mathbf{b} \in \mathbb{R}^S$ are weights and bias parameters for Bayesian linear regression. $\mathbf{S} \in \mathbb{R}^{N \times S}$ indicates the observed rehabilitation scores, which can be seen as the rehabilitation outcome $\mathbf{B}$ masked by boolean observation indicator $\mathbf{I}'$, described similarly to the previous section. For a patient who has gone through the rehabilitation monitoring, our model can be used to predict the missing scores. For a new patient who has just received treatment, our model can predict the future healing outcome. In the case of the predictors (Figure 2(*b*)), we make use of the observed values so that the input $\mathbf{X}$ is $\hat{\mathbf{P}} = \mathbf{I} * \mathbf{P} + (1 - \mathbf{I}) * \mathbf{A}$, where $\mathbf{I}$ is the $N \times M$ measurement observation indicator described in Section 4.1. In the case of the patient traits (Figure 2(*c*)), we simply use $\hat{\mathbf{U}}$ as the input $\mathbf{X}$. In fact, predictors $\hat{\mathbf{P}}$ contain more information but also more noise, and $\hat{\mathbf{U}}$ can be seen as a summary of each patient's characteristics. Therefore, we do our experiments with either $\hat{\mathbf{P}}$ or $\hat{\mathbf{U}}$ as inputs for the second component. Figure 2 displays the graphical model in these two cases.

Bayesian neural network. We also consider a BNN, i.e. a neural network with probabilistic distributions on its weights and biases. In this case, we have the following conditional distribution of the scores

$$p(\mathbf{S} \mid \theta, \mathbf{X}) = \prod_{n=1}^{N} \prod_{s=1}^{S} \left[\mathcal{N}(S_{ns} \mid \mathbf{NN}(\mathbf{x}_n; \theta), \sigma^2_{\mathbf{S}}) \right]^{\mathbf{I}'_{ns}}, \tag{4}$$

where $\mathbf{NN}$ is a Bayesian neural network parameterized by θ, the collection of all weights and biases of the network. Typically, we consider fully connected layers with hyperbolic tangent activations. For a network of L layers, we have

$$\mathbf{H}_l = \tanh\Big(\mathbf{H}_{l-1}\mathbf{W}_l + \mathbf{b}_l\Big) \quad \text{for } l = 1, ..., L, \tag{5}$$

where $\mathbf{H}_l$ is the output of layer l ($\mathbf{H}_0 = \mathbf{X}$), $\mathbf{W}_l$ is the matrix of weights from neurons of layer $l-1$ to neurons of layer l, and $\mathbf{b}_l$ is the bias vector for layer l. We start our experiments by setting up priors on weights according to Xavier's initialization (Glorot and Bengio, 2010). That is, the prior variance of a weight w that feeds into the j-th neuron of layer l depends on $n_{lj,\text{in}}$, the number of neurons feeding into this neuron, and $n_{lj,\text{out}}$, the number of neurons which the result is fed to. For a weight w_{lij} going from layer $l-1$, neuron i to layer l, neuron j, we have

$$p(w_{lij}) = \mathcal{N}(w_{lij} \mid 0, \sigma^2_{w_{lij}}), \tag{6}$$
$$\sigma^2_{w_{lij}} = \frac{2}{n_{lj,\text{in}} + n_{lj,\text{out}}}. \tag{7}$$

We limit the complexity of the networks that we evaluate a small number of hidden layers, since there is a limited amount of data. A model with increased complexity would be more prone to overfitting. The graphical model resembles the one in Figure 2, except that instead of the weights $\mathbf{W}$ and biases $\mathbf{b}$, we have the set of parameters of the network θ. The exact shape of the network is described in the experiments section.

Rehabilitation outcome prediction at various timestamps. As we discussed before, the patient returns to hospital for rehabilitation monitoring after 3, 6, and 12 months. Thus at different timestamps, we have different amounts of observed data. We move further in the patient's journey and apply changes to the previous model so that it can be used at the 3-month or 6-month mark. In the first case, we rearrange our inputs and move the scores at 3 months from $\mathbf{S}$ to $\mathbf{P}$. We denote these rearranged inputs $\mathbf{P}_3$ and $\mathbf{S}_3$. We apply the same procedure for the 6-month mark and define $\mathbf{P}_6$ and $\mathbf{S}_6$. We demonstrate in the next section that if we add more information from the healing monitoring, the performance of the final healing outcome prediction (at 12 months) is clearly improved.

Inference. We run inference on this whole model in a end-to-end manner. We use variational inference with the KL divergence (Blei et al., 2017; Zhang et al., 2017). We implemented all our models with the Edward library (Tran et al., 2016), a probabilistic programming library. We use Gaussian distributions to approximate all posteriors.

5. Experiments

We evaluate our method in this section. We first verify our model and inference algorithm using a synthetic dataset. We then focus on the real-world ATR rehabilitation cohort and present preprocessing details. We compare our proposed method with multiple baselines. Finally, we discuss all experimental results. The experimental results show that our proposed end-to-end model clearly improves the predictive performance in comparison to the baselines. Additionally, we evaluate the rehabilitation outcome prediction at various timestamps and show that the accuracy of the rehabilitation outcome prediction increases with more observations.

5.1. Inference verification with synthetic data

We test our model and inference algorithm with a synthetic dataset first. We build this synthetic dataset based on the generative process of the model and infer the latent parameters. We observe that our algorithms can successfully recover the latent parameters in all different settings.

More precisely, we use $N = 100$, $M = 30$, $P = 10$, and a latent space of size $D = 10$. We generate the true patient and measurement traits by sampling from a normal distribution with mean 0.5 and variance 0.5. As discussed in (Zhang et al., 2015), this model is symmetric: parameters can rotate and inference can yield multiple valid solutions for $\mathbf{U}$ and $\mathbf{V}$. To ensure that we recover the true latent traits, we fix the upper square of $\mathbf{V}$ to the $D \times D$ identity matrix to avoid parameter rotation. Also, we add Gaussian noise with variance 0.1 to the resulting $\mathbf{P}$ matrix. We set priors matching this generative process. For evaluation, we randomly split the data into a training and a testing set with proportions 80%-20%.

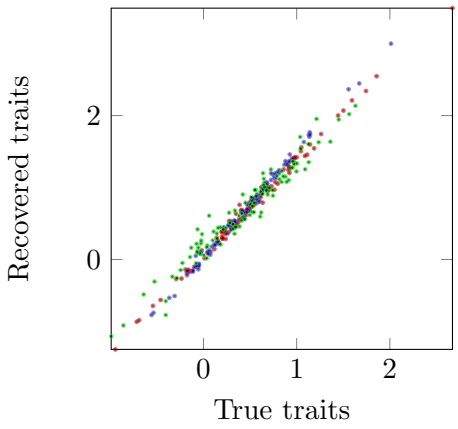


Figure 3: Synthetic data experiment. The plot demonstrates the learned latent traits with respect to the ground-truth. For a point in the figure, its x-coordinate is the true trait value, and y-coordinate is the corresponding recovered trait value. Different colors represent different latent traits. For brevity, we don't show all the recovered latent traits.

By comparing the inferred latent variables with their ground-truth values, we see that we can recover all of them. As an example, we show the ability of our end-to-end model with linear regression on $\mathbf{P}$ (Figure 2(b)) to recover the patients' latent traits and to predict missing values in $\mathbf{P}$ and $\mathbf{S}$. Figure 3 depicts examples of the recovered patient traits. We can see that recovered values are close to true values because points in Figure 3 are close to the diagonal of the square figure. Additionally, we evaluate the training and testing error with the Mean Absolute Error (MAE). We obtain an average training error of 0.042 for $\mathbf{P}$ and 0.027 for $\mathbf{S}$, and an average testing error of 0.056 and 0.037 in the same order. These results validate our model and inference algorithm. Next, we evaluate our model on the real-life ATR dataset.

5.2. Preprocessing of the Achilles tendon rupture rehabilitation cohort

The cohort comes from clinical records with various formats, thus preprocessing is needed before we evaluate our method. First, we convert the whole dataset to numerical values, for example we convert the starting and ending time of the surgery to surgery duration. This process is done under the supervision of medical experts and the whole list of variables that we use are presented in the appendix. The ranges of measurements differ significantly due to the various units in use. We normalize every variable to be in the range of $[0, 1]$. These are affine transformations and the original value can be easily recovered. We do not fill in the missing data in the preprocessing steps as it is one of the goals of our model.

5.3. Baselines

We compare our proposed method with seven variations of our proposed model with two types of baselines. The first type of baseline uses traditional data imputation methods to impute the missing values in $\mathbf{P}$ and predict $\mathbf{S}$. The second one is a two-stage version of our proposed model where data imputation and rehabilitation outcome prediction are performed in a sequential manner.

Traditional data imputation. We first consider imputing the per-patient mean (Schefer, 2002) to all missing values. For each patient, the mean value of their observations belonging to the training set is imputed to all their missing measurements. We also apply

Component 2 Input	BLR $\mathbf{P}$	BLR $\mathbf{S}$	BLR $\mathbf{S}_{ATRS}$
$\hat{\mathbf{P}}$ 2-stage (mean)	0.228 ± 0.0014	0.230± 0.008	0.200 ± 0.010
$\hat{\mathbf{P}}$ 2-stage (OptSpace)	0.224± 0.0028	0.207 ± 0.009	0.193 ± 0.010
$\hat{\mathbf{P}}$ 2-stage (SoftImpute)	0.2049 ± 0.002	0.206 ± 0.008	0.192 ± 0.010
$\hat{\mathbf{P}}$ 2-stage (SVP)	0.316 ± 0.003	0.205 ± 0.012	0.200 ± 0.014
$\hat{\mathbf{P}}$ 2-stage (IALM)	0.237 ± 0.008	0.201 ± 0.011	0.201 ± 0.010
$\hat{\mathbf{P}}$ 2-stage (PMF)	0.164 ± 0.002	0.220 ± 0.006	0.201 ± 0.007
$\hat{\mathbf{U}}$ 2-stage (PMF)	0.164 ± 0.002	0.237 ± 0.006	0.208 ± 0.006
$\hat{\mathbf{P}}$ EE (proposed)	0.181 ± 0.001	0.202± 0.003	0.195± 0.005
$\hat{\mathbf{U}}$ EE (proposed)	0.178± 0.001	**0.164**± 0.004	**0.146**±0.005
Component 2 Input	BNN $\mathbf{P}$	BNN $\mathbf{S}$	BNN $\mathbf{S}_{ATRS}$
$\hat{\mathbf{P}}$ 2-stage (mean)	0.228 ± 0.0014	0.233 ± 0.005	0.202 ± 0.005
$\hat{\mathbf{P}}$ 2-stage (OptSpace)	0.224± 0.0028	0.203 ± 0.008	0.187± 0.009
$\hat{\mathbf{P}}$ 2-stage (SoftImpute)	0.2049 ± 0.002	0.201 ± 0.007	0.186 ± 0.008
$\hat{\mathbf{P}}$ 2-stage (SVP)	0.316 ± 0.003	0.194 ± 0.010	0.187 ± 0.010
$\hat{\mathbf{P}}$ 2-stage (IALM)	0.237 ± 0.008	0.187± 0.011	0.187 ± 0.009
$\hat{\mathbf{P}}$ 2-stage (PMF)	0.164 ± 0.002	0.207 ± 0.007	0.190 ± 0.007
$\hat{\mathbf{U}}$ 2-stage (PMF)	0.164 ± 0.002	0.208± 0.007	0.190 ± 0.007
$\hat{\mathbf{P}}$ EE (proposed)	0.158 ± 0.001	**0.152**±0.004	**0.143**±0.004
$\hat{\mathbf{U}}$ EE (proposed)	0.167±0.001	0.174±0.003	0.152± 0.003

Table 1: Mean Absolute Error (MAE) and standard deviation over 5 runs for outcome prediction. Each time, we use random splits of the data with 80% data for training and 20% data for testing. "EE" indicates end-to-end which is our proposed model. "2-stage" is the baseline model where data imputation and rehabilitation outcome prediction are performed in a sequential manner. BLR stands for Bayesian Linear Regression and BNN stands for Bayesian Neural Network. For the 2-stage models, the error on $\mathbf{P}$ remains the same because the matrix $\mathbf{P}$ is imputed once with the mean imputation or the matrix factorization based methods. In addition, we report the MAE for the ATRS separately. The target is that the MAE of $\mathbf{S}_{ATRS}$ gets smaller than 0.1, because only a difference larger than 0.1 is considered to be clinical different.

traditional matrix factorization based methods: OptSpace (Keshavan et al., 2010), SoftImpute (Mazumder et al., 2010), Singular Value Projection (Jain et al., 2010) and Inexact Augmented Lagrange Multiplier (Lin et al., 2010), to missing values. The predicted values based on observations are imputed to all the missing values. We then use the imputed data to predict rehabilitation outcomes using Bayesian linear regression and Bayesian neural network.

Two-stage version of the proposed model. We run inference on the probabilistic matrix factorization part and only retrieve predictions for the first part of the dataset, $\hat{\mathbf{P}}$, and the patient trait matrix $\hat{\mathbf{U}}$. Then, we define the second model which uses either linear regression or a neural network on this output to give the scores predictions, $\hat{\mathbf{S}}$. Inference is run separately on each component. The intent is to compare our end-to-end model with its direct multi-staged equivalent.

5.4. Results

We split the training and testing set to reflect the treatment journey. In all of our experiments, we first pick training and testing data for $\mathbf{P}$ and $\mathbf{S}$ with the following strategy: we randomly pick 80% of the patients, take all their available data for training and leave the remaining 20% of patients for testing. In other words, we split the dataset on a per-patient basis. We do this since the goal of our work is to predict the rehabilitation outcome after a patient receives the initial treatment. All experiments are repeated 5 times, with all the learned variables getting reset at each run.

We use grid search for hyper-parameter tuning, starting with the matrix factorization part. We observe that the prior mean on the traits has little effect on the end performance. However, the prior variance on both traits and scores has a big impact on how the model fits the training data. We evaluate latent space sizes $D \in [1, 20]$ as well as latent trait variances $\sigma^2_{\mathbf{U}}$ and $\sigma^2_{\mathbf{V}}$, ranging from 0.1 to 0.9 by steps of 0.2. We find the optimal D to be 8 and the optimal variance to be 0.5. Next, we tune the linear regression and neural network. We start by tuning the linear regression then turn it into a neural network with growing complexity by adding activation functions and layers as soon as we find that the model lacks expressive power. We run grid searches on the prior means and variances of the weights and biases as well as the observation noise on $\mathbf{S}$. We notice that for the model to properly fit the training data, weights need to have a very small variance. This is expected since the data is very high-dimensional and the results of dot products need to be constrained in $[0, 1]$). Doing so, we find that dividing the weight variance computed with Xavier's initialization by 10^3 and the prior observation noise by 10^4 yields the best results.

We report the performance of the rehabilitation outcome prediction in Table 1. The Mean Absolute Error (MAE) on the testing set is used as metric for our results. In practice, only a difference larger than 0.1 is considered to be clinically different. Thus, results with MAE within 0.1 are ideal. To compare with this standard, we evaluate prediction methods only with 11 ATRS (10 criteria and the sum), whose results are shown in $\mathbf{S}_{ATRS}$ columns of Table 1 and Table 2. We can see that our proposed end-to-end model with neural network applied on the whole $\mathbf{P}$ matrix achieves the best performance for predicting rehabilitation outcomes and its $\mathbf{S}_{ATRS}$ result is close to the ideal MAE target 0.1. We see that for predicting $\mathbf{S}$, using the patient traits $\hat{\mathbf{U}}$ works better in the case of linear regression, and using the whole predictors matrix $\hat{\mathbf{P}}$ works better in the case of neural network. This is certainly due to the fact that the dimensionality of $\hat{\mathbf{P}}$ makes it difficult for a simple model such as linear regression to extract the key features; a task that a more complex neural network would manage better. In these experiments, we start with a neural network that basically replicates the linear regression then gradually add complexity until we can't

improve the performance without overfitting. The optimal network we found has 1 hidden layer with P (the number of columns of $\mathbf{S}$) hidden units and a hyperbolic tangent activation.

Our proposed method shows clear improvement on the rehabilitation outcome prediction over baselines. We can also see that latent variable models have a good performance on the missing value imputation. Our proposed model is trained for the rehabilitation outcome prediction, so SVP and IALM could have the better performance on the missing value imputation than ours.

	Discharge $\hat{\mathbf{P}}$	3 Month $\hat{\mathbf{P}}_{\mathbf{3}}$	6 Month $\hat{\mathbf{P}}_{\mathbf{6}}$
MAE $\mathbf{S}_3$	0.177 ± 0.006		
MAE $\mathbf{S}_{ATRS-3}$	0.173 ± 0.005		
MAE $\mathbf{S}_6$	0.172± 0.007	0.178± 0.006	
MAE $\mathbf{S}_{ATRS-6}$	0.167 ± 0.009	0.169±0.010	
MAE $\mathbf{S}_{12}$	0.138 ± 0.006	0.140 ± 0.006	**0.132**± 0.006
MAE $\mathbf{S}_{ATRS-12}$	0.111 ± 0.003	0.114 ± 0.004	**0.108**±0.003

Table 2: Rehabilitation outcome prediction performance comparison at various timestamps. Our proposed model with Bayesian neural network is used for this evaluation. We show that the final rehabilitation outcome prediction accuracy increases with time and our model can be used for the rehabilitation outcome prediction at various rehabilitation stages. $\mathbf{S}_{ATRS}$ is evaluated only with ATRS.

Evaluation of the rehabilitation outcome prediction at different timestamps. Here we evaluate the ability of our model to predict scores at different timestamps when we extend $\mathbf{P}$ to include scores at 3 months (yielding $\mathbf{P}_3$) and 6 months (yielding $\mathbf{P}_6$). We report the performance per-timestamp of our model with $\mathbf{P}$, $\mathbf{P}_3$ and $\mathbf{P}_6$ in Table 2.

We observe that including future measurements helps predicting the final scores. Including all the previously observed data in the predictors helps improving the accuracy of future score predictions. Moreover, results of the ATRS prediction at 12 months are close to 0.1 which is our target value.

Per-variable analysis. We further evaluate the prediction accuracy of our best performing model by looking the mean error for each variable. Taking the model with $\hat{\mathbf{P}}$ with BNN as an example, Figures 4 and 5 display the errors and the number of data points available for each variable.

In Figure 5, we show that the number of scores per period varies. In fact, each period has at least 11 ATRS (10 criteria and the sum in blue) and 5 FAOS scores (in red). On top of that, scores at 6 and 12 months both include additional tests such as the evaluation of the heel rise angle (in green). The clinical practice uses scores at 12 months more because they can reflect rehabilitation states better. Figure 5 shows that our model is able to predict the rehabilitation outcome at 12 months better comparing to 3 and 6 months.

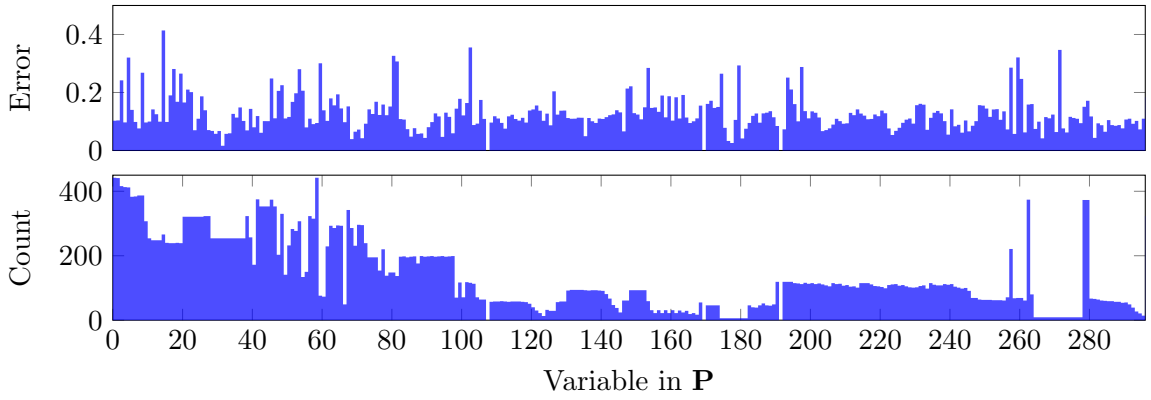


Figure 4: Per-variable mean MAE and number of data points available for training for the predictors $\mathbf{P}$.

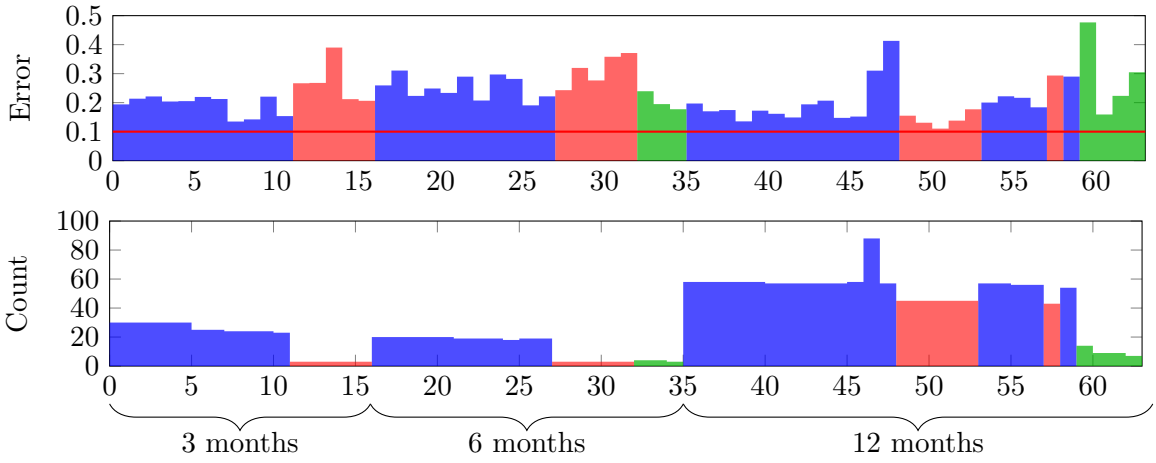


Figure 5: Per-variable mean absolute error (MAE) and number of data points available for training for different components of scores at different rehabilitation time. Blue bars represent ATRS. Red bars represent FAOS. Green bars represent other test scores.

6. Conclusions

We developed a probabilistic end-to-end framework to simultaneously predict the rehabilitation outcome and impute the missing entries in data cohort in the context of Achilles Tendon Rupture (ATR) rehabilitation. We evaluated our model and compared its performance with multiple baselines. We demonstrated a clear improvement in the accuracy of the predicted outcomes in comparison with traditional data imputation methods. Additionally, the performance of our method on rehabilitation outcome prediction is close to the ideal clinical result.

Future work. There is still considerable work to be done in the interpretation of our results, in a clinical sense. An analysis of the impact of each predictor in each model as in Popkes et al. (2019) and a discussion on how these relate to the ATR clinical experiences are desirable to strengthen this work from a medical point of view. Additionally, we are keen to work closely with practitioners to validate our method in a real-life clinical context. We would work on improving the accuracy and interpretability of our model to make it beneficial for both patients and practitioners in real-life health-care process.

A computational aspect to be considered is to investigate in depth on the uncertainty estimation of our model. The Bayesian framework offers a way to compute uncertainties when predicting outcome scores, however, in many tasks, these models have shown to be over confident (Nalisnick et al., 2018). It would be useful for to know when a new patient arrives, how well - with which certainty - we can predict their outcome scores.

The proposed method is a general framework that can be applied to numerous health-care applications involving a long-term healing process after the treatment. In the future, we would collaborate with more health-care departments, test and improve our method in these applications.

References

Md Abdul Alim, Simon Svedman, Gunnar Edman, and Paul Ackermann. Procollagen markers in microdialysate can predict patient outcome after achilles tendon rupture. 2016.

E Domeij Arverud, Per Anundsson, Eva Hardell, Gunilla Barreng, Gunnar Edman, Ali Latifi, Fausto Labruto, and PW Ackermann. Ageing, deep vein thrombosis and male gender predict poor outcome after acute achilles tendon rupture. *The bone & joint journal*, 98(12):1635–1641, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

Geoff P Bostick, Nadr M Jomha, Amar A Suchak, and Lauren A Beaupré. Factors associated with calf muscle endurance recovery 1 year after achilles tendon rupture repair. *journal of orthopaedic & sports physical therapy*, 40(6):345–351, 2010.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.

Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. Bidirectional lstm-crf for clinical concept extraction. *arXiv preprint arXiv:1611.08373*, 2016.

E Domeij-Arverud, P Anundsson, E Hardell, G Barreng, G Edman, A Latifi, F Labruto, and PW Ackermann. Ageing, deep vein thrombosis and male gender predict poor outcome after acute achilles tendon rupture. *Bone Joint J*, 2016.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010.

Thomas Horstmann, C Lukas, J Merk, Torsten Brauner, and Annegret Mndermann. Deficits 10-years after achilles tendon repair. 2012.

Tuomas T. Huttunen, Pekka A Kannus, Christer Gustav Rolf, Li Felländer-Tsai, and Ville M. Mattila. Acute achilles tendon ruptures: incidence of injury and surgery in sweden between 2001 and 2012. *The American journal of sports medicine*, 2014.

Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, 2010.

Yohan Jo, Lisa Lee, and Shruti Palaskar. Combining lstm and latent topic modeling for mortality prediction. *arXiv preprint arXiv:1709.02842*, 2017.

Raghunandan H Keshavan and Sewoong Oh. A gradient descent algorithm on the grassman manifold for matrix completion. *arXiv preprint arXiv:0910.5260*, 2009.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 2010.

Thomas A Lasko. Efficient inference of gaussian-process-modulated renewal processes with application to medical event data. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, 2014.

Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 2011.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 2010.

Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, 2008.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Radford M Neal. *Bayesian learning for neural networks*. 2012.

Uroš Ocepek, Jože Rugelj, and Zoran Bosnić. Improving matrix factorization recommendations for examples in cold start. *Expert Systems with Applications*, 2015.

N Olsson, J Karlsson, BI Eriksson, A Brorsson, M Lundberg, and KG Silbernagel. Ability to perform a single heel-rise is significantly related to patient-reported outcome after achilles tendon rupture. *Scandinavian journal of medicine & science in sports*, 2014.

Anna-Lena Popkes, Hiske Overweg, Ari Ercole, Yingzhen Li, José Miguel Hernández-Lobato, Yordan Zaykov, and Cheng Zhang. Interpretable outcome prediction with sparse bayesian neural networks in intensive care. *arXiv preprint arXiv:1905.02599*, 2019.

Praxitelis Praxitelous, Gunnar Edman, and Paul W Ackermann. Microcirculation after achilles tendon rupture correlates with functional and patient-reported outcome. *Scandinavian journal of medicine & science in sports*, 2017.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets. *arXiv preprint arXiv:1710.08531*, 2017.

Judi Scheffer. Dealing with missing data. 2002.

Peter Schulam and Suchi Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

Weiwei Shi, Yongxin Zhu, S Yu Philip, Tian Huang, Chang Wang, Yishu Mao, and Yufeng Chen. Temporal dynamic matrix factorization for missing data prediction in large scale coevolving time series. *IEEE Access*, 2016.

David Stern, Ralf Herbrich, and Thore Graepel. Matchbox: Large scale bayesian recommendations. In *Proceedings of the 18th International World Wide Web Conference*, 2009.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337, 2017.

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: A library for probabilistic modeling, inference, and criticism, 2016.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 2001.

Kars P Valkering, Susanna Aufwerber, Francesco Ranuccio, Enricomaria Lunini, Gunnar Edman, and Paul W Ackermann. Functional weight-bearing mobilization after achilles tendon rupture enhances early healing response: a single-blinded randomized controlled trial. *Knee Surgery, Sports Traumatology, Arthroscopy*, 2017.

Cheng Zhang, Mike Gartrell, Thomas Minka, Yordan Zaykov, and John Guiver. Groupbox: A generative model for group recommendation. Technical Report MSR-TR-2015-61, 2015.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *arXiv preprint arXiv:1711.05597*, 2017.

Appendix A. Variables

A.1. Predictors

0	ID
1	Study
2	Gender
3	Age
4	DIC_age_40
5	Length
6	Weight
7	BMI
8	DIC_BMI_27
9	Smoker
10	ln_Age
11	ln_Lenght
12	ln_Weight
13	ln_BMI
14	Inj_side
15	Complication
16	Paratenon
17	Fascia
18	PDS
19	Surg_comp
20	Treatment_Group
21	Ort_B
22	Plast
23	Healthy_control
24	Plast_foot
25	Vacoped
26	VTIS
27	TTS
28	ln_TTS
29	TTS_no_of_24h_cycles
30	ln_TTS_no_of_24h_cycles
31	TEST_TTS_48h_POL
32	TEST_TTS_24h_POL
33	TEST_TTS_12h_POL
34	DIC_TTS_by_VTIS_median
35	TRICH_48_84_TTS
36	TRICH_48_96_TTS
37	TRICH_48_72_TTS
38	NEW_Time_er_to_op_start
39	DIC_NEW_Time_er_to_op_start

40	DIC3_NEW_Time_er_to_op_start
41	OLD_Time_er_to_op_start
42	Op_time
43	DIC_op_34min
44	Op_B_Dic
45	OP_GBG_dic
46	Op_time_dic
47	OP_NR
48	EXP
49	Q_RANK_B
50	Q_RANK_ABCD
51	NR_of_Op
52	DIC_nr_OP
53	ASS_Y_N
54	DIK_SPEC
55	PP_IPC_Study_B
56	Pump_pat_reg
57	Pump_reg
58	Highest_pump_reg
59	DIC_86h_highest_pump_reg
60	Pump_comp
61	Incl_Excl
62	DVT_2
63	DVT_6w
64	DVT_2w_and_6w
65	DVT_2w_or_6w
66	DVT_8w
67	Any_dvt
68	Inf_2w
69	Inf_6w
70	Any_inf
71	Rerupture
72	Adeverse_events_1
73	Adeverse_events_2
74	Adeverse_events_3
75	Adeverse_events_4
76	Preinjury
77	Post_op
78	D_PAS
79	Preinj_2cl
80	Preinj_3cl
81	Post_op_2cl
82	Con_Power_I

83	Con_Power_U
84	LSI_Con_Power
85	Total_work_I
86	Total_work_U
87	NEW_LSI_Total_work
88	LSI_Total_work
89	Repetition_I
90	Repetition_U
91	LSI_Repetitions
92	Height_Max_I
93	Height_Max_U
94	LSI_Height
95	Height_A_I
96	Height_A_U
97	LSI_Height_Ave
98	Ecc_Power_I
99	Height_Min_I
100	Ecc_Power_U
101	Height_Min_U
102	LSI_Height_2cl
103	LSI_Height_Min
104	LSI_Ecc_Power
105	Muscle_vein_thrombosis_2
106	Thompson_2
107	Wound_2
108	Podometer_day1
109	Podometer_day2
110	Podometer_day3
111	Podometer_day4
112	Podometer_day5
113	Podometer_day6
114	Podometer_day7
115	Podometer_day8
116	Podometer_day9
117	Podometer_day10
118	Podometer_day11
119	Podometer_day12
120	Podometer_day13
121	Podometer_day14
122	Podometer_day15
123	Podometer_day16
124	Mean_podometer
125	Total_podometer

126	DIC_podometer_16500
127	Podometer_on_day_of_microdialysis
128	Podometer_on_day_minus_1
129	Podometer_on_day_minus_2
130	Subjective_load_day1
131	Subjective_load_day2
132	Subjective_load_day3
133	Subjective_load_day4
134	Subjective_load_day5
135	Subjective_load_day6
136	Subjective_load_day7
137	Subjective_load_day8
138	Subjective_load_day9
139	Subjective_load_day10
140	Subjective_load_day11
141	Subjective_load_day12
142	Subjective_load_day13
143	Subjective_load_day14
144	Subjective_load_day15
145	Subjective_load_day16
146	Mean_subjective_load
147	DIC_43_Mean_subjective_load
148	Days_until_microdialysis
149	Load_on_day_of_microdialysis
150	Load_on_day_minus_1
151	Load_on_day_minus_2
152	Number_of_days_with_load_prior_to_microdialysis
153	DIC_13_days_with_load
154	VAS_day1_act
155	VAS_day1_pas
156	VAS_day2_act
157	VAS_day2_pas
158	VAS_day3_act
159	VAS_day3_pas
160	VAS_day4_act
161	VAS_day4_pas
162	VAS_day5_act
163	VAS_day5_pas
164	VAS_day6_act
165	VAS_day6_pas
166	VAS_day7_act
167	VAS_day7_pas
168	VAS_injured_2weeks

169	VAS_control_2weeks
170	Calf_circumference_injured_1
171	Calf_circumference_injured_mean
172	Calf_circumference_control_1
173	Calf_circumference_control_mean
174	Plantar_flexion_injured_1
175	Plantar_flexion_injured_2
176	Plantar_flexion_injured_3
177	Plantar_flexion_injured_mean
178	Plantar_flexion_control_1
179	Plantar_flexion_control_2
180	Plantar_flexion_control_3
181	Plantar_flexion_control_mean
182	Dorsal_flexion_injured_1
183	Dorsal_flexion_injured_2
184	Dorsal_flexion_injured_3
185	Dorsal_flexion_injured_mean
186	Dorsal_flexion_control_1
187	Dorsal_flexion_control_2
188	Dorsal_flexion_control_3
189	Dorsal_flexion_control_mean
190	Q1
191	Q2
192	Q3
193	Q4
194	Q5
195	EQ5D_ix
196	VAS
197	VAS_2
198	Gluc2_2_i
199	Gluc2_3_i
200	Gluc2_4_i
201	GLUC_injured_mean
202	Gluc2_2_c
203	Gluc2_3_c
204	Gluc2_4_c
205	GLUC_control_mean
206	Lact2_2_i
207	Lact2_3_i
208	Lact2_4_i
209	LAC_injured_mean
210	Lact2_2_c
211	Lact2_3_c

212	Lact2_4_c
213	LAC_control_mean
214	Pyr2_2_i
215	Pyr2_3_i
216	Pyr2_4_i
217	PYR_injured_mean
218	Pyr2_2_c
219	Pyr2_3_c
220	Pyr2_4_c
221	PYR_control_mean
222	Glyc2_2_i
223	Glyc2_3_i
224	Glyc2_4_i
225	GLY_injured_mean
226	Glyc2_2_c
227	Glyc2_3_c
228	Glyc2_4_c
229	GLY_control_mean
230	Glut2_2_i
231	Glut2_3_i
232	Glut2_4_i
233	GLUT_injured_mean
234	Glut2_2_c
235	Glut2_3_c
236	Glut2_4_c
237	GLUT_control_mean
238	Lac2_Pyr2_ratio_2_i
239	Lac2_Pyr2_ratio_3_i
240	Lac2_Pyr2_ratio_4_i
241	LAC2_PYR2_ratio_injured_mean
242	Lac2_Pyr2_ratio_2_c
243	Lac2_Pyr2_ratio_3_c
244	Lac2_Pyr2_ratio_4_c
245	LAC2_PYR2_ratio_control_mean
246	PINP_injured
247	PIIINP_injured
248	Bradford_injured
249	PINP_normalized_Injured
250	PIIINP_normalized_injured
251	PINP_uninjured
252	PIIINP_uninjured
253	Bradford_uninjured
254	PINP_normalized_Uninjured

255	PIIINP_normalized_uninjured
256	Collagen
257	Glut_2_inj_values
258	P_ratio_inj
259	DIC_PIIINP
260	DIC_PIIINP_3
261	P_ratio_uninj
262	FIL_OP_STUDY
263	Gly_inv
264	RF_injured
265	RF_uninjured
266	BZ_injured
267	BZ_uninjured
268	MF_injured
269	MF_uninjured
270	T_RF_injured
271	T_RF_uninjured
272	T_MF_injured
273	T_MF_uninjured
274	T_HR_injured
275	T_HR_uninjured
276	Ratio_MF_RF_injured
277	Ratio_MF_RF_uninjured
278	B1_D66
279	Sthlm_gbg
280	stepsxload_day1
281	stepsxload_day2
282	stepsxload_day3
283	stepsxload_day4
284	stepsxload_day5
285	stepsxload_day6
286	stepsxload_day7
287	stepsxload_day8
288	stepsxload_day9
289	stepsxload_day10
290	stepsxload_day11
291	stepsxload_day12
292	stepsxload_day13
293	stepsxload_day14
294	stepsxload_day15
295	stepsxload_day16
296	INC_A42

A.2. Scores

0	DIC_TTS_by_valid_ATRS80_median
1	Control_1yr
2	Heel_rise_average_height_injured_6mo
3	Heel_rise_average_height_control_6mo
4	difference_heel_raise_6mo
5	Heel_rise_average_height_injured_1yr
6	Heel_rise_average_height_control_1yr
7	difference_heel_raise_1yr
8	ATRS_3_Strenght
9	ATRS_3_tired
10	ATRS_3_stiff
11	ATRS_3_pain
12	ATRS_3_ADL
13	ATRS_3_Surface
14	ATRS_3_stairs
15	ATRS_3_run
16	ATRS_3_jump
17	ATRS_3_phys
18	ATRS_3_Sum
19	ATRS_item1_6month
20	ATRS_item2_6month
21	ATRS_item3_6month
22	ATRS_item4_6month
23	ATRS_item5_6month
24	ATRS_item6_6month
25	ATRS_item7_6month
26	ATRS_item8_6month
27	ATRS_item9_6month
28	ATRS_item10_6month
29	ATRS_total_score_6month
30	ATRS_12_strength
31	ATRS_12_tired
32	ATRS_12_stiff
33	ATRS_12_pain
34	ATRS_12_ADL
35	ATRS_12_Surface
36	ATRS_12_stairs
37	ATRS_12_run
38	ATRS_12_jump
39	ATRS_12_phys
40	ATRS_12m
41	valid_ATRS_12m

42	ATRS_2cl
43	FAOS_3_Pain
44	FAOS_3_Symptom
45	FAOS_3_ADL
46	FAOS_3_sport_rec
47	FAOS_3_QOL
48	FAOS_6_Symptom
49	FAOS_6_Pain
50	FAOS_6_ADL
51	FAOS_6_Sport_Rec
52	FAOS_6_QOL
53	FAOS_12_Pain
54	FAOS_12_Symptom
55	FAOS_12_ADL
56	FAOS_12_Sport_Rec
57	FAOS_12_QOL
58	ATRS_12_pain_log
59	ATRS_12_ADL_log
60	ATRS_12_surface_log
61	ATRS_12_phys_log
62	FAOS_12_pain_log