Simultaneous Measurement Imputation and Outcome Prediction for Achilles Tendon Rupture Rehabilitation

Charles Hamesse[1], Paul Ackermann[2], Hedvig Kjellström[1], and Cheng Zhang[3]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
[2] Karolinska University Hospital, Stockholm, Sweden
[3] Microsoft Research, Cambridge, UK

Abstract. Achilles Tendon Rupture (ATR) is one of the typical soft tissue injuries. Accurately predicting the rehabilitation outcome of ATR using noisy measurements with missing entries is crucial for treatment decision support. In this work, we design a probabilistic model that simultaneously predicts the missing measurements and the rehabilitation outcome in an end-to-end manner. We evaluate our model and compare it with multiple baselines including multi-stage methods using an ATR clinical cohort. Experimental results demonstrate the superiority of our model for ATR rehabilitation outcome prediction.

1 Introduction

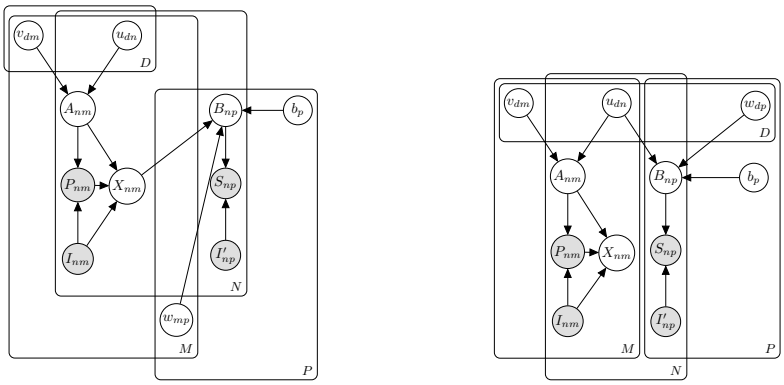
Soft tissue injuries, such as Achilles Tendon Rupture (ATR), are increasing in recent decades [3]. These injuries require lengthy healing processes with abundant complications which can cause severe incapacity to individuals. Numerous measurements are not carried out for all patients since they can be costly and/or painful. Thus, accurately predicting the rehabilitation outcome at different stages using the existing measurements is a crucial problem. Leveraging the predictive power of data-driven approaches, it is of great interest to find out whether we can predict potential outcomes for new patients with sparse and noisy data, and thus provide decision support for practitioners. In this work, we focus on predicting the rehabilitation outcome of ATR, but our framework can be further applied to a wider domain of conditions. In particular, we develop a generic, end-to-end model to tackle two problems at once: imputing the missing measurements and test values for patients during their stay at the hospital, and predicting the outcome of their rehabilitation after 3, 6 and 12 months.

2 Cohort

We use a real-life dataset from the Orthopedic Research Group, aggregated from multiple previous studies [8, 2]. It consists of a longitudinal cohort with 442 patients described by 360 variables. A snapshot of the dataset is shown in Fig. 1. We split all variables into two categories based on the patient's journey. The first category contains patients' demographics and measurements realized during their stay at hospital; variables in this category are referred to as *predictors* $\mathbf{P}$ in

	Age	Length	Weight	...	DVT_2	...	ATRS_12_stiff	ATRS_12_pain
1	27	190	79.8	...	1	...	8	10
2	36	×	×	...	×	...	×	8
3	41	172	×	...	0	...	10	10

Fig. 1. Snapshot of the ATR dataset. Each row represents a patient's medical record and each column represents a measurement. × indicates the entry is missing.



(a) Using the measurement $\hat{\mathbf{P}}$ (b) Using patient traits $\hat{\mathbf{U}}$

Fig. 2. Graphical representation of the models with matrix factorization and linear regression. Panel (a) shows the model using imputed measures to predict the rehabilitation outcome, Panel (b) shows the model using patient traits.

the following. The second category is the *scores* $\mathbf{S}$, and includes all rehabilitation outcome tests, such as ATRS [5] or FAOS taken at 3, 6 and 12 months. ATRS and FAOS assess the function and symptom of the tendon by a number of patient-reported criteria. At the time when the patient is discharged from hospital, the number of measurements is $M = 297$, and the number of the scores to predict in the next three visits is $P = 63$, for $N = 442$ patients. 69.5% entries are missing for the predictors, and 64.2% for the scores.

3 Methods

We design an end-to-end probabilistic model to simultaneously impute the missing entries and predict the rehabilitation outcome as shown in Fig. 2.

We formulate the missing data imputation problem into a collaborative filtering or matrix factorization problem [6]. The model assumes that the patient measurement affinity matrix $\mathbf{A}$ is generated from the patient traits $\mathbf{U}$, which reflect the health status of the patient, and predictor traits $\mathbf{V}$, which map different health status to measurements from various medical instruments. We use Gaussian distributions to model these entries in the same way as [6]: $p(\mathbf{U}|\sigma_{\mathbf{U}}^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{u}_n|\mu_{\mathbf{U}}, \sigma_{\mathbf{U}}^2\mathbf{1})$, $p(\mathbf{V}|\sigma_{\mathbf{V}}^2) = \prod_{m=1}^{M} \mathcal{N}(\mathbf{v}_m|\mu_{\mathbf{V}}, \sigma_{\mathbf{V}}^2\mathbf{1})$. The measurement imputation model is:

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

where I_{nm} is an indicator set to 1 if P_{nm} is observed and 0 otherwise.

We then predict the score matrix $\mathbf{S}$ using the patient information, which can be either the imputed measurement matrix as shown in Fig. 2(a) or the patient trait vector which summaries the patient state as shown in Fig. 2(b).

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

$$p(\mathbf{S} \mid \mathbf{W}, \mathbf{b}, \mathbf{X}) = \prod_{n=1}^{N} \prod_{p=1}^{P} \left[\mathcal{N}(S_{np} \mid \mathbf{x}_n \mathbf{w}_p + b_p, \sigma_{\mathbf{S}}^2) \right]^{I'_{np}}, \tag{2}$$

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{W} \mid \mathbf{0}, \sigma_w^2 \mathbf{1}), \qquad p(\mathbf{b}) = \mathcal{N}(\mathbf{b} \mid \mathbf{0}, \sigma_{\mathbf{b}}^2), \tag{3}$$

where the input $\mathbf{X}$ is either the predictors or the patient traits, $\mathbf{W}$ and $\mathbf{b}$ are the weights and bias parameters for Bayesian linear regression. $\mathbf{S}$ indicates the observed rehabilitation scores, that is the rehabilitation outcome $\mathbf{B}$, masked by the boolean observation indicator $\mathbf{I}'$. In the case of the predictors (Fig. 2(a)), we use the observed values so that $\mathbf{X} = \hat{\mathbf{P}} = \mathbf{I} * \mathbf{P} + (1 - \mathbf{I}) * \mathbf{A}$, where $\mathbf{I}$ is the $N \times M$ measurement observation indicator. In the case of the patient traits (Fig. 2(b)), we have $\mathbf{X} = \hat{\mathbf{U}}$.

Bayesian neural network We also consider a Bayesian neural network (BNN) [4]. In this case, we have the following conditional distribution of the scores:

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

where $\mathbf{NN}$ is a BNN parameterized by θ, the collection of all weights and biases of the network. We consider fully connected layers with hyperbolic tangent activations. The graphical model resembles the one in Fig. 2, with the exception that instead of the weights $\mathbf{W}$ and biases $\mathbf{b}$, we have BNN parameters θ.

Inference We run inference on the entire model in an end-to-end manner. We use variational inference with the KL divergence [1, 9]. We implement all our models using Edward [7], a Python library for probabilistic programming that offers various inference choices including variational inference with KL divergence.

4 Experiments

We present our experimental results, comparing our proposed model with multiple baselines using the ATR dataset. We convert the whole dataset to numerical values and normalize all variables to fit in the range of $[0, 1]$.

Baselines We compare four variations of our proposed model against two baselines: mean data imputation, and two-stage model where inference is run in a sequential manner. In the case of mean imputation, the per-patient mean of observations that belong to the training set is imputed to all of their missing measurements. We then use the completed measurement data in the second component of the model. The second baseline is a two-stage version of our model. We first use the measurement imputation part of our method to impute missing data. Then, we use this completed predictor matrix for the outcome prediction in the second stage. Inference is run separately on each component.

Component 2 Input	BLR **P**	BLR **S**	BNN **P**	BNN **S**
Whole $\hat{\mathbf{P}}$ 2-stage mean imputation	0.338	0.236	0.338	0.219
Whole $\hat{\mathbf{P}}$ 2-stage	0.163	0.225	0.163	0.185
Patient traits $\hat{\mathbf{U}}$ 2-stage	0.163	0.210	0.163	0.184
Whole $\hat{\mathbf{P}}$ EE (proposed)	0.166	0.201	0.168	**0.156**
Patient traits $\hat{\mathbf{U}}$ EE (proposed)	0.180	**0.161**	0.166	0.172

Table 1. MAE for all models and baselines. 0.1 is the target MAE, which is the clinical resolution. "EE" indicates end-to-end, our proposed model. "2-stage" is the baseline model where data imputation and 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 **P** remains the same since it's computed once, using the mean of observations in the first case and the matrix factorization in the two others.

Results We split the training and testing set to reflect the treatment journey. In all of our experiments. For **P**, we randomly pick 80% of the available data for training and leave the rest for testing. For **S**, we randomly pick 80% of the patients, take all of their available scores for training and leave all scores of the remaining 20% of patients for testing, since the goal of our work is to predict the outcome using patients' incomplete measurements. We use grid search for hyper-parameter tuning and report the best result for each method in Table 1, using the Mean Absolute Error (MAE).

We can see that our proposed end-to-end model with Bayesian neural network applied on **P** achieves the best performance for the prediction of rehabilitation outcomes. Our proposed method shows clear improvement over the baselines. We can see that using latent variable models for data imputation gives better performance than the traditional mean imputation method. Additionally, it was established that a difference of 0.1 is significant in the case of ATR in clinical practice. Thus, our result is close to the ideal MAE target 0.1.

5 Conclusions

We developed a probabilistic framework to simultaneously predict the rehabilitation outcome and impute the missing entries in a clinical cohort in the context of Achilles Tendon Rupture rehabilitation. We demonstrated a clear improvement in the accuracy of the predicted outcomes in comparison with traditional data imputation methods. Using the proposed model, we obtained a mean absolute error of 0.156. This result is close to the target 0.1 which is the clinical difference step. We will thus continue to explore modeling choices to improve the outcome prediction accuracy. Additionally, the proposed method is a general framework that can be used in numerous health-care applications involving a long-term healing process after the treatment. In the future, we would also collaborate with more health-care departments to test and improve our method in these applications.

References

1. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American Statistical Association **112**(518), 859–877 (2017)
2. Domeij-Arverud, E., Anundsson, P., Hardell, E., Barreng, G., Edman, G., Latifi, A., Labruto, F., Ackermann, P.: Ageing, deep vein thrombosis and male gender predict poor outcome after acute achilles tendon rupture. Bone Joint J **98**(12), 1635–1641 (2016)
3. Huttunen, T.T., Kannus, P.A., Rolf, C.G., Felländer-Tsai, L., Mattila, V.M.: Acute achilles tendon ruptures: incidence of injury and surgery in sweden between 2001 and 2012. The American journal of sports medicine **42 10**, 2419–23 (2014)
4. Neal, R.M.: Bayesian learning for neural networks, vol. 118. Springer Science & Business Media (2012)
5. S Kearney, R., Achten, J., E Lamb, S., Parsons, N., L Costa, M.: The achilles tendon total rupture score: A study of responsiveness, internal consistency and convergent validity on patients with acute achilles tendon ruptures **10**, 24 (02 2012)
6. Stern, D., Herbrich, R., Graepel, T.: Matchbox: Large scale bayesian recommendations. In: Proceedings of the 18th International World Wide Web Conference (January 2009), https://www.microsoft.com/en-us/research/publication/matchbox-large-scale-bayesian-recommendations/
7. Tran, D., Kucukelbir, A., Dieng, A.B., Rudolph, M., Liang, D., Blei, D.M.: Edward: A library for probabilistic modeling, inference, and criticism (2016), http://arxiv.org/abs/1610.09787, cite arxiv:1610.09787
8. Valkering, K.P., Aufwerber, S., Ranuccio, F., Lunini, E., Edman, G., Ackermann, P.W.: Functional weight-bearing mobilization after achilles tendon rupture enhances early healing response: a single-blinded randomized controlled trial. Knee Surgery, Sports Traumatology, Arthroscopy **25**(6), 1807–1816 (2017)
9. Zhang, C., Butepage, J., Kjellstrom, H., Mandt, S.: Advances in variational inference. arXiv preprint arXiv:1711.05597 (2017)