Estimating Missing Data in Temporal Data Streams Using Multi-directional Recurrent Neural Networks

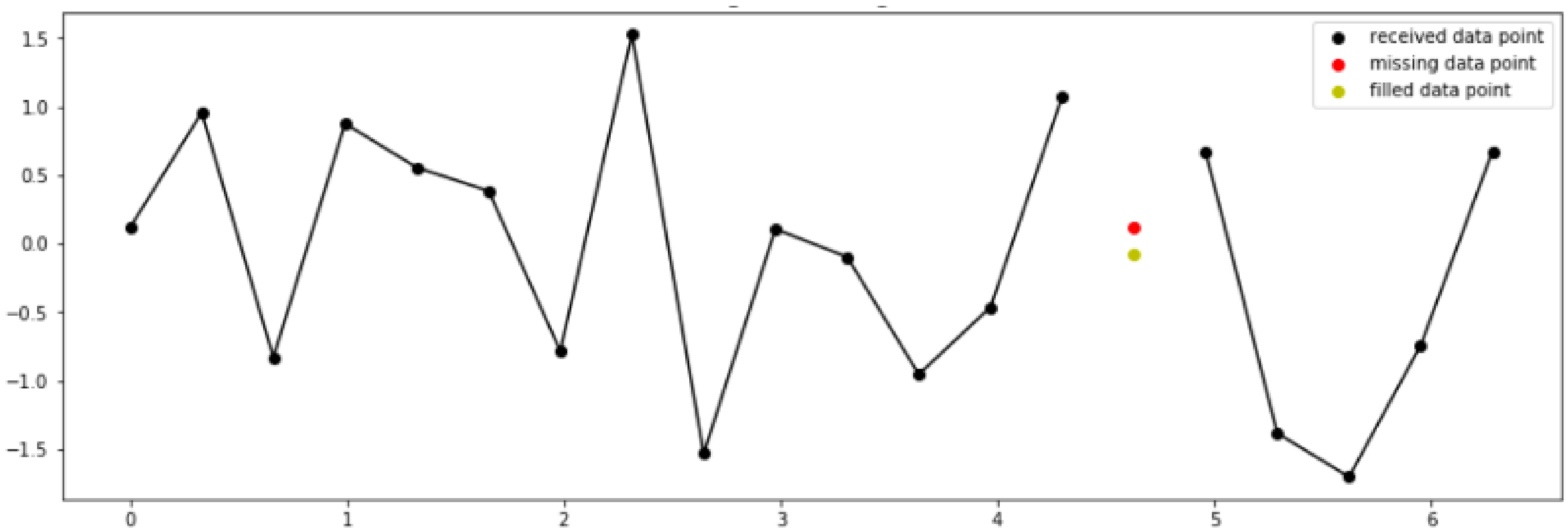
고려대학교 산업경영공학과
DAHS 연구실
김홍범

CONTENTS

Estimating missing value methods

- Interpolation – within each stream
- Imputation – across streams
- Matrix completion – within and across streams(ignore temporal aspect)

Interpolation method



- Advantages: Reflects the amount of data changes over time
- Disadvantage: Unable to determine association between variables

01 Methods

03. Imputation

Imputation method

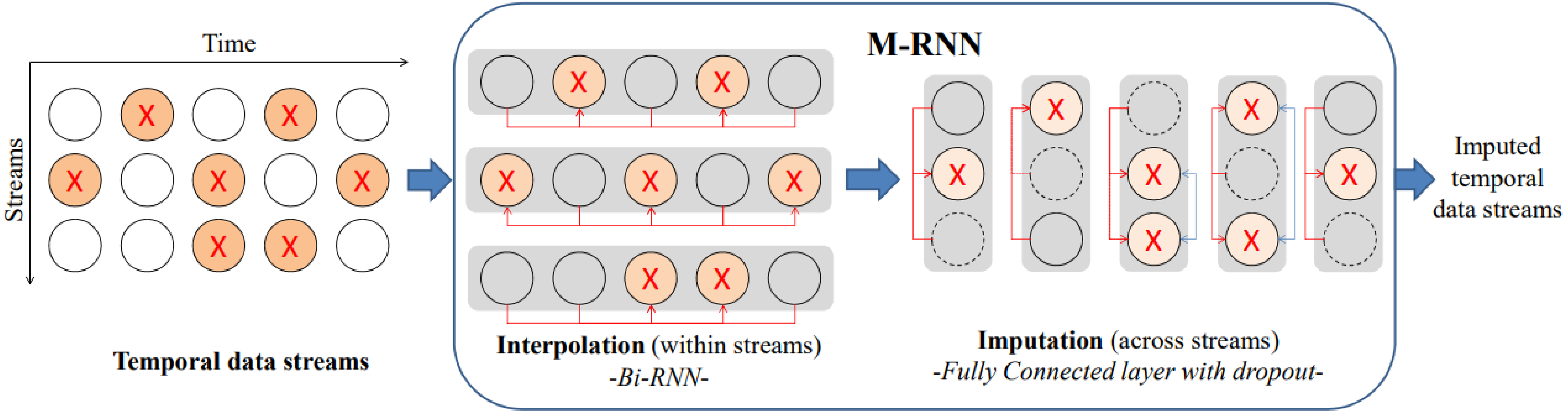
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean() →

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

- Advantages: Easy and fast
- Disadvantage: It is impossible to identify the time series association.
  Sensitive to outlier

M-RNN



- X = missing values, red lines = connections between observed values and missing values in each layer, blue lines = connections between interpolated values, dashed line = dropout

1. Notation

Datasets notations

- Dataset consists of N patients with D Channels and length T

$$X_n = \begin{bmatrix} 2 & 4 & 8 & * \\ 5 & * & 9 & 10 \end{bmatrix} \qquad D=2, T=4$$

$$x_t^d = * \qquad \text{missing value}$$

- Binary mask is defined to mask missing value (1 if data is observed, 0 if missing)

$$m = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

- Delta to be the actual amount of time that has elapse from $s_t(normalized)$, $\delta_1^d = 0, s_1 = 0$

$$\delta = \begin{bmatrix} 0 & 1 & 1.2 & 1 \\ 0 & 1 & 3 & 1.4 \end{bmatrix}, \qquad \delta_t^d = \begin{cases} s_t - s_{t-1} + \delta_{t-1}^d & \text{if } t > 1, m_{t-1}^d = 0. \\ s_t - s_{t-1} & \text{if } t > 1, m_{t-1}^d = 1 \end{cases}$$

Objective function

- Also y_t represents vector of outcomes for this patients. Such as discharge, death etc. y_t =0, 1

- The entire dataset consists of all above the patients $D = \{(S(n), X(n), y(n)\}$
- (Time stamps = S, measurements = X, lables = Y)

- **Objective function** ($\hat{x}_t^d = f_t^d(S, X)$, $\mathcal{L}(\hat{x}_t^d, x_t^d) = (\hat{x}_t^d - x_t^d)^2$)

$$\min_{\mathbf{f}} \mathbb{E}_{\mathcal{F}}\left[\sum_{t=1}^{T}\sum_{d=1}^{D}(1 - m_t^d)\mathcal{L}(\hat{x}_t^d, x_t^d)\right]$$

$$= \min_{\mathbf{f}} \mathbb{E}_{\mathcal{F}}\left[\sum_{t=1}^{T}\sum_{d=1}^{D}(1 - m_t^d)(f_t^d(\mathcal{S}, \mathcal{X}, \mathcal{Y}) - x_t^d)^2\right].$$

- We do not observe the true data, so we will minimize the empirical loss.

Error/Loss

- $\boldsymbol{LOSS(total) = mean\ squared\ error\ (mse) =}$

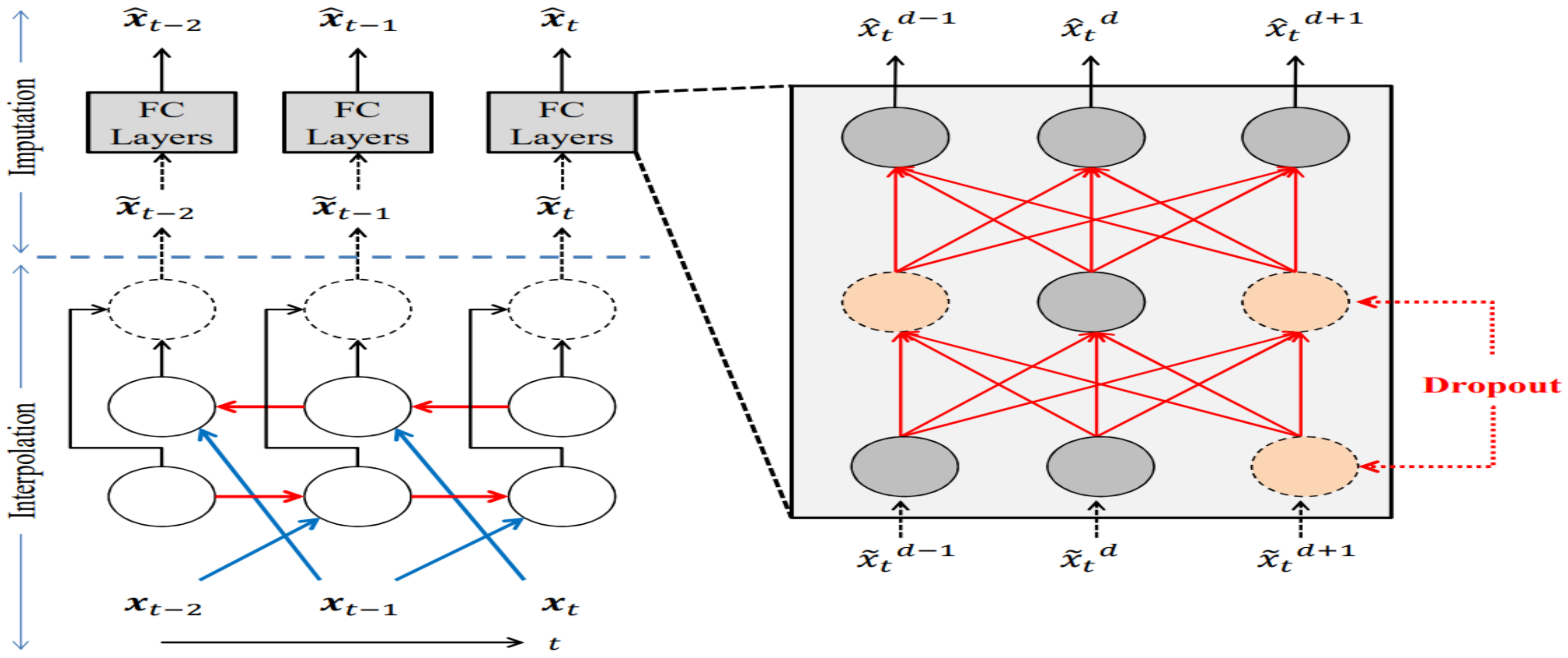
$$\mathcal{L}(\{\hat{x}_t^d, x_t^d\}) = \sum_{n=1}^{N} \left[\frac{\sum_{t=1}^{T_n} \sum_{d=1}^{D} m_t^d(n) \times (\hat{x}_t^d(n) - x_t^d(n))^2}{\sum_{t=1}^{T_n} \sum_{d=1}^{D} m_t^d(n)} \right]$$

- If we have missing data in x_t, then we have two options.
  1. if x_t is missing, use x_{t-n} to compute $\hat{x}_t$.
  2. Use a simple Linear interpolation(backward or forward fill for first and last value) - > authors use this implementation.

- Note : this is the empirical error, which is actually achievable.

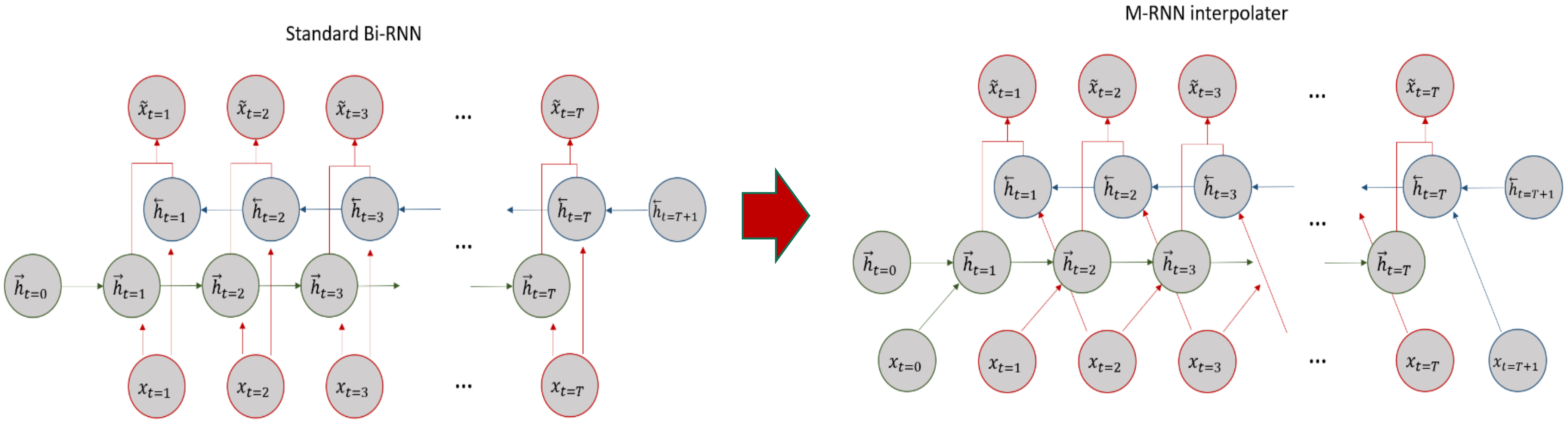
2. M-RNN

M-RNN Architecture



2. M-RNN

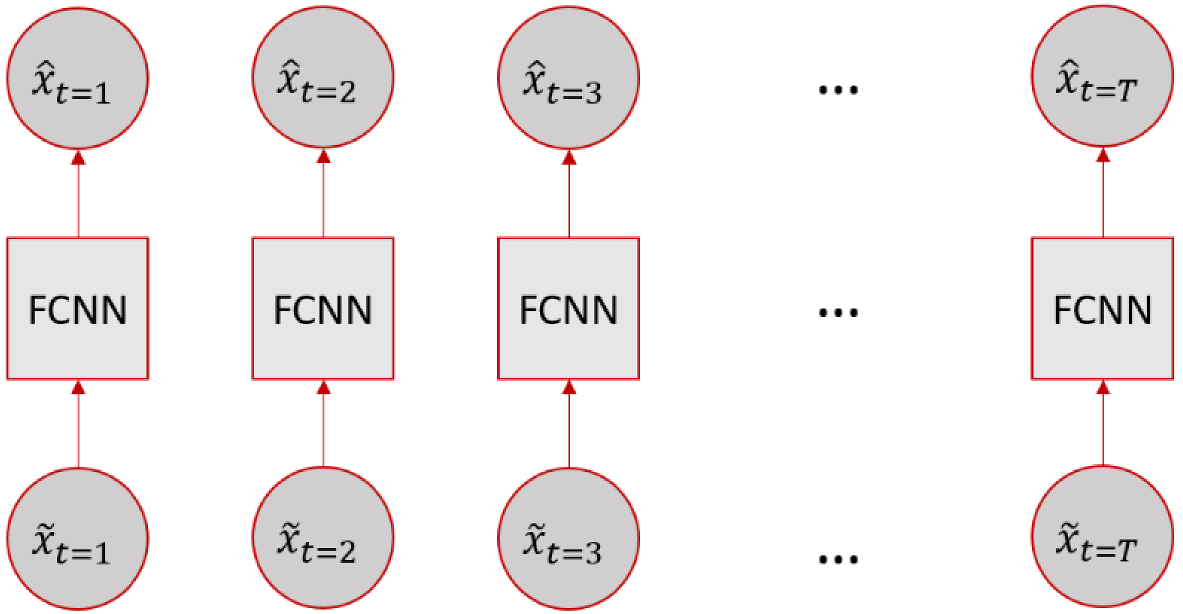
M-RNN Interpolator



- This procedure ensure that the actual value x_t^d is automatically not used in the estimation $\tilde{x}_t^d$.

2. M-RNN

M-RNN imputer



- Note : Always the same FCNN for each timestap, also we do not use x_t^d in every step.
- In this process, we use dropout process for multiple imputation.

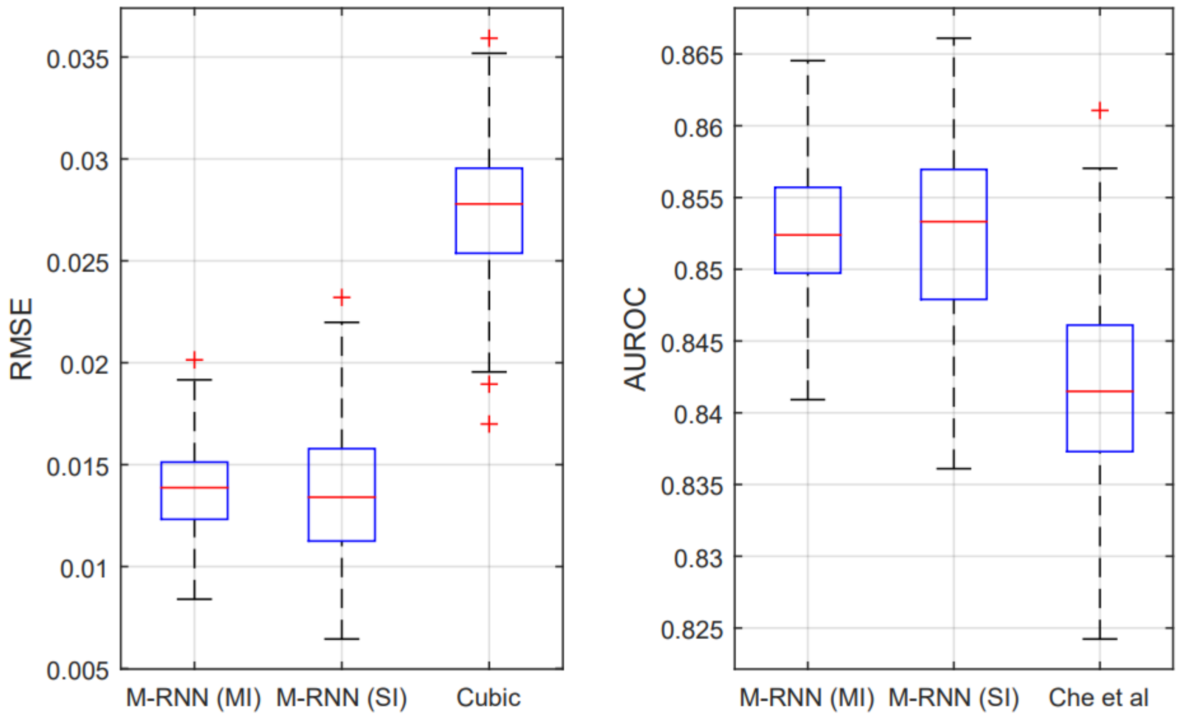
Imputation Accuracy on the Given Datasets

Table 2: Performance comparison for missing data estimation

Category	**Algorithm**	**Mean RMSE (% Gain of M-RNN (Multiple Imputations))**				
		MIMIC-III	Deterioration	UNOS-Heart	UNOS-Lung	Biobank
M-RNN	**M-RNN (MI)**	**0.0141 (-)**	**0.0105 (-)**	**0.0479 (-)**	**0.0606 (-)**	**0.0637 (-)**
	M-RNN (SI)	**0.0144 (-)**	**0.0108 (-)**	**0.0477 (-)**	**0.0609 (-)**	**0.0629 (-)**
RNN-based	[23]	0.0337 (58.2%)	0.0258 (59.3%)	0.1352 (64.6%)	0.1343 (54.9%)	0.0812 (21.6%)
	[24]	0.0295 (52.2%)	0.0241 (56.4%)	0.1179 (59.4%)	0.1264 (52.1%)	0.0801 (20.5%)
	[25]	0.0292 (51.7%)	0.0233 (54.9%)	0.1057 (54.7%)	0.1172 (48.3%)	0.0778 (18.1%)
Interpolation	Spline	0.0735 (80.8%)	0.0215 (51.2%)	0.1102 (56.5%)	0.1199 (49.5%)	0.0845 (24.6%)
	Cubic	0.0279 (49.5%)	0.0223 (52.9%)	0.1072 (55.3%)	0.1177 (48.5%)	0.0887 (28.2%)
Imputation	MICE	0.0611 (76.9%)	0.0319 (67.1%)	0.1147 (58.2%)	0.1151 (47.4%)	0.0915 (30.4%)
	MissForest	0.0293 (51.9%)	0.0264 (60.2%)	0.0489 (2.0%)	0.0652 (7.1%)	0.0892 (28.6%)
	EM	0.0467 (69.8%)	0.0355 (70.4%)	-	-	0.0978 (34.9%)
Others	Matrix Completion	0.0311 (54.7%)	0.0264 (60.2%)	0.0974 (50.8%)	0.0942 (35.7%)	0.0886 (28.1%)
	Auto-encoder	0.0412 (66.0%)	0.0309 (65.0%)	0.0589 (18.7%)	0.0712 (14.9%)	0.0805 (20.9%)
	MCMC	0.0437 (67.7%)	0.0364 (71.2%)	0.1091 (56.1%)	0.1124 (46.1%)	0.0936 (31.9%)

1. Results

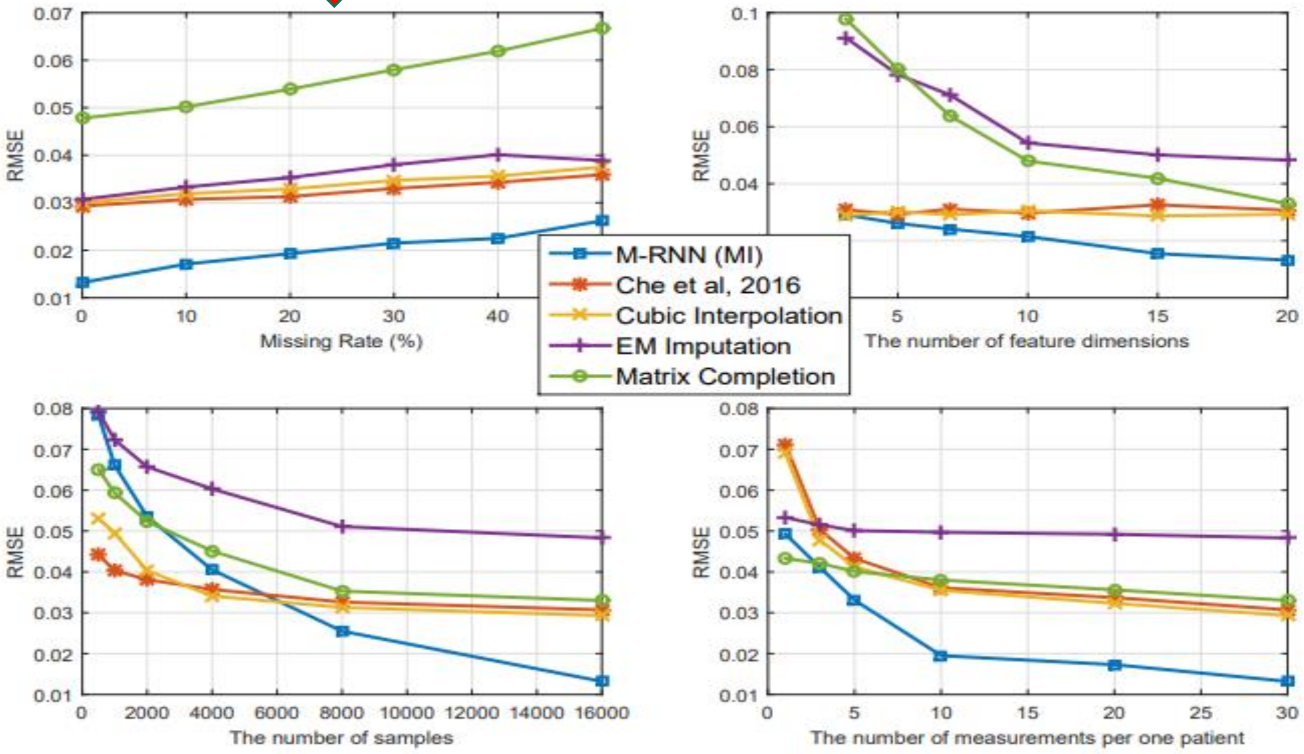
Combining models of interpolation and imputation



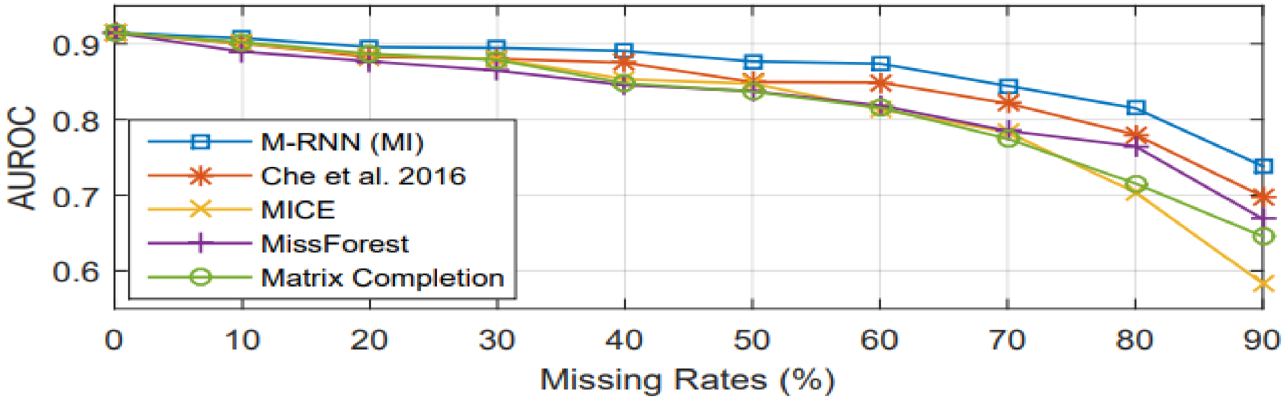
- The purpose of conducting multiple imputation is to reduce uncertainty/shrink confidence intervals(rather than to improve performance)

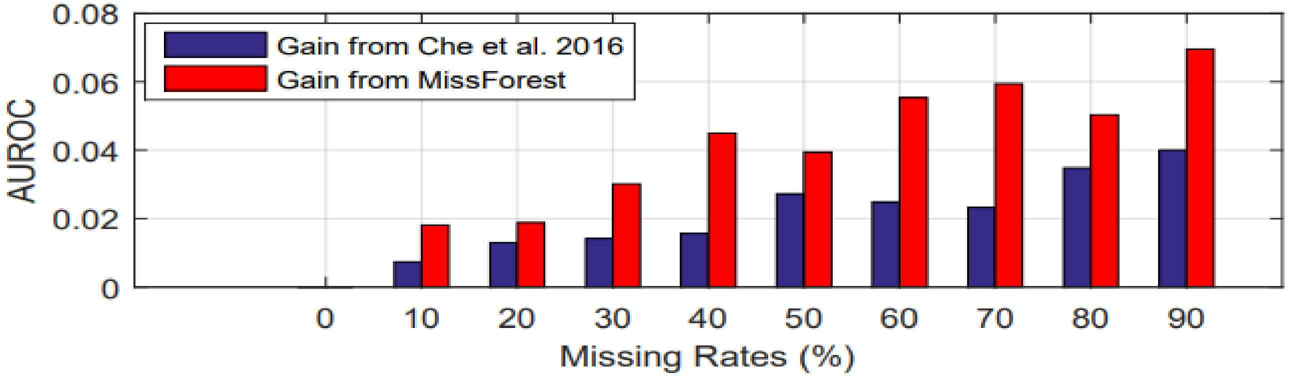
Source of Gain of M-RNN

Our dataset(25%)



Prediction accuracy on the original datasets





Congeniality of the model

Table 6: Congeniality of imputation models

Algorithm	Mean Bias ($\|\|\mathbf{w} - \hat{\mathbf{w}}\|\|_1$)	Root Mean Square Error ($\|\|\mathbf{w} - \hat{\mathbf{w}}\|\|_2$)
M-RNN (MI)	0.0814 ± 0.0098	0.1229 ± 0.0151
[25]	0.1097 ± 0.0104	0.1649 ± 0.0212
Cubic Interpolation	0.1169 ± 0.01075	0.1816 ± 0.0201
MissForest	0.0842 ± 0.0103	0.1312 ± 0.0139
Matrix Completion	0.1001 ± 0.0125	0.1551 ± 0.0230

- Congeniality of an imputation model can be evaluated by specified metric.(mean bias, RMSE)
- For comparison, we delete 20% of the data.
- Result shows our methods is mor congenial than the benchmarks.

FIRM DATASET

- **Percentage of missing value - 20%**

- **Dataset (132 rows, 14 feature)**

PROMs(patients report outcome measurements :

FAC(보행/이동성), KOVAL(이동성), EQ5D(삶의 질), IADL(일상생활수행능력), FRAIL(노쇠 지수)

PBOMS(performance based outcome measurements) :

FIM(보행), MRMI(운동성), MBI(일상생활수행능력), BBS(낙상위험도), GDS(정동), MMSE(인지기능), HGS_R, HGS_L (악력)

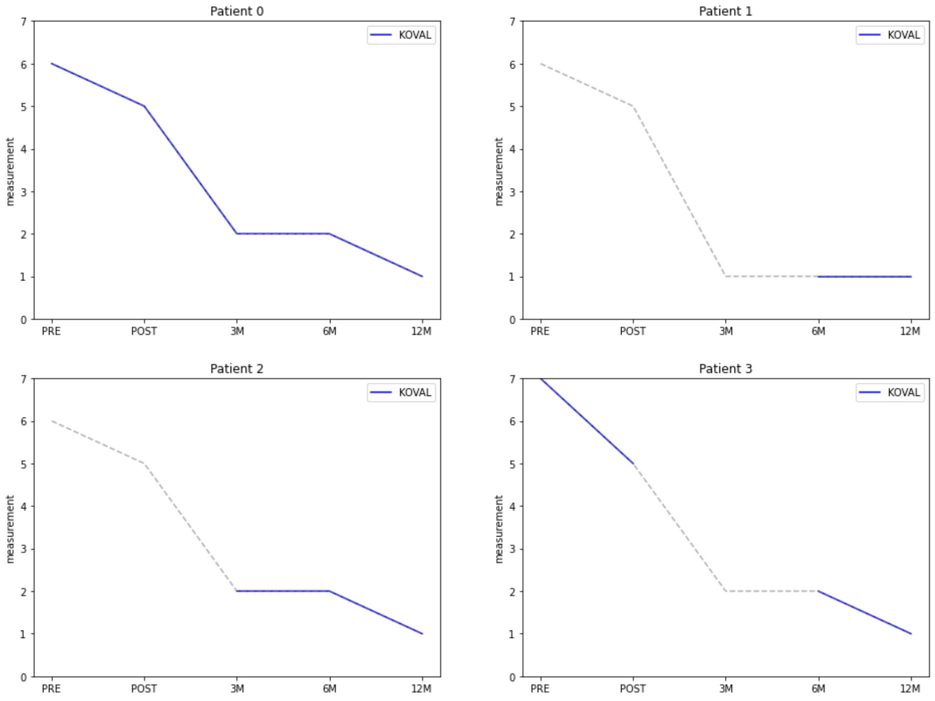
FIRM DATASET

Congeniality of the FIRM data

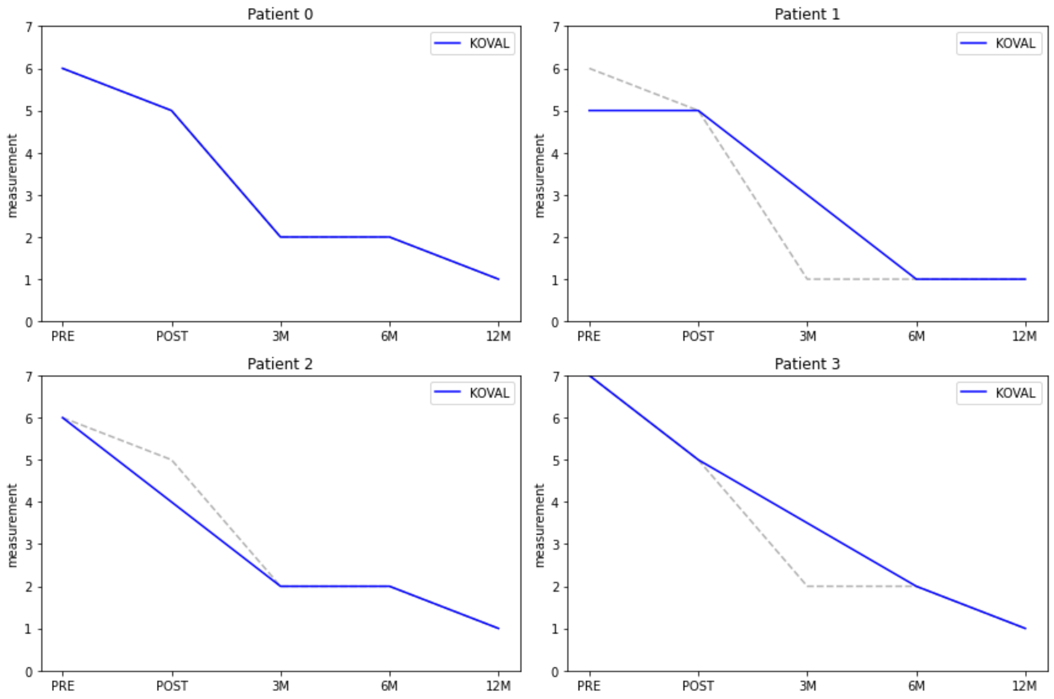
1. 23 patients with no missing values were randomly generated (20%)
2. Checked the difference between the generated data and the actual value through the model M-RNN
3. After that, we checked the performance of 132 people.

Congeniality of the FIRM data (KOVAL)

Before

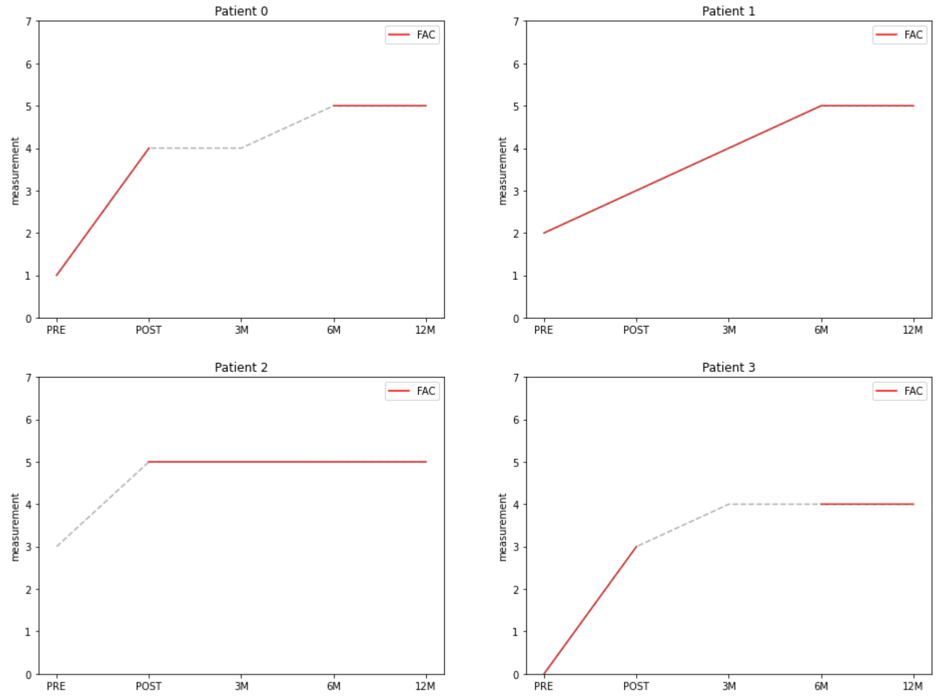


After

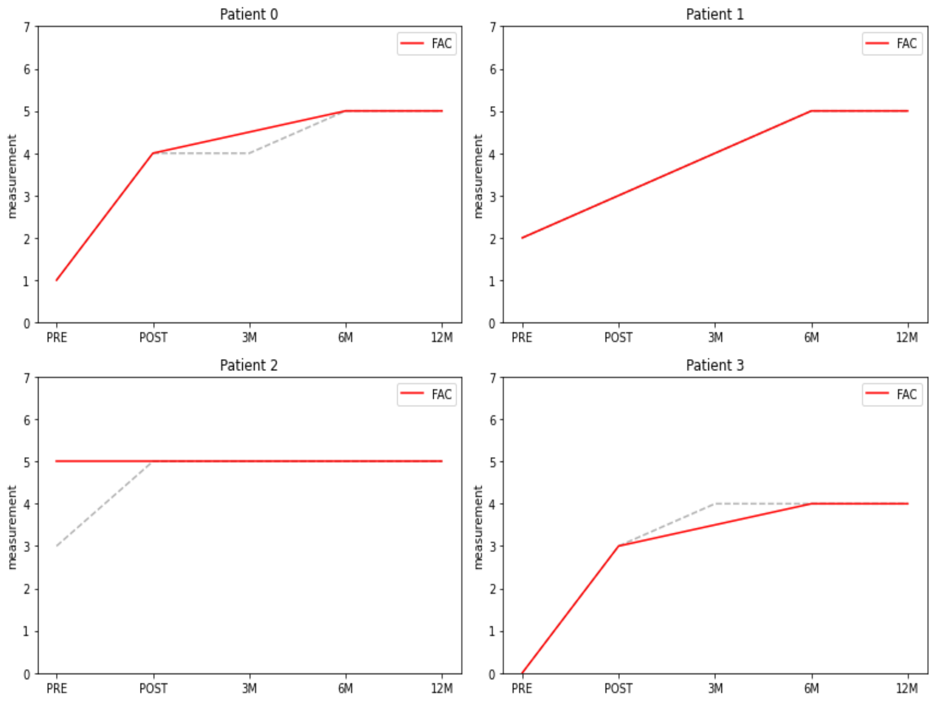


Congeniality of the FIRM data (FAC)

Before



After

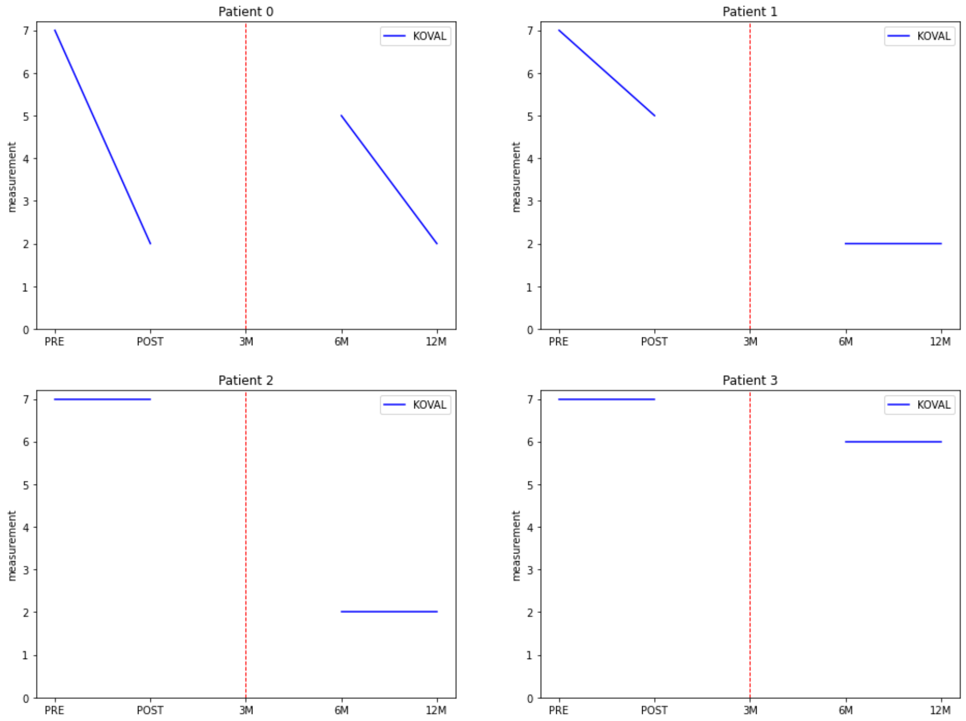


04 Results and Discussions

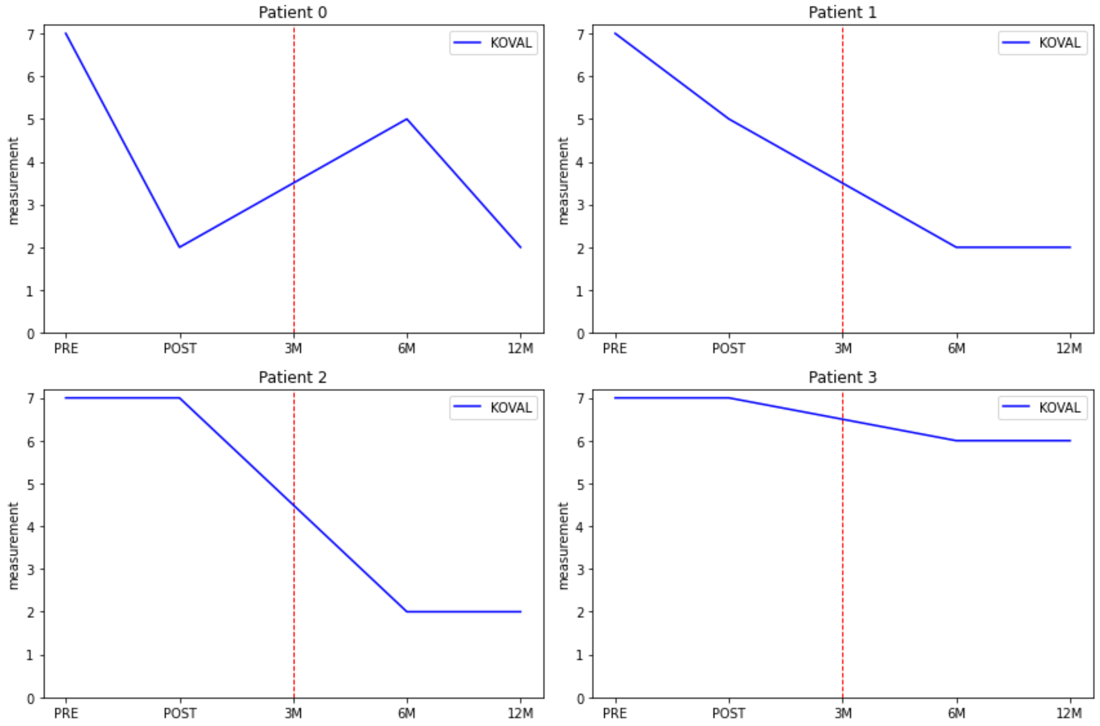
3. FIRM DATA

FRIM DATA(Using M-RNN)

Before

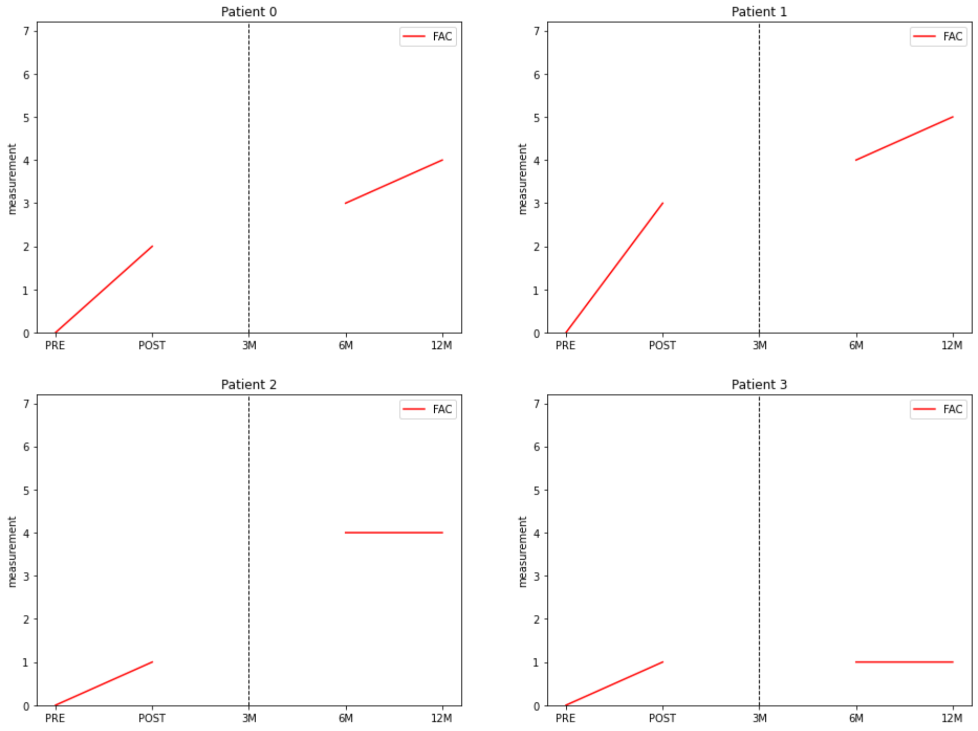


After

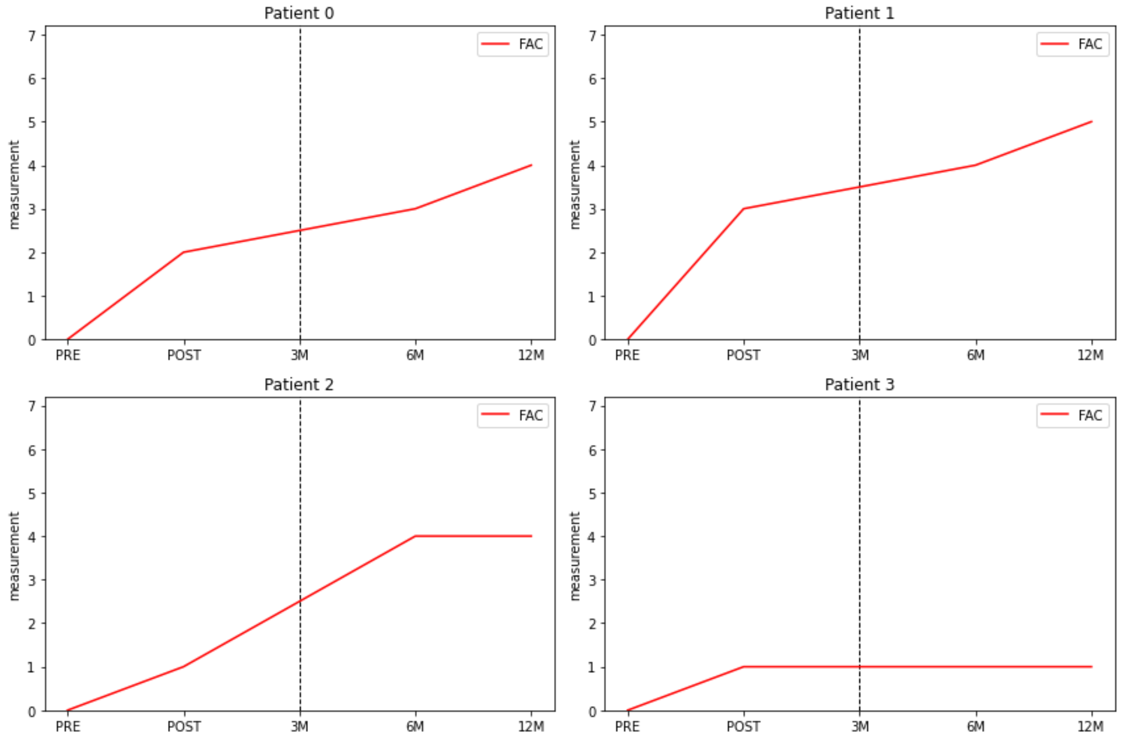


FRIM DATA(Using M-RNN)

Before



After

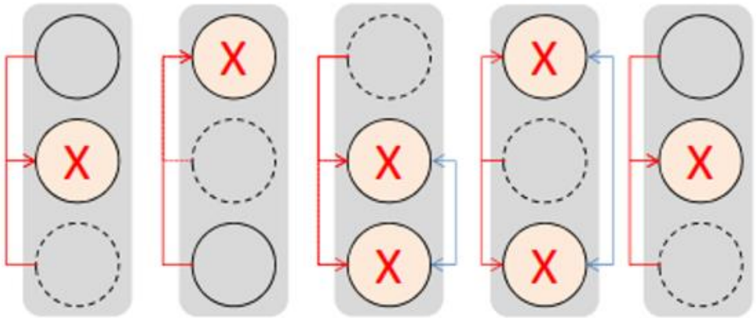


Additional study

Increase data accuracy by adding patient clinical characteristics (During Imputation Procedure)

Clinical Characteristics

NAME	age	height	…	edu	Blood test
DATA TYPE	Continuous	Continuous	…	Nominal	Continuous
Ex)	80, 79	143, 160	…	1,2,3, 4,	1.12, 1.13



Imputation (across streams)

3. References

- https://ieeexplore.ieee.org/abstract/document/8485748
- https://www.kaggle.com/code/markwallbang/m-rnn-estimate-missing-values-in-time-series/notebook
- https://github.com/jsyoon0823/MRNN

Q&A

감사합니다

billkim418@naver.com