CS598 DLH Reproducibility Project - Final Report (Spring 2022)

Shirley Li and Stanley Ho
{qiuyuli2, smho2}@illinois.edu

Group ID: 22, Paper ID: 189
Presentation link: https://uofi.box.com/s/zrpprtxedxl0mi4kshwdu1webz1x7d56
Code link: https://github.com/shirleyli2015/CS-598-DLH-Final-Project

1 Introduction

Our reproduction study is based on this paper:

Barbieri, Sebastiano, et al. "Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk." Scientific reports 10.1 (2020): 1-10. https://www.nature.com/articles/s41598-020-58053-z

The general problem which the original paper wanted to tackle was to predict the risk of patient's readmission within 30 days of discharge from the ICU, as readmission to the ICU is costly in the healthcare system and is mostly avoidable. The original paper attempted to identify the best deep neural architecture for addressing this problem, by comparing variations of architectures for predicting patient's ICU readmission based on time-series data, and utilizes attention layers for more interpretable feature detection.

2 Scope of reproducibility

There were three claims in the original paper we choose to verify:

1. *Claim 1*: Prediction accuracy of deep neural networks based on *ODEs* (ordinary differential equations) for modeling the predictive relevance of medical codes over time is considerably higher than for the logistic regression baseline model.
2. *Claim 2*: Applying attention layers to deep neural networks does not degrade average precision but can improve model interpretability.
3. *Claim 3*: Models with a recurrent component perform slightly better than models based solely on attention layers.

These claims were not only central contributions of the original paper, but their scopes were also specific that we could either support or reject based on our reproduced results.

3 Methodology

The original authors implemented and compared 14 deep neural network architectures for predicting patient's ICU readmission. Our approach was to use the existing code from the original authors as starting point to reproduce and compare results of all the models. We used the MIMIC-III data, and pre-processed it the same way as the original paper. We then trained and evaluated our models on a *2022 MacBook Pro* with *M1 Max* processor with 64 GB memory with 80 epochs for each model.

3.1 Model descriptions

The original paper described 14 deep neural network architectures for predicting patient's readmission to the ICU. The high level structures in all architectures were similar: 1) computed embeddings based on timestamped code, 2) computed scores based on procedure/diagnosis/vital sign/medication codes using attention and/or recurrent layers, 3) concatenated the scores with static variables and passed to the logistic regression layer. These architectures had different variations of attention mechanism, recurrent layers, neural ordinary differential equations (*ODEs*), and medical concept embeddings (*MCEs*) with time-aware attention:

- **ODE + RNN + Attention**: Use neural *ODEs* to model dynamics in time of embeddings, embeddings are then passed to *RNN* layers, with attention applied to the *RNN* outputs.
- **ODE + RNN**: Use neural *ODEs* to model dynamics in time of embeddings, embeddings are then passed to *RNN* layers.
- **RNN (ODE time decay) + Attention**: Use neural *ODEs* to model embeddings which are

passed to *RNN* layers with dynamics in time, with attention applied to *RNN* outputs.

- **RNN (ODE time decay)**: Use neural *ODEs* to model embeddings which are passed to *RNN* layers with dynamics in time.
- **RNN (exp time decay) + Attention**: Embeddings are passed to *RNN* layers with dynamics in time which decays exponentially over time, with attention applied to *RNN* outputs.
- **RNN (exp time decay)**: Embeddings are passed to *RNN* layers with dynamics in time which decays exponentially over time.
- **RNN (concatenated time delta) + Attention**: Time differences between observations are concatenated with embeddings which are passed to *RNN* layers, with attention applied to *RNN* outputs.
- **RNN (expconcatenated time delta)**: Time differences between observations are concatenated with embeddings which are passed to *RNN* layers.
- **ODE + Attention**: Use neural *ODEs* to model dynamics in time of embeddings, with attention applied to embeddings.
- **Attention (concatenated time)**: Elapsed times are concatenated with embeddings, with attention applied to the embeddings.
- **MCE + RNN + Attention**: Compute *MCE* (Medical concept embeddings) and pass to the *RNN* layers, with attention applied to *RNN* outputs.
- **MCE + RNN**: Compute *MCE* and pass to *RNN* layers.
- **MCE + Attention**: Compute *MCE*, with attention applied to the embeddings.
- **Logistic Regression**: Use logistic regression as a baseline model for comparison against neural network models.

3.2 Data descriptions

The dataset is the MIMIC-III data which is available at *PhysioNet* upon request. It contains patients health data at Beth Israel Deaconess Medical Center in Boston, with 61,532 ICU stays as well as 46,476 critical care stays, between year 2001 and 2012. The patients health data contains no PII (Personal Identifiable Information) for privacy reason.

3.3 Hyperparameters

There are several hyperparameters which are configurable in *hyperparameters.py* in the code:

Name	Description	Default
batch_size	Batch size in training	128
num_epochs	Number of epochs	80
dropout_rate	Dropout rate	0.5
patience	# of patience	10
sigma1	Sigma#1 in BN	e-0
sigma2	Sigma#2 in BN	e-6

We used the same original hyperparameters' values in this reproduction study.

3.4 Implementation

Our implementation is located at https://github.com/shirleyli2015/CS-598-DLH-Final-Project, and it was based on the code from the original authors at https://github.com/sebbarb/time_aware_attention. Generally, we kept our code modifications to a minimum to avoid unexpected side effect, and only modified the code out of necessity. Our implementation also includes two ablations as described in Section 4.3, and the code is located at https://github.com/shirleyli2015/CS-598-DLH-Final-Project/blob/master/related_code/modules.py#L1171.

3.5 Computational requirements

Each of the 14 models in the original paper took 80 epochs to train. Our original computational estimate for each model expected less than 24 hours to train on a modern system with 64 GB memory, equivalent to no more than 18 mins per epoch. We also confirmed with the original authors that our computational estimate was indeed realistic.

To reproduce the results, we used a *2022 Apple MacBook Pro* with *M1 Max* processor with 64GB memory to train all 14 models each with 80 epochs. The actual training time of each model ranged from 1.5 to 20 hours, equivalent to 1 to 15 mins per epoch. The actual computation requirements were well within our original estimation.

4 Results

In total, the models were trained with 25 static variables, 992 unique ICD-9 diagnosis codes, 298 unique ICD-9 procedure codes, 586 unique medications, and 32 chart codes related to vital signs. We

were able to reproduce the models within the same 95th percentile for average precision, meaning that they had comparable performance in predicting true positives to the models in the original paper (Table 1 in Section 4.1).

As described in Section 2, there were three claims in the original paper we aimed to verify. Generally, our reproduced results supported these claims, and we observed similar model performance in our results.

4.1 Result 1

We reproduced the results of all 14 models with average precision as follows:

	Average Precision	
Model	Reproduced	Original
ODE + RNN + Attention	0.325 [0.317,0.334]	0.314 [0.306,0.321]
ODE + RNN	0.319 [0.311,0.328]	0.331 [0.323,0.339]
RNN (ODE time decay) + Attention	0.321 [0.312,0.33]	0.316 [0.307,0.324]
RNN (ODE time decay)	0.308 [0.301,0.315]	0.300 [0.293,0.308]
RNN (exp time decay) + Attention	0.309 [0.302,0.317]	0.320 [0.312,0.328]
RNN (exp time decay)	0.311 [0.304,0.319]	0.304 [0.297,0.311]
RNN (concat time delta) + Attention	0.307 [0.299,0.316]	0.312 [0.303,0.320]
RNN (concat time delta)	0.314 [0.306,0.323]	0.311 [0.303,0.320]
ODE + Attention	0.292 [0.284,0.3]	0.294 [0.285,0.302]
Attention (concat time)	0.287 [0.278,0.295]	0.286 [0.277,0.295]
MCE + RNN + Attention	0.304 [0.296,0.313]	0.317 [0.308,0.325]
MCE + RNN	0.297 [0.29,0.305]	0.298 [0.291,0.306]
MCE + Attention	0.282 [0.273,0.291]	0.269 [0.261,0.278]
Logistic Regression	0.257 [0.248,0.266]	0.257 [0.248,0.266]

Table 1: Average precision of the evaluated deep neural architectures.

Of the 5 models with *ODE*, all performed significantly better than the logistic regression model which only used the most recent patient chart information as covariates.The logistic regression model achieved an average precision of 0.257 and the *ODE* models achieved average precisions between 0.292 to 0.325 (Table 1). Our results supported *Claim #1* that neural *ODEs* can model the predictive relevance of medical codes better than baseline logistic regression models.

We also observed that adding an attention layer on top of *RNN* models did not decrease average precision. In our reproduced results in comparing attention models with their non-attention counterparts, the average precision of the *ODE + RNN + Attention* model was 0.6% higher, the *RNN (ODE time decay) + Attention* model was 1.3% higher, the *RNN (exp time decay) + Attention* was 0.2% lower, the *RNN (concat time delta) + Attention* was 0.7% lower, and the *MCE + RNN + Attention* model was 0.7% higher than the average precision of their counterparts without attention layers. The highest average precision of 0.325 was obtained by the *ODE + RNN + Attention* model, which is contrary to the paper's original highest precision of 0.331 obtained by the *ODE + RNN* model, although both models have overlapping 95% confidence intervals for true precision. In fact, all models with *RNN* and attention layers have overlapping 95% confidence intervals with the corresponding models without attention layers, demonstrating that our results supported *Claim #2* that additional attention layer had marginal impact to prediction precision, but having an attention layer itself could improve interpretability.

In addition, we observed that adding a recurrent component to the model significantly improved model performance. Models with a recurrent component had average precision ranging from 0.297-0.325, and generally performed better than models based only on attention layers, which had average precision ranging from 0.282-0.292, which is also consistent with the findings of the paper. Our results supported *Claim #3* that models with a recurrent component performed slightly better than models based solely on attention layers.

4.2 Result 2

We generated the odds ratios from the last fully connected layer of the *Attention concatenated time* model (Table 2) as described in the original paper:

	OR [95% CI]	
Covariate	Reproduced	Original
ICU Length of Stay (days)	0.998 [0.996, 1]	1.000 [0.998, 1.002]
Gender: Male	1.078 [1.054, 1.103]	1.114 [1.092, 1.136]*
Number of Recent Admissions	1.123 [1.106, 1.141]	1.187 [1.170, 1.205]*
Age (years)	1.005 [1.005, 1.005]	1.009 [1.009, 1.010]*
Pre-ICU Length of Stay (days)	1.006 [1.003, 1.009]	0.994 [0.993, 0.996]*
Elective Surgery	0.892 [0.852, 0.934]	0.941 [0.891, 0.993]*
Admission Location: Clinic Referral/Premature Delivery	0.943 [0.904, 0.983]	0.998 [0.992, 1.003]
Admission Location: Other/Unknown	1 [0.995, 1.005]	1.639 [1.146, 2.345]*
Admission Location: Physician Referral/Normal Delivery	0.927 [0.893, 0.962]	0.882 [0.844, 0.922]*
Admission Location: Transfer from Hospital/Extramural	1.154 [1.111, 1.198]	1.115 [1.074, 1.157]*
Admission Location: Transfer from Skilled Nursing Facility	1.216 [0.962, 1.537]	1.001 [0.996, 1.006]
Insurance: Government	0.761 [0.692, 0.837]	0.775 [0.694, 0.865]*
Insurance: Medicaid	1.002 [0.996, 1.009]	0.997 [0.992, 1.002]
Insurance: Private	0.854 [0.825, 0.883]	0.820 [0.798, 0.843]*
Insurance: Self Pay	0.464 [0.384, 0.561]	0.559 [0.447, 0.700]*
Marital Status: Other/Unknown	0.921 [0.855, 0.993]	0.918 [0.845, 0.997]*
Marital Status: Single	1 [0.995, 1.006]	1.000 [0.995, 1.005]
Marital Status: Widowed/Divorced/Separated	0.974 [0.938, 1.012]	0.996 [0.991, 1.001]
Ethnicity: Asian	0.875 [0.783, 0.978]	0.772 [0.694, 0.858]*
Ethnicity: Black/African American	1 [0.995, 1.005]	1.165 [1.118, 1.215]*
Ethnicity: Hispanic/Latino	1.001 [0.996, 1.007]	1.001 [0.996, 1.006]
Ethnicity: Other/Unknown	0.913 [0.873, 0.955]	0.873 [0.832, 0.916]*
Ethnicity: Unable to Obtain	0.852 [0.742, 0.978]	1.000 [0.995, 1.004]
Score: Diagnoses and Procedures	1.73 [1.715, 1.745]	3.780 [3.663, 3.902]*
Score: Medications and Vital Signs	1.8 [1.744, 1.858]	2.044 [1.979, 2.110]*

Table 2: Odds ratios from the last connected layer from the *Attention (concatenated time)* model

Length of stay in ICU was in our results associated with a slightly lower OR of readmission in 30 days (OR 0.998, 95% credible interval of [0.996,

0.9997], but was very close to the original authors' results of OR of 1. We found that a longer stay before admission to ICU had a slightly higher risk of experiencing readmission (OR 1.006 credible interval [1.003,1.009] as opposed to the original authors' finding of a slightly protective effect of 0.994 (95% credible interval of [0.993, 0.996]). We also identified number of recent hospital admissions, higher age, and male gender to have higher odds of readmission (OR: 1.123 [1.106, 1.141], 1.005 [1.005, 1.005], and 1.006 [1.003, 1.009], respectively) which is in agreement with the original authors' findings (OR: 1.114 [1.092, 1.136], 1.187 [1.170, 1.205], and 1.009 [1.009, 1.010]). Admission for elective surgery were found to have lower odds for readmission (OR 0.8920386, credible interval [0.85191107, 0.9340564]), in agreement with the original authors (OR: 0.941 [0.891, 0.993]). Patients with admission via referral or normal delivery had lower odds of readmission compared to those transferred from hospitals or nursing facilities. Patients insured through government, private, or self-pay had lower odds of readmission than those insured with Medicaid. Marital status was not associated with difference in odds. Notably, we did not find that Black/African American patients had a higher odds of experiencing readmission (0.99979645 [0.9948968, 1.0047201]) whereas the original authors did (OR: 1.165 [1.118, 1.215]). Overall, our generated odds ratios were similar to the original authors' for the static variables, with 18/25 of the variables having overlapping 95th percentile credible intervals.

Scores for diagnosis, procedure, and mediation codes for readmission are also reproduced from the *Attention concatenated time* model (Table 3). In total 5/10 of the medications, 2/10 of the procedures, and 0/10 ICD-9 diagnoses of the top 10 scores in each category of our reproduced results were also in the top 10 for each category in the original paper.

In summary, while we were able to reproduce scores from the *Attention concatenated time* model, our top scores were only comparable to the original paper's for prescriptions but not for diagnoses and procedures. Since the relevancy of these scores were not covered by the three claims we aimed the verify, this discrepancy neither supported nor rejected the claims.

Reproduced	
ICD-9 Diagnoses	Score [95% CI]
Goiter, unspecified	5 [4, 5.9]
Foreign body in larynx	4.2 [2.8, 5.7]
Acute and subacute bacterial endocarditis	4.1 [3.6, 4.5]
Disruption of internal operation (surgical) wound	3.9 [3.3, 4.5]
Acquired coagulation factor deficiency	3.8 [2.9, 4.8]
Other specified disorders of biliary tract	3.7 [3, 4.4]
Acute and chronic respiratory failure	3.7 [3.1, 4.3]
Subdural hemorrhage	3.5 [2.9, 4.1]
Hemorrhage, unspecified	3.1 [2.2, 4.1]
Other pulmonary insufficiency, not elsewhere classified	3 [2.5, 3.5]
ICD-9 Procedures	Score [95% CI]
Percutaneous abdominal drainage	4.3 [4, 4.5]
Pericardiotomy	4.1 [3.3, 5]
Temporary tracheostomy	4 [3.8, 4.3]
Cardiopulmonary resuscitation, not otherwise specified	3.9 [3.2, 4.6]
Thoracentesis	3.5 [3.2, 3.9]
Ventricular shunt to abdominal cavity and organs	3.4 [1.6, 5]
Therapeutic plasmapheresis	3.2 [2.3, 4.2]
Enteral infusion of concentrated nutritional substances	3.1 [2.9, 3.3]
Extraction of other tooth	3.1 [2.1, 4.1]
Continuous invasive mechanical ventilation for 96 consecutive hours or more	3 [2.7, 3.3]
Medications	Score [95% CI]
heparinsodium	4.5 [2.9, 6.3]
d5w	4.1 [3, 5.2]
bag	3 [1.4, 4.6]
albuterol0.083%nebsoln	2.1 [0.8, 3.6]
ondansetron	1.9 [-0.2, 4]
furosemide	1.8 [1.2, 2.6]
acetylcysteine20%	1.6 [0.1, 3.2]
vial	1.5 [0.3, 2.7]
lorazepam	1.4 [0.7, 2.2]
Original	
ICD-9 Diagnoses	Score [95% CI]
Infection and inflammatory reaction due to cardiac device, implant, and graft	7.5 [4.8, 10.3]
Other and unspecified infection due to central venous catheter	6.9 [4.7, 9.1]
Need for desensitization to allergens	6.5 [5.3, 7.6]
Hepatorenal syndrome	6.2 [4.5, 7.9]
Diabetes with renal manifestations, type I [juvenile type], uncontrolled	5.8 [3.9, 7.7]
Hydantoin derivatives causing adverse effects in therapeutic use	5.4 [3.1, 7.9]
Encounter for palliative care	5.4 [3.4, 7.4]
Dysphagia, oropharyngeal phase	5.3 [3.0, 7.7]
Spontaneous bacterial peritonitis	5.2 [2.7, 8.2]
Other sequelae of chronic liver disease	5.0 [2.4, 7.4]
ICD-9 Procedures	Score [95% CI]
Other gastrostomy	6.9 [5.0, 9.0]
Therapeutic plasmapheresis	6.1 [4.8, 7.4]
Incision of abdominal wall	5.6 [2.7, 8.4]
Transcatheter embolization for gastric or duodenal bleeding	4.9 [3.0, 7.0]
Transfusion of coagulation factors	4.8 [3.0, 6.7]
Graft of muscle or fascia	4.4 [2.4, 6.4]
Cardiopulmonary resuscitation, not otherwise specified	4.3 [2.7, 6.0]
Endovascular implantation of other graft in abdominal aorta	4.2 [2.7, 5.6]
Reopening of recent thoracotomy site	4.1 [3.0, 5.2]
Other percutaneous procedures on biliary tract	4.0 [1.6, 6.3]
Medications	Score [95% CI]
D5W	4.7 [3.3, 6.2]
Phytonadione	4.6 [2.1, 7.2]
5% Dextrose	4.2 [2.1, 6.4]
Furosemide	3.8 [2.2, 5.3]
Albuterol 0.083% neb soln	3.4 [1.7, 5.1]
Heparin Sodium	3.3 [1.8, 4.7]
Lorazepam	3.2 [1.7, 4.8]
Hydralazine	3.2 [1.3, 5.0]
0.9% Sodium Chloride	3.1 [1.5, 4.8]

Table 3: Top scores from the *Attention (concatenated time)* model

4.3 Additional Results

We added two ablations to find out the impact on removing timestamped code from the embeddings in the model:

- **RNN**: Embeddings are passed to *RNN* layers.
- **RNN + Attention**: Embeddings are passed to *RNN* layers, with attention applied to *RNN* outputs.

Both models did not use timestamped code, and they obtained average precision (0.302, 95% CI [0.294,0.31]) and (0.297, 95% CI [0.289,0.305]) (Table 4) respectively, in-line with other models including *RNN with ODE time decay*, *Exponential time decay*, *exponential time decay + Attention*, *concat time delta + Attention*, *concat time delta*, *MCE*, and *MCE + Attention* models.

	Average Precision	
Model	Reproduced	Original
ODE + RNN + Attention	0.325 [0.317,0.334]	0.314 [0.306,0.321]
ODE + RNN	0.319 [0.311,0.328]	0.331 [0.323,0.339]
RNN (ODE time decay) + Attention	0.321 [0.312,0.33]	0.316 [0.307,0.324]
RNN (ODE time decay)	0.308 [0.301,0.315]	0.300 [0.293,0.308]
RNN (exp time decay) + Attention	0.309 [0.302,0.317]	0.320 [0.312,0.328]
RNN (exp time decay)	0.311 [0.304,0.319]	0.304 [0.297,0.311]
RNN (concat time delta) + Attention	0.307 [0.299,0.316]	0.312 [0.303,0.320]
RNN (concat time delta)	0.314 [0.306,0.323]	0.311 [0.303,0.320]
RNN + Attention	0.297 [0.289,0.305]	N/A
RNN	0.302 [0.294,0.31]	N/A
ODE + Attention	0.292 [0.284,0.3]	0.294 [0.285,0.302]
Attention (concat time)	0.287 [0.278,0.295]	0.286 [0.277,0.295]
MCE + RNN + Attention	0.304 [0.296,0.313]	0.317 [0.308,0.325]
MCE + RNN	0.297 [0.29,0.305]	0.298 [0.291,0.306]
MCE + Attention	0.282 [0.273,0.291]	0.269 [0.261,0.278]
Logistic Regression	0.257 [0.248,0.266]	0.257 [0.248,0.266]

Table 4: Average precision of all models, including ablations.

In comparing attention model with non-attention counterpart, the average precision of the *RNN + Attention* model was 1.6% lower than the *RNN* model. Both models had overlapping 95% confidence interviews, demonstrating that our ablations supported *Claim #2* that additional attention layer had marginal impact on prediction precision but could improve interpretability. Moreover, considering all the reproduced and ablation results together, the average precision range of the models with recurrent component remained the same from 0.297-0.325, and performing better than models based on attention layers, which had average precision ranging from 0.282-0.292. This means our ablations also supported *Claim #3* that models with a recurrent component performed slightly better than models based solely on attention layers.

5 Discussion

Our approach covered all major experiments in the original paper. More specifically, we reproduced all 14 deep neural network models that the original authors evaluated (including a baseline logistic regression model), and calculated odds ratios and coefficient scores for the *Attention concatenated time* model as described in the original paper.

For the most part, our reproduced results were in-line with the results reported in the paper, and we could verify the three major claims which the paper stated. That said, our results were not 100% identical to what the original authors reported even we leveraged the their code. One factor was that the original authors did not document their environment and machine configuration, making it difficult to reproduce results under the exact same environment. Another factor was that the original authors did not use fixed seed everywhere for training the neural network models, introducing some non-determinism in the model and producing different results every time a model was re-trained.

For the three claims we aimed to verify, based on our reproduced results in Section 4.1, we validated *Claim #1* that *ODE* models performed significantly better than the logistic regression model, demonstrating that training an *ODE* neural network with data over time had some predictive advantage in ICU readmission compared to training a simpler logistic regression model with data at a static timepoint. We also observed *Claim #2* that adding an attention layer to various neural network models did not decrease average precision, demonstrating that adding an attention layer is attractive for model interpretability without having to trade-off predictive ability. However, because our reproduced scores for covariates of the model in Section 4.2 were vastly different from the original authors', this could indicate that while interpretable scores can be generated from an attention layer, the scores and weights are not necessarily a true representation of the factors for ICU readmission. Indeed, even the best models in the original paper could only achieve around 30% average precision, meaning that the coefficients evaluated are generally poor predictors for ICU readmission, and therefore the scores can be somewhat random. Since the odds ratios for static variables and also scores for medication were more comparable to the values obtained in the original paper, these values may be more robust predictive factors for ICU readmission than others. Moreover, we validated *Claim #3* in Section 4.1 that a recurrent layer was beneficial for predictive performance, outperforming neural network models without it. This demonstrates the benefit of adding a recurrent component in neural network models on time series data and demonstrates its benefits even with irregular-time data.

We also added two ablations as described in Section 4.3. While our ablations were not applicable to *Claim #1*, they did validate *Claim #2* and *Claim #3* and strengthened these claims in the original paper. When compared to all of the models using some method of accounting for irregular time-intervals, our ablation models performed significantly worse (with 95% CI) than models using ODE time de-

cay. Using pre-trained *MCE* (medical concept embeddings) did not seem to enhance model precision. Our ablation models without timestamped embeddings did perform slightly worse than models which accounted for them using concatenated time delta, or exponential time decay on average, and performed significantly worse than models using ODE time decay. This seems to indicate that using irregular-time data is useful for model prediction, but effect is highly dependent on how the time intervals are modeled.

5.1 What was easy

1. Once the environment was set up, getting original authors' code to work was easy.
2. The original authors' code was written clearly, and it was easy to understand and follow.
3. The original authors were communicative over emails, and promptly answered any questions and uploaded missing files when asked.

5.2 What was difficult

1. The original authors' code was written in *Python*, but the required *Python* libraries and their versions were undocumented. It took us quite a while to set up a working environment. While our environment was functional, it was clearly not identical to the environment which the original authors had. The differences between our environment and the original authors' might impact our reproduced results.
2. There were 14 models in the original paper plus two ablations of our own, and majority of them took 12 - 20 hours to run. This was very time consuming without GPU acceleration. We tried to use free GPU compute resources available through *Google Colab*, but this did not significantly decrease training time and in fact was slower than running the scripts on a *Macbook Pro* with a *M1 Max* processor.
3. We were unable to run 3 of the models initially due to missing precomputed weighted embeddings, and there was no instruction in the original authors' code on how to generate them. After contacting the original authors, they eventually provided the missing files to unblock us.
4. The preliminary results (e.g. average precision, etc) we obtained from the models were close, but not exactly identical to what's in the original paper. This took us quite a long time to debug and rerun the models multiple times, until we realized there were fundamental issues in the original authors' code which did not produce deterministic results.
5. The prediction scores we obtained from the models were very different from what's in the paper. This took us quite a long time to debug and analyze the issues.

5.3 Recommendations for reproducibility

To improve reproducibility of the original paper, here is a set of recommendations we have:

1. Use *Dockerfile* to set up the environment with required OS, libraries, and toolchain for running the model. Make the *Dockerfile* as part of the code repository such that others would be able to create the exact environment later.
2. Use and document the fixed seed for randomization and sampling.
3. When splitting the dataset into train/dev/test, sort the dataset in some deterministic order before sampling, to ensure the dataset is splitted in exactly the same way for reproducibility.
4. Document exact system configuration which the models were trained and tested.
5. Run the model on multiple machines to ensure the results are consistent and deterministic.

6 Communication with original authors

We reached out to the original authors in the following extent:

1. Several precomputed weighted embeddings were missing from the original authors' code that blocked us from reproducing the results of three models, and we asked the authors to provide the missing files.
2. We asked the original authors to help us better understand the computational requirements.
3. We asked the original authors on pointers to generate the odd ratios in the original paper.
4. We asked the original authors to help us better understand why *ODE + RNN* model achieved the highest precision in the original paper.
5. We discussed with the original authors about some non-deterministic behavior we observed in our reproduction study.

References

[1] Barbieri, Sebastiano, et al. "Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk." Scientific reports 10.1 (2020): 1-10.

[2] Paszke, Adam, et al. "Pytorch: An imperative style, high-performance deep learning library." Advances in neural information processing systems 32 (2019).

[3] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3.1 (2016): 1-9.

[4] Rubanova, Y., R. T. Chen, and D. Duvenaud. "Latent odes for irregularly-sampled time series (2019)." arXiv preprint arXiv:1907.03907 (1907).

[5] Blundell, Charles, et al. "Weight uncertainty in neural network." International conference on machine learning. PMLR, 2015.