# Prediction Model for Multiple Readmission on Congestive Heart Failure Patients

Yiting Xiao

Fangyun Shi

Yilin Li

# Abstract

In order to improve the effectiveness of treatment and reduce readmission of congestive heart failure in hospitals, our project did regression modeling to figure out the relationship between multiple readmission of CHF patients and their socio-demographic characteristics, medical and health conditions.

We first used logistic regression to give us an idea of how readmission is related to these three types of factors. And then reduced the model using both AIC and BIC stepwise search. Reading the results we have several ideas about how race, source of patients and severity affect patients' risk of readmission. We accessed our reduced model on C-statistics and then did similar analysis on other link functions like probit, cloglog, cauchit, and loglog. We compared these models using classification rate and selected the best binomial regression model for our problem—the loglog regression model.

We future applied other nonlinear models including decision tree models and neural network models to see whether these models can give us a better classification result. The decision tree model gives us more factors to consider when deciding factors that affect patients' risk of readmission. And we successfully reduced the dimension of factors in neural network without sacrificing classification rate using the most important factors in our previous models.

The final analysis showed that there was only slightly difference among those prediction models, yet we chose the decision tree model as the recommended one. Furthermore, we suggested that hospitals and doctors should pay especially attention to CHF patients who is admitted for different diagnosis and with musculoskeletal system and connective tissue problems. All in all, to reduce the readmission of CHF, both hospitals, doctors and patients must work together!

# Introduction

## ➢ Motivation

Hospital readmission shortly after discharge is increasingly recognized as a marker of inpatient quality of care and a significant contributor to rising healthcare costs. Though it remains unclear whether such readmissions are entirely preventable, there is good evidence that targeted interventions initiated before and/or shortly after discharge can decrease the likelihood of readmission. Identifying patients at risk of readmission can guide efficient resource utilization and permit valid comparisons of hospital quality across institutions.

Thus we did research about hospital readmission based on congestive heart failure (CHF), one of the common heart failures occurred in situations of high output. According to Emory Healthcare and CDC report, nearly 5 million Americans are currently living with CHF. This disease affects people of all ages, from children and young adults to the middle-aged and the elderly. Heart failure is responsible for 11 million physician visits each year, and more hospitalizations than all forms of cancer combined. CHF is the first-listed diagnosis in 875,000 hospitalizations, and the most common diagnosis in hospital patients age 65 years and older. More than half of those who develop CHF die within 5 years of diagnosis.

## ➢ Problem Description

The motivation verifies the importance of effective treatment of congestive heart failure in hospitals. However, there were 2500 readmission cases in just one hospitals, EUHM (Emory University Hospital Midtown) from Jan 2011 to Sep 2013. And 30% of the readmitted patients went through multiple readmissions, some were even readmitted more than 5 times, which not only increase the cost of hospitals but also reflect the ineffective treatment of CHF patients.

We used the knowledge of regression and machine learning to figure out the relationship between multiple readmission of congestive heart failure patients and the characteristic and pathological diagnosis. Based on the mathematical outcomes, we could offer both clinical and managerial suggestions to hospitals. It can help CHF patients as well.

## ➢ Challenges

Before building the model, the most challenge thing is the pre-processing data. First, the measurements that are essential for modeling may not be presented directly in the dataset. Therefore, we need to understand what each column means, then extract that information from the data set and decide how to organize them in a way that most benefits the clarification of our model. This can be time-consuming and we should answer following questions.

1) How to assembly and organize the raw data so that we can be able to provide a useful model to measure the rate of readmission?
2) How to build the model so that the sparse nature of diagnosis won't affect the quality of our

modeling?

3) How to transfer the text format data into the format that could be ran in R?

After building the models, we would like to see how well it describes the data and how well they predicts. We have to answer:

1) Are the models significant?
2) How well does the models predict?
3) What methods are more suitable for testing models?
4) Are the models robust?

# Data processing

## ➢ Data Sources

We got the data of De-identified EUHM (Emory University Hospital Midtown) CHF Readmissions (Any Diagnosis) from Professor Ayer. It contains 2500 EMR records and 868 same patients' readmission records. Time period last from Jan 2011 to Sep 2013. Report patient type is inpatient. The readmit cases excludes chemotherapy, radiation therapy, rehabilitation, death 1st Admit, dialysis, delivery / birth and mental patients. Specific part of the raw data can be found in table A1 of the appendix.

## ➢ Initial data pre-process

Diagnosis Related Groups (DRGs) are a patient classification scheme which provides a means of relating the type of patients a hospital treats (i.e., its case mix) to the costs incurred by the hospital. There are three types of DRG listed in EMR, we only choose the All Patient Refined DRGs (APR-DRG) which incorporate severity of illness subclasses and is exactly classified based on the detailed data in different diagnosis listed at the end of each row. So we cut off the detailed diagnosis data and generate the new columns *Heart failure diagnosis, Diagnosis Stability, Major Diagnostic Categories, Medical Service* based on DRG numbers.

From the original EMR records, we deleted the Emergency Room Patient, as 429 patients lack records; Ethnicity as half of the patient are not reported; UHC Secondary Payer as most secondary payers are self-pay.

Some parts of factors like age, cost, and length of stay are consecutive numbers; we divided them into several categories which is more reasonable for model analysis and interpretation. In case that it would impact the prediction of model, we used Weka software to build statistical models, yet the result was not better than the model that divided them into categories. And some factors contained too many minors, some of which may just cover few patients, we also processed data based on medical logic. For example, we combined *Medicare/Managed Care* and *Medicare Traditional/Indemnity* into the same category-*Medicare*.

## ➢ Variables

**Socio-demographic factors**

- **Age**

  Based on Heart Failure Statistics from Emory Healthcare, CHF is present in 2 percent of persons age 40 to 59. More than 5 percent of persons age 60 to 69 have CHF. Besides, CHF annual incidence approaches 10 per 1,000 population after 69 years of age. Thus, we divided the age of patients into four categories based on the distribution of incidence of CHF.

  18-39 years  40-59 years  60-69 years  >69 years

- **Sex**

  Female  Male

- **Race**

  Black  Other (Asian 2, White 145, other 7)

- **Primary insurance** *(UHC Primary Payer)*

  Private insurance *(Health Maintenance Organization (HMO); Point-of-Service (POS); Preferred Provider Organization (PPO); Traditional/Indemnity; Transplant Network)*

  Medicaid *(include military and research indemnity)*

  Medicare

  Uninsured (*Self-Pay)*

**Medical condition**

- **Admission Day**

  It is possible that patients are more likely to be taken care less comprehensively at weekend since there are less workforce, and thereby are under higher risk of readmission, because the first day in hospitals should be the intensive period, and later is the time for recovering.

  Weekday  Weekend

- **Admission Source**

  Clinic referral/Non-Facility Point of Origin/Transfer from a different hospital/Transfer from skilled nursing facility or ICF

- **Admission Status**

  Elective  Emergency  Urgent

- **Discharge MD Specialty**

  Hospitalist  None-Hospitalist

- **Discharge Status**

  Routine discharge *(Discharged to home care or self-care (routine discharge))*

  Home health service *(Discharged/transferred to home under care of organized home health service)*

  SNF *(Discharged/transferred to skilled nursing facility (SNF) with Medicare certification)*

  Hospice

  Self-left *(Left against medical advice or discontinued care)*

  Other *(Other discharged/transferred)*

- **Charges Observed (in dollars)**

  <=15000  15001-30000  30001-60000  >60000

**Health Condition** (determines the types of patients treated – disease and diagnosis related)

- **Heart failure diagnosis**

Heart failure *(DRG number 194)* Other

- **Major Diagnostic Categories (MDCs)**
  Specific categories can be found in table A2 of the appendix.
- **Medical Service** *(considering APR-DRG table, each DRG number will reflect if certain patient is treated by procedure or medicine. Procedure groups means patients have a procedure performed which requires the use of the operating room.)*
  Procedure groups Medicine groups
- **Diagnosis Stability**
  Same diagnosis *(MDC number stay the same, all the admission diagnosis category are the same)*
  Multiple diagnosis *(multiple MDC numbers, patients have different admission diagnosis categories)*
- **Admit Severity of illness**
  Extreme Major Minor Moderate
- **Admit Risk of Mortality**
  Extreme Major Minor Moderate
- **ICU stayed**
  None-ICU stayed patients *(0 day stay in ICU)*
  ICU stayed patients
- **LOS Observed**
  <=3 days 4-7 days 8-14 days >=15 days

**Patient Characteristics Summary**

| Characteristic | Entire cohort n=868 | Characteristic | Entire cohort n=868 |
|---|---|---|---|
| Multiple Readmitted, n (%) | 263 (30%) | | |
| **Socio-demographic factors** | | | |
| Sex | | Race | |
| Female | 462 (53%) | Black | 714 (82%) |
| Male | 406 (47%) | Other | 154 (18%) |
| Age group | | Primary insurance | |
| 18-39 years | 51 (6%) | Private insurance | 88 (10%) |
| 40-59 years | 285 (33%) | Medicaid | 123 (14%) |
| 60-69 years | 189 (22%) | Medicare | 635 (73%) |
| >69 years | 343 (39%) | Uninsured | 22 (3%) |
| **Medical condition** | | | |
| Admission Day | | Discharge MD Specialty | |
| Weekday | 660 (76%) | Hospitalist | 449 (52%) |
| Weekend | 208 (24%) | None-Hospitalist | 419 (48%) |
| Admission Source | | Admission Status | |
| Clinic referral | 38 (4%) | Elective | 81 (9%) |
| Non-Facility Point of Origin | 744 (86%) | Emergency | 651 (75%) |
| Different hospital | 53 (6%) | Urgent | 136 (16%) |
| Skilled nursing facility | 33 (4%) | | |
| Discharge Status | | Charges Observed | |
| Routine discharge | 528 (61%) | (in dollars) | |

| | | | |
|---|---|---|---|
| Home health service | 192 (22%) | <=15000 | 228 (26%) |
| SNF | 112 (13%) | 15001-30000 | 223 (26%) |
| Hospice | 8 (1%) | 30001-60000 | 206 (24%) |
| Self-left | 10 (1%) | >60000 | 211 (24%) |
| Other | 18 (2%) | | |
| **Health Condition** | | | |
| Heart failure diagnosis | | Diagnosis Stability | |
| Heart failure (0) | 175 (20%) | Same diagnosis (0) | 598 (69%) |
| None-heart failure (1) | 693 (80%) | Multiple diagnosis (1) | 270 (31%) |
| Major Diagnostic Categories | Table A2 | | |
| Admit Severity of illness | | Admit Risk of Mortality | |
| Extreme | 155 (18%) | Extreme | 93 (11%) |
| Major | 540 (62%) | Major | 368 (42%) |
| Moderate | 152 (18%) | Moderate | 377 (43%) |
| Minor | 21 (2%) | Minor | 30 (4%) |
| ICU stayed | | Medical Service | |
| None-ICU stayed patients (0) | 582 (67%) | Surgical groups | 591 (68%) |
| ICU stayed patients (1) | 286 (33%) | Medicine groups | 277 (32%) |
| LOS Observed | | | |
| <=3 days | 273 (32%) | | |
| 4-7 days | 288 (33%) | | |
| 8-14 days | 177 (20%) | | |
| >=15 days | 130 (15%) | | |

# Proposed Methodology

## ➢ Data process

We first used Excel to do basic data analysis:

1) use Pivot Table to see the distribution of each factors of dataset;
2) use IF ELSE and LOOKUP function to divide consecutive data

We also used R to make boxplot and did correlation analysis to assess the relationship between different factors, which may help us prevent co-linearity problems and avoid big mistakes.

## ➢ Logistic regression

The patient was the unit of analysis. We chose a split-sample design to derive and internally validate our prediction model. We randomly selected 70% of patients from each site and combined them to create a derivation cohort (training data) and subsequently combined the remaining 30% of patients from each site to create a validation cohort (testing data).

To assess whether the candidate patient factors were significantly associated with hospital readmission, we fitted ***multivariable logistic regression*** models for each patient factors using data from the

derivation cohort. Then we build other four regression models: ***probit, cloglog, cauchit, loglog***. We also used ***AIC criteria*** to shrink factors and included significant factors in selected models.

We tested the performance of our models using data from the validation cohort. And we calculated the ***error rates*** for those five models in order to do comparison. We especially assessed the logistic model discrimination by measuring the ***C-statistic***, which is the area under the receiver operating characteristic (ROC) curve, is defined as the proportion of times the model correctly discriminates a pair of single-readmission and multiple-readmission patients. A c statistic of 0.50 indicates that the model performs no better than chance; a c statistic of 0.60 to 0.80 indicates modest or acceptable discriminative ability; and a c statistic of greater than 0.80 indicates good discriminative ability.

### ➢ Other classification models

Considering the none-linear and complicated unknown relationship under variables, we used ***neural network*** and ***decision tree*** to let R intelligently learn the relationship between multiple readmission and different patients' characteristics. Those two methods also provide us models of classification.

# Analysis and Results

This section displays the results of model analysis. The detailed running processes by R is referred to Appendix B.

### ➢ Correlation analysis

Collinearity may arise when the predictor variables are strong correlated among themselves. In such a case, collniearity inflates the errors. Thus, we examined the correlation matrix of predictor variables even if they are not measured in continuous scales and see whether their correlation coefficients are way too high. We excluded the examination of factors in the same predictor categories (such as Age), the results showed no obvious correlation. So it is fair to do further regression with all predictor variables.

### ➢ Logistic regression

- Full logistic regression model

We first choose logistic regression to model the times of readmission against all the EMR variables. One simple reason is that it is the most popular and applicable way to model binary response (in our case if the patient is readmitted once then Y=1, otherwise Y=0). The technical reason would be that logistic regression uses the canonical link function. We first model the outcome against all 55 EMR variables. A proportion of results is shown below, see Appendix for details.

**Association of Patient Characteristics with Multiple Readmission in the Derivation Cohort**

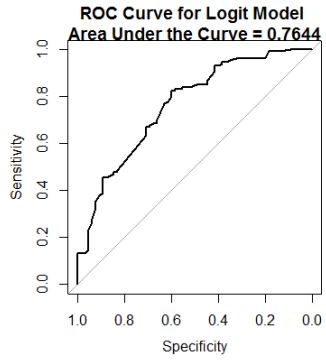
| Category | Characteristic | Odds ratio (95% CI) | P value |
|---|---|---|---|
| | Race | | |
| **Socio-demographic factors** | Black | 0.42(0.20-0.83) | 0.015831 * |
| | Other | Reference | |
| **Medical** | Admission Source | | |

| | | | |
|---|---|---|---|
| **condition** | Clinic referral | Reference | |
| | Non-Facility Point of Origin | 1.14(0.35-3.62) | 0.455388 |
| | Different hospital | 1.92(0.36-11.9) | 0.820664 |
| | Skilled nursing facility | 1.51(0.83-4.22) | 0.085949 • |
| | Admission Status | | |
| | Elective | Reference | |
| | Emergency | 0.39(0.14-1.02) | 0.064120 • |
| | Urgent | 0.99(0.31-3.12) | 0.992478 |
| **Health Condition** | Heart failure diagnosis | | |
| | Heart failure | 0.30(0.15-0.598) | 0.000723*** |
| | None-heart failure | Reference | |
| | Major Diagnostic Categories | | |
| | Musculoskeletal System And Connective Tissue | 0.15(0.02-1.02) | 0.059069 • |
| | Diagnosis Stability | | |
| | Same diagnosis | Reference | |
| | Multiple diagnosis | 0.058(0.03-0.096) | < 2e-16 *** |

As expected, health condition factors has most impact over readmission. Being diagnosed with heart disease and having multiple different diagnosis contributes to the possibility of being readmitted. This makes sense because a diagnosis of heart disease indicates serious health issue and having multiple diagnosis indicates comorbidity.

For a hospital manager, one may want to strengthen business relationship with skilled nursing facilities. Patients transferred from there have a significantly lower possibility of being readmitted multiple times thus potentially saves hospital lots of money. Managers should also warn patients about leaving hospital without being discharged (self.left) since this kind of behavior may lead to readmission.
Other than what we expected, most social and demographic factors like age and gender has little impact on readmission. Therefore regardless of how old the patient is, all patients should be treated equally. However, it's shown that black people are significantly more susceptible to CHF. This matches the medical research of the illness on of genes. Physicians and hospitalists should pay special attention to them.

- Reduced logistic model using AIC stepwise search



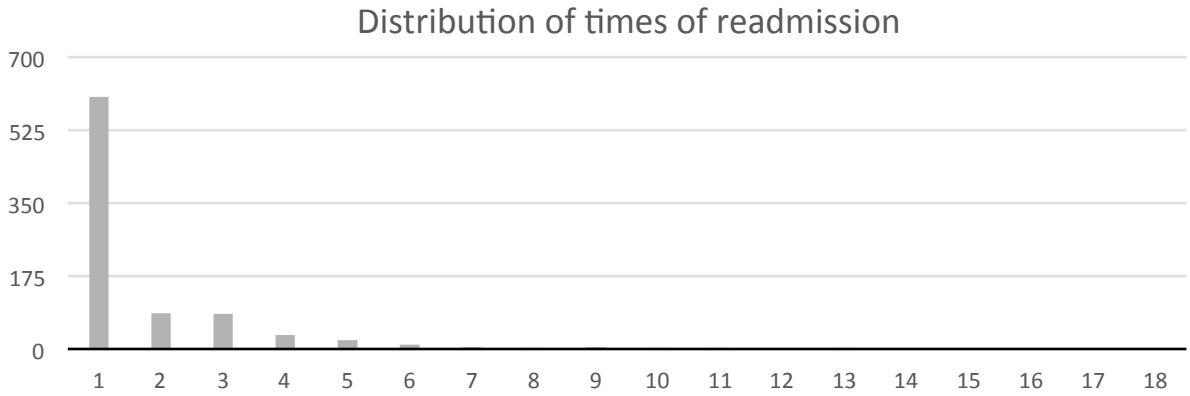

| logit | Error rate = 23.08% | |
|---|---|---|
| Y | False | True |
| 0 | 39 | 26 |
| 1 | 34 | 161 |

While the full model gives an insight into the relationships among different patient record variables and

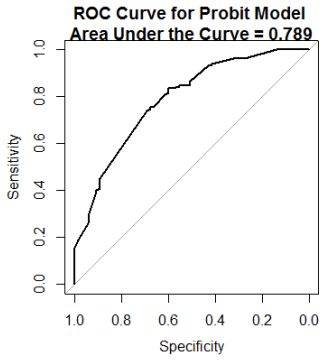
readmission, the model is neither accurate nor efficient enough to predict readmission. In order to have a better prediction, we use AIC as criteria for a stepwise variable reduction. The shrunk model has C-statistics of 0.766, pretty well over 0.5, which means that it is an acceptable classifier for our problem. Only 13 variables were kept in this model and the classification error rate decreased from 27.3% of the original model to 23.1%. Details about this model is shown in the appendix.

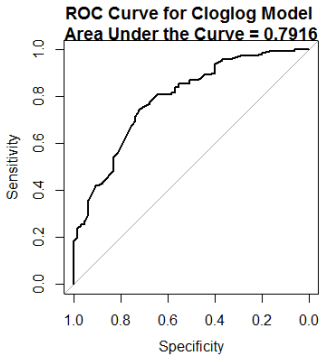
- Regression model with other link functions

The reduced logistic model was a good classifier for our problem. But still, we would like to see if other binomial models can deliver a better result. When choosing a link function there are several things to consider. Knowledge of the response distribution, theoretical considerations and empirical fit to the data should all be taken into consideration.



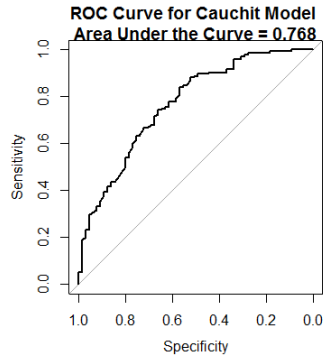

Since the distribution of the response is heavy tailed to the right, log-log link function seems to be the most promising one to give a better result. However, since all the other links, probit, cloglog, cauchit are all imbedded in the glm package of R, there's no harm in trying more links out. Similarly with the logistic regression, we first regressed on the full model and then performed AIC stepwise search to find out the best set of variables. And then calculated the C-statistics of each model to see if they are acceptable classifiers.

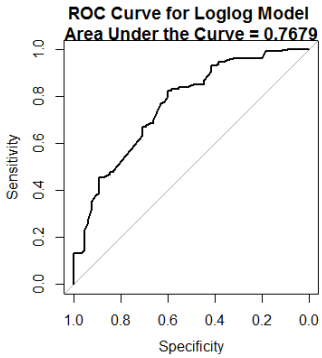



| | | |
|---|---|---|
| Probit | Error rate = 23.84% | |
| Y | False | True |
| 0 | 39 | 26 |
| 1 | 36 | 159 |
| Cloglog | Error rate = 40.38% | |
| Y | False | True |
| 0 | 54 | 11 |
| 1 | 94 | 101 |
| Cauchit | Error rate = 26.92% | |
| Y | False | True |
| 0 | 40 | 25 |
| 1 | 45 | 150 |
| Loglog | Error rate = 23.08% | |
| Y | False | True |
| 0 | 35 | 30 |
| 1 | 30 | 165 |

Table of Confusion Matrices



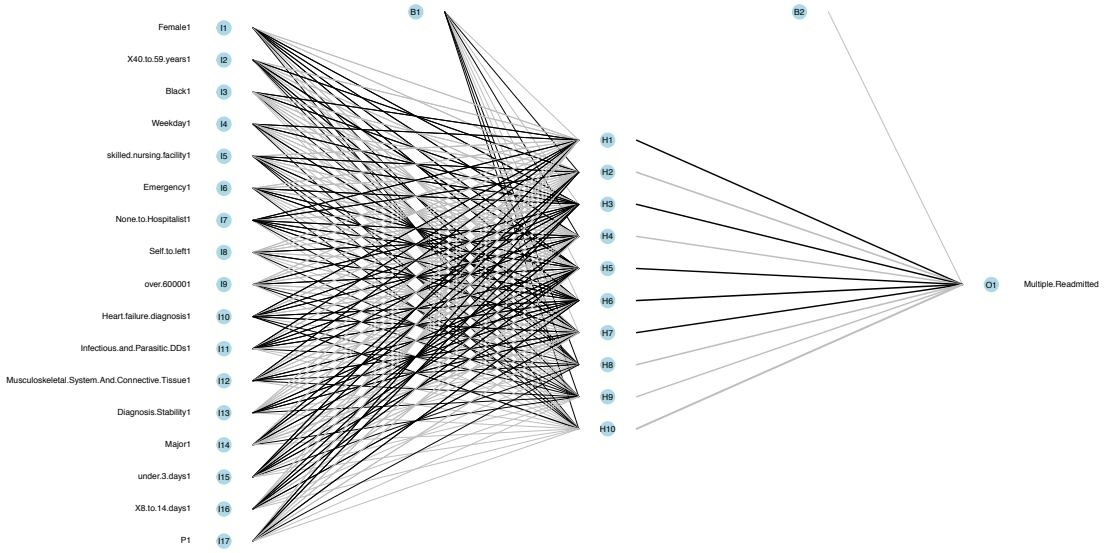

- Model comparison

From the C-statistics we can see that all five models proves to be acceptable classifiers. Comparing the error rate on testing data, the logistic and loglog regression model both showed a low 23.08%. But since CHF is a type of serious illness, having a heart attack is already traumatic enough. Clearly one may not want to suffer from more than one heart attack. Therefore, loglog model which has a lower false positive rate is a more logical choice given the circumstances.

For hospital service quality engineer, if he/she is required to instruct IT technicians to incorporate predictions about readmission into the hospital EMR system so that physicians, hospitalists and nurses can be warned about patients' readmission risks before they are discharged, loglog regression model is definitely his/her best choice. Since loglog regression model has the best classification rate and lowest false positive rate, it will best aid physicians and hospitalist make decisions about discharging patients from hospital.

## ➢ Neural network

We utilized nnet package in R and get the following result.



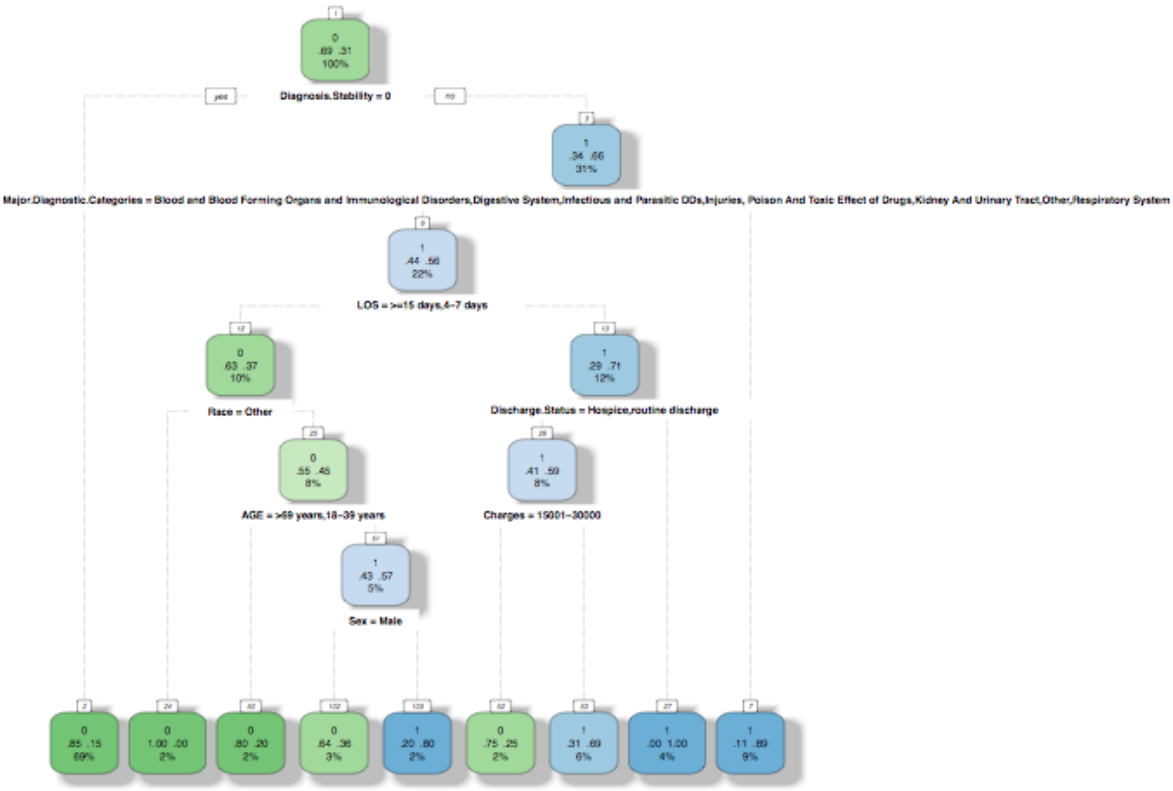

We randomly select about half of our data (400 records) as training data and the rest as test data. And

then use tune.nnet() to estimate the optimal size and decay by using 10-fold cross-validation. It shows that the optimal size is 10 and the optimal decay is 0.1.

To simplify our model, we only consider the significant variables from our previous analysis. Compared to our original model using all variables, the simplified model gives a pretty good result. The classification rate of our neural network model on the training data is 91.25%, and the classification rate on the test data is 75.64%.

## ➢ Decision tree

We utilized rpart package in R and get the following result.




As is shown in the above chart, Diagnosis Stability stands out to be the most important factors for our classification. To be more specific, patients diagnosed unstable at first time are far less likely to be readmitted. (In the training data, 85% of patients with Diagnosis.Stability = 0 were not readmitted.) A possible reason for this could be doctors pay more attention to patients whose conditions are unstable. In other word, these patients get a more thorough treatment, and thus there is no need for them to go to the hospital again.

Another important factor is major diagnostic categories. The one listed above represents a more severe type of diagnostic compared to the ones not listed, like Eye, Ear, Nose, Mouth And Throat, etc. The classification result showed that patients with less severe major diagnostic categories are more likely to

be readmitted. (In the training data, 89% of such patients were readmitted.) This result is also reasonable in that patients with less severe conditions may not be paid much attention to and as time goes by, it's very likely that their conditions get worse and have to be readmitted.

Other results are also intuitive. Non-black patients, elderly and young patients have a lower chance to be readmitted. Maybe it's because they are paid more attention to; and male patients are also less likely to be readmitted than female patients.
We randomly select about half of our data (400 records) as training data and the rest as test data. The classification rate of our tree-based model on the training data is 83.25% and the classification rate on the test data is 77.35%.

# Conclusions

## ➢ Recommended model

| Model | Classification Rate |
|---|---|
| Logistic Regression | 76.92% |
| Loglog Regression | 76.92% |
| Decision Tree | 77.35% |
| Neural Network | 75.64% |

From analysis of all the different models it seems clear that performance is quite similar across the board. For this particular dataset, data manipulation, extraction, and imputation seem much more important tasks for improving accuracy than model selection. That being said, the decision tree did produce the best training and test dataset results, and thus is the recommended model for this dataset. From the decision tree model, we can see that the most important factors we need to take into account when predicting whether a patients will be readmitted or not are Diagnosis Stability, Major Diagnostic Categories, Race and Age.

## ➢ Research conclusions

As there is only slightly difference among those prediction models. We offer our suggestions to hospitals and patients based on the observations of all the four models.

1） Predictor variable Diagnosis Stability is significant in all models. So Hospitals and doctors should pay especially attention to CHF patients who is admitted for different diagnosis. This may because CHF has caused other harmful complications to patients, like diabetes and lungs problem.
2） Patients with musculoskeletal system and connective tissue problems should be worried more. IF patients is admitted for this major diagnosis, hospitals may let them stay in hospitals for longer time and discharge them after removing the most risks. This can also be an observable diagnosis. So patients themselves with heart failure history should be careful if the symptoms of visible swelling of the ankles and legs take place.
3） Emory health report has showed that African-Americans are 1.5 times more likely to develop heart failure than Caucasians, which agrees with our study models. So black, male patients should lead a healthy life and avoid the risk factors that normally increase the chances of CHF, such as smoking

and obesity. Besides, it is no wonder that the self-left patients are more likely to be readmitted. Therefore, to reduce the readmission of CHF, both hospitals, doctors and patients must work together!

## ➢ Future work

First, our sample is still too small. We should gain larger data from more general recourse such as Centers for Medicare & Medicaid Services (CMS) rather than just one single hospital if we want to convince other.

Secondly, our research lacks of data of none-readmission CHF patients. After all, it is more common to build prediction model for hospital readmission. Limited to dataset, we made prediction for multiple readmission; yet the idea is similar and our project can easily turn to extended prediction models.

Readmission assessment could be used to help target the delivery of these resource intensive interventions to the patients at greatest risk. Ideally, models designed for this purpose would provide clinically relevant stratification of readmission risk and give information early enough during the hospitalization to trigger a transitional care intervention, many of which involve discharge planning and begin well before hospital discharge. Thus research could further involve with sensitivity analysis and test whether intervention can positively reduce the multiple readmission of certain patient.

# Appendix

In this part, we list all the details about our modelling of readmission.

In Appendix A, raw data and summaries of our data set are given.

In Appendix B, important R codes are listed.

In Appendix C, R output are listed.

# Appendix A

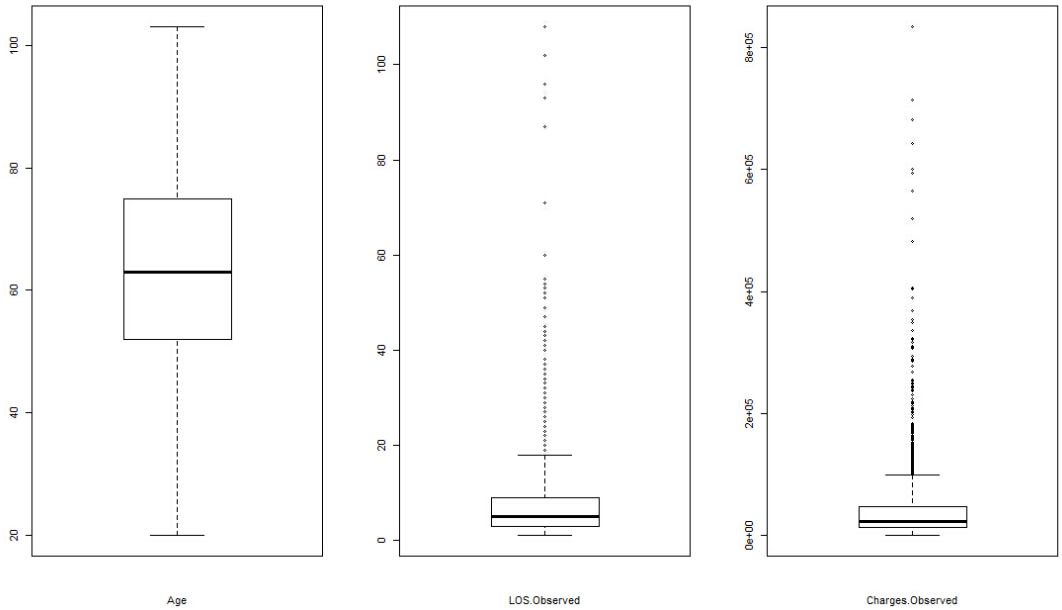
**Table A1-Excel table of raw data**

| Deidentified ID | Encounter Number | Admission Date | Admission Day | Admission Source | Admission Status | Age |
|---|---|---|---|---|---|---|
| 1 | C0002123310262 | 09/19/2010 | Sunday | Transfer from a different hospital | Urgent | 57 |
| 1 | C0002123311031 | 01/31/2011 | Monday | Transfer from a different hospital | Urgent | 58 |
| 2 | C0008375230326 | 11/22/2010 | Monday | Non-Facility Point of Origin | Emergency | 76 |
| 2 | C0008375231038 | 02/08/2011 | Tuesday | Non-Facility Point of Origin | Emergency | 76 |
| 3 | C0013586290351 | 12/20/2010 | Monday | Non-Facility Point of Origin | Elective | 74 |
| 3 | C0013586291038 | 02/07/2011 | Monday | Transfer from a different hospital | Urgent | 75 |
| 4 | C0010649320358 | 12/24/2010 | Friday | Non-Facility Point of Origin | Emergency | 46 |
| 4 | C0010649321022 | 01/22/2011 | Saturday | Non-Facility Point of Origin | Emergency | 46 |
| 4 | C0010649321112 | 04/22/2011 | Friday | Non-Facility Point of Origin | Emergency | 46 |
| 4 | C0010649321122 | 05/02/2011 | Monday | Non-Facility Point of Origin | Emergency | 46 |
| 4 | C0010649321159 | 06/08/2011 | Wednesday | Non-Facility Point of Origin | Emergency | 47 |
| 4 | C0010649321182 | 07/02/2011 | Saturday | Non-Facility Point of Origin | Emergency | 47 |
| 4 | C0010649321211 | 07/30/2011 | Saturday | Non-Facility Point of Origin | Emergency | 47 |
| 4 | C0010649321218 | 08/06/2011 | Saturday | Non-Facility Point of Origin | Emergency | 47 |

| HCO Primary Payer | HCO Secondary Payer | Diagnosis1 | Diagnosis2 |
|---|---|---|---|
| MCA - MEDICARE A | SPU - SELF-PAY UNINSURED | 03849 - gram-neg septicemia nec | 70724 - stage iv pressure ulcer |
| MCA - MEDICARE A | SPU - SELF-PAY UNINSURED | 03849 - gram-neg septicemia nec | 4275 - cardiac arrest |
| MCA - MEDICARE A | SPU - SELF-PAY UNINSURED | 8248 - clsd fx ankle nos | 99667 - infect d/t orth dev nec |
| MCA - MEDICARE A | SPU - SELF-PAY UNINSURED | 8438 - hip & thigh sprain nec | 5856 - esrd |
| MGI - MEDICARE HMO GENERIC INPT | SPU - SELF-PAY UNINSURED | 4241 - aortic valve disorder | 42823 - ac & chr systolic hf |
| M6I - COVENTRY AVANTRA MCR IP | SPU - SELF-PAY UNINSURED | 0389 - septicemia nos | 78552 - septic shock |
| MCA - MEDICARE A | MCA - MEDICARE A | 486 - pneumonia organism nos | 42823 - ac & chr systolic hf |
| MCA - MEDICARE A | MCA - MEDICARE A | 42823 - ac & chr systolic hf | 486 - pneumonia organism nos |
| MCA - MEDICARE A | SPU - SELF-PAY UNINSURED | 42823 - ac & chr systolic hf | 4280 - chf nos |

*The raw data contains details about each hospitalization, one patient may have multiple hospitalizations as shown below.

**Picture A1-boxplot for consecutive factors**




*The boxplot of continuous variables help us to segment it into intervals.

**Table A2-Major Diagnostic Categories (MDCs)**

| MDC | Description | APR-DRG | Number of patients | Percentage |
|---|---|---|---|---|
| 0 | Pre-MDC | 020-058 | 23 | 3% |
| 1 | Nervous System | 070-082 | 1 | 0% |
| 2 | Eye | 089-115 | 4 | 0% |
| 3 | Ear, Nose, Mouth And Throat | 120-144 | 98 | 11% |
| 4 | Respiratory System | 160-207 | 399 | 46% |
| 5 | Circulatory System | 220-229 | 15 | 2% |
| 6 | Digestive System | 240-254 | 43 | 5% |
| 7 | Hepatobiliary System And Pancreas | 260-284 | 11 | 1% |
| 8 | Musculoskeletal System And Connective Tissue | 301-351 | 29 | 3% |
| 9 | Skin, Subcutaneous Tissue And Breast | 361-385 | 10 | 1% |
| 10 | Endocrine, Nutritional And Metabolic System | 401-425 | 50 | 6% |
| 11 | Kidney And Urinary Tract | 440-468 | 77 | 9% |
| 12 | Male Reproductive System | 480-501 | 2 | 0% |
| 13 | Female Reproductive System | 510-532 | 3 | 0% |
| 14 | Pregnancy, Childbirth And Puerperium | 540-566 | 1 | 0% |
| 15 | Blood and Blood Forming Organs and Immunological Disorders | 650-663 | 18 | 2% |
| 16 | Myeloproliferative DDs (Poorly Differentiated Neoplasms) | 680-694 | 7 | 1% |
| 17 | Infectious and Parasitic DDs | 710-724 | 38 | 4% |
| 18 | Mental Diseases and Disorders | 740-760 | 2 | 0% |
| 19 | Injuries, Poison And Toxic Effect of Drugs | 791-816 | 17 | 2% |

| | | | | |
|---|---|---|---|---|
| 20 | Factors Influencing Health Status | 850-863 | 4 | 0% |
| 21 | Multiple Significant Trauma | 890-894 | 7 | 0% |
| 22 | Other | 950-956 | 9 | 0% |

**Table A4-parts of correlation analysis**

| | Readmit.b | Female | X40.59.ye | X.69.year | X60.69.ye | Black | Medicare | Medicaid | Private.i | Weekday | different |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Readmit.b | 1 | -0.06 | -0.06 | 0.09 | -0.05 | -0.15 | 0.07 | -0.1 | 0.01 | -0.02 | 0.13 |
| Female | -0.06 | 1 | -0.06 | 0.09 | -0.03 | 0.07 | 0.05 | 0.06 | -0.12 | 0.02 | -0.16 |
| X40.59.ye | -0.06 | -0.06 | 1 | -0.57 | -0.36 | 0.09 | -0.32 | 0.31 | 0.08 | -0.04 | 0.02 |
| X.69.year | 0.09 | 0.09 | -0.57 | 1 | -0.44 | -0.1 | 0.41 | -0.33 | -0.18 | -0.01 | -0.04 |
| X60.69.ye | -0.05 | -0.03 | -0.36 | -0.44 | 1 | 0.02 | -0.04 | -0.01 | 0.09 | 0 | 0.04 |
| Black | -0.15 | 0.07 | 0.09 | -0.1 | 0.02 | 1 | -0.03 | 0.09 | -0.07 | -0.08 | -0.21 |
| Medicare | 0.07 | 0.05 | -0.32 | 0.41 | -0.04 | -0.03 | 1 | -0.68 | -0.56 | -0.01 | -0.01 |
| Medicaid | -0.1 | 0.06 | 0.31 | -0.33 | -0.01 | 0.09 | -0.68 | 1 | -0.13 | 0 | -0.04 |
| Private.i | 0.01 | -0.12 | 0.08 | -0.18 | 0.09 | -0.07 | -0.56 | -0.13 | 1 | 0.02 | 0.04 |
| Weekday | -0.02 | 0.02 | -0.04 | -0.01 | 0 | -0.08 | -0.01 | 0 | 0.02 | 1 | 0.03 |
| different | 0.13 | -0.16 | 0.02 | -0.04 | 0.04 | -0.21 | -0.01 | -0.04 | 0.04 | 0.03 | 1 |

**Table A5- full result for logistic regression**

| Association of Patient Characteristics with Multiple Readmission in the Derivation Cohort | | | |
|---|---|---|---|
| **Category** | **Characteristic** | **Odds ratio (95% CI)** | **P value** |
| **Socio-demographic factors** | Sex | | |
| | Female | 0.75 (0.45-1.20) | 0.236343 |
| | Male | Reference | |
| | Age group | | |
| | 18-39 years | Reference | |
| | 40-59 years | 0.51(0.16-1.48) | 0.231145 |
| | 60-69 years | 0.47(0.14-1.43) | 0.569289 |
| | >69 years | 0.71(0.21-2.22) | 0.195847 |
| | Race | | |
| | Black | 0.42(0.20-0.83) | 0.015831 * |
| | Other | Reference | |
| | Primary insurance | | |
| | Private insurance | 0.29(0.049-1.57) | 0.156678 |
| | Medicaid | 0.33(0.072-1.66) | 0.183040 |
| | Medicare | 0.37(0.06-1.75) | 0.220473 |
| | Uninsured | Reference | |
| **Medical condition** | Admission Day | | |
| | Weekday | 0.72(0.42-1.22) | 0.228599 |
| | Weekend | Reference | |
| | Admission Source | | |
| | Clinic referral | Reference | |
| | Non-Facility Point of Origin | 1.14(0.35-3.62) | 0.455388 |
| | Different hospital | 1.92(0.36-11.9) | 0.820664 |
| | Skilled nursing facility | 1.51(0.83-4.22) | 0.085949 • |
| | Admission Status | | |
| | Elective | Reference | |
| | Emergency | 0.39(0.14-1.02) | 0.064120 • |
| | Urgent | 0.99(0.31-3.12) | 0.992478 |

| | | | |
|---|---|---|---|
| | Discharge MD Specialty | | |
| | Hospitalist | Reference | |
| | None-Hospitalist | 0.71(0.43-1.17) | 0.184393 |
| | Discharge Status | | |
| | Routine discharge | 0.38(0.05-1.83) | 0.276156 |
| | Home health service | 0.47(0.06-2.32) | 0.398022 |
| | SNF | 0.35(0.04-1.90) | 0.265229 |
| | Hospice | Reference | |
| | Self-left | 0.11(0.008-1.24) | 0.078494 • |
| | Charges Observed (in dollars) | | |
| | <=15000 | Reference | |
| | 15001-30000 | 1.18(0.56-2.49) | 0.669287 |
| | 30001-60000 | 1.29(0.51-3.27) | 0.589695 |
| | >60000 | 1.82(0.53-6.30) | 0.337708 |
| **Health Condition** | Heart failure diagnosis | | |
| | Heart failure | 0.30(0.15-0.598) | 0.000723*** |
| | None-heart failure | Reference | |
| | Major Diagnostic Categories | | |
| | Musculoskeletal System And Connective Tissue | 0.15(0.02-1.02) | 0.059069 • |
| | Medical Service | | |
| | Procedure groups | 1.12(0.63-1.97) | 0.697134 |
| | Medicine groups | Reference | |
| | Diagnosis Stability | | |
| | Same diagnosis | Reference | |
| | Multiple diagnosis | 0.058(0.03-0.096) | < 2e-16 *** |
| | Admit Severity of illness | | |
| | Extreme | 1.50(0.25-9.19) | 0.655089 |
| | Major | 0.90(0.18-4.33) | 0.655089 |
| | Moderate | 0.55(0.11-2.67) | 0.459928 |
| | Minor | Reference | |
| | Admit Risk of Mortality | | |
| | Extreme | 0.78(0.16-9.20) | 0.751795 |
| | Major | 1.04(0.29-3.62) | 0.953774 |
| | Moderate | 1.54(0.45-2.67) | 0.486887 |
| | Minor | Reference | |
| | ICU stayed | | |
| | None-ICU stayed patients | Reference | |
| | ICU stayed patients | 0.95(0.52-1.73) | 0.866840 |
| | LOS Observed | | |
| | <=3 days | Reference | |
| | 4-7 days | 0.75(0.37-1.51) | 0.425638 |
| | 8-14 days | 0.43(0.16-1.73) | 0.866840 • |
| | >=15 days | 0.92(0.25-3.37) | 0.896547 |

# Appendix B

**Program B1-Definition of loglog link**

```
### create a loglog link
loglogfun <- function() {
   ## link
   linkfun <- function(mu) -log(-log(mu))
   ## inverse link
   linkinv <- function(eta) exp(-exp(-eta))
   ## derivative of invlink wrt eta
   mu.eta <- function(eta) exp(-exp(-eta)-eta)
   valideta <- function(eta) all(eta != 0)
   link <- "-log(-log(mu))"
   structure(list(linkfun = linkfun, linkinv = linkinv,
                  mu.eta = mu.eta, valideta = valideta,
                  name = link),
             class = "link-glm")
}
loglog <- loglogfun()
glm(Readmit.bin~., family=binomial(link=loglog), data=data0train)
```

**Program B2-C-statistics**

```
### ROC of models
## 1. logit
testpred.logit <- predict(logitmod, newdata=data0test,type=c("response"))
data0test1 <- data0test
data0test1$prob <- testpred.logit
g1 <- roc(Readmit.bin ~ testpred.logit, data = data0test1)
auc(g1)
plot(g1, main=paste("ROC Curve for Logit Model", "\nArea Under the Curve = 0.7644"))
```

**Program B3-Converting categorical data to binary format**

```
chfu <- lapply(chf, unique)
chf.reg <- data.frame(Readdmisson=chf$Readdmisson, Multiple.Readmitted = chf
$Multiple.Readmitted)

ini <- rep(0,868)

for (m in 3:12){
   col.new <- as(chfu[[m]],"character")
   col.old <- as(attr(chfu[m],"names"),"character")
   #initialize admission source
   for (k in seq_along(col.new)) {chf.reg[,col.new[k]] <- ini}
   for (i in 1:868){
      for (j in seq_along(col.new)){
```

```
        if ( chf[i,col.old]==col.new[j]) {chf.reg[i, col.new[j]] <- 1}
      }
   }
   rm(col.new)
   rm(col.old)
}
```

**Program B4-Decision Tree**

```
library(rpart)
library(rattle)
DF <- read.csv("CHF.csv")
DF[1:15] <- lapply(DF[1:15], as.factor)
set.seed(123)
sub <- c(sample(1:217, 100), sample(218:434, 100), sample(435:651, 100),  sample(652:868, 100));
fit1 <- rpart(Multiple.Readmitted ~ ., data= DF, subset = sub, method="class", parms=list(split="information"));
table(predict(fit1, DF[sub,], type="class"), DF[sub, "Multiple.Readmitted"])
table(predict(fit1, DF[-sub,], type="class"), DF[-sub, "Multiple.Readmitted"])
fancyRpartPlot(fit1, sub="")
print(fit1)
```

**Program B5-Neural Network**

```
library(nnet)
library(e1071)
DS <- read.csv("CHF_bin_nnet.csv")
DS[1:18] <- lapply(DS[1:18], as.factor)
set.seed(123)
sub <- c(sample(1:217, 100), sample(218:434, 100), sample(435:651, 100),  sample(652:868, 100));
model1 <- tune.nnet(Multiple.Readmitted ~ ., data= DS, size =10:15, decay = c(0,0.1, 0.01, 0.001));
summary(model1)
plot(model1);
model1$best.model
fit2 <- nnet(Multiple.Readmitted ~ ., data= DS, subset = sub,  size = 10, decay = 0.1)
table(predict(fit2, DS[sub,], type="class"), DS[sub, "Multiple.Readmitted"])
table(predict(fit2, DS[-sub,], type="class"), DS[-sub, "Multiple.Readmitted"])
#import the function from Github
library(devtools)
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff044412703516c34f1a4684a5/nnet_plot_update.r')
plot.nnet(fit2)
```

# Appendix C

**Output C1- Best sub-loglog model using AIC**

```
Call:
glm(formula = Readmit.bin ~ Black + Weekday + skilled.nursing.facility +
    Emergency + None.Hospitalist + Self.left + Heart.failure.diagnosis +
    Infectious.and.Parasitic.DDs + Musculoskeletal.System.And.Connective.Tissue +
    Diagnosis.Stability, family = binomial(link = loglog), data = data0train)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.5050 -0.5975  0.4397  0.6405  1.9331

Coefficients:
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       4.2155     0.4599   9.166  < 2e-16 ***
Black                                            -0.7786     0.2631  -2.959 0.003089 **
Weekday                                          -0.3041     0.1856  -1.638 0.101347
skilled.nursing.facility                          1.2272     0.5458   2.249 0.024542 *
Emergency                                        -0.7964     0.2221  -3.585 0.000337 ***
None.Hospitalist                                 -0.3294     0.1644  -2.004 0.045119 *
Self.left                                        -1.1945     0.6486  -1.842 0.065525 .
Heart.failure.diagnosis                          -0.7524     0.1943  -3.873 0.000108 ***
Infectious.and.Parasitic.DDs                     -0.4542     0.3056  -1.486 0.137216
Musculoskeletal.System.And.Connective.Tissue     -1.2365     0.3673  -3.367 0.000761 ***
Diagnosis.Stability                              -2.1780     0.1733 -12.569  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 767.37  on 607  degrees of freedom
Residual deviance: 548.90  on 597  degrees of freedom
AIC: 570.9

Number of Fisher Scoring iterations: 6
```

**Output C2-Best sub-decision tree model**

```
n= 400

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 400 123 0 (0.6925000 0.3075000)
    2) Diagnosis.Stability=0 277  42 0 (0.8483755 0.1516245) *
    3) Diagnosis.Stability=1 123  42 1 (0.3414634 0.6585366)
      6) Major.Diagnostic.Categories=Blood and Blood Forming Organs and Immunological Disorders,Digestive System,Infectious and Parasitic DDs,Injuries, Poison And Toxic Effect of Drugs,Kidney And Urinary Tract,Other,Respiratory System 86  38 1 (0.4418605 0.5581395)
        12) LOS=>=15 days,4-7 days 38  14 0 (0.6315789 0.3684211)
          24) Race=Other 7  0 0 (1.0000000 0.0000000) *
          25) Race=Black 31  14 0 (0.5483871 0.4516129)
            50) AGE=>69 years,18-39 years 10  2 0 (0.8000000 0.2000000) *
            51) AGE=40-59 years,60-69 years 21  9 1 (0.4285714 0.5714286)
              102) Sex=Male 11  4 0 (0.6363636 0.3636364) *
              103) Sex=Female 10  2 1 (0.2000000 0.8000000) *
        13) LOS=<=3 days,8-14 days 48  14 1 (0.2916667 0.7083333)
          26) Discharge.Status=Hospice,routine discharge 34  14 1 (0.4117647 0.5882353)
            52) Charges=15001-30000 8  2 0 (0.7500000 0.2500000) *
            53) Charges=<=15000,30001-60000 26  8 1 (0.3076923 0.6923077) *
          27) Discharge.Status=home health service,Self-left,SNF 14  0 1 (0.0000000 1.0000000) *
      7) Major.Diagnostic.Categories=Ear, Nose, Mouth And Throat,Endocrine, Nutritional And Metabolic System,Eye,Factors Influencing Health Status,Female Reproductive System,Hepatobiliary System And Pancreas,Multiple Significant Trauma,Musculoskeletal System And Connective Tissue,Pre-MDC 37  4 1 (0.1081081 0.8918919) *
```

**Output C3-Best sub-neural network model**

```
Parameter tuning of 'nnet':

- sampling method: 10-fold cross validation

- best parameters:
  size decay
    10  0.01

- best performance: 0.2292302
```

## Reference

[1] Knoblauch, Kenneth, and Laurence T. Maloney. Modeling psychophysical data in R. Vol. 32. London: Springer, 2012.

[2] Kansagara, Devan, et al. "Risk prediction models for hospital readmission: a systematic review." Jama 306.15 (2011): 1688-1698.

[3] Strack, Beata, et al. "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records." BioMed research international 2014 (2014).

[4] Hasan, Omar, et al. "Hospital readmission in general medicine patients: a prediction model." Journal of general internal medicine 25.3 (2010): 211-219.