# PREDICTING THE PROGRESSION OF HEART DISEASE USING MCMC

by-Erik Nylander, Tulasi Ramarao, Youqing Xiang | IS604 | Spring 2016

## Abstract

Close to 600 thousand people die of heart disease every year in the U.S which equates to 1 out of every 4 deaths. The lowest death rates were in the southwest and the highest in the south [Fig 1]. Several tools are now available to estimate one's risk of getting a heart disease. However, the limited availability of good predictive tools is one of the reasons that the death rate due to heart disease hasn't slowed down. In this paper, a Monte Carlo simulation model is designed to evaluate the impact of risk factors on life expectancy of a patient.

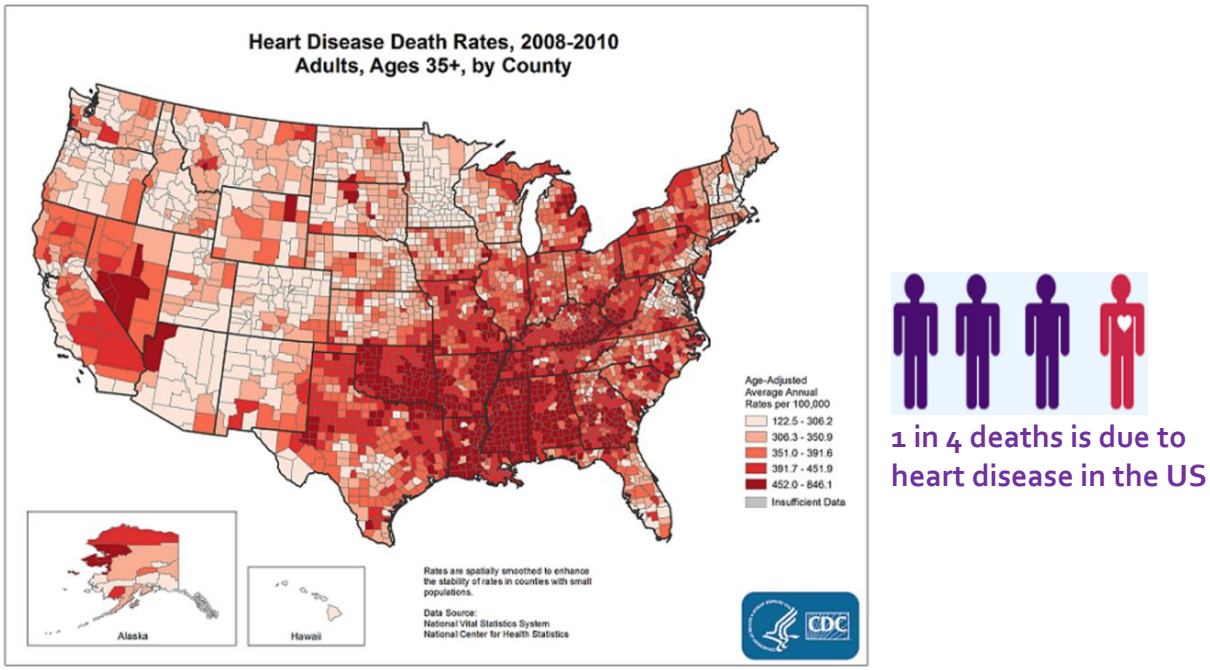



**Fig 1: Results posted on cdc.gov**

**Keywords:** Cardiovascular disease, Markov Chain Monte Carlo, Diagnosis, Simulation, Intervention.

## Introduction and Literature Review

If heart disease is detected at an earlier stage, a treatment regime can not only avoid the progression of this disease, but can also reduce the medical costs associated with treatment of this disease in later stages. Also, if this diagnosis is automatic, then it can reduce costs related to doctor visits and the charges associated with treatment of this disease in later stages. The authors of Journal Ref# 1 used decision tree and Naïve Bayes to develop a system to diagnose and predict heart disease and the authors of Journal Ref #2 used Markov Transition model to determine the overall as well as per patient cost of illness due to Type 2 Diabetes.

## Methodology

In this proposal, a Markov model [Fig 6] is used to simulate the progression of heart disease by using the cardiovascular disease risk factors downloaded from a clinical database from UCI Machine Learning Repository. This data was collected for patients from Cleveland, OH and a Time series was added to this dataset to fit the data to a Monte Carlo Simulation model.

The model was created by reducing the description of the disease into a five state model as described by NY Heart Association (NYHA). The five states are listed below and the progression of the disease is as shown in Fig 2.

- State 0: No-heart disease
- State 1: Mild
- State 2: Moderate
- State 3: Severe
- State 4: Death

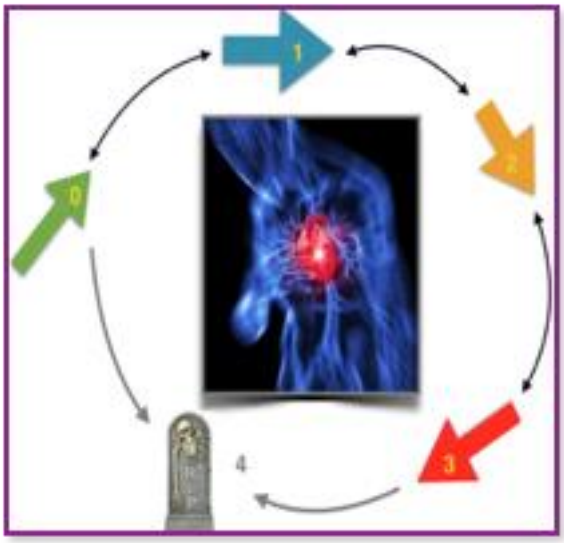



*Fig 2: Stages in the progression of heart disease*

A simulation model was developed to predict the progression of heart disease based on predictive models generated from the UCI dataset. The dataset has information such as symptoms and risk factors and these predictors were used to predict a person's susceptible to a heart disease

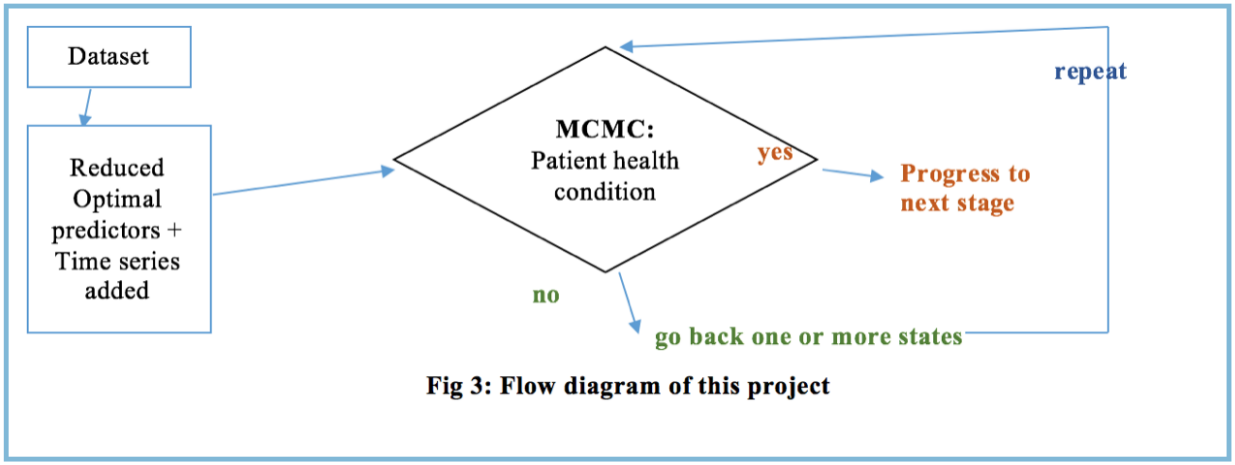



**Fig 3: Flow diagram of this project**

The model was used to predict events that occur over time with data from screening for the heart disease at fixed intervals. Disease progression is described probabilistically as a set of transitions among the states in fixed duration(months). The likelihood of making a transition is defined as s set of transition probabilities. Average cost involved in staying in class 1 for a year will be assessed.

Effects of an Intervention will be assessed mathematically by using:

1. The transition probabilities among the states (by reducing the probability of death)

## Assumptions

A patient in a Markov model is always in one of a finite number of discrete states called Markov states. The events are represented by transitions from one state to the other. As people progress through the simulation model from a healthy state to the onset of heart disease and to death, the patients can develop symptoms that lead to more and more complications represented by states. A patient can belong to 2 or more states in real life, but for simplicity sake, that scenario is not considered in this model. Also, an initial distribution was assigned in order for the simulation to start. A patient can advance from state 1 to any states as long as the patient is alive. A patient can die in any state. Transitions can occur from lower state to a much higher state, example from state 1 to state 3.

**MCMC steps:**

The steps implemented to develop the Markov model are:

1. Construct a decision tree
   - Enumerate the possible states
   - Define the allowable state transitions
2. Identify the probabilities
   - Associate the probabilities with the transitions
   - Identify the cycle length

Identify the number of cycles
Identify an initial distribution of patients within the states

3. Identify the outcome values
4. Calculate the expected values

The data description is as shown in Table 2 in the Appendix.
The attribute called *num* will be used as a response variable. Its value range from 0 to 4 indicating the states of heart disease.

**Decision tree for the Heart Disease:**
This represents the logic used in creating a decision tree for the progression of the heart disease.

**Associating probabilities with the transitions:**
Cleveland dataset was analyzed for the distribution of these States. There were 164 patients were classified as State 0(no heart disease), 55 as State 1 and so on.

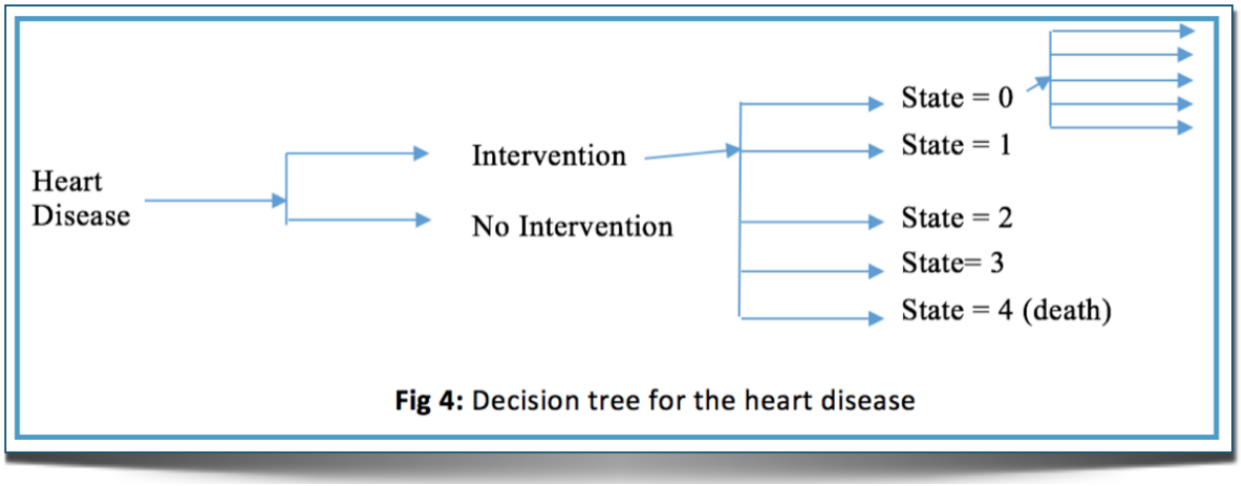



**Fig 4:** Decision tree for the heart disease

| States | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| Cleveland | 164 | 55 | 36 | 35 | 13 | 303 |

Next, the transition probabilities were assigned arbitrarily and adjusted based on the readings from Ref 4.

State 0 transition probabilities:(total-164):
Suppose in the following year: 105 spent in State 0, 59 in State 1, none in State 2-4

State 1 transition probabilities:(total-55):

Suppose in the following year: 8 spent in State 0, 35 in State 1, 10 in State 2, 2 died.

State 2 transition probabilities:(total-36):
Suppose in the following year: 0 spent in State 0, 9 in State 1, 15 in State 2, 10 in State 3, 2 died

State 3 transition probabilities:(total-35):
Suppose in the following year: 0 spent in State 0, 0 in State 1, 10 in State 2, 20 in State 3, 5 died

State 4 transition probabilities: (total-13):
Suppose in the following year: 0 spent in State 0, 0 in State 1, 0 in State 2, 0 in State 3, 13 died

This data was then converted into a transition table [Table 1].

## Probability Transition Table:

| Transition | Data | Probability |
|---|---|---|
| 0 to 0 | 70/164 | 0.423 |
| 0 to I | 35/164 | 0.21 |
| 0 to II | 59/164 | 0.36 |
| 0 to III | 0 | 0 |
| 0 to IV | 0 | 0 |
| I to 0 | 8/55 | 0.15 |
| I to I | 35/55 | 0.63 |
| I to II | 10/55 | 0.18 |
| I to III | 0 | 0 |
| I to IV | 2/55 | 0.04 |
| II to 0 | 0 | 0 |
| II to I | 9/36 | 0.25 |
| II to II | 15/36 | 0.42 |
| II to III | 10/36 | 0.27 |
| II to IV | 2/36 | 0.06 |
| III to 0 | 0 | 0 |
| III to I | 0 | 0 |
| III to II | 10/35 | 0.29 |
| III to III | 20/35 | 0.57 |
| III to IV | 5/35 | 0.14 |
| IV to 0 | 0 | 0 |
| IV to I | 0 | 0 |
| IV to II | 0 | 0 |
| IV to III | 8/13 | 0 |

| | | |
|---|---|---|
| IV to IV | 5/13 | 0.38 |

**Fig 3: Probability Transition Table**

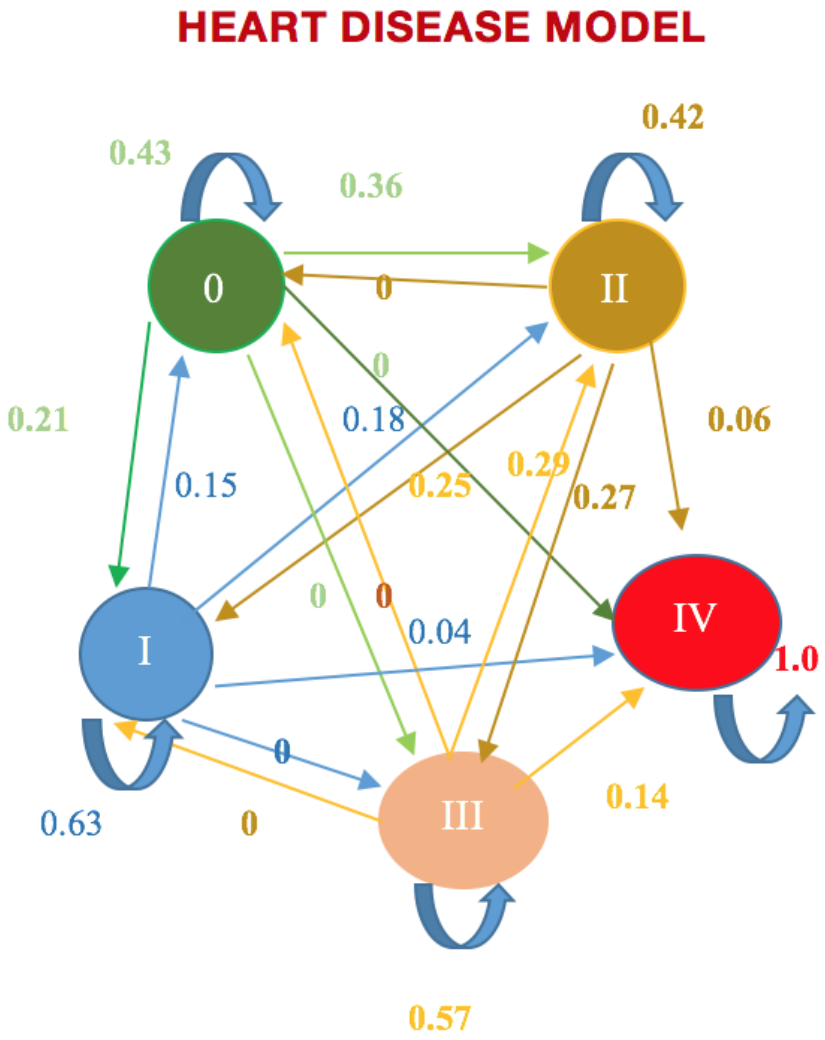



**Fig 6:** Markov model

**Initial Distribution**:

An arbitrary initial distribution was assigned to start the Simulation [Fig 7a]. To fit the data, a multistate model with a transition intensity matrix Q is defined [Fig 7b and Fig 7c]

State 1: 0.43
State 2: 0.15
State 3: 0
State 4: 0
State 5: 0

$$Q = \begin{pmatrix} q11 & q12 & q13 & q14 & q15 \\ q21 & q22 & q23 & q24 & q25 \\ q31 & q32 & q33 & q34 & q35 \\ q41 & q42 & q43 & q44 & q45 \\ q51 & q52 & q53 & q54 & q55 \end{pmatrix}$$

**Fig 7a**: Initial distribution

**Fig 7b**: General model for disease progression

$$Q = \begin{pmatrix} -(q12+q13+q15) & q12 & q13 & q14 & q15 \\ 0 & -(q21+q23+q25) & q23 & 0 & q25 \\ 0 & q32 & -(q32+q34+q35) & q34 & q35 \\ 0 & 0 & 0 & -(q44+q45) & q45 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

**Fig 7c:** Matrix 2

```
> Q
     [,1] [,2] [,3] [,4] [,5]
[1,] 0.43 0.21 0.36 0.00 0.00
[2,] 0.15 0.63 0.18 0.00 0.04
[3,] 0.00 0.25 0.42 0.27 0.06
[4,] 0.00 0.00 0.29 0.57 0.14
[5,] 0.00 0.00 0.00 0.00 1.00
```

Now, a matrix of the same size as in Fig 7b is defined, with zeroes in corresponding positions in Fig 7c, where there are zeroes. The other positions have their initial values filled from the probability Transition table [Fig 3]. This matrix will be used in the qmatrix for the MSM function. So, finding the unknown values represented in the matrix (the non-zero values) will result in fitting the model.

**Fig 7d:** Matrix 3

To find the true maximum likelihood estimates, models should be run repeatedly using the initial values. To define allowed transitions in this model, a matrix which has the same size of Q is defined. Since there are not many changes between between the observation times, this crude method will work [Fig 10] [Ref#: 5]

```
> Q.crude
           [,1]        [,2]        [,3]       [,4]       [,5]
[1,] -0.3414634  0.31707317  0.02439024  0.0000000 0.00000000
[2,]  0.0000000 -0.12834225  0.08556150  0.0000000 0.04278075
[3,]  0.0000000  0.02702703 -0.37837838  0.2432432 0.10810811
[4,]  0.0000000  0.00000000  0.00000000 -0.1538462 0.15384615
[5,]  0.0000000  0.00000000  0.00000000  0.0000000 0.00000000
```

**Fig 10:** Matrix 4

## Results and experimentation

A fitted model object called MSM in R is used with dataset [Ref # 6], the transition matrix Q and the initial values as its parameters.

```
> printold.msm(heart.msm)

Call:
msm(formula = state ~ years, subject = PTNUM, data = heartdata2,     qmatrix = Q)

Maximum likelihood estimates:
Transition intensity matrix

        State 1 State 2 State 3 State 4 State 5
State 1 -0.57    0.21    0.36    0.00    0.00
State 2  0.15   -0.37    0.18    0.00    0.04
State 3  0.00    0.25   -0.58    0.27    0.06
State 4  0.00    0.00    0.29   -0.43    0.14
State 5  0.00    0.00    0.00    0.00    0.00
```

**Fig 11:** Resulting model using the msm call in R

From the table above, we can see that patients are more likely to die with symptoms than without symptoms. (That is, transition from State 1 to State 5).
Once the disease is onset, the transition from State 2 to State is more likely (0.18) than staying at State 2(0.-37).
Once in State 4 (with severe heart disease), that its highly unlikely (-0.43) to stay in that state.
Patient spends a mean of 2.32 years (-1/0.43) in State 4 before death or even recovery.

**Transition probability Matrices:**

For a given period, say 10 years, if transition probability is needed, then Pmatrix extracts that information.

```
> pmatrix.msm(heart.msm,t=10)
           State 1   State 2   State 3   State 4   State 5
State 1 0.06332304 0.2047094 0.1930663 0.1372114 0.4016898
State 2 0.06437092 0.2064460 0.1870594 0.1276541 0.4144696
State 3 0.04774590 0.1643129 0.1617283 0.1236948 0.5025181
State 4 0.03345060 0.1235293 0.1328574 0.1122494 0.5979133
State 5 0.00000000 0.0000000 0.0000000 0.0000000 1.0000000
```

A person without heart disease has a probability of
0.40 of being dead in 10 years from today,
0.06 of staying without heart problems
0.2 of getting into a slightly mild state of disease
0.19 of getting into a mild state of disease
0.13 of getting into a serious state of this disease.

**Probability of the next state:**

This information shows how Markov model works. This gives a more intuitive view of a continuous-time Markov model rather than the transition matrix we saw earlier. This matrix kind of gives the mean of sojourn times. The mean sojourn times is the average time in a single stay and forecasts the the total time before the death in other states.

```
> pnext.msm(heart.msm)
        State 1 State 2 State 3 State 4 State 5
State 1  0.0000  0.3684  0.6316  0.0000  0.0000
State 2  0.4054  0.0000  0.4865  0.0000  0.1081
State 3  0.0000  0.4310  0.0000  0.4655  0.1034
State 4  0.0000  0.0000  0.6744  0.0000  0.3256
State 5  0.0000  0.0000  0.0000  0.0000  0.0000
```

The total length of stay is given below. The R call totlos.msm estimates the total length of time spent in each transient state between State 1 and State 5. This is calculated by the formula:

$$L_s = \int_{t_1}^{t_2} P(t)_{r,s} dt$$

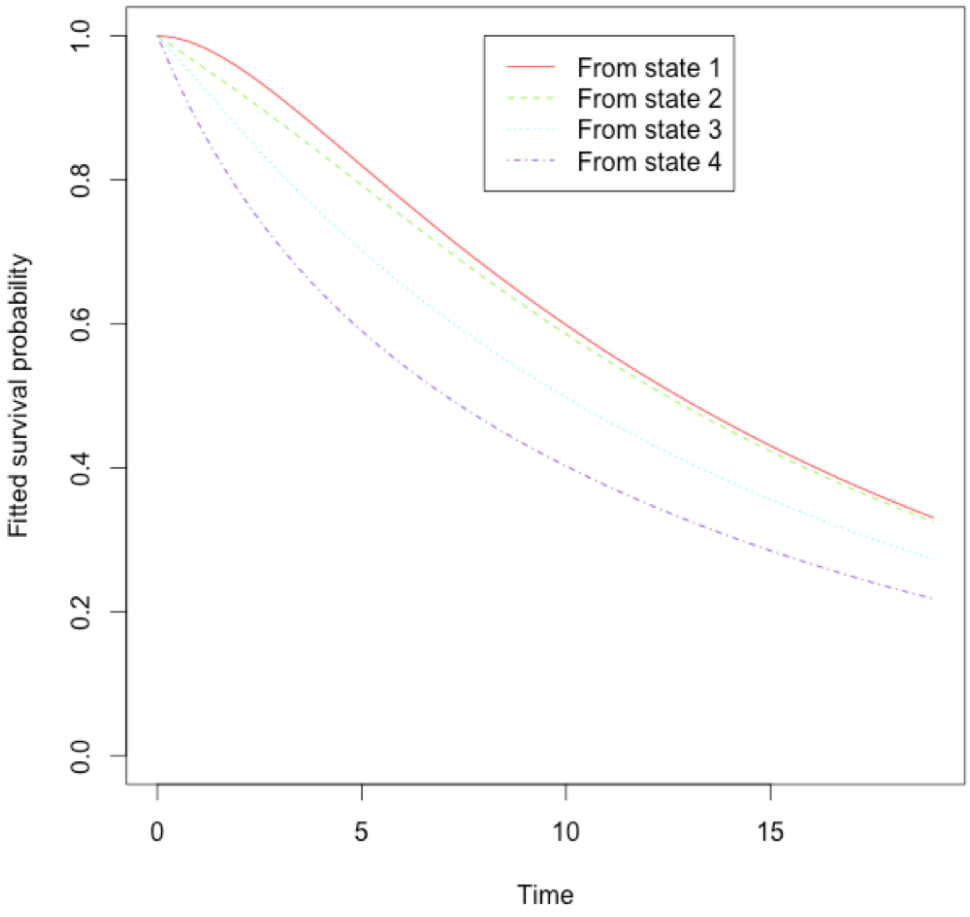
where r represents the begin state(State 1). The results are as shown below.

```
> totlos.msm(heart.msm)
  State 1   State 2   State 3   State 4   State 5
3.171309 5.384306 5.304873 3.330967       Inf
```

So the model [Fig 11] shows that each patient can be expected to spend
3.17 years in State 1,
5.38 years in State 2,
5.30 years in State 3 and
3.3 years in State 4

**Survival probability:**

If Expected(probability) of survival against time is plotted for each of these states, then this plot shows that a 10 year (survival) probability is near to 0.2 for State 3, close to 0.3 for state 2 and close to 0.4 for state 1.



**Simulation:**

After analyzing the Markov Chain Monte Carlo simulation for heart disease in R, we decided to create a simulation that shows the progression of heart disease in Simio. Screenshots of the model as well as the results are shown in the Appendix. To simulate this system in Simio we elected to use the following procedure. Patients are created and placed in State 0. These patients then select a path back to State 0 or to other States in the system based on the transition probability from the transition matrix that was developed above. We then designed the patients and the paths in such a way that it takes a single day for the patients to travel their assigned paths. In this way we were able to create a one simulation that has transitions that occur every one unit of time. The servers in our model have an infinite capacity so that patients can then transition at the end of each unit of time to a State of the disease. Once a patient transitions to State 4 we remove them from the simulation.

We found the Simio simulation to be a very natural way to experiment with our model and used this to experiment with the different interventions by changing the probability of patients selecting a given path. Youqing can you please add what you did in your experiments here.

**Results from the Simulation:**

We have constructed the simulation to measure the effects of intervention on the number of deaths from heart disease. Given this use for the simulation we need to confirm that the simulation is doing a good job of calculating the average number of deaths due to heart disease. We know from our above analysis of the Markov Chain Monte Carlo simulation that we should expect 40% of all individuals with heart disease to have reached Stage 4 after a 10 year period. To validate our simulation against this finding we have constructed an experiment to run the simulation 30 times and calculate the mean and standard deviation for the number of individuals reaching Stage 4. We will then construct a statistical test to determine if our results are consistent with the MCMC model. Using the experiment function in Simio we found the following:

$$\bar{x} = 38.2667$$
$$s = 4.6456$$

We then construct the following hypothesis test at an α = 0.05 level:

$$H_0: \bar{x} = 40$$
$$H_A: \bar{x} \neq 40$$

We can use these two values to calculate our t-statistic as follows:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{38.2667 - 40}{4.6456/\sqrt{30}} = -2.436$$

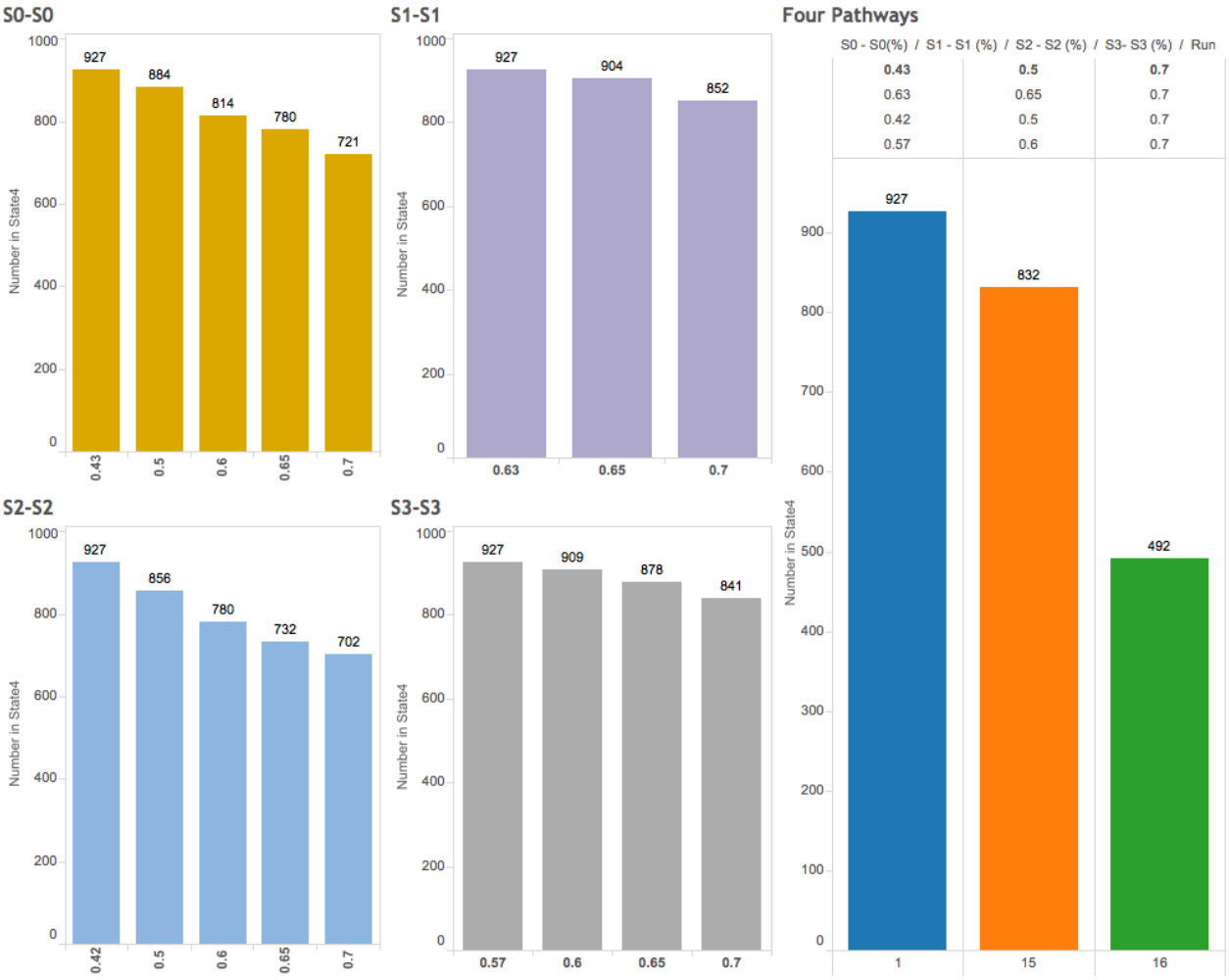
Calculating the p-value associated with this test statistic we get a p-value = 0.051. While this result is borderline we fail to find enough evidence to reject the null hypothesis and conclude that there is not enough evidence to reject the null hypothesis and conclude that our simulation gives us a reasonable approximation for the number of deaths after 10 years. While

not an ideal p-value we will move forward with this simulation to experiment on the effects of interventions at different stages of the disease.

Youqing can you please add what you found in your experiments here.

**Outcome:** Erik/Youqing – need more explantion - paragraph

Prediction of a heart disease
Health outcomes (life expectancy)




**Summary and future improvements:**
More work will need to be done to increase this model's consistency and efficiency. Including weights associated with each symptom would make a realistic model. Also including cost involved in patient care would have made a good proposal.

**References:**


1. Issn (Online) : 2319 - 8753, and Issn (Print) : 2347 - 6710. *K.L.N. College of Engineering, Madurai, Tamil Nadu, India Heart Disease Diagnosis Using Predictive Data Mining* (n.d.): n. pag. Web.

2. "Cost-of-illness Study of Type 2 Diabetes Mellitus in Colombia." *Cost-of-illness Study of Type 2 Diabetes Mellitus in Colombia*. N.p., n.d. Web. 05 May 2016.

3. 550, Epi, and 2012 March 2. *Introduction to Markov Models (part 1)* (n.d.): n. pag. Web.

4. http://cvdrisk.nhlbi.nih.gov/

5. "A Discrete Time Markov Chain (DTMC) SIR Model in R." *Rbloggers*. N.p., n.d. Web. 14 May 2016.

6. Abstract. *Multi-state Modelling with R: The Msm Package* (n.d.): n. pag. Web.

**Source of the Dataset**: UCI Machine learning

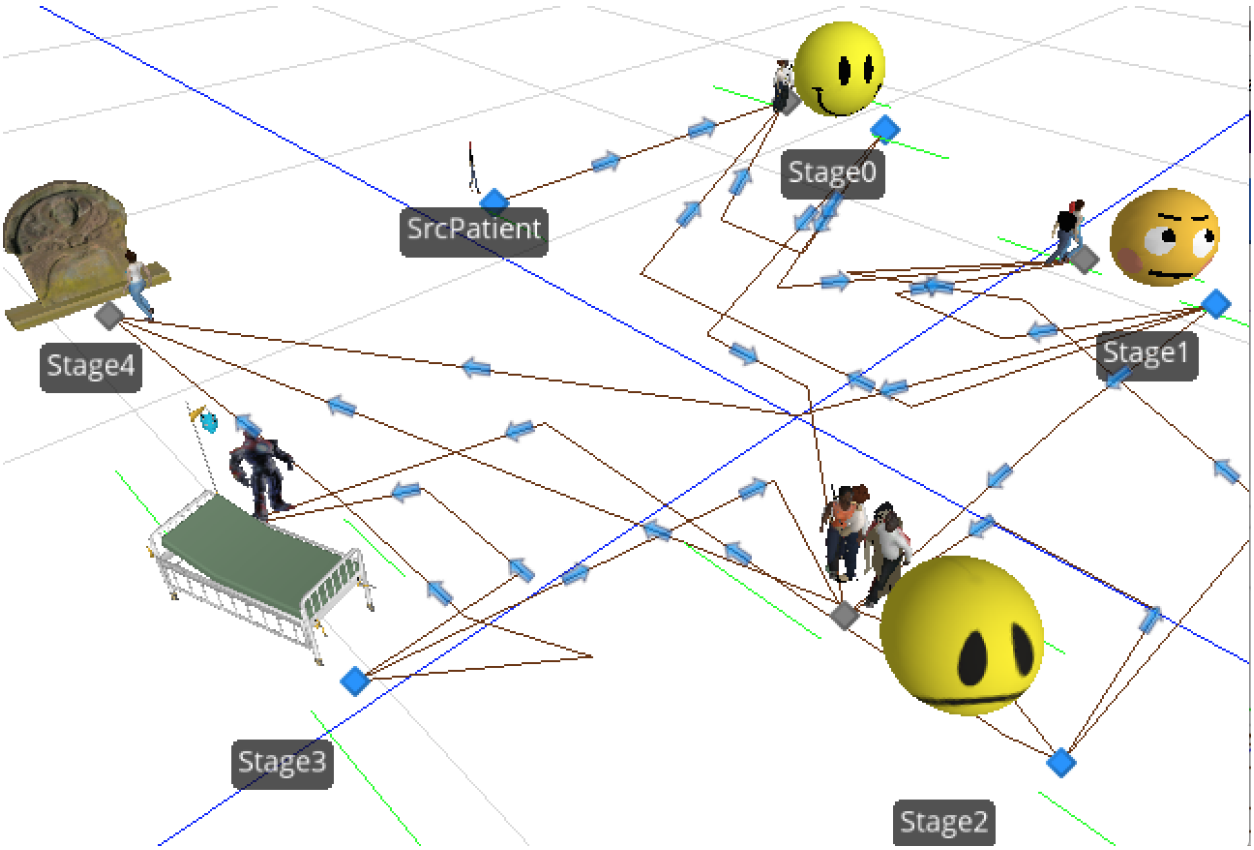
7. http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ (Cleveland.csv) [ Missing Attribute Values in the dataset: Several – replaced with os.]

**APPENDIX:**

| **Attribute** | **Range** | **Attribute name** |
|---|---|---|
| Age | [29-77] | Age |
| Sex | [0,1] | Sex |
| Resting blood pressure | | Trestbps |
| Cholesterol | [126,564] | Chol |
| Fasting blood sugar | 1 = true > 120)<br>0 = false | Fbs |
| Resting electrocardiographic results | 0 – normal<br>1 – having ST-T abnormality<br>2 – probable left ventricular hypertrophy | Restecg |
| Maximum heart rate achieved | | thalach |
| Exercise induced angina | 1=yes, 0 = no | Exang |
| ST depression induced by exercise relative to rest | | Oldpeak |
| Slope of the peak exercise ST segment | 1 = upsloping<br>2= flat<br>3 = downsloping | slope |
| Number of major vessels(0-3) colored by fluoroscopy | | ca |
| Thal | 3 = normal<br>6 = fixed defect<br>7 = reversible defect | thal |

| | | |
|---|---|---|
| Diagnosis of heart disease | 0 < 50% diameter narrowing<br>1 > 50% diameter narrowing | num |

**Table 2:** Data description




**Snapshot 1:** model simulation using the Simio software

REMINDER to Youqing: Do you have the latest snapshot for this?

| Object Type | Object Name | Data Source | Category | Data Item | Statistic | Average Total |
|---|---|---|---|---|---|---|
| Source | SrcPatient | OutputBuffer | Throughput | NumberEntered | Total | 2,500.0000 |
| | | | | NumberExited | Total | 2,500.0000 |
| Sink | State4 | [DestroyedObjects] | FlowTime | TimeInSystem | Average (Hours) | 5.6726 |
| | | | | | Maximum (Hours) | 9.0192 |
| | | | | | Minimum (Hours) | 2.0172 |
| | | | | | Observations | 868.0000 |
| | | InputBuffer | Throughput | NumberEntered | Total | 868.0000 |
| | | | | NumberExited | Total | 868.0000 |
| Server | State0 | [Resource] | Capacity | ScheduledUtilization | Percent | 0.0000 |
| | | | | UnitsAllocated | Total | 5,743.0000 |
| | | | | UnitsScheduled | Average | Infinity |
| | | | | | Maximum | Infinity |
| | | | | UnitsUtilized | Average | 0.1595 |
| | | | | | Maximum | 2,500.0000 |
| | | | ResourceState | TimeProcessing | Average (Hours) | 0.0003 |
| | | | | | Occurrences | 10.0000 |
| | | | | | Percent | 0.0278 |
| | | | | | Total (Hours) | 0.0028 |
| | | | | TimeStarved | Average (Hours) | 0.9088 |
| | | | | | Occurrences | 11.0000 |
| | | | | | Percent | 99.9722 |
| | | | | | Total (Hours) | 9.9972 |

**Snapshot 2:** Results from the simulation model

| Run | S0 - S0 | S0 - S1 | S0 - S2 | S1 - S1 | S1 - S0 | S1 - S2 | S1 - S4 |
|---|---|---|---|---|---|---|---|
| 1 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 2 | 0.57 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 3 | 0.86 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 4 | 1.06 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 5 | 1.33 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 6 | 0.43 | 0.21 | 0.36 | 0.69 | 0.15 | 0.18 | 0.04 |
| 7 | 0.43 | 0.21 | 0.36 | 0.86 | 0.15 | 0.18 | 0.04 |
| 8 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 9 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 10 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 11 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 12 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 13 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 14 | 0.43 | 0.21 | 0.36 | 0.63 | 0.15 | 0.18 | 0.04 |
| 15 | 0.57 | 0.21 | 0.36 | 0.69 | 0.15 | 0.18 | 0.04 |
| 16 | 1.33 | 0.21 | 0.36 | 0.86 | 0.15 | 0.18 | 0.04 |

**Snapshot 3:** Setup 1 of 2 from the simulation model

| S2 - S2 | S2 - S1 | S2 - S3 | S2 - S4 | S3 - S3 | S3 - S2 | S3 - S4 | Number of entities in State4 |
|---|---|---|---|---|---|---|---|
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 927 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 884 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 814 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 780 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 721 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 904 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 852 |
| 0.58 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 856 |
| 0.87 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 780 |
| 1.08 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 732 |
| 1.35 | 0.25 | 0.27 | 0.06 | 0.57 | 0.29 | 0.14 | 702 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.65 | 0.29 | 0.14 | 909 |
| 0.42 | 0.25 | 0.27 | 0.06 | 0.8 | 0.29 | 0.14 | 878 |
| 0.42 | 0.25 | 0.27 | 0.06 | 1.0 | 0.29 | 0.14 | 841 |
| 0.58 | 0.25 | 0.27 | 0.06 | 0.65 | 0.29 | 0.14 | 832 |
| 1.35 | 0.25 | 0.27 | 0.06 | 1.0 | 0.29 | 0.14 | 492 |

**Snapshot 4:** Setup 2 of 2 from the simulation model

| Run | S0 - S0(%) | S1 - S1 (%) | S2 - S2 (%) | S3- S3 (%) | Number in State4 |
|---|---|---|---|---|---|
| 1 | 43% | 63% | 42% | 57% | 927 |
| 2 | 50% | 63% | 42% | 57% | 884 |
| 3 | 60% | 63% | 42% | 57% | 814 |
| 4 | 65% | 63% | 42% | 57% | 780 |
| 5 | 70% | 63% | 42% | 57% | 721 |
| 6 | 43% | 65% | 42% | 57% | 904 |
| 7 | 43% | 70% | 42% | 57% | 852 |
| 8 | 43% | 63% | 50% | 57% | 856 |
| 9 | 43% | 63% | 60% | 57% | 780 |
| 10 | 43% | 63% | 65% | 57% | 732 |
| 11 | 43% | 63% | 70% | 57% | 702 |
| 12 | 43% | 63% | 42% | 60% | 909 |
| 13 | 43% | 63% | 42% | 65% | 878 |
| 14 | 43% | 63% | 42% | 70% | 841 |
| 15 | 50% | 65% | 50% | 60% | 832 |
| 16 | 70% | 70% | 70% | 70% | 492 |

**Snapshot 5:** Final results from the simulation model

**APPENDIX:**

**R Code:**

```{r}

#install.packages("msm")
setwd("/Users/tulasiramarao/Documents/CUNY-SPRING2016/IS604-Simulation/FinalProjectHeart")

# read the patient's heart disease data from the UCI machine repository
pdata <- read.csv("cleveland.csv",header=TRUE, sep=",",stringsAsFactors=FALSE)

# attach the dataset
head(pdata)
nrow(pdata)
attach(pdata)
unique(Num)

require(msm)
head(cav)
print(cav[1:10,])
 statetable.msm(state, PTNUM, data=cav)
#write.csv(cav,file="cav.csv")

# get first 303 rows and first 3 columns
predata <-cav[1:303,1:8]

# combine with the patientdata from UCL
newdata <- cbind(predata,pdata)
head(newdata)

#drop columns
df <- subset(newdata, select = -c(Age,dage,Sex,pdiag,cumrej) )
head(df)
df$age <- round(df$age,0)
df$years <- round(df$years,0)

head(df)
df$state[df$state == 1]
write.csv(df,file="heartData.csv")
# manually change heartData.csv to add zero states
# that is ( choose 1s and replace with 0 for state column)
# load that file here

```

```{r}
```

```
# load the modified data
heartdata2 <- read.csv("heartData2.csv",header=TRUE, sep=",",stringsAsFactors=FALSE)
head(heartdata2,30)
unique(heartdata2$state)

# Now change the states from 0 to 4 to 1 to 5
heartdata2$state[heartdata2$state==0] <- 10
heartdata2$state[heartdata2$state==1] <- 11
heartdata2$state[heartdata2$state==2] <- 12
heartdata2$state[heartdata2$state==3] <- 13
heartdata2$state[heartdata2$state==4] <- 14

heartdata2$state[heartdata2$state==10] <- 1
heartdata2$state[heartdata2$state==11] <- 2
heartdata2$state[heartdata2$state==12] <- 3
heartdata2$state[heartdata2$state==13] <- 4
heartdata2$state[heartdata2$state==14] <- 5
unique(heartdata2$state)

# This is the probability matrix
# Q <- rbind ( c(0.42, 0.21, 0.35, 0, 0),
# + c(0.14, 0.63, 0.18, 0, 0.36),
# + c(0, 0.25, 0.41, 0.27, 0.05),
# + c(0, 0, 0.28,0.57,0.14 ),
# + c(0, 0, 0,0.61,0.38 ))

Q <- rbind ( c(0.43, 0.21, 0.36, 0, 0),
+ c(0.15, 0.63, 0.18, 0, 0.04),
+ c(0, 0.25, 0.42, 0.27, 0.06),
+ c(0, 0, 0.29,0.57,0.14 ),
+ c(0, 0, 0, 0, 1 ))

Q

# Do the simulation using the crude method since there are not many changes between the
# observations
Q.crude <- crudeinits.msm(state ~ years, PTNUM, data=heartdata2, qmatrix=Q)

Q.crude

```
```{r}
head(heartdata2,121)
#show the maximum likelihood estimates and 95% confidence intervals
```

```
heart.msm <-msm(formula = state ~ years, PTNUM, data = heartdata2, qmatrix = Q,
deathexact = 5,method = "BFGS",control = list(fnscale = 4000, maxit = 10000))

heart.msm
printold.msm(heart.msm)


# transition probability matrix
# 10 year transition probability ->
pmatrix.msm(heart.msm,t=10)

# probability that each state is next
pnext.msm(heart.msm)

totlos.msm(heart.msm)

#Survival plots
plot(heart.msm, legend.pos=c(8, 1))
```

```
```