# Markov Chains as a Predictive Analytics Technique Using SAS/IML

Gregory McKinney, Gerard Group

## Abstract

As a predictive analytics approach, Markov Chains provide a powerful tool for modeling complex multi-state dynamic systems. Because of this power, Markov Chain models have been used to address many real-world problems including disease progression, loan portfolio risk, measuring marketing campaign effectiveness, evaluation of clinical drug trials, etc. The list of potential applications of Markov chains is endless. SAS/IML provides a potent framework for implementing Markov chain models in your organization. Sample code for the use of SAS/IML for Markov chain models is included.

## Introduction

A stochastic process (a random process) is a mathematical model of a system that changes over time in a probabilistic manner. The status of the system is represented by a set of random variables. For comparison, a deterministic process can represent a system that will evolve in only one way. Given a starting condition, a deterministic process will always evolve in exact same way. Not so with a stochastic process.

A Markov Chain is a special type of stochastic process where the system evolves between a finite set of states and the possible changes are described by a transition probability matrix. The "Markov Property" implies that the conditional probabilities for transitions to the next state only depend upon the current state. For this reason, Markov Chains are referred to a "memoryless".

Markov Chains can be used to describe processes in a wide variety of fields including economics, sociology, finance, physics, chemistry, biology, medicine, etc. This paper looks the use of Markov Chains to describe disease progression,

## Markov Chains

### State Space

A Markov Chain model begins with a finite set of states that are mutually exclusive and exhaustive. At any point in time, the process is in one and only one state. The model described in this paper is a discrete time process. It is possible to define a Markov Chain as a continuous time process, but that is outside the scope of this paper.

### Transition Probability Matrix

The potential transitions from one state to another in a unit of time are described by a matrix of conditional probabilities. Each row represents a current state while a column represents the next state.

The probability that the process whenever in state i moves next (one unit of time) into state j is denoted by $\boldsymbol{p_{ij}}$. This is called a one-step transition probability.

$$p_{ij} = Pr\{X_{n+1} = j \mid X_n = i\}$$

These transition probabilities can be arranged into a matrix where each row represent the current state and columns represent the future state.

$$\boldsymbol{P}^{(1)} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

Because the state space is mutually exclusive and exhaustive, the probabilities in each row of the transition matrix sum to 1. For each i

$$\sum_{j \in S} p_{ij} = 1$$

Once we have a transition matrix containing one-step probabilities, we can calculate the probabilities of transition over multiple time periods. The probability of transitioning from state I to state j in 2 time periods is given by:

$$p_{ij}^{(2)} = \sum_k p_{ik}\, p_{kj}$$

Generalizing this we calculate the probability of transitioning from I to j in n time periods as:

$$p_{ij}^{(n)} = \sum_k p_{ik}^{(n-1)}\, p_{kj}$$

Using matrix multiplication, we have this:

$$P^{(n)} = P^{(n-1)}\, P^{(1)}$$

Because these calculations can be treated as Matrix calculations, SAS/IML is a useful and powerful tool for building and manipulating Markov Chains.

In some cases, it is impossible to transition out of a state once it is reached. These states are called "absorbing states." If it is possible for every state to eventually reach an absorbing state, then we have an "absorbing Markov chain."

# Disease Progression Model

## Healthcare Applications

A potentially useful application of Markov chains is to model disease progression. For numerous planning purposes, it is useful for healthcare plans and providers to predict the future numbers of individuals with specific medical conditions.

We define disease progression as the change in a medical condition as it moves from early (less severe) stages to later (more severe) stages.

Healthcare Plans and Providers have a need to estimate the number of members with specific conditions for medical utilization and capacity planning purposes. Additionally, other healthcare uses of Markov Chain models include;

- Effectiveness of medical treatments
- Quality-adjusted life years
- Effectiveness of lifestyle programs:
    - Smoking cessation
    - Alcohol / substance abuse
    - Exercise
    - Diabetes management

# Model Specification

## State Space

A Markov Chain model requires a state space that is exhaustive and mutually exclusive. Defining the disease-progression states based upon ICD-10 (International Statistical Classification of Disease) codes would result in huge number of states that are not mutually exclusive.

The U.S. Centers for Medicare and Medicaid Services (CMS) maintains a Hierarchal Condition Category model that maps tens of thousands of diagnosis codes into condition categories. These condition categories are then grouped into disease families (Neoplasms, Renal Failure, Diabetes, etc.).

To illustrate, here are the categories within the Neoplasm (Cancer) disease family:

| | |
|---|---|
| HCC8 | Metastatic Cancer and Acute Leukemia |
| HCC9 | Lung and Other Severe Cancers |
| HCC10 | Lymphoma and Other Cancers |
| HCC11 | Colorectal, Bladder, and Other Cancers |
| HCC12 | Breast, Prostate, and Other Cancers and Tumors |

HCC8 is considered to be the most severe (later stage) condition category while HCC12 is the early stage (less severe) condition category.

For example, here are some of the ICD-10 diagnosis codes that map to HCC8 (Condition Category 8 or Metastatic Cancer and Acute Cancer):

*Table 1 ICD-10 Diagnosis Codes in HCC8*

| | |
|---|---|
| … | |
| C800 | Disseminated malignant neoplasm, unspecified |
| C9100 | Acute lymphoblastic leukemia not having achieved remission |
| C9101 | Acute lymphoblastic leukemia, in remission |
| C9102 | Acute lymphoblastic leukemia, in relapse |
| C9200 | Acute myeloblastic leukemia, not having achieved remission |
| C9201 | Acute myeloblastic leukemia, in remission |
| C9202 | Acute myeloblastic leukemia, in relapse |
| C9240 | Acute promyelocytic leukemia, not having achieved remission |
| C9241 | Acute promyelocytic leukemia, in remission |
| C9242 | Acute promyelocytic leukemia, in relapse |
| C9250 | Acute myelomonocytic leukemia, not having achieved remission |
| C9251 | Acute myelomonocytic leukemia, in remission |
| C9252 | Acute myelomonocytic leukemia, in relapse |
| C9260 | Acute myeloid leukemia with 11q23-abnormality not having achieved remission |
| C9261 | Acute myeloid leukemia with 11q23-abnormality in remission |
| C9262 | Acute myeloid leukemia with 11q23-abnormality in relapse |
| C92A0 | Acute myeloid leukemia with multilineage dysplasia, not having achieved remission |
| C92A1 | Acute myeloid leukemia with multilineage dysplasia, in remission |
| C92A2 | Acute myeloid leukemia with multilineage dysplasia, in relapse |
| C9300 | Acute monoblastic/monocytic leukemia, not having achieved remission |
| … | |

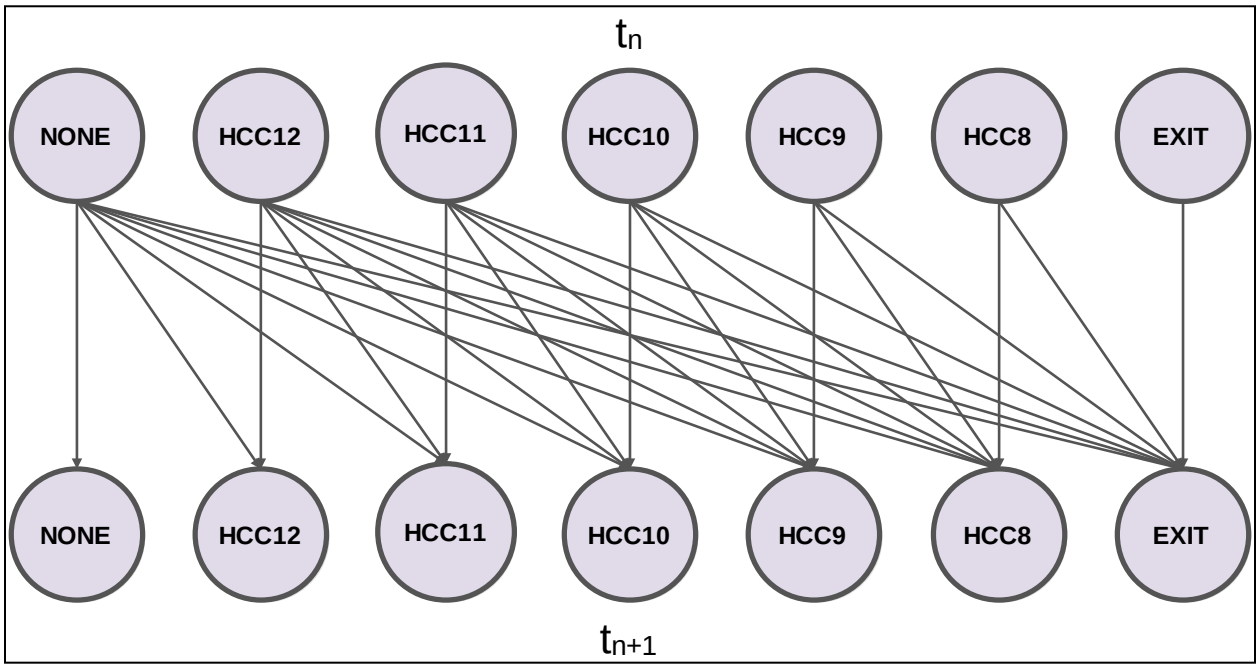
Hierarchy rules are then applied ensuring that an individual is in no more than one category within a disease family.

| Hierarchical Condition Category (HCC) | If individual has this HCC… | …then drop these HCC(s) |
|---|---|---|
| … | … | … |
| **8** | **Metastatic Cancer and Acute Leukemia** | **9, 10, 11, 12** |
| **9** | **Lung and Other Severe Cancers** | **10, 11, 12** |
| **10** | **Lymphoma and Other Cancers** | **11, 12** |
| **11** | **Colorectal, Bladder, and Other Cancers** | **12** |
| … | … | … |

The CMS HCC structure provides a convenient and widely accepted method for grouping diagnoses into meaningful condition categories.

## Transition Probabilities

This model deals with chronic conditions, we do not expect to see transactions from more severe to less severe. Therefore, the transition diagram for this disease family looks something like this.




Translating the transition patterns into matrix form, the transition matrix looks something like this:

| | | To State -→ None | HCC12 | HCC11 | HCC10 | HCC9 | HCC8 | Exit |
|---|---|---|---|---|---|---|---|---|
| ← From State | None | $p_{11}$ | $p_{12}$ | $p_{13}$ | $p_{14}$ | $p_{15}$ | $p_{16}$ | $p_{17}$ |
| | HCC12 | 0 | $p_{22}$ | $p_{23}$ | $p_{24}$ | $p_{25}$ | $p_{26}$ | $p_{27}$ |
| | HCC11 | 0 | 0 | $p_{33}$ | $p_{34}$ | $p_{35}$ | $p_{36}$ | $p_{37}$ |
| | HCC10 | 0 | 0 | 0 | $p_{44}$ | $p_{45}$ | $p_{46}$ | $p_{47}$ |
| | HCC9 | 0 | 0 | 0 | 0 | $p_{55}$ | $p_{56}$ | $p_{57}$ |
| | HCC8 | 0 | 0 | 0 | 0 | 0 | $p_{66}$ | $p_{67}$ |
| | EXIT | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

A data warehouse with diagnosis histories for millions of individuals covering multiple years of history was used to calculate the individual transition probabilities.

## Simulation

The expected number of individuals in each state in the future can be estimated by simply multiplying the current distribution vector by the transition matrix.

While useful, point estimates provide no information about the uncertainty around that estimate. Therefore, a simulation approach was used to distributions around each future state.

*Figure 1 Markov Chain Simulation for Disease Progression*

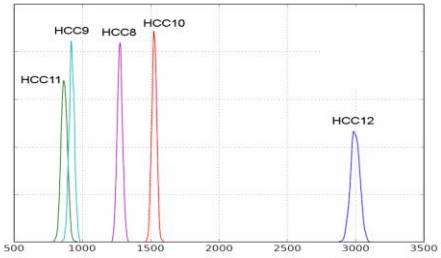
| | |
|---|---|
| 1 | Number of Time Periods |
| 1000 | Number of Indepenent Trials |

| | | NONE | HCC12 | HCC11 | HCC10 | HCC9 | HCC8 | EXIT |
|---|---|---|---|---|---|---|---|---|
| NONE | 34,755 | 0.9834 | 0.0060 | 0.0031 | 0.0019 | 0.0017 | 0.0016 | 0.0024 |
| HCC12 | 2,435 | 0.0000 | 0.8381 | 0.0424 | 0.0222 | 0.0481 | 0.0331 | 0.0160 |
| HCC11 | 415 | 0.0000 | 0.0000 | 0.8652 | 0.0510 | 0.0391 | 0.0312 | 0.0135 |
| HCC10 | 1,030 | 0.0000 | 0.0000 | 0.0000 | 0.9694 | 0.0154 | 0.0111 | 0.0041 |
| HCC9 | 475 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9801 | 0.0151 | 0.0048 |
| HCC8 | 890 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.9161 | 0.0839 |
| EXIT | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 1.0000 |
| Total | 40,000 | | | | | | | |

| Trial # | NONE | HCC12 | HCC11 | HCC10 | HCC9 | HCC8 | EXIT | Total |
|---|---|---|---|---|---|---|---|---|
| 1 | 32,632 | 2,792 | 799 | 1,650 | 855 | 1,217 | 55 | 40,000 |
| 2 | 33,036 | 2,668 | 731 | 1,560 | 792 | 1,173 | 40 | 40,000 |
| 3 | 32,535 | 2,821 | 817 | 1,668 | 870 | 1,230 | 59 | 40,000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1000 | 33,033 | 2,653 | 742 | 1,549 | 797 | 1,193 | 33 | 40,000 |

Simulation Results Summarized

| %-tile | NONE | HCC12 | HCC11 | HCC10 | HCC9 | HCC8 | EXIT |
|---|---|---|---|---|---|---|---|
| 1 | 32,753 | 2,646 | 722 | 1,544 | 778 | 1,144 | 29 |
| 25 | 32,817 | 2,703 | 753 | 1,587 | 809 | 1,173 | 40 |
| 50 | 32,847 | 2,730 | 767 | 1,604 | 822 | 1,186 | 44 |
| 75 | 32,878 | 2,755 | 780 | 1,622 | 835 | 1,195 | 49 |
| 99 | 32,957 | 2,821 | 817 | 1,668 | 870 | 1,230 | 59 |




Please note, the data used in this paper is purely fictitious, created solely for the purposes of illustration. No business conclusions should be drawn from this data.

## SAS/IML Implementation

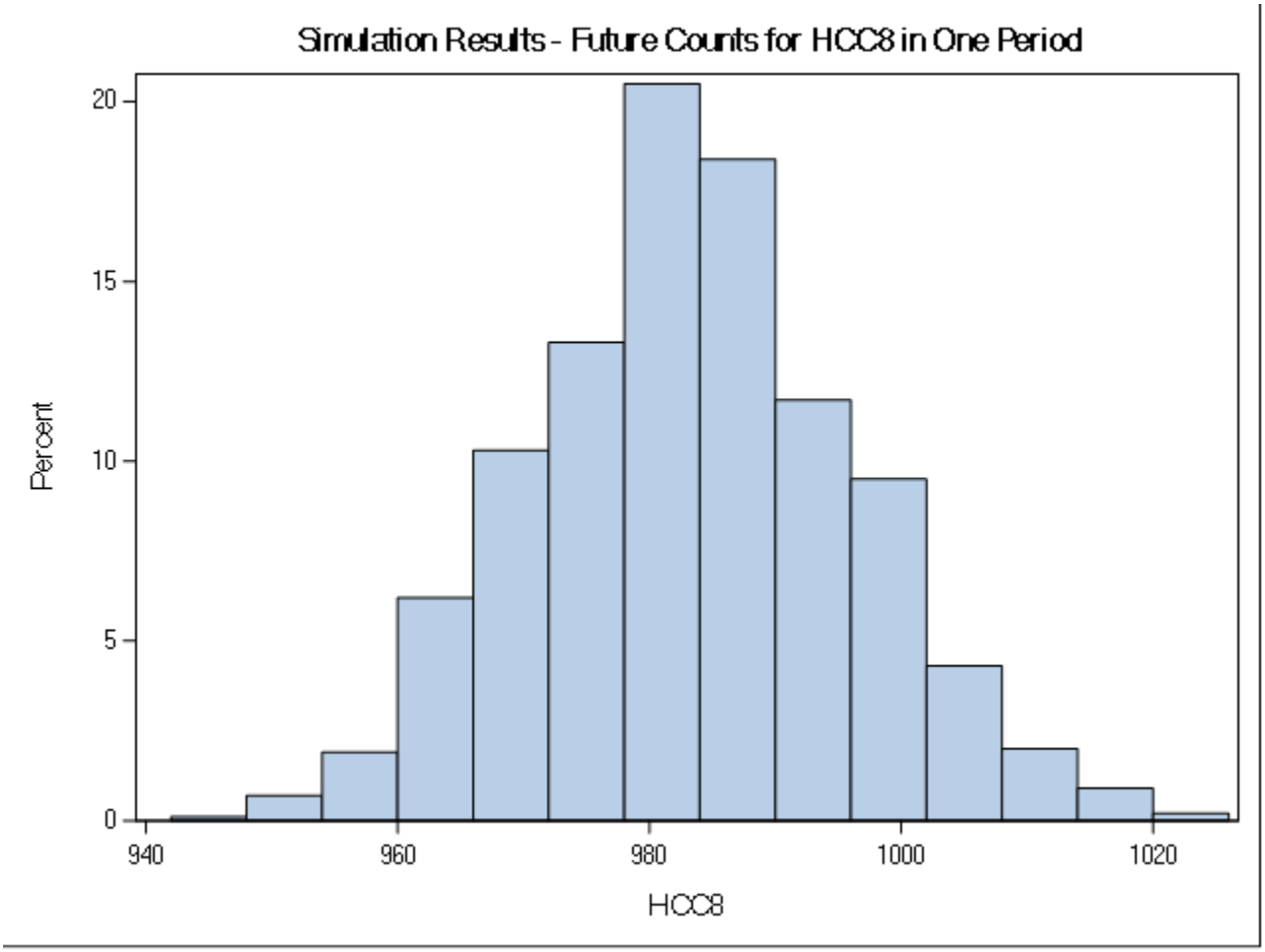
SAS/IML is a powerful high-level matrix programming language making it an ideal tool for implementing Markov Chain models. The code below implants a simple simulation of a population of 40,000 individuals with a transition matrix covering seven states (e.g., NONE, HCC12, HCC11, HCC10, HCC9, HCC8, EXIT).

```
proc iml;
/*
        P1 is single period transition matrix...
        Rows are current state, Columns are next state
*/

/*      NONE    HCC12   HCC11   HCC10   HCC9    HCC8    EXIT */
P1 = {  0.9834  0.0060  0.0030  0.0019  0.0017  0.0016  0.0024, /* NONE  */
        0       0.8381  0.0424  0.0222  0.0481  0.0331  0.0161, /* HCC12 */
        0       0       0.8652  0.0510  0.0391  0.0312  0.0135, /* HCC11 */
        0       0       0       0.9694  0.0154  0.0111  0.0041, /* HCC10 */
        0       0       0       0       0.9801  0.0151  0.0048, /* HCC9  */
        0       0       0       0       0       0.9161  0.0839, /* HCC8  */
        0       0       0       0       0       0       1 };     /* EXIT  */


states= {"NONE", "HCC12", "HCC11", "HCC10", "HCC9", "HCC8", "EXIT"};
print P1[r= states c= states l= "Transition Matrix - 1 Period"];

/* verify that each row in the transition matrix sums to 1 */

SumCheck= P1[ ,+];
print SumCheck[r=states l="Transition Matrix Rows Should Sum to 1"];

II= {34755 2435 415 1030 475 890 0}; /* start conditions for 40K member health plan */
```

```
ni= II[+]; /* number of individuals to simulate */

E= II*P1;
print E[c= states l= "Expected Counts by State - 1 Period in Future"];

dim= dimension(P1);
nstates= dim[2]; /* number of states */

/* Simulation */

itr= 1000; /* number of iterations */
Rslts= j(itr, nstates, 0); /* results matrix to accumulate counts from simulation */
CP= j(nstates, nstates, 0); /* matrix will hold cumulative probabilities */

/* Calculate cumulative probabilities along each row */
do i= 1 to nstates; /* rows */
       cx= 0;
       do j= 1 to nstates; /* columns */
              CP[i,j]= P1[i,j]+cx;
              cx= CP[i,j];
       end; /* columns */
end; /* rows */

call randseed(250339); /* seed random number generator */

do i= 1 to itr; /* iteration loop */
       xs= j(itr*ni, 1);
       call randgen(xs, "Uniform"); /* vector of random numbers */

       do j= 1 to nstates; /* current state */
              do k= 1 to II[j]; /* individual in current state m */
                     xx= xs[k];
                     do m= nstates to 1 by -1; /* future state */
                            if xx <= CP[j,m] then fs= m;
                            else m= 0;
                     end; /* future state */
                     Rslts[i,fs]= Rslts[i,fs]+1;
              end; /* individual */
       end; /* current state */
end; /* iteration */

/* produce SAS dataset with simulation results */
create WORK.SimRslts  var{NONE HCC12 HCC11 HCC10 HCC9 HCC8 EXIT};
append from Rslts;

quit;

proc univariate DATA=WORK.SimRslts; run;

proc sgplot DATA= WORK.SimRslts;
       title 'Simulation Results - Counts for Future States';
       histogram HCC12;
       histogram HCC11;
       histogram HCC10;
       histogram HCC9;
       histogram HCC8;
run;

proc sgplot DATA= WORK.SimRslts;
       title 'Simulation Results - Future Counts for HCC8 in One Period';
       histogram HCC8;
run;

proc sgplot DATA= WORK.SimRslts;
       title 'Simulation Results - Future Counts for HCC11 in One Period';
       histogram HCC11;
run;
```

Once the simulation has produced that data, all of the SAS analytics tools are available to analyze, display, and summarize the results.

The Simulation provides us with distributions around future states.




Those distributions are displayed graphically as a histogram but also in tabular form.

| | FUTURE STATES | | | | |
|---|---|---|---|---|---|
| **%-tile** | **HCC12** | **HCC11** | **HCC10** | **HCC9** | **HCC8** |
| **95%** | 2,286 | 592 | 1,162 | 699 | 1,004 |
| **75%** | 2,264 | 577 | 1,149 | 684 | 991 |
| **50%** | 2,248 | 567 | 1,139 | 674 | 983 |
| **25%** | 2,233 | 556 | 1,131 | 664 | 974 |
| **5%** | 2,212 | 542 | 1,119 | 650 | 962 |

## Some Other Applications for Markov Chains

Besides healthcare applications, the possible applications for Markov Chains is endless. For example, Markov Chains have been used to model these situations:

- Commercial Loan (Corporate Debt) Portfolio Risk

- Consumer Loan Portfolio Risk
- System Reliability (System Availability)
- Sales Cycle / Marketing Campaign Effectiveness
- Nonprofit Donor Lifecycle
- Social Worker (or Probation Officer) Workload

The above list is just a small cross-section of the practical applications for Markov Chains.

## Conclusions

Markov Chains are well suited for modeling dynamic systems that evolve through a finite set of states. These models can provide useful estimates of future conditions.

By using Markov Chain simulation, distributions for each state can be estimated. These distributions are more insightful than point estimates since they provide information regarding uncertainty around the expected value.

For disease progression models, the CMS HCC model provides a convenient and widely accepted structure for grouping diagnoses into meaningful condition categories. These categories are then the state space for the Markov Chain Model.

A data warehouse sufficient granularity and history is required to calculate the transition probabilities.

## References


- Çinlar, Erhan. *Introduction to Stochastic Processes*, Dover Books on Mathematics (2012)
- Wicklin, Rick. Getting Started with the SAS/IML Language (2013)
- Pope, G.C., R.P. Ellis, A.S. Ash, et al. "Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment." Final Report for Health Care Financing Administration. (Dec 2000).
- Pope, G.C., J. Kautter, M.J. Ingber, et al., "Evaluation of the CMS-HCC Risk Adjustment Model." Final Report to CMS, Contract HHSM-500-2005-000291 TO 0006
- ;pPalmer, S., M Sculpher, Z Phillips, M Robinson, L Ginnelly, A Bakhai, et al. "Management of non-ST Elevation Acute Coronary Syndrome: How Cost Effective are Glycoprotein IIb/IIIa Antagonists in the UK National Health Service?". International Cardiology 100 (2005) 229-240.


## Contact Information

Your comments and questions are welcomed. Contact the author:


Greg McKinney
gmckinne@pacbell.net



SAS and SAS/IML are trademarks of SAS Institute Inc. in the USA and other countries.