**Data Analytics Project Proposal**

Tricia M. Lang

Southern New Hampshire University

DAT 510: Foundations of Data

Prof. Domingo Genao

6 March 2022

**Introduction**

**Background**

A health insurance program director has approached this data analytics team seeking assistance in identifying a specific population of their members. The program director is looking to create programs targeting individuals who have had a heart attack that have a higher risk of having a second heart attack. The insurance company is currently storing lakes of member data, according to Perficient (2022), including information on those who have had at least one heart attack.

Health insurance companies currently utilize data analytics for the purposes of pricing and reducing costs (Center for Insurance Policy and Research, 2021). The data also identifies individuals with higher risks overall to pool those individuals by risk of high utilization of insurance (Society of Actuaries, 2022). While there are predictive analytics involved in the actuarial analyses, the application is for a different purpose than is being proposed at this time.

**Data Sources**

The health insurance company has supplied data that is generated from their medical claims database (North, 2012). The benefit of using this database is that there is an abundance of data available, including diagnoses, demographic information, medical records, and co-morbid medical or mental health issues. The database also shows which interventions have been billed through the insurance company. Limitations to this dataset are individuals who may have had a heart attack prior to having this insurance not being included. Additionally, if an individual started on medications to bring down cholesterol levels following the first heart attack, it is not clear whether this risk factor has been mitigated. This data may not capture individuals who have sought treatment for anxiety without utilizing insurance. Some information contained in the insurance database can also be outdated, such as marital status. Individuals who have divorced or are widowed may show as “spouse” on the policy but no longer be married. Overall, the data that is being used can be informative in a preliminary and exploratory manner, but not as accurate as with additional or better data.

**Data Needs**

Additional data that is available to the health insurance company is ongoing treatments following the first heart attack. Information such as cholesterol levels in subsequent blood draws, not just on the day of the first heart attack, may indicate whether there is a relationship between a decreased cholesterol level and subsequent heart attacks. Individuals may also have heart attacks who are on cholesterol-reducing medications and this information would be available through pharmacy claims. Medications alone may be identified as being sufficient, or insufficient, in reducing the probability of a second heart attack, as well as the possibility that specific medications are more effective than others given various risk factors. The same may be true for weight, as a weight loss following the first heart attack may also be related to the prevalence of second heart attacks.

Another data point that may be available and useful is activity level. This is often, but not always, collected by the insurance company can include step monitoring through wearable tracking devices. Some insurance companies incentivize the completion of a certain number of steps with a reduced premium cost or other reward, which is automatically synchronized between the tracking app and the insurance company (Ingraham, 2018).

Should anxiety and stress management also play a role, as is hypothesized by the program director, co-occurring mental health diagnoses may also be a useful datapoint. A diagnosed anxiety or depressive disorder or a substance use disorder may impact numerous factors related to having a second heart attack. Compliance with medications, ability to initiate exercise, missing follow-up appointments due to other co-morbid factors, and numerous other confounds may impact the likelihood of a second heart attack (Martin et al., 2005).

Support systems are a factor in overall health as well. If individuals are engaged in mental health treatment or case management services may be sufficient support, others may have family and friends that are supportive, and some may not identify any supports. While capturing this data is not necessarily straightforward for the insurance company, there are medical records associated with the first heart attack that can be made available with consent of the patient. During the course of treatment for the first heart attack, a psychosocial assessment is completed which identifies what supports an individual has.

**Data Analytics Initiative**

This project aims to use predictive analytics to identify the higher risk individuals to provide targeted programs to reduce that risk. Utilizing logistic regression, the probability of whether a person will have a second heart attack can be predicted based on the predictor variables presented by the health insurance company of age, marital status, gender, weight category, cholesterol level at time of first heart attack, attendance of a stress management course, and natural anxiety scores (North, 2012). Predictions will be based on the training data for individuals who have and have not had second heart attacks, which will be used to train the model on which individuals have had second heart attacks and which have not with the same predictor variables. The outcome is a binary response of “yes” or “no” for likelihood of a second heart attack and a confidence level for that response. This output will help the program director identify those who are likely to have a second heart attack, as well as those who are identified as not likely but the confidence level for that outcome is low.

Some of the factors that have been identified as key risk factors are measured only at the time of the first heart attack. An analysis of the same data points at the time of the second heart attack may help refine the risk factors and interventions as the data may change from the first to the second heart attack for individuals. These data points also being taken for those who did not have a second heart attack, at the average time from initial for those who did have a second heart attack, would also lend further insight into whether there were changes in data that impacted their cardiac health.

Analyzing information around the time of the second heart attack in the training data set will add business value as it may be determined that certain factors or a combination of factors are more effective in reducing the likelihood of a second heart attack. The treatment programs can then focus on these areas. Having this information will allow best use of the more effective targeted interventions to reduce the likelihood of subsequent heart attacks.

**Proposal**

**Goals**

The goal of this initiative is to have a tool to identify individuals for a healthcare program. Specifically, the individuals who the health insurance program manager wants to target are those who have had one heart attack and have an increased risk of possibly having a second. To do this, the risk factors associated with individuals who have had two heart attacks are identified and used to train a predictive model. This predictive model is used to evaluate the risk factors of the individuals being screened after one heart attack and identify the probability of having a second heart attack for each person. The individuals with greater probabilities will be identified for enrollment in the programs to reduce the risk of a second heart attack occurring.

Health insurance companies aim to improve the health and wellness of the individuals they serve. This is both to promote wellness in the community and reduce costs. This project is in support of improving the health and wellness of the individuals served by attempting to identify those who are at risk of a grave medical event and providing supports and guidance to reduce that risk. At the same time, the programs implemented would reduce cost of care for the individuals which also maximizes profits for the insurance company. By preventing or reducing the incidence of second heart attacks, this project supports the mission of the health insurance company.

The threshold of the confidence level identified by the project manager to identify participates is 0.450. This falls slightly below the typical 0.500 mark for a positive outcome. The reasoning for this is to err on the side of caution when providing preventative supports and assist those who may have marginally less than an identified 50% chance of having a second heart attack. It is expected that the model will not be perfect in predicting and there will be individuals who fall above 0.500 who never have a second heart attack, as well as those who fall below 0.500 who do. Ideally this will be minimized by using the appropriate predictor variables for the model.

The success of this project will be measured by creating a predictive model that can accurately identify individuals who are at a higher risk of having a second heart attack. This can be determined by having the training data partitioned into a training subset and a test subset. Since the training data contains the outcomes of having or not having a second heart attack as well as the measures of different risk factors, the model that is created can be tested with a subset of the data that has known outcomes for comparison. A successful predictive model will have high concordance with the test data.

**Data Analytics Life Cycle**

There are six steps to the data analytics life cycle: understanding business issues, understanding the data, preparing the data, performing exploratory analysis and modeling, validating the data, and visualizing and presenting the findings (Data Science Process Alliance, 2021). Following these six steps in an iterative approach is important to provide the structure and apply standards to the data project. As the project progresses through these stages, it may become apparent that previous stages need to be revisited to ensure the project is continuing in a successful direction.

***Understand Business Issues***

The life cycle begins with identifying the needs of the project. This includes identifying the goals of the project, the information required to perform the analysis of the data, as well as the methods that will best achieve the desired output (Data Science Process Alliance, 2021). It will be very important for the needs of the project to be clear before beginning, to ensure that the analysis will be aligned with what the project was designed to achieve (Northeastern University Graduate Programs, 2021). Once this groundwork has been outlined, the project will have a clearly defined framework that can be referred to throughout the project to ensure that the goals and outputs remain aligned as the project progresses.

***Understand the Data Set***

The next step in the life cycle of this project is to understand the data set. There are 690 rows of data in the scoring data set and 138 in the testing data set. The data is categorized by 7 variables in the scoring data and 8 in the testing data, with the difference being the variable of whether the individual had a second heart attack or not, indicated by “yes” or “no”. The other seven variables are all represented by integers and include the categories of: age, marital status, gender, weight category, cholesterol level, stress management course, and anxiety trait.

Age and cholesterol level are integers by nature and reported as such for these data sets. The data for marital status is coded from 0 to 3 for single, married, divorced, and widowed. Gender is categorized as 0 for female and 1 for male. The weight category is identified as 0 for a normal weight range, 1 for overweight, and 2 for obese. The attendance of a stress management course is represented by 0 for no and 1 for yes. The anxiety trait is a score from 0 to 100 generated by a test of natural anxiety that was performed a short time after the individual had their first heart attack (North, 2018).

***Prepare the Data***

Once the business objectives and the data are understood, the preparation of the data will begin. The data preparation stage includes cleaning the data. Removing any duplicates, identifying missing data, and making the data useable when imported into a programming environment. As these data sets are csv files with a relatively small number of rows, they can be cleaned using Excel. Many of the different functions in Excel can identify data errors and correct some very easily, such as TRIM to remove spaces before and after entries and UPPER or LOWER to ensure the string data is uniform (Felvegi et al., n.d.). One of the noticeable omissions in this data is the use of a key for identifying the line of data for a corresponding subject. Adding this key will also allow the output provided to correspond to an individual known to the insurance company.

***Perform Exploratory Analysis and Modeling***

Once the data has been cleaned, the models to test the data will be created. For this predictive model, a linear regression will likely be the best analysis, but the model chosen may change based on the outcomes produced. Using a linear regression for this project will identify statistically significant risk factors and create a predictive model for the probability of a second heart attack with future data.

***Validate the Data***

The data validation phase of the life cycle is when the output of the models will be evaluated to determine if any of the data appears to be incorrect and needs further cleaning. It also will identify whether the model is producing the desired output or if a different method should be chosen and a new model built. This is the final check of the project before it is presented to the insurance company. It is likely that this will be the step that prompts returning to previous steps, possibly multiple times, to correct, refine, and clarify the data set and model to ensure it is meeting the objective before completion.

***Visualize and Present Findings***

The final step of the data analysis life cycle will be to visualize and present the findings to the insurance company. Using visualizations to present the data to a client allows for a presentation that is easier to understand. The test data set will produce an output in chart form, as in Figure 1. It clearly outlines with a “yes” or “no” whether someone is predicted to have a second heart attack. It also lists the confidence of that prediction.

Showing the relationship between different risk factors and the probability of having a second heart attack with a chart or graph will make obvious the trend and the strengths of relationships without having to compare numbers. Figure 2 illustrates an example of how the risk factors will show the probability of having a second heart attack. In conjunction with the data from the chart in Figure 1, it will be clear to see how the confidence can be traced along the curve, or Sigmoid, showing the data point falls closer to the cluster of individuals at 0 for no second heart attac) or 1 for those who had a second heart attack. This will help to communicate the findings from this project to the company with clarity and accessibility.

**Value of Life Cycle**

The life cycle for this data project will ensure that the results are what the insurance company needs to move forward with their initiative. The data life cycle, in general, allows for the quality of the data to be maintained during the project and the output to be valuable to the business that is requesting it (Marget, 2021). The data that is utilized will be understood and prepared so it is known to be accurate at the outset. Through the iterative process of exploratory analysis and modeling, the predictive model will be tested and assessed to ensure it is producing accurate results. Once the data has been produced it will be validated, to ensure there were no errors through the process so far, and should any indicators of unexpected results be present, the process of evaluating the data and the model will be revisited.

Once this process of creating the results occurs, the data will be presented to the insurance company for their use. Identifying the incidence of higher risk individuals as well as the risk factors associated with those individuals will provide the company the information needed to proceed. The presentation of this data will be in a way that can be easily observed by the insurance company without need for significant interpretation.

**Data**

The data that has been provided contains all the elements needed to create a predictive model of the risk factors. The program that will be developed targets improvements in the areas that are reported as risk factors. The age range of identified individuals is from 42 to 81 years old in each of the testing and scoring sets. Each data set also presents with similar diversity of risk factors, with frequencies for each of the measured factors closely distributed in each data set. There are also no obvious errors in the data, such as missing data or wrongly-coded data. The data sets are useable for the purposes of this project.

The current data is useful for this project, but also limits the scope of this project. In the training data, there is no paired data for the risk factors that are being evaluated at the time of a second heart attack or after a specified timeframe for those who did not have a second heart attack. This information can refine the weight of the risk factors in the predictive model as the difference in values from the time of the first to the second heart attack may impact the outcome. It would also be useful for the purpose of the project to identify if there are positive effects from mitigating the higher risk factors on the outcome of having a second heart attack.

The data collection is done by searching through medical claims information that is aggregated by the insurance company. Patients can consent and withdraw consent to be contacted for participation in voluntary programs, such as the one that is intended to be developed through this project. Since the data for the creation of the model does not contain any personally identifiable information, it would not be a violation of the Health Insurance Portability and Accountability Act (HIPAA), according to the National Institutes of Health (2005). The scoring data set will have a key added for each subject which will correspond to a patient. The patient information is maintained by the company and not known to this project as it is not necessary for the completion of the analysis. The key will allow the interpretation of significant results to be associated with individuals for the outreach to participate in targeted programs by the insurance company. The collection and use of this data appears to be within legal and ethical standards.

**Tool Applicability to Initiative**

The aim of the project is to create a tool that will predict the individuals who are higher risk of having a second heart attack to identify those who may benefit from an additional healthcare program to reduce that risk. This project will create a predictive algorithm that can take the data and produce the probability of the individual having a second heart attack. The outcome will also be presented to the stakeholders, including the health insurance company and the program manager for this program.

The use of the R programming language will be an effective tool for the manipulation of the data as well as a visualization (Kabacoff, 2020). With the capacity to perform the necessary analyses as well as the ability to create the tables and graphs that will demonstrate the results for the stakeholders, R is a very accessible and appropriate tool available for this initiative. For the determination of risk factors to include, multivariate graphs may help visualize the strength of predictor variables on the outcome variable.

**Tool Applicability to Data**

Analyzing the data that is presented by the health insurance company is straightforward, as the data is structured. The data can be extracted using SQL to query the database and clean the data (Li, 2021), regardless of the size of the database. According to Perficient (2022), health insurance companies are storing lakes of data and often with a Hadoop Data Warehouse. Utilizing a Hadoop framework in conjunction with SQL and R, the data will be able to be extracted, cleaned, transformed, and presented.

**Tool Recommendations**

The outcome of this project is one of predictive analytics. Predictive analytics is used in healthcare frequently to anticipate the needs of patients based on their current presentation (Philips, 2020). Accomplishing the goal of utilizing predictive analytics will require the tools capable of doing so easily. For this project, the tools to use with this technology that seem most appropriate are SQL and R.

Exploratory analysis may show that there is a need to expand the number of risk factors and data from other sources within the health insurance company may be helpful, including demographic information that is not included in claims, or services provided in other areas from medical, such as behavioral health. This would involve the technologies of data virtualization and integration, which allows for a greater reach of sources to extract the data as well as integrate data from multiple sources (Maruti Techlabs, 2018). In anticipation of this, tools to manage the greater data set rather than Excel would be preferred. The use of SQL may be more efficient and capable at cleaning the data and presenting in the format that can easily be fed into the predictive analytics algorithm of R (Li, 2021). Utilization of SQL in R (Errickson, n.d.) appears to be an option that would be useful for the extraction and manipulation of the data and creation of a predictive algorithm.

The data to be analyzed is structured and the analysis is not utilizing streaming data that will need to be analyzed in real-time. For these reasons, it is unnecessary to utilize more complex analytic tools, such as Apache Storm or Apache Spark (Vohra, 2021). However, the data set size may prove to be greater than anticipated. A Hadoop framework would be able to manage the enormity of the data and is scalable (SAS, 2022), which would be essential should the vast data of a healthcare insurance company need to be accessed. With the incorporation of Apache Hive, SQL would be able to be used as a SQL-on-Hadoop tool (Smallcombe, 2018). This would allow for batch processing of the data.

**Conclusion**

**Value**

This project is designed to identify and help individuals served by the health insurance company who have had a heart attack to provide programs that may reduce the risk of having a second heart attack. Having this early intervention with measures that would promote a healthier outcome would help the insurance company with reducing costs associated with care for a second heart attack, and more importantly, improve the lives of their members through prevention of a serious adverse health event.

The data stored by the insurance company is vast and includes information on both the potential factors influencing the medical events and the interventions that have been tried by the members to improve their health status. Analysis of this data for the correlated risk factors may yield opportunities for refining what interventions have been helpful for others in preventing a second heart attack. This will help the health insurance industry overall in these endeavors as predictive analytics is already shown to be useful in hospital settings (Philips, 2020). Having this information available to the aggregated source of an individual’s health history allows the expansion of preventative care.

Data analysis and predictive analytics, specifically, in the health insurance industry can improve the lives of their members while reducing the cost of care. Given the cost of having a heart attack is estimated at being approximately $1 million, according to United Policy Holders (2013). This includes the emergency, rehabilitative, and follow-up care. This drives the cost of care per patient who has a second heart attack to over $2 million. The cost of preventative care is a fraction of the cost of a heart attack, according to NCCDPHP (2014), costing very little to implement compared to the savings to the provider as well as the patient for more serious events. Figure 3 shows how the cost of a heart attack compares to the cost for prevention programs.

The data to be used in this project is already available to the insurance company from their existing data collection sources. The cleaning and organizing of the data can become an automated process utilizing the tools from this project. Once a predictive model is built, information on newly identified members can be analyzed quickly and efficiently. Should other factors be considered for the model in the future, with the clear documentation throughout the CRISP-DM process of building this model, additional variables can be tested and analyzed easily. The outcome of this project will be beneficial for the insurance company beyond these specific results.

**Insights**

Utilizing the data that is generated by a health insurance company historically has been underutilized. With numerous available sources within the company, such as claims, provider search history, and communications with the members, the insurance products offered to individuals can be tailored to their needs (McGinley et al., 2020). This insurance products can also include services that are preventative and supportive.

Collaborations with the community providers based on the insights gained from this project may be available as well. Increasing the use of preventative programs targeting health, wellness, mental health, and socialization may prove to reduce the negative health events in the lives of the individual members. Expanding the health insurance industry to a model that is more wellness-encouraging and less payment-focused.

The data analytic tools utilized for extracting, cleaning, and then creating models from the data will include SQL and R. The health insurance data is stored in a Hadoop Data Warehouse. Incorporating Apache Hive will benefit the project and make future data analysis easier since it will be utilized as a SQL-on-Hadoop tool (Smallcombe, 2018). SQL is the most used language for data analysis, according to a survey noted by Grupman (2021). Due to this, the health insurance company data analysts are likely well-versed in SQL and would be able to analyze further data with facility using these tools and applying the model created by this project.

As a data analyst, utilizing these tools and creating this model will reinforce the use of many skills including the programming languages of R and SQL. Working with a Hadoop Data Warehouse and Apache Hive to access the data from that warehouse, as well as learn about the ways some of the largest data generators store their data and how those data lakes are accessed and utilized, is an opportunity to expand experience working with big data. Additionally, the overall use of a CRISP-DM process model will demonstrate the benefits of using a clear process and strengthen the reliability of the outcome.

**References**

Center for Insurance Policy and Research. (2021, May 27). *Big Data*. NAIC.Org. Retrieved March 4, 2022, from https://content.naic.org/cipr\_topics/topic\_big\_data.htm

Errickson, J. (n.d.). *SQL in R*. University of Michigan. Retrieved February 13, 2022, from https://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/sql.html

Grupman, C. (2021, November 23). *Want a Job in Data? Learn SQL.* Dataquest. Retrieved February 13, 2022, from https://www.dataquest.io/blog/why-sql-is-the-most-important-language-to-learn/

Ingraham, C. (2018, September 25). *An insurance company wants you to hand over your Fitbit data so it can make more money. Should you?* Washington Post. Retrieved March 4, 2022, from https://www.washingtonpost.com/business/2018/09/25/an-insurance-company-wants-you-hand-over-your-fitbit-data-so-they-can-make-more-money-should-you/

Kabacoff, R. (2020, December 1). *Data Visualization with R*. Github. Retrieved February 13, 2022, from https://rkabacoff.github.io/datavis/

Li, L. (2021, December 12). *Cleaning and Transforming Data with SQL - Towards Data Science*. Medium. Retrieved February 13, 2022, from https://towardsdatascience.com/cleaning-and-transforming-data-with-sql-f93c4de0d2fc

Martin, L. R., Williams, S. L., Haskard, K. B., & DiMatteo, M. R. (2005, September). *The challenge of patient adherence*. NIH. Retrieved March 4, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1661624/

Maruti Techlabs. (2018, June 4). *10 Key Technologies that enable Big Data Analytics for businesses*. Medium. Retrieved February 13, 2022, from https://towardsdatascience.com/10-key-technologies-that-enable-big-data-analytics-for-businesses-d82703891e2f

McGinley, D., MacDonnell, B., & Comerford, E. (2020). *Use of Internal Data in Insurance*. Milliman. Retrieved February 13, 2022, from https://us.milliman.com/-/media/milliman/pdfs/articles/uses-of-internal-data-in-insurance\_v2.ashx

National Center for Chronic Disease Prevention and Health Promotion, Division for Heart Disease and Stroke Prevention (NCCDPHP). (2014, July 22). *The Stroke and Heart Attack Prevention Program (SHAPP) Fact Sheet*. CDC. Retrieved March 5, 2022, from https://www.cdc.gov/dhdsp/data\_statistics/fact\_sheets/fs\_shapp.htm

North, M. A. (2012). *Data Mining for the Masses*. Global Text Project. https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf

Perficient. (2022). *The Role of Data Lakes in Healthcare*. Retrieved February 13, 2022, from https://www.perficient.com/-/media/files/guide-pdf-links/the-role-of-data-lakes-in-healthcare.pdf

Philips. (2020, June 12). *Predictive analytics in healthcare: three real-world examples*. Retrieved February 13, 2022, from https://www.philips.com/a-w/about/news/archive/features/20200604-predictive-analytics-in-healthcare-three-real-world-examples.html

SAS. (2022). *What is Hadoop?* SAS India. Retrieved February 13, 2022, from https://www.sas.com/en\_in/insights/big-data/hadoop.html

Smallcombe, M. (2018, November 27). *12 SQL-on-Hadoop Tools*. Integrate.Io. Retrieved February 13, 2022, from https://www.integrate.io/blog/12-sql-on-hadoop-tools/

Society of Actuaries. (2022). *Actuarial Toolkit.* SOA.org. Retrieved March 4, 2022, from https://actuarialtoolkit.soa.org/tool/glossary/pooling

United Policy Holders. (2013, March 12). *The costs of surviving cardiovascular disease: It’s enough to give you a heart attack!* Retrieved March 5, 2022, from https://uphelp.org/the-costs-of-surviving-cardiovascular-disease-its-enough-to-give-you-a-heart-attack/

University of Cincinnati. (n.d.). *Logistic Regression · UC Business Analytics R Programming Guide*. Retrieved March 4, 2022, from https://uc-r.github.io/logistic\_regression

Vohra, G. (2021, November 8). *25 Most Popular Data Analytics Tools To Know In 2021*. Jigsaw Academy. Retrieved February 13, 2022, from https://www.jigsawacademy.com/10-most-popular-analytic-tools-in-business/

**Figure 1**

*Sample Table of Predictions of a Second Heart Attack*

Table

Description automatically generated

*Note.* Example of possible data output from a logistic regression. Columns of confidence(Yes) and confidence(No) indicate the confidence level predicting whether a second heart attack would occur. The variable prediction(2nd\_Heart\_Attack) indicates “Yes” for the values that have met the threshold determined by the program manager for confidence(Yes), 0.450. Adapted from *Data Mining for the Masses* by North, M.A. (2012). Global Text Project. https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf. p.149.

**Figure 2**

*Sample Plot of Logistic Regression Model*

*Chart, line chart

Description automatically generated*

*Note.* Example of a model fit for individual variable, such as weight, cholesterol, or gender, as predictor for probability of having a second heart attack. This graph is generated from the training data for individuals who have or not had a second heart attack. Where new predictor data falls on the curve shows the probability for that new person to have a second heart attack. Adapted from *Logistic Regression - UC Business Analytics R Programming Guide* by University of Cincinnati. (n.d.)

Figure 3

*Cost of Care for Patients Receiving Different Levels of Care*

*Note.* A comparison of the different costs of care for patients based on the level of care that is needed. The cost of having a heart attack is much higher than the cost of prevention programs. Mitigating the risk of cost of care going from $1 Million to $2 Million with only a slight increase in cost of preventative care seems an obvious option. The prevention program, which will identify participants using the predictive analytics from this project, will reduce the incidence of patients incurring the added cost of care. Data for this is based on information from NCCDPHP (2014), which places the cost of a prevention program as a small fraction of the cost of care, and United Policy Holders (2013), which estimates the cost of care for a heart attack to be $1 million.