

Data Mining Episode Groupers

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

It is the purpose of this study to develop a method to define sequential episodes of patient care. We will use data from a cohort of patients with heart problems and diabetes. We will focus on the condition of congestive heart failure, a co-morbid disease of diabetes that is progressive and irreversible. The biggest problem is to determine where one episode ends and another begins. We start with time series methods to order the claims sequentially. Then we use path analysis in SAS Enterprise Miner to see which episodes are related sequentially. Domain knowledge is also required to define the episodes. Once the episodes are defined, we use decision trees to examine the relationship between treatment and outcome. We want to determine whether different treatments lead to different outcomes.

INTRODUCTION

Physicians make many different decisions to treat patients, especially those with multiple chronic illnesses. For example, there are many different medications for the treatment of Type II diabetes, and the physician chooses one or more of them for their patients. There is also the decision to start a patient on insulin; moreover, there is now a choice between insulin injections and inhaled insulin. A patient with blocked arteries can receive angioplasty, or bypass surgery. It is the accumulated consequences of these decisions that result in differing patient outcomes. In a continuum of treatment for chronic conditions, it is difficult to determine where one decision starts and ends. It is the purpose of this project to examine claims data to investigate sequential patterns of physician decision making by defining episodes of patient care.

We first need to preprocess the data to create treatment episodes to construct a sequence of care. We assume that episodes of treatment can be defined. Some treatments, for example, chemotherapy for cancer, can have a start date and an ending period with follow up so that recurrence begins a new treatment episode. However, for chronic conditions such as congestive heart failure and diabetic foot ulcers with chronic osteomyelitis, it is not clear when one treatment ends and another begins. Instead, the continuum of care should be considered.

We then define events in the sequence of treatment that suggest disease changes. We assume that the illness will change over time, for the better or for the worse, and that these markers can be used to examine treatment differences related to outcome. Consider, for example, Type II diabetes. One marker is the initial disease diagnosis followed by drug treatment. A second marker is a change in the type of medication, or the dose. A third marker would be a transition to insulin. We will use survival data mining to investigate the relationship between treatments and time to events for chronic diseases.

The next step is to construct a decision tree based on the analysis of the treatment sequence. We make the assumption that decision trees can be so constructed to examine the competing risks of different treatment sequences. We will use data from several sources to examine the data longitudinally.

EPISODE GROUPERS

Solutions under the general category of episode grouper have been developed specifically to fuse claims data. The methodology is difficult to find since it is mostly proprietary and little exists in the research literature.¹⁻³ A brief summary is given in Forthman, Dove and Wooster.³ The main purpose of these groupers is to identify homogeneous groups of patients so that cost comparisons and summaries can be made. These "episode groupers" are used in analysis with little understanding as to how episodes are defined or how patients are grouped.⁴⁻⁸ However, it is known that the groupers do not take into consideration the severity of an individual patient's condition.⁶

One method of grouping is to examine medications of a similar nature, and to define the end of an episode if there is at least one day between claims.⁹ The Medicare Claims Processing Manual defines an episode of care as having a maximum time period of 60 days or until discharge, although episodes of care can be overlapping.¹⁰ Another study defined episodes as 30-day periods while a third considered a 4-month to 9-month absence of treatment as the end of an episode.^{11,12} There are still other definitions of episodes, including one per year.¹³ However, the main method used to define an episode of care is a variable timeframe, or "washout" period, with a continuous time period with an absence of treatment; that time period changes with the definition of the patient's condition.¹⁴⁻¹⁶

Unfortunately, it is not always clear just what that time period should be. For example, when a bone gets infected with a superbug known as MRSA, recurrence can occur up to a year after treatment is completed. Should this year be the definition of an episode, or should a period of say, six months be used to end the episode? One study that attempted to define an episode concluded that the duration was approximately 5 weeks for treatment of diabetic foot ulcers, excluding all patients who had a bone infection or amputation.¹⁷ Yet most clinical studies of the same

problem consider 8-12 weeks as a minimum for healing of the wounds, almost twice the length of the defined episode.¹⁸⁻²⁰ As another example, chronic diseases that are physician managed will have ongoing treatment if periodic testing and monitoring occur. In that case, an episode has to be defined differently for different treatments for the same patient condition.

Once a patient episode is defined, it is usually examined independently of other episodes for the same patient.²¹ The main measure of an episode is its total cost.²² However, that means that the likelihood that a treatment choice in episode one leads to episode two is not examined.²³ In particular, we want to determine whether treatment choices lead to additional episodes of care. For example, suppose a treatment standard decreases inpatient stays from 5 days to 4 days, but at the cost of doubling the readmission rate.²⁴ Without examining the sequence of admissions, the 1-day reduction would be considered a cost effective outcome, especially if an episode is defined as the time from admission to discharge.²⁵ In addition to variability in patient response to treatments, there are competing risks that result in different choices of treatment made either by the physician or the patient.²⁶ ²¹ Treatment variability is very characteristic of psychiatric treatment, even more so than for physician medicine.²⁷

Another consideration is the pathway itself, defined by compliance with treatment and the continuity of care, especially to determine the effectiveness of disease management.^{28, 29} We need to create a definition of compliance with care, and to rank compliance with treatment. We also need to ensure that all treatments (including prescribed medications) are included in the pathway, and are used to define episodes of care.

Relying on claims data, which is combined into one database from multiple sources, a date of care is included in each claim. However, if a patient is treated in the hospital, there can be several different physicians giving different types of care. It becomes a major challenge to relate these together into one episode. At a minimum, claims from the same episode should have the same diagnosis related group (DRG) code. However, this code may be entered inaccurately. Claims for medications may not contain this code at all. Each claim will have a service date. We start by creating a clustering for each patient based upon date and DRG codes. Not every patient claim will be clustered successfully. From there, predictive modeling will be used for the unclustered values to predict membership into each cluster.^{30, 31}

To examine the sequence of episodes, we will define a time series with multiple time endpoints. The initial time point will define the initial treatment and beginning of a chronic problem. The additional time points will be defined as either the end of the episode, or a change in condition, where the chronic illness gets better, or worse. We will use both fixed and dynamic regressors to investigate the patient outcomes. These regressors can represent a different medication, or a decision to perform surgery, or a change from outpatient to inpatient status. They can also represent a new, ongoing treatment. The fixed regressors will represent patient demographic information, and the initial severity of the patient's condition. The time series will be transactional in nature as the changes in treatment will not necessarily occur at fixed intervals. We will start by defining a time series for each patient, and then consolidating them into a series of outcomes. Once we have the likelihood of various outcomes defined by the time series, we can create a decision tree to look at the probability of each outcome given treatment choices.

In addition, it will be important to detect outliers either as they occur, or before they occur in terms of both cost and outcomes. Therefore, the claims data can also be considered streaming data, with changes in treatments indicative of future outcomes that can be costly either to payer or patient.

Data Mining Methodology

Data mining and statistics have generally developed in different domains. Statisticians are primarily interested in inference; data miners in exploratory data analysis. Nevertheless, there are some instances where data mining and statistics have blended. Many statisticians remain dubious about the data mining process.³² Others are concerned with the lack of a theoretical framework similar to the one for inferential statistics, especially since data mining tends to be algorithmic-based.³³⁻³⁵

Statistics and data mining differ in the use of machine learning methods, the volume of data, and the role of computational complexity. Our need for analysis is exceeding our abilities to handle the complexity.^{36, 37} Preprocessing is far more important with large datasets, especially as we approach the petabyte level.³⁸ However, there are indications that data mining is focused on the data mining process itself with little emphasis on the knowledge actually extracted.³⁹ We need to know whether the extracted pattern is real or spurious, meaningful or meaningless. Will the extracted knowledge motivate positive action? Will it motivate decision making? Can the extracted information be interpreted?

While some of the methodologies are similar in both data mining and statistical analysis, the desired outcomes can differ substantially. For example, market segmentation is a problem of clustering; however, in the data mining approach, the clustering is acceptable if the result is increased sales or better prediction.^{40, 41} In the statistical approach, the clustering is good if there is homogeneity within clusters and heterogeneity across clusters.⁴¹ On the other hand, an association rule or market basket analysis is a technique of data mining used almost exclusively in marketing applications.^{42, 43} The primary concern of this type of analysis is sales, and more recently to distinguish

between customers with higher levels of sales.⁴⁴ However, in other, non-marketing applications, the optimal goal might be to change behavior rather than to just model customer behavior.^{45, 46} Therefore, the potential of market basket analysis still needs to be exploited statistically.

Another difference in approach occurs with binary or ordinal outcomes. Typically in a logistic regression analysis from a statistical perspective, the sample size is too small to allow us to over-sample rare occurrences.⁴⁷ This over-sampling is necessary because logistic regression performs poorly if the group sizes are not similar. Yet, especially in medical studies, logistic regression is used frequently to predict rare occurrences. Often, high rates of accuracy are not examined in terms of differing false positive and false negative rates, resulting in a very inflated outcome.⁴⁸ Sometimes, attempts are made to find matching cohorts; however, they are only matched on parameters defined by the investigator; the rare occurrence remains rare.^{49, 50} While there are concerns about the use of statistical models in medicine, the issue of sampling rare occurrences is not considered important.⁵¹ High risk versus low risk is often the binary outcome under consideration. In statistical models, linear and logistic models are used to distinguish between population groups. Often risk, particularly patient risk, is assumed uniform across the population base,⁵²⁻⁵⁴ for example, when we consider the risk of polio when the disease now occurs from a vaccine, or from the potential risk of bioterrorism.⁵⁵ We still assume that everyone is equally at risk.⁵⁶ Pooled risk, too, assumes that risk is uniform throughout the pool.⁵⁷ The use of more input variables allows for individual assessment so that in data mining, risk is defined by individuals in the population base.

While statistical software simplifies the development of predictive models, there is danger in the inapplicability of models that must be clearly understood.⁵⁸ In the data mining approach, the number of rare occurrences is sufficiently large so that over-sampling still results in a sufficiently large sample.⁵⁹ Therefore, we can change the focus from prediction of risk to prediction of diagnosis. Data mining procedures can also rank observations to determine those most likely to predict accurately.⁶⁰

One of the major problems with either data mining or statistical analysis is the requirement of preprocessing data.^{60, 61} Often, different pieces of the databases are located at different sources that are not necessarily compatible. This is particularly true in healthcare. Information publicly available, but located at different web locations, is also problematic.

There are indications that 80-90% of available data are in text form.⁶² For too long, such data have been largely ignored, or used to define simple frequency counts. Text mining can now be used to analyze smaller and smaller pieces of text, allowing it to be used to compress large, categorical variables.^{63, 64} Text mining can also be used to find a natural ordering in the data for the purpose of ranking clusters.⁶⁵

Much of the data collected in databases nowadays is incomplete and noisy. This may be deliberate as, for example, when a customer refuses to provide an accurate date of birth or accidental as due to input error. Also, there is always the danger that data may be old or redundant. Thus, it is essential to researchers to base their analysis on what is described as "*clean data*". Cleaning data or preprocessing the data prior to mining is designed to eliminate the following anomalies:

1. Missing field values.
2. Outliers.
3. Obsolete and/or redundant data.
4. Data in clear contradiction of common sense or well established industry norms.
5. Data in inconsistent state or format.

It is estimated the 50-60% of researchers' time is spent in data preprocessing to create databases suitable for data mining. Thus, it is no surprise that data preparation is an integral phase of the data mining process as a whole.

It is also the case that data preprocessing requires an understanding of the data and of the statistical analysis that is necessary to manipulate the data in order to remove any anomalies.⁶⁶⁻⁶⁸ Another issue in preprocessing is the need to define the observational unit. For example, the dataset might focus on individual claims from one inpatient hospital stay. However, there would be separate claims for the hospital, the physicians, the medications prescribed on discharge, and any home health care required. In order to examine the entire cost of one visit, the observational unit must be changed from claim to inpatient process.

Data Fusion

Data fusion has been a trend in the field of imaging, text and signal analyses, and it is a combination of many disciplines. Communication and data management technologies focus on the organization, storage, preservation,

and distribution of data. Mathematics, computer science, and artificial intelligence all contribute to the development of automatic and principled methods for combining, restructuring and summarizing diverse, incomplete and conflicting information. Data fusion covers an entire process: data gathering from multiple sources, data format conversion, data combination, conflict resolution, data summarization and distribution.^{69, 70} The process takes input from heterogeneous sources and produces a coherent representation. Although multi-sensor data fusion is still not regarded as a formal professional discipline, tremendous progress has been made. The success of data fusion, and later data mining, depends as much on the adoption of appropriate methodologies and processes as it does on the availability of suitable data and the use of appropriate technology.

Medical data fusion is an emerging field which has recently experienced a tremendous reduction of innovation cycles. Progress and advances in medical imaging, medical signals, and an unstructured text format, have an immediate impact on commercial products and clinical practice. Today, various data modalities with completely different capabilities are available for diagnosis, intervention, surgery, or monitoring.⁷¹ In multi-modal data registration, data of different modalities are transformed into a single coordinate system. Physicians get simultaneous access to the patient's data.

SAS CODE FOR PREPROCESSING DATA

We first preprocess the data. The data were already in the form of SAS datasets. They were merged together so that the longitudinal progression for individual patients was maintained. The data were previously de-identified according to HIPAA (Health Insurance Portability and Accountability Act) requirements, using randomly selected keys to substitute for actual patient identifiers.

PROC HPF (for high performance forecasting) is used to bin the patient claims by month, with each month defining an episode. Any month that does not exceed a minimum specified cost is subsequently filtered out of the dataset to exclude consideration of routine, follow up visits to the physician.

```
proc hpf data=sasuser.dataset out=sasuser.episodegroup;  
id treatment_date interval=month accumulate=total;  
by patient_id;  
run;
```

We next use PROC Transpose to shift the values so that each patient identifier has just one observation in the dataset.

```
proc transpose data=sasuser.episodegroup out=sasuser.transposedataset  
prefix=procedure_  
var treatment_date;  
by patient_id;  
run;
```

METHOD

We applied the proposed methodology to patients with blocked arteries, requiring either angioplasty, angioplasty with a stent, or bypass surgery. There are three major steps to the creation of a sequential episode grouper.

Step 1.

We first isolate patient identifiers with surgical procedures involving bypass or angioplasty. We use DRG (Diagnosis Related Group, or primary procedure code) to filter these patients from the database. A DRG is a diagnosis related group that is used as an identifier to determine insurance reimbursement for the procedure. The DRG codes that define these procedures have been changed regularly, particularly with the introduction of a new type of stent. Once filtered, we collected the series of treatments for each patient that are related to bypass and angioplasty.

The DRG codes used to filter the values are given in Table 1. Table 2 gives the changes to codes during the period under study.

Table 1. DRG Codes for Bypass and Angioplasty

DRG Code	DRG Description
106	Coronary Bypass W PTCA
547	Coronary Bypass W/Cardiac Cath W/Major CV Dx
548	Coronary Bypass w/cardiac Cath W/O Major CV Dx
549	Coronary Bypass W/O Cardiac cath W/Major CV Dx
550	Coronary Bypass W/O Cardiac Cath W/O Major CV Dx
555	Percutaneous Cardiovascular Procedure W/ major CV Dx
556	Percutaneous Cardiovascular Procedure W/non-drug-eluting stent W/O major CV Dx
557	Percutaneous Cardiovascular Procedure W/drug-eluting stent W/major CV Dx
558	Percutaneous Cardiovascular procedure W/drug-eluting stent W/O major CV Dx
518	Percutaneous Cardiovascular proc w/o AMI w/o coronary artery stent implant

Table 2. Changes to DRG Coding

DRG Codes before October, 2005	DRG codes after October, 2005
547 and 548	107 coronary bypass w/cardiac cath
549 and 550	109 coronary bypass w/o major cath
555	516 percutaneous cardiovascular proc w/ AMI
556	DRG 517 percutaneous cardiovascular proc w/o AMI, w/coronary artery stent implant
557	DRG 526 percutaneous cardiovascular proc w/drug-eluting stent w/ AMI
558	527 percutaneous cardiovascular proc w/drug-eluting stent w/o AMI.

Step 2.

Once filtered, we need to determine the potential length of an episode. Fortunately, both angioplasty and bypass tend to have relatively short duration, with the probability of exceeding 30 days as inpatient so small that it can be discarded. Other problems will have more of a continuum of treatment, and the episode will be more difficult to define. Therefore, we define an episode as occurring within a 30-day period.

Step 3

We divide identifiers into two subsets; those with only one inpatient stay and those with more than one procedure. Once separated, we place a code of "1" on the first subset to represent censored data. Similarly, we place a code of "0" on the second subset to represent uncensored data. The analysis differs from the standard survival analysis in that there can be multiple events occurring over time, with multiple recurrences. We can extend this process indefinitely to find the third, fourth, and so on, episode of care.

To ensure that we are capturing events rather than follow up, we will also exclude all defined episodes below a threshold amount. In this example, we use the value of \$20,000.

```
data sasuser.transposecensor;
set sasuser.transposedataset;
lastdate='31dec2004'd;
if (procedure_2 = '.') then censor=0;
else censor=1;
if (censor=1) then time=datdif(procedure_1,procedure_2,'act/act');
```

```

else time=datdif(procedure_1,lastdate,'act/act');
if (censor=1 and procedure_3='.') then censor2=0;
if (censor=1 and procedure_3 ne '.') then censor2=1;
if (censor2=1) then time2=datdif(procedure_2,procedure_3,'act/act');
if (censor2=0) then time2=datdif(procedure_2,lastdate,'act/act');
run;

```

As traditional survival analysis cannot be used, we turn toward survival data mining. The technique has been developed primarily to examine the concept of customer churn, again where multiple end points exist.^{72, 73} However, medical use of survival is still generally limited to one defined event, although some researchers are experimenting with the use of predictive modeling rather than survival analysis.⁷⁴⁻⁸⁰ Nevertheless, in a progressive disease, the event markers of that progression should be considered.

Assuming a proportional hazard rate, we can use PROC PHREG to estimate the hazard function for each new event. We can then compare the different initial procedures more easily. We use the SAS code:

```

data sasuser.datasetforphreg;
  set sasuser.olddataset;
  by patient_id treatment_data;
  where accum_amt>20000;
censor=0;
if last.patient_id then censor=1;
  if first.patient_id then sequence_num=0;
  sequence_num+1;
run;
proc phreg covs(aggregate) data=sasuser.datasetforphreg;
model start_date*sequence_num(0)=z11 z12 z13 z14;
strata sequence_num;
id patient_id;
z11=drg*(sequence_num=1);
z12=drg*(sequence_num=2);
z13=drg*(sequence_num=3);
z14=drg*(sequence_num=4);
Equaldrg: test z11=z12 z11=z13 z11=z14;
run;

```

RESULTS

Once the data are filtered by DRG, there remain approximately 270,000 total claims. Once binned, table 3 gives the number of episodes by DRG code. We also specify that an episode has to have a minimum cost so that we exclude all follow up patient events that are related to the initial episode. There are almost 22,000 total episodes, of which

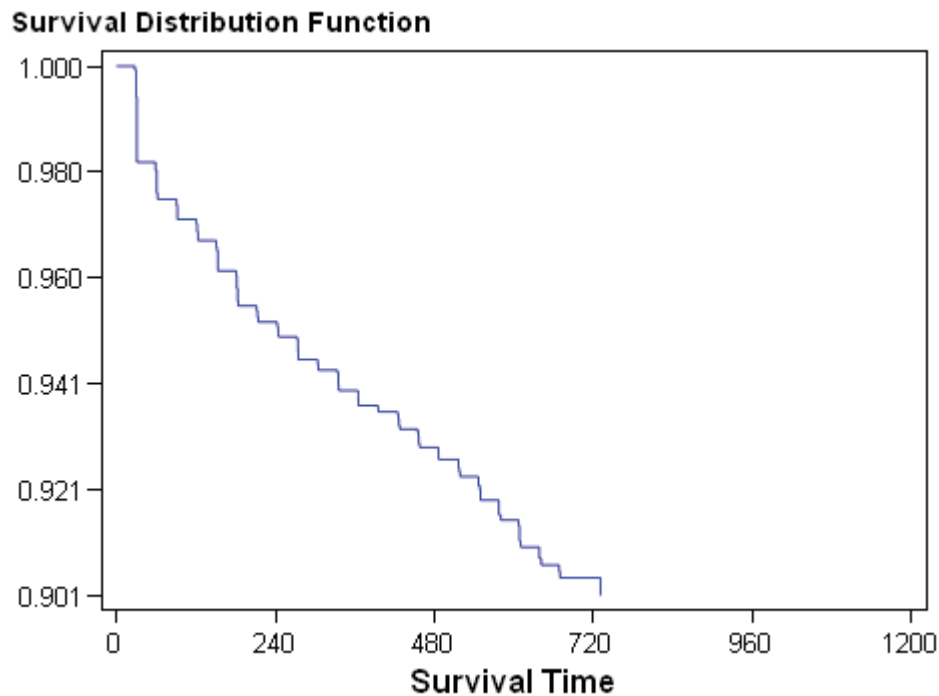
14,006 (approximately 2/3) are initial events without any recurrences. It also gives the percentage by DRG for repeat episodes.

Table 3. Episodes by DRG

DRG	Frequency	Percent	Percentage of Non-Repeat Patients
106	95	0.43	98.94
107	1895	8.62	96.59
109	1961	8.92	97.92
516	1745	7.94	97.92
517	3769	17.15	92.21
518	4348	19.79	92.36
526	1149	5.23	98.38
527	5336	24.28	91.02
547	92	0.42	100.00
548	72	0.33	100.00
549	75	0.34	100.00
550	143	0.65	100.00
555	205	0.93	99.51
556	43	0.20	100.00
557	390	1.77	98.27
558	655	2.98	96.61

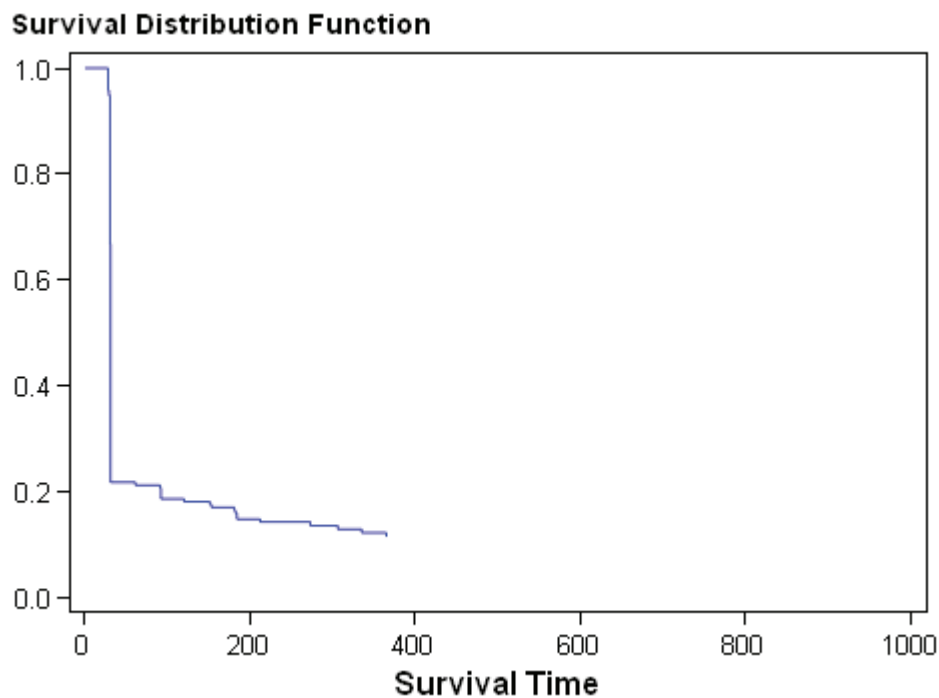
As expected, patients with DRG codes 517 and 518, angioplasty with and without a traditional stent have a high rate of repeats. However, patients with DRG code 527 with the newly developed drug-eluting stent also have a high rate of repeat episodes. This was unexpected. In the first study, we filter the data down to code 527 and construct survival functions for the second and third episode, assuming the first episode as the initial time=0 point. Figure 1 gives the survival curve from the first episode to the second. Figure 2 gives the survival curve from the second episode to the third, for those who have a second episode.

Figure 1. Time to Second Treatment Event with Drug Eluting Stent



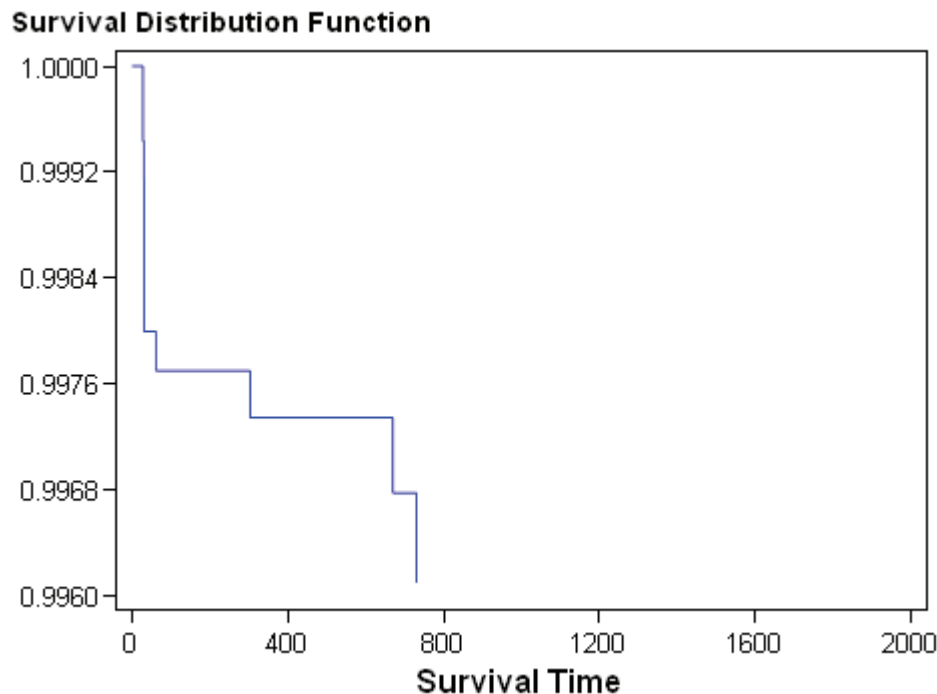
As it turns out, almost 9% of the patients have a second episode with the drug eluting stent. Approximately 12% of this 9% have a third episode.

Figure 2. Time to Third Treatment Event with Drug Eluting Stent



A total of six patients who had an initial angioplasty with a drug-eluting stent also failed, and had a bypass procedure. We also considered the initial episode of a bypass using DRG codes 106, 107, 109, 547, 548, 549, and 550. The survival curve from the initial bypass to a second episode is given in Figure 3.

Figure 3. Time to Second Treatment with Bypass Surgery



Less than 1% of the bypass patients have a second event with bypass surgery.

PROC PHREG, however, did not find that the differences between initial procedures was statistically significant, and found a hazard rate of 1.006.

DISCUSSION

Physicians tend to be autonomous in their decision making, especially in the absence of treatment guidelines. Variability in decision making can lead to variability in patient outcomes. Only by comparing outcomes across physicians can optimal treatment pathways be discovered.

In a future study, we will next examine the more complex issue of osteomyelitis in patients with diabetes. Patients with diabetes are at high risk of developing diabetic foot ulcers. If the ulcers get infected, especially with the bacteria, MRSA, the risk of developing osteomyelitis and subsequent amputation are also quite high. The longer it takes to heal the ulcers; the greater the probability of infection. Moreover, the choice antibiotic treatment, and its duration are also directly related to success in healing. Because the treatment of foot ulcers and osteomyelitis are ongoing as the conditions, once started, often become chronic, we need to examine the totality of care. Unfortunately, in the past, each episode has been considered independent of other episodes. Therefore, the development of a sequential treatment pathway is invaluable to determine which pathways have a higher risk of amputation compared to others.

We will also use the survival data mining to find the probability values needed to construct the decision tree. The results of this study indicate that bypass surgery will result in fewer repeat procedures. However, the eluting stent is far less costly. It was surprising that this newer technology resulted in a higher rate of repeat procedures, even if the difference was not statistically significant.

ACKNOWLEDGEMENT

We want to thank John Cerrito, PharmD and Glenn Lambert, MD, for their support in the development of this paper, which was supported in part by NIH grant #1R15RR017285-01A1, *Data Mining to Enhance Medical Research of Clinical Data*.

REFERENCES

1. Rosen A, Mayer-Oakes A. Episodes of care: theoretical frameworks versus current operational realities. *Joint Commission on Quality Improvement*. 1999;25(3):111-138.
2. Claus P, Carpenter P, Chute C, Mohr D, Gibbons P. Clinical care management and workflow by

- episodes. *Proceedings AMIA Annual Fall Symposium*. 1997;1997:91-95.
3. Forthman MT, Dove HG, Wooster LD. Episode treatment groups (ETGs): a patient classification system for measuring outcomes performance by episode of illness. *Top Health Information Management*. 2000;21(2):51-61.
4. Wan G, Crown W, Berndt E, Finkelstein S, Ling D. Healthcare expenditure in patients treated with vanaifaxine or selective serotonin reuptake inhibitors for depression and anxiety. *Internation Journal of Clinical Practice*. 2002;56(6):434-439.
5. Kerr E, McGlynn E, Vorst KV, Wickstrom S. Measuring antidepressant prescribing practice in a healthcare system using administrative data: implications for quality measurement and improvement. *The Joint Commission Journal on Quality Improvement*. 2000;265(4):203-216.
6. Thomas JW. Should episode-based economic profiles be risk adjusted to account for differences in patients' health risks? *Health Research and Educational Trust*. 2005;April, 2006:581-590.
7. Currie CJ, Morgan CL, Dixon S, et al. The financial costs of hospital care for people with diabetes who have single and multiple macrovascular complications. *diabetes Research and Clinical Practices*. 2005;67:144-151.
8. Bassin E. Episodes of care: a tool for measuring the impact of healthcare services on cost and quality. *Disease Management & Health Outcomes*. 1999;6:319-325.
9. Bonetto C, Nose M, Barbui C. Generating psychotropic drug exposure data from computer-based medical records. *Computer Methods and Programs in Biomedicine*. 2006;83:120-124.
10. Anonymous. Medicare Claims Processing Manual: Chapter 10, Home Health Agency Billing. *Health and Human Services*. Available at: <http://www.cms.hhs.gov/manuals/downloads/clm104c10.pdf>, 2007.
11. Ritzwoller DP, Crounse L, Shetterly S, Rublee D. The association of comorbidities, utilization and costs for patients identified with low back pain. *BMC Musculoskeletal Disorders*. 2006;7:1-10.
12. Thomas JW. Economic profiling of physicians: does omission of pharmacy claims bias performance measurement? *American Journal of Managed Care*. 2006;12:341-351.
13. Hong W, Rak I, Ciuryia V, Wilson A, Kylstra J, Meltzer H. Medical-claims databases in the design of a health-outcomes comparison of quetiapine. *Schizophrenia Research*. 1998;32(1):51-58.
14. Anonymous. episode treatment groups; 2006:1-8.
15. Claus PL, Carpenter PC, Chute CG, Mohr DN, Gibbons PS. Clinical care management and workflow by episodes. Available at: <http://www.amia.org/pubs/symposia/D004137.PDF>, 2007.
16. Hall DL, Llinas J. *Handbook of Multisensor Data Fusion*. Cleveland: CRC; 2001.
17. Mehta S, Suzuki S, Glick H, Schulman K. Determining an episode of care using claims data: diabetic foot ulcer. *Diabetes Care*. 1999;22(7):1110-1115.
18. Ling X, McLennan SV, Lo L, et al. Bacterial load predicts healing rate in neuropathic diabetic foot ulcers. *Diabetes Care*. 2007;30(2):378-380.
19. Sheehan P, Jones P, Giurini JM, Caselli A, Veves A. Percent change in wound area of diabetic foot ulcers over a 4-week period is a robust predictor of complete healing in a 12-week prospective trial. *Plastic and Reconstructive Surgery*. 2006;117(Suppl):239S-244S.
20. Jude E, Apelqvist J, Spraul M, Martini J. Prospective randomized controlled study of Hydrofiber dressing containing ionic silver or calcium alginate dressings in non-ischaemic diabetic foot ulcers. *Diabetic Medicine*. 2006;24:280-288.
21. Jonsson L, Bolinder B, Lundkvist J. Cost of hypoglycemia in patients with Type 2 diabetes in Sweden. *Value in Health*. 2006;9(1):193-198.
22. Peltokorpi A, Kujala J. Time-based analysis of total cost of patient episodes. *International Journal of Health Care Quality Assurance*. 2006;19(2):136-143.
23. Horn SD. Quality, clinical practice improvement, and the episode of care. *Managed Care Quarterly*. 2001;9(3):10-24.
24. Koh H, Leong S. Data mining applications in the context of casemix. *Annals of the Academy of Medicine, Singapore*. 2001;30(4 Suppl):41-49.
25. Kujala J, Lillrank P, Kronstrom V, Peltokorpi A. Time-based management of patient processes. *Journal of Health Organization and Management*. 2006;20(6):512-524.
26. Keen J, Moore J, West R. Pathways, networks and choice in health care. *International Journal of Health Care Quality Assurance*. 2006;19(1):316-327.
27. Singh SP, Grange T. Measuring pathways to care in first-episode psychosis: a systematic review.

- Schizophrenia Research*. 2005;81:75-82.
28. Greenberg GA, Rosenheck RA. Continuity of care and clinical outcomes in a national health system. *Psychiatric Services*. 2005;56(4):427-433.
29. Solz H, Gilbert K. Health claims data as a strategy and tool in disease management. *Journal of Ambulatory Care Management*. 2001;24(2):69-85.
30. Xue S. A fault diagnosis system based on data fusion algorithm. Paper presented at: First international conference on innovative computing information and control, 2006; Beijing, China.
31. Putten Pvd, Kok JN, Gupta A. Data fusion through statistical matching. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=297501#, 2007.
32. Lee S. Predicting atmospheric ozone using neural networks as compared to some statistical methods. Paper presented at: Northcon 95. I IEEE Technical Applications Conference and Workshops Northcon95, 1995; Portland, Oregon.
33. Hand DJ, Bolton RJ. Pattern discovery and detection: a unified statistical methodology. *Journal of Applied Statistics*. 2004;8:885-924.
34. Giudiei P, Passerone G. Data mining of association structures to model consumer behaviour. *Computational Statistics & Data Analysis*. 2002;38:533-541.
35. Sargan JD. Model building and data mining. *Econometric Reviews*. 2001;20(2):159-170.
36. Hosking JR, Pednault EP, Sudan M. Statistical perspective on data mining. *Future Generation Computer Systems*. 1997;13(2-3):117-134.
37. Keim DA, Mansmann F, Schneidewind J, Ziegler H. Challenges in visual data analysis. *Information Visualization*. 2006;2006(9-16).
38. Mannila H. Data mining: machine learning, statistics and databases. Paper presented at: Eighth International Conference on Scientific and Statistical Database Systems, 1996. Proceedings, 1996; Stockholm.
39. Pazzani MJ. Knowledge discovery from data? *IEEE Intelligent Systems*. 2000;March/April:10-13.
40. Bruin JSd, Cocx TK, Kusters WA, Laros JF, Kok JN. Data mining approaches to criminal career analysis. Paper presented at: Proceedings of the Sixth International Conference on Data Mining, 2006; Hong Kong.
41. Jiang T, Tuxhilin A. Improving personalization solutions through optimal segmentation of customer bases. Paper presented at: Proceedings of the Sixth International Conference on Data Mining, 2006; Hong Kong.
42. Wong RC-W, Fu AW-C. Data mining for inventory item selection with cross-selling considerations. *Data Mining and Knowledge Discovery*. 2005;11:81-112.
43. Wang K, Zhou X. Mining customer value: from association rules to direct marketing. *Data Mining and Knowledge Discovery*. 2005;11:57-79.
44. Brus T, Swinnen G, Vanhoof K, Wets G. Building an association rules framework to improve produce assortment decisions. *Data Mining and Knowledge Discovery*. 2004;8(7-23).
45. Cerrito PB, Cerrito JC. Data and text mining the electronic medical record to improve care and to lower costs. Paper presented at: SUGI31, 2006; San Francisco.
46. Giudier P, Passerone G. Data mining of association structures to model consumer behavior. *Computational Statistics & Data Analysis*. 2002;38(4):533-541.
47. Foster DP, Stine RA. Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association*. 2004;99(466):303-313.
48. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*. 2006;98(17):1204-1214.
49. Ried R, Kierk Nd, Ambrosini G, Berry G, Musk A. The risk of lung cancer with increasing time since ceasing exposure to asbestos and quitting smoking. *Occupational and Environmental Medicine*. 2006;63(8):509-512.
50. Claus EB. Risk models used to counsel women for breast and ovarian cancer: a guide for clinicians. *Familial Cancer*. 2001;1:197-206.
51. Freedman AN, Seminara D, Mitchell H, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. *Journal of the National Cancer Institute*. 2005;97(10):715-723.
52. Louis Anthony Cox J. Some limitations of a proposed linear model for antimicrobial risk

- management. *Risk Analysis*. 2005;25(6):1327-1332.
53. Thompson KM, Tebbins RJD. Retrospective cost-effectiveness analyses for polio vaccination in the United States. *Risk Analysis*. 2006;26(6):1423-1449.
54. Gaylor DW. Risk/benefit assessments of human diseases: optimum dose for intervention. *Risk Analysis*. 2005;25(1):161-168.
55. Tebbins RJD, Pallansch MA, Kew OM, et al. Risks of Paralytic disease due to wild or vaccine-derived poliovirus after eradication. *Risk Analysis*. 2006;26(6):1471-1505.
56. Siegrist M, Keller C, Kiers HA. A new look at the psychometric paradigm of perception of hazards. *Risk Analysis*. 2005;25(1):211-222.
57. Tsanakas A, Desli E. Measurement and pricing of risk in insurance markets. *Risk Analysis*. 2005;23(6):1653-1668.
58. CHi-Ming C, Hsu-Sung K, Shu-Hui C, et al. Computer-aided disease prediction system: development of application software with SAS component language. *Journal of Evaluation in Clinical Practice*. 2005;11(2):139-159.
59. Xiangchun, Kim X, Back Y, Rhee DW, Kim S-H. Analysis of breast cancer using data mining & statistical techniques. Paper presented at: Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2005; Las Vegas.
60. Sokol L, Garcia B, West M, Rodriguez J, Johnson K. Precursory steps to mining HCFA health care claims. Paper presented at: 34th Hawaii International Conference on System Sciences, 2001; Hawaii.
61. Popescul A, Lawrence S, Ungar LH, Pennock DM. Statistical relational learning for document mining. Paper presented at: Proceedings of the Third IEEE International Conference on Data Mining, 2003; Melbourne, Florida.
62. Menon R, Tong LH, Sathiyakeerthi S, Brombacher A, Leong C. The needs and benefits of applying textual data mining within the product development process. *Quality and Reliability Engineering International*. 2004;20:1-15.
63. Cerrito P, Badia A, Cerrito JC. Data Mining Medication Prescriptions for a Representative National Sample. Paper presented at: Pharmasug 2005, 2005; Phoenix, Arizona.
64. Yuhua Li DM, Bandar ZA, O'Shea JD, Crockett K. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*. 2006;18(6):1138-1148.
65. Moches TA. Text data mining applied to clustering with cost effective tools. Paper presented at: IEEE International Conference on Systems, Man, and Cybernetics, 2005; Waikoloa, HI.
66. Zhu X, Wu X, Chen Q. Bridging local and global data cleansing: identifying class noise in large, distributed data datasets. *Data Mining and Knowledge Discovery*. 2006;12(2-3):275.
67. Wong K, Byoung-ju C, Bui-Kyeong H, Soo-Kyung K, Doheon L. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*. 2003;7:81-99.
68. Hernandez MA, Stolfo SJ. Real-world data is dirty: data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*. 1998;2:9-17.
69. Makela T. *Data registration and fusion for cardiac applications*. Helsinki, University of Helsinki; 2003.
70. Upstill C, Addis M, Choi F, Taylor S, Watkins R. *Infectious diseases: preparing for the future*. United Kingdom: Foresight Science Reviews; 2006.
71. Denzler J. Sensor data and information fusion in computer vision and medicine, Executive Summary. Paper presented at: Dagstuhl Seminar Proceedings, 2007; Germany.
72. Potts W. Survival Data Mining. Available at: <http://www.data-miners.com/resources/Will%20Survival.pdf>, 2007.
73. Linoff GS. Survival Data Mining for Customer Insight. *Intelligent Enterprise*. Available at: www.intelligententerprise.com/showArticle.jhtml?articleID=26100528, 2007.
74. Xie H, Chausalet TJ, Millard PH. A model-based approach to the analysis of patterns of length of stay in institutional long-term care. *IEEE Transactions on information technology in biomedicine*. 2006;10(3):512-518.
75. Shaw B, Marshall AH. Modeling the health care costs of geriatric inpatients. *IEEE Transactions on information technology in biomedicine*. 2006;10(3):526-532.

76. Pinna G, Maestri R, Capomolla S, et al. Determinant role of short-term heart rate variability in the prediction of mortality in patients with chronic heart failure. *IEEE Computers in Cardiology*. 2000;27:735-738.
77. Berzuini C, Larizza C. A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *IEEE Transactions on pattern analysis and machine intelligence*. 1996;16(2):109-123.
78. Eleuteri A, Tagliaferri R, Milano L, et al. Survival analysis and neural networks. Paper presented at: 2003 Conference on Neural Networks, 2003; Portland, Oregon.
79. Seker H, Odetayo M, Petrovic D, et al. An artificial neural network based feature evaluation index for the assessment of clinical factors in breast cancer survival analysis. Paper presented at: IEEE Canadian Conference on Electrical & Computer Engineering, 2002; Winnipeg, Manitoba.
80. John TT, Chen P. Lognormal selection with applications to lifetime data. *IEEE Transactions on reliability*. 2006;55(1):135-148. *Antimicrobial agents and chemotherapy*. 2005;49(3):1029-1038.

CONTACT INFORMATION

Patricia B. Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6010
Fax: 502-852-7132
E-mail: pcerrito@louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.