ABSTRACTS

Informatics Capstone Presentations: 9:00 – noon

1. A Model for Health Response Assessment (HRA)

Swati Gupta

The goal of my Capstone project is to develop an extended model for Health Response Assessment (HRA) for a local insurance agency (Priority Health). The project is mainly based on targeting diverse segments of consumers with the most relevant products and services. The main objective of this project is to predict members with higher cost on the basis of a health questionnaire. So we could find consumers before they might be having higher cost and we could provide them good care, which would help reduce the company’s overall cost. In this project I used the UPMC model as a starting point on our members’ dataset to evaluate whether it is at all applicable to our members or to what extent. Another objective was, if the first approach was not successful, to develop a new model to predict members with higher cost.

The project is based on data of the initial Health Questionnaire from the year 2013. It contains a total of 388 questions and 8968 observations. The data types are only are categorical and numeric. I used a frequency distribution graph to identify members in the dataset that potentially can generate high cost expenditures in the future. According to this dataset I decided to have the top 10% of all expenditures defined as high cost and rest as low cost. The UPMC model, developed by the University of Pittsburgh Medical Center to predict high cost members, is able to predict a total of 505 members and out of those 160 members are true positive. The final model was developed using a different set of questions. It is able to predict a total of 328 members and out of that 146 members are true positive. To develop this model linear regression, multiple regression, and a decision tree algorithm was used. Only those questions whose impact is significant on the total cost were selected from the Health Questionnaire. Generalization of the developed model was assessed by a validation technique.

Because the developed model is based only on Priority Health data, it is not necessarily generalizable to other insurers or health agencies, while the UPMC model is considered a universal model that can be used in the entire US.
2. Challenges in Clinical Research Informatics:
Data Quality and Transferability in Publically Available Databases.

Michelle Padley

**Background:** Publically-available databases were established with the purpose of making a large amount of patient information readily accessible to researchers and clinicians who would otherwise not be able to obtain it. Some of these databases number in the tens or even hundreds of thousands of datapoints, and are formatted in order to be easily analyzed using programs like R, SPSS and Excel. Through these methods, sub-analysis can be conducted on existing information, adding to the overall knowledge base, and providing new insights on a wide range of indications, from genetic conditions to many different forms of cancer.

As a researcher, both academically and professionally, I have utilized both public and privately held databases. Both have their advantages and disadvantages. Electronic medical records have in-depth, very individualized information, essential for high-quality manuscripts, rated at Level II in terms of evidence. However, these records are limited by the number of patients within a single practice or hospital system, and can be lacking in terms of specific information of interest to the investigator. These measures can include subjective outcomes, and specialty-specific scoring tools.

Alternatively, I used the Chemical Effects on Biological Systems (CEBS) for a group research paper in a previous semester. Initially, after deciding our population of interest, our plan was to perform a statistical analysis on basic patient demographic data. These include the most straightforward information, including gender, age and ethnicity. We found to our surprise, that the database of nearly 86,000 patient was severely lacking in this regard. Ultimately, only ethnicity was considered to be useable, as we were able to demonstrate that the locations that we chose tended to be culturally homogenous.

I found this to be a considerable issue, especially given my professional experience will study protocol and manuscript preparation. While I’m fully aware of the requirements and regulations set in place by HIPPA (Health Insurance Privacy and Portability Act), and the importance of protecting the confidentiality of sensitive patient information, the datapoints I was concerned with fell well outside the realm of these parameters. I chose this topic for my capstone as it coincides well with new challenges with clinical research information systems, and as well as future professional and scholarly projects.

**Purpose:** The purpose of this project is to closely analyze the number and quality of manuscripts published from select publically-available databases.
**Methods:** Using the following databases (Chemical Effects in Biological Systems (CEBS), National Database for Clinical Trials Related to Mental Illness (NDCT), Clinicaltrials.gov and The National Cardiovascular Data Registry (NCRD). I selected a random sample of 20 publications which used data from these publically-available databases for four specialties (this was reduced to three groups due to a low response rate). Experienced raters were selected and given abstracts and literature excerpts at two intervals in February and March of 2016. A total of 20 sets were given to each subject. In all cases, sets were randomized to avoid bias.

**Results:** Using SPSS (v20), a Cohen’s Kappa coefficient analysis was conducted to test interrater reliability. Cardiovascular (clinical research) subset consisted of poor reliability in terms of my testing parameters. K values for rounds 1 and 2 were 0.274 and 0.178. Sociobehavioral (K = 0.379 and 0.437) and basic research (K = 0.555 and 0.531) fared somewhat better, with fair to good reliability results. However, the questionnaire system needs to be modified to make the comparisons that I hoped to make.

**Conclusion:** I learned quite a bit more about the complexities of publishing from publically-available data. Many of the studies I used in my examples laid the groundwork for more important prospective work. It is difficult to create a scale to cover this many broad topics. I hope to modify the scales, and retest in a larger population over a longer period of time to help address the shortfalls of this preliminary project.
3. Exploring Medicare Costs using Machine Learning

Alec Ashburn

As various forms of technology become more ubiquitous in the field of health care, an enormous amount of data is being collected in hope of making new scientific discoveries and reforming the way we understand health care as a society. Specifically, data mining has opened up a portal to discovery and comprehension of otherwise meaningless information. Preprocessing and cleaning techniques, advanced machine learning algorithms, and data visualization tools can be of extraordinary use when trying to make sense of the vast amount of health information at our fingertips. One area of health care that is always undergoing reform and debate is Medicare. I decided to analyze inpatient Medicare coverage data in R for the years 2011 through 2013 to get a better idea of how Medicare dollars are being spent in recent years, how they compare to past spending rates, and what future rates may look like. I was able to determine that among the medical procedures that receive the most Medicare coverage, the top causes of death among the elderly were not included. Many other factors contribute to Medicare costs and were not explored during my research, but the data that I was able to analyze using data mining techniques provides a great deal of insight into an area of much discussion and controversy.
4. Investigation of the Relationship of Sleep/Rest to Different Diseases

Garima Vohra

Background: The emergence of devices, gadgets and applications is a popular trend of today’s technology focused society. Those devices not only monitor the hours of sleep of their owners, but also display the quality of sleep. Various industries are competing to come up with more features to make the devices unique and more advanced in terms of technology. However, the main concern behind the development of such gadgets and devices is insufficient sleep and increasing diseases associated with it.

Insufficient rest/sleep and sleep loss are characteristics of modern society. Studies show that nearly one third of adults report sleeping less than 7 hours (compared to 7-9 hours as the recommended sleep length according to the National Sleep Foundation). There can be many factors responsible to the sleep deprivation such as longer commuting time to workplace, introduction of electric light, different shifts and night work, expansion of industries and different sectors to 24 hours operation, internet, television, and people in America working on multiple jobs, financial concerns and the like.

This study is based on the question on insufficient rest or sleep introduced for the first time in the Behavioral Risk Factor Surveillance System (BRFSS) by Center for Disease Control in year 2008, for all the 50 states and three US territories and expands on a previous study.

Method: The BRFSS 2008 survey data used for the study considers only subjects with over the age of twenty years with 60.47% women and 39.53 % men. The self-reported days of insufficient sleep per month were divided into four groups: zero, 1–13, 14–29, and 30 days. There were four outcomes considered: 1) any CVD, 2) coronary heart disease (CHD), 3) stroke, and 4) diabetes mellitus. I used multivariable logistic regression to calculate odds ratio (OR), (95% confidence interval (CI), considering zero days of insufficient rest/ sleep as the referent category.

Conclusion and findings: All four diseases were found to be associated with insufficient sleep or rest. The Odd’s Ratio (95% CI) corresponding to all 30 days of insufficient sleep was 1.98 (1.51–1.24) for any cardiovascular disease, 1.26 (1.21–1.31) for CHD, for stroke, 1.28 (1.22-1.34), and 1.18(1.52–1.22) for diabetes. Hence, in a multiethnic sample in USA, insufficient sleep was found to be associated independently with CHD, CVD, Diabetes and Stroke.
5. A Cancer Risk Study

Raveena Pendyam

Cancer is a complex disease and it remains the second leading cause of death in the United States. Cancer statistics depict what happens in extensive groups of individuals and to provide a picture in time of the burden of cancer on society. Statistics can give us details such as how many people die from each year, the number of people who are currently living after cancer diagnosis and more. I analyzed cancer data provided by The American Cancer Society.

The aim of the project was twofold: to extract most impactful cancers from the data and to explore the pharmacogenetics and pharmacokinetics of those cancers. Statistical tests such as ANOVA and t-test were performed to identify the top 5 significant cancers, which contribute to the highest death rates in the US. Those were breast, colorectal, prostrate, lung cancers and myeloma. Programs in Python and Plotly were developed to analyze death trends in male and females, risk estimates for new cases, and death rates in the United States.

The above analysis was performed to identify the rate of increase in cancer incidence. To explore further the above stated cancer groups, the following analysis was performed at the level of pharmacogenetics and pharmacokinetics: cancer causing genes were collected from several biological sources and also several drugs acting upon these cancer genes. The genes were filtered to extract the most common cancer genes for the study. There are various approaches in which genes can influence reaction to certain medications depending upon whether they impact the pharmacokinetic drug reaction pathways. For instance, alterations in genes in the PK pathway may influence the absorption, distribution, metabolism or elimination of the drug. The VisANT software was used to create a gene network to see if these genes interact with other genes. As other researchers found in similar situations, it could be shown that a drug can have alternative gene targets.
6. 3-D Modeling of Diffusion Limited Aggregation (DLA) in Hydraulics of Urine

Krishna Nadiminti

Diffusion Limited Aggregation (DLA) clusters are aggregates of particles, and the shape of the cluster is controlled by the possibility of particles to associate with other particles. The aggregates typically grow as long as there are particles moving around. During diffusion of a particle through a solution it is more likely, that it attaches to the outer regions of the cluster. Thus, a solid shape with many dendritic structures, like corals or trees, is generated. The volume is not filled in its entirety, causing many gaps. The premise is that you have particles moving randomly (Brownian motion). For crystals all biological processes are controlled in a semi solid environment. Hence, diffusion plays a vital role in various chemical compositions, temperature of the body, formation of tissues, tumors and more importantly formation of certain crystals like oxalate crystals and fibrinogen crystals. Tracking the growth of such a cluster is challenging because the surrounding medium is the controlling parameter for the growth or movement of the particle that has been present. The project has tracked the random movement of a particle in one and two-dimensional projections. However, the random walk just gives a preliminary idea of the hydraulics of the particle in the lower dimensions.

The goal of the study was to implement existing simulation algorithms for modeling the formation of crystals of urine in the programming languages C++ and openGL. Because of the computational complexity of those more advanced models, existing Python implementations are of limited value for high performance (parallel) computing.

For example, the processing time to animate 20000 particles in 2D using openGL and C++ within the Cinder framework on a laptop computer was 5 minutes. A GPU parallel computing environment would reduce the processing time significantly. Visualizing and modeling such complex crystals could help in medical technology to predict the growth of crystals with respect to time. This can help in informing the therapy from surgery to drug dosage.
7. Prevalence and Severity of Asthmatic Symptoms in Grenada

Rohit Kandalkar

**Background:** Asthma is the most common childhood disease. Asthma causes inflammation in the airways interrupting the airflow in the bronchi and causes suffocation and wheeling of the chest while breathing. Recent studies suggest that there is no longer an increase in asthmatic patients, but a review study of 2010 suggest that there is an increase in the prevalence of asthma in developing nations.

**Objective:** The aim of this study was to investigate the connection of asthma and the different regions in Grenada.

**Methods:** The dataset was obtained from datadryad.org and consisted of data of 1374 children between 6 and 7 years of age with 32 attributes each. The analysis was performed using both the SAS and RStudio statistics software. To categorize an individual into the group asthmatic “wheeling in last 12 month” and the physician’s diagnosis were chosen as the parameters. The following procedures were applied: ANOVA, decision tree analysis, logistic regression, and artificial neural networks.

**Results:** A total number of 1088 cases were used for the calculations with 305 considered as asthmatic and 783 as normal cases. The total patient female and male percentage were 28.61% and 33.72%. Using the “last 12 month of wheeling” attribute the parishes with the highest prevalence were St. George, St. David, and St. Andrew and the lowest was Petite Martinique.

Using logistic regression these factors could be established to trigger asthma the most: Burning Bush, exercise, and a pet at home. On the other hand an ANOVA analysis suggested landfill as a reason for the trigger. A classification tree analysis found dust and cigarette smoke as primary result.

The classification tree analysis on the basis of area and severity showed that St. David, St. George, and Carriacou patients belong to severity level III, while Petite Martinique and St. Patrick patient belong to severity level II and St. Andrew, St. John and St. Mark belong to severity level IV. Of all patients, only 230 saw a physician for their asthma, with an average of 2.82 %.
Internship Presentations: Noon – 1:00 pm

8. Priority Health Internship

Alec Ashburn

Priority Health is a health insurance agency associated with Spectrum Health system that provides coverage to over 600,000 Michigan residents. As Michigan’s fastest growing health plan, Priority Health operates as a non-profit health plan that spends 90 cents of every dollar on customer medical care. This different approach to health care also comes with perks such as PriorityRewards, which offers cash incentives to members based off particular health goals. In short, Priority Health strives to provide dependable health care to a closely-knit community of people.

My internship with Priority Health began in May of 2015 and came to a close in September of that same year. I had the privilege of working in the Data Analytics department with Jason Gillikin as my supervisor and a small team with a diverse skillset. The main bulk of my internship focused on the 2014 Clinician and Group Consumer Assessment of Healthcare Providers and Systems, abbreviated as CG-CAHPS. This survey was presented in one of two formats: an event-based survey that patients completed immediately after their visit and a lookback-based survey that patients completed a few months after their visit. Although the timing differed in survey completion, the 52 question/response format was basically identical for both surveys. My job was to clean the survey data (originally recorded in Excel spreadsheets) and make sense of what the responses meant to Spectrum Health system. Analysis was completed using the boxplot tool within the statistical data visualization software Tableau.

Out of the 52 initial survey questions, 12 were chosen as essential to research:

1) Appointment Scheduling  7) Prescription Advantages
2) Appointment Availability  8) Prescription Disadvantages
3) Provider Communication  9) Prescriptions – Patient Input
4) Provider Listening  10) Knowledge of Specialist
5) Provider Respect  11) Provider Rating
6) Time with Provider  12) Health Goals
9. Internship Report

Garima Vohra

Not available.