9:00 AM	Chris Willock?	Internship
9:12 AM	Chris Willock?	Capstone
9:24 AM	Cornelius	Capstone
9:36 AM	Cornelius	Internship
9:48 AM	Spencer Vandecar	Capstone
10:00 AM	Break	
10:12 AM	Caris Newton	Capstone
10:24 AM	Rachel Dzapo	Internship
10:36 AM	Sreevarsha potluri	Capstone
10:48 AM	Kritika Sijapati	Capstone
11:00 AM	Break	
11:12 AM	Neema Maharjan	Capstone
11:24 AM	Rachel Dzapo	Capstone
11:36 AM		
11:48 AM	Dhynasah	Capstone
12:00 PM	End	

Internship Presentations

Internship at MiHIN

Rachel A. Dzapo

Michigan Health Information Network Shared Services (MiHIN) is a public and private nonprofit organization located in East Lansing, Michigan. MiHIN is a health information exchange devoted to improving the healthcare experience, improving quality and decreasing costs for Michigan residents by supporting the statewide exchange of health information and making valuable data available at the point of care. MiHIN's goal is to streamline the flow of healthcare information to ensure that patients can easily access their personal medical records.
During the course of my internship at MiHIN I was able to gain hands on experience working on two projects aimed at improving public health.

- **Project 1**: Interstate Immunizations
    - Public Health project aimed at leveraging technology to provide patients and providers easier access to obtain immunization records that may be scattered across various states in the country
- **Project 2:** Knowledge Grid
    - Predictive modeling project aimed at decreasing opioid abuse in healthcare patients

◄► ◄► ◄► ◄► ◄►

Internship at Great Lakes Health Connect

Cornelius Scott

Great Lakes Health Connect (GLHC) is a health information exchange located in Grand Rapids, Michigan. I have spent the last semester interning in their health integration analytics department. During my internship at Great Lakes I worked in a team and had a multitude of task. My daily job was to build, maintain and test hospital result delivery interfaces to community practices' electronic health records (EHRs). This is done by validating and testing HL7 messages created through physician patient interactions. My experience at GLHC has given me a chance to put my degree to work. I have gained a lot of knowledge when it comes to SQL querying, data transformations and EHR experience. It also has helped me improve my skills with the non-clinical side of a hospital as well. We communicate directly to hospital IT as well as the EHR companies themselves for vital feedback.

◄► ◄► ◄► ◄► ◄►

Capstone Presentations

Visualizing Opioid Overdose Deaths in the United States

Rachel A. Dzapo

Drug overdoses currently kill more people in the United States than car accidents. The number of drug overdose deaths has steadily increased every year following the pharmaceutical industries push for prescription opioid pain relievers in the late 1990s. This is a national crisis that affects public health as well as social and economic welfare. This study utilizes modern data-mining and information visualization techniques to accurately visualize areas in the United States where opioid overdoses are most prevalent. This study will help organizations such as the U.S. Department of Health and Human Services and the National Institute on Drug Abuse determine areas in the United States where prevention efforts should be heavily focused.

◄► ◄► ◄► ◄► ◄►

Comparison of Machine Learning Classification Algorithms - Obesity Dataset with Categorical Variables

Neema Maharjan

OBJECTIVE: To compare the performance of Machine Learning Classification Algorithms on an Obesity dataset with categorical variables and visualize the dataset.

METHODS AND MATERIALS: The dataset, Behavioral risk factor surveillance system (BRFSS), has been obtained from Kaggle. Only fifteen categorical variables were selected from the dataset. Independent variables were Age, Sex, Marital Status, Year, Education, Race, Income, Smoking, Drinking alcohol and Mental Health. Four different classification models i.e. Random Forest, Decision Tree, Logistic Regression and Naïve Bayes were built. 70-30% data-splitting was used for Training-Testing process. The performances of models were compared in terms of accuracy. R statistical programming and Tableau software were used for model building and visualization respectively. Data from 2011 to 2015 was merged by using sqldb library in R. Python was used to replace numbers with text and save data frame.
RESULTS: All four classification models showed accuracy around 70% i.e. Random Forest (70.8%), Decision Tree (70.3%), Logistic Regression (70.6%) and Naïve Bayes (69.9%)
CONCLUSION: None of these models has potential to classify obese dataset with categorical variables. Classification of data is not improved even if some variables are removed from the model. Obesity is high in those people who do not drink alcohol in comparison to those who drink.

◄► ◄► ◄► ◄► ◄►

Information Services Role in Infection Prevention within Healthcare

Caris Newton

This paper deals with the specific problem of Information Services infection prevention. A literature search revealed a significant number of reports on the spread of Clostridium difficile (C. Diff) in the healthcare environment, especially regarding computer technology use. Sterilization and cleaning of information technology within healthcare is an issue that needs special attention.
Here are some of the findings from the literature research: Higher totals of reported hospital onset of C. Diff did not necessarily correlate to higher percentages of advanced primary providers electronic health record (EHR) usage. The following are the highest concentrations observed for C. Diff onset reported throughout U.S. for each type:

- Acute Care Hospitals: California, Texas
- Critical Access Hospitals: Washington, S. Dakota, Kansas, Iowa, Wisconsin, Illinois, Indiana
- Inpatient Rehabilitation Facilities: Texas
- Long-Term Acute Care Hospitals: Texas

We compare the C. Diff infection spread recorded in 2017 against the 2016 EHR Incentives Programs (Meaningful Use and Percentage of physicians, physician assistants, and nurse practitioners). We investigate if higher EHR utilization correlates to higher rates of C. Diff contamination and discuss best practice sterilization opportunities for information technology within patient care areas. One of the challenges is that the technology used in hospitals is not quantified nor publicly available.
We conclude that the use of technology in healthcare will continue to expand. IS has an integral role in preventing spread of bacterium such as C. Diff. Safety standards for cleaning technology including keyboards and laptops should be actively applied and reviewed with IS teams in healthcare.

◄► ◄► ◄► ◄► ◄►

Health Insurance Marketplace – Interactive Analytics and Infographic (Decision Made Easy)

Sreevarsha Potluri

For many enrolling in a marketplace was their first experience making a health insurance decision. Due to the complex structure of health insurance information, many individuals struggle to make an informed choice. Without support, many struggles to select an Insurance plan that meet their financial and health needs. So, through this project, I tried to build decision support tools such as Interactive Health care dashboard and Infographic by using different software technologies which facilitates the decision process. Datasets of qualified health plans from HealthCare.gov were used for this project. Analytical and Visualization techniques used for building this dashboard were presented in the paper.

◄► ◄► ◄► ◄► ◄►

A Statistical Analysis of Pregnancy Mortality Rate Based on Race

Cornelius Scott

Currently the United States of America has the second highest pregnancy mortality rate of developed countries in the world. A more detailed analysis shows a huge discrepancy when it comes to the basis of race for pregnancy mortality. According to previous mortality analysis white women die at a rate of 12.4 out of 1000 live births, 17.8 for Asian and Native American women and a staggering 40.0 for black women. Using mortality files available from the CDC's vital statistics website a look into finding a possible reason why the rates are so high. Race was used and 10 variables were selected relating to health issues to perform analysis using R. Logit regression models and chi-square statistics showed no real significance between the races and the selected health issues.

◄► ◄► ◄► ◄► ◄►

A Comparative Study On Patient-Level Factors Influencing Readmission Within Dialysis Facilities in the US Using ML Algorithms

Kritika Sijapati

The skyrocketing health care cost in America combined with yet unsatisfactory quality in care is significantly associated with readmission to health care facility soon after discharge. In 2011, the United States health system had to incur about $41.3 billion in hospital costs associated with readmissions. As reported by the USRDS, on average, more than one in three hospital discharges

among patients with end stage renal disease (ESRD) are followed by a readmission within 30 days. Approximate to 33% of total Medicare expenditures that goes towards ESRD inpatient treatment clearly represents the significant societal and financial burden of readmission. Results from studies on dialysis facility readmission entail different combination of patient-level, treatment and clinical as well as facility-level factors to directly impact readmission of dialysis patients within 4-30 days after discharge. The goal of this study was to identify specific patient level factors associated with dialysis facility readmission using appropriate machine learning algorithms. The datasets used for this study was extracted from the Dialysis Facility Report Data for FY 2018. It focuses on one critical outcome measure -the Standardized Readmission Ratio (SRR), which is the measures of a facility's actual over expected hospital readmission within 30-days of discharge. Thus, this study was based on the readmission measure that directly contributes in determining the SRR. Each record in the data contains list of demographic, socioeconomic, and clinical factors about a facility; from which 30 variables were selected for the purpose of this study. The Dialysis Facility Compare Calendar Year (CA) 2018 was used to lookup the total number of patients for each facility by matching on the CMS Certification Number (CCN) present on both datasets. The statistical package used for this study was R software version 3.5.1. Raw data was first cleaned and filtered for unmatched records in Excel; and then all the numeric variables were imputed for missing values in R using the K-Nearest Neighbor algorithm. The significance of attributes for model fitting was tested by plotting Correlogram chart and Correlation plot in R that resulted in showing some significant attributes to consider in relation to the main response variable, i.e. readmission. Boruta analysis was also performed to further test attribute significance, which recommended high significance of few attributes like index hospital stay than others. The main machine learning algorithms used to study the correlation of readmission with other patient-level attributes were Beta regression and Simple linear regression. After the transformations and standardization performed for different attribute types, both the algorithms performed well in displaying the correlation of patient readmission with the factors considered. The Pseudo R-squared values from Beta regression ranged from 88-90% with train and test sets. While the R-squared values from Linear regression ranged just between 78-80%. Hence, it was concluded that Beta Regression model was the better fit to correlate readmission with patient-level factors for the dialysis facility data. Both the algorithms identified specific attributes affecting patient readmission like years since ESRD diagnosis, index hospital stay, etc. and showing the correlation. This project was substantial in applying data analytics and visualization skills in the field of health care. However, it leaves space for further study into the correlation while encompassing various other factors.

◄► ◄► ◄► ◄► ◄►

Population Health – Targeted Support

Spencer Vandecar

Social and physical determinants of health make massive impacts to the health of a community. Seeking out the communities in the need of the most support will enable us to help them increase the health of their populations. This study attempts to pinpoint these communities based on publicly available data. For the purposed of this study, Excel and Tableau were used to aggregate, manipulate, analyze, and visualize five social and physical determinants of health: poverty,

population change, housing vacancy, unemployment, and crime rate. Through the use of these programs, hotspot areas of poor health were determined to be southern Mississippi River counties, many of the southwestern border counties, and the state of South Carolina into southern Georgia. This study demonstrates that by analyzing key social determinants of health we can aid the areas of the US in the direst need of support. Supporting these counties through targeted programs and policymaking will improve the health of their residents.