**DS 785: Capstone Course**

**Project Proposal**

**• Project Description**: This will be a client based project. It will be with Dr. Peggy Peissig, who is the Director of the Bioinformatics Research Center at the Marshfield Clinic Research Institute.

The deaths and injuries resulting from suicide attempts are regarded as very preventable with proper early intervention. A recent NIH funded study appearing in JAMA Psychiatry (April 29 ,2017) found that a screening program implemented in emergency departments lowered attempts by 30 percent in a 52-week follow-up period, compared to those who received standard department care. In addition, there is some evidence that indicating patients who have attempted suicide make use of an ER prior to the attempt (Cerel, 2016). This may indicate that ER visits could be a factor in identifying potential suicide attempts.

This project will examine the data associated with behavioral health and suicide and attempt to make a predictive model to see if patients at risk for suicide can be identified. This would potentially enable health care providers to identify at risk patients and provide necessary care to prevent the suicide attempt.

**• Rationale (list of reasons) for undertaking this project.**

Identifying the high-risk patients and providing necessary care could result in significant reduction of the social and economic costs of suicide.

Expand knowledge of information that can be inferred from Marshfield Clinic Electronic Health Record(EHR)

Explore the possible benefits of predictive models using electronic health record data in assisting research and patient care.

**• Proposed Project Title**: Development of Behavioral Health Predictive Models of High Risk (Individuals) for ER Services and Suicidality/Suicide.

**• Proposed Project Purpose**:

**Client-based project**:

**Organization**: Marshfield Clinic Research Institute

**Contact**: Peggy Peissig, PhD Research Scientist and Director of BIRC.

# Purpose: The goal of this research is to Develop a predictive model for behavioral health patients to inform health care operations.

**• Capstone Project Objectives:**

**AIM 1: Exploratory data analysis of behavioral health (BH) patients with respect to suicidality/suicides/suicide attempts and/or ER visits;**

1. Characterize BH patients with suicide-related diagnoses.
2. Characterize BH patient’s utilization of emergency room services.

**AIM 2: Develop a predictive model to identify BH patients at high risk of future suicide/suicide attempts**

**AIM 3: Apply knowledge gained in previous data science courses to a non- academic research project.**

**AIM 4: Develop predictive model for identifying BH patients at high risk of future emergency room utilization (if time).**

**• Timeline:**

**Week 1:** Project proposal idea. Due June 5, 2017.

**Week 2:** Project proposal. Due June 12, 2017.

**Week 3:** Begin gathering data and cleanup (look for unrealistic values, such as adults that are 12cm tall. These will be removed). Also examine a few records manually to ensure that the data pull is accurate.

**Week 4:** Data cleaning, separate testing and training sets, final data prep, begin exploratory analysis. Activity update 1 due June 19, 2017.

**Week 5:** Complete exploratory analysis. This may result in changes in type of predictive models that will be tried. For now, the following will be considered. Interview 1 due June 30, 2017.

**Week 6:** Model testing – Logistic regression Activity update 2 due July 3, 2017.

**Week 7:** Model testing - SVM

**Week 8:** Model testing – Neural networks Activity update 3 due July 17, 2017. Peer Interaction Section – July 19, 2017

**Week 9:** Model testing – Clustering (maybe, not sure will use this one) – Interview 2 July 28, 2017.

**Week 10:** Model evaluation – choose best performing model and build predictive model. Test against withheld testing set. Begin final document Activity update 4 due July 31, 2017.

**Week 11:** Finish final document Activity update 5 due August 14, 2017.

**Week 12:** Final document due August 18, 2017.

This timeline is subject to change.

**• Interviewees**

Peggy Peissig, PhD. Director of Biomedical Informatics Research Center at Marshfield Clinic Research Institution. She does a lot of work in machine learning and is the project leader for this project.

Ahmad Pahlavan Tafti, PhD Associate Research Scientist in BIRC also active in the field of machine learning, he is interested in discussing deep learning.

Zhan (Harold) Ye, PhD Biostatistician at BIRC (will be interviewed if one of the two above are unavailable).

**• Application of Data Science Concepts**

This project will use several techniques learned in the data science program. Data cleansing, text analysis, machine learning, and predictive modelling will all play a role in this project.

# Exploratory data analysis:

Exploratory data analysis will be performed to characterize behavioral health patients who are at increased risk of ER visits and/or suicides (including suicide attempts, and suicidal ideation). Data for this aim will come from the EHR. We will examine frequency counts for various sub-classifications of BH patients; characterize BH patients by summarizing median age, sex, age at first BH diagnosis, number of ER visits before and after first BH diagnosis, and other relevant variables. In addition, prevalence and incidence rates within the Marshfield Clinic Health System will be determined.

# Predictive modeling:

We plan to use machine learning approaches to develop models to predict the risk of a BH patient either using ER services or committing suicide. We will investigate the model effectiveness using several prediction time periods (i.e. 30 days prior to the prediction, 60 days, 90 days, etc.). Several modeling approaches will be considered: logistic regression, SVM, random forest, artificial neural networks and perhaps others. The models will be evaluated for performance to identify, which if any, are effective at predicting suicides based on historic data. Ten-fold cross validation will be used. After identifying the best model, this model will be tested on a set of withheld data from the exploratory phase to measure the effectiveness of the model against data that played no role in the training or validation.

•**Description of Final Document.**

The final document will be a report on the results of the project. It will include a description of the training and testing data sets, the methods used, and the findings. While it’s unlikely a manuscript for publication in a peer reviewed journal will be produced by the end of the project, the final document will serve as the template for the publication that may eventually result.

# References

Cerel J, Singleton MD, Brown MM, Brown SV, Bush HM, Brancado CJ., [Emergency Department Visits Prior to Suicide and Homicide: Linking Statewide Surveillance Systems.](https://www.ncbi.nlm.nih.gov/pubmed/26620917) Crisis. 2016;37(1):5-12. doi: 10.1027/0227-5910/a000354. Epub 2015 Dec 1.

**Appendix**

Previously submitted project idea is below. This is essentially what was submitted to the Internal Review Board(IRB) for approval.

**Study Proposal for Medical Record Review Research**

# Title: Development of Behavioral Health Predictive Models of High Risk (Individuals) for ER Services and Suicidality/Suicide

# Principal Investigator & Other Investigators: Peggy Peissig PhD & John Mayer PhD

# Purpose: The goals of this research are to:

1. Develop a predictive model for behavioral health patients to inform health care operations

# Specific aims of the research:

**AIM 1: Exploratory data analysis of behavioral health (BH) patients with respect to suicidality/suicides/suicide attempts and/or ER visits;**

1. Characterize BH patients with suicide-related diagnoses.
2. Characterize BH patient’s utilization of emergency room services.

**AIM 2: Develop a predictive model to identify BH patients at high risk of future suicide/suicide attempts**

**AIM 3: Develop predictive model for identifying BH patients at high risk of future emergency room utilization**

# Background information

**Suicide:** Roughly 40,000 people a year die in the United States from suicide. This number has increased about 24% over the last fifteen years. The Center for Disease Control estimates this has an annual cost of 44 billion dollars. Wisconsin Department of Health Services reports that there are greater than 700 suicides per year in Wisconsin, with an additional 5,500 hospitalizations due to self-inflicted injury. The hospitalization cost alone of these self-inflicted injuries is estimated to be $369 million.

The deaths and injuries resulting from suicide attempts are regarded as very preventable with proper early intervention. A recent NIH funded study appearing in JAMA Psychiatry (April 29 ,2017) found that a screening program implemented in emergency departments lowered attempts by 30 percent in a 52-week follow-up period, compared to those who received standard department care. In addition, there is some evidence that indicating patients who have attempted suicide make use of an ER prior to the attempt (Cerel, 2016). This may indicate that ER visits could be a factor in identifying potential suicide attempts. Identifying the high-risk patients and providing necessary care could result in significant reduction of the social and economic costs of suicide.

A second area of inquiry of this study will be to assess and predict the ER utilization of behavior health patients. There is literature suggesting increased use of ER departments by behavioral health patients. This is resulting in ER providers treating behavioral health issues, which they may not have much experience or training to do. If a model can detect and predict these patients, it may be possible to identify undiagnosed behavioral health patients and refer them to a behavior health provider.

# Brief summary of statistical analysis that will be conducted on the data collected.

We plan to use data from the Marshfield Clinic EHR, obtained via the research data warehouse. Patient demographics, diagnoses, procedures, medications, labs, emergency room visits, and observations such as: blood pressure, height and weight (for BMI calculations) may be considered. There may be some use of natural language processing of clinic office notes to determine if the ICD coded patients match the NLP results. Record dates will be explored to verify a cause and effect relationship between the suicide/attempt and the conditions examined.

All data will be stored on Marshfield Clinic password protected encrypted laptops or on servers maintained by the Marshfield Clinic in the BIRC. Only the principal investigators and a designated data validation person will have access to patient data.

Behavioral health issues will be identified by ICD coding (see attached spreadsheet). Suicide attempts are identified by E95# in ICD 9, and X60-X84 in ICD 10. These codes identify the method, not the result. Suicide ideation is identified by V62.84 in ICD 9 and R45.85 and R45.851 in ICD 10. A small subset of each type of record (suicide, suicide attempt, suicide ideation, and behavior health diagnoses) will be reviewed in CMR to verify the codes are identifying the patients of interest. This is especially necessary for suicide, as there is no code to explicitly identify it. Some additional sections of the EHR (ER notes, hospital admission/discharge summary, etc.) that may substantiate the cause and effect relationship, and/or provide additional socioeconomic and demographic data, may be reviewed by the research programmer and/or research project abstractor to validate a sample of cases

# Exploratory data analysis:

We will perform exploratory data analysis to characterize behavioral health patients who are at increased risk of ER visits and/or suicides (including suicide attempts, and suicidal ideation). Data for this aim will come from the EHR. We will examine frequency counts for various sub-classifications of BH patients; characterize BH patients by summarizing median age, sex, age at first BH diagnosis, number of ER visits before and after first BH diagnosis, and other relevant variables. In addition, prevalence and incidence rates within the Marshfield Clinic Health System will be determined.

# Predictive modeling:

We plan to use machine learning approaches to develop models to predict the risk of a BH patient either using ER services or committing suicide. We will investigate the model effectiveness using several prediction time periods (i.e. 30 days prior to the prediction, 60 days, 90 days, etc.). Several modeling approaches will be considered: logistic regression, SVM, random forest, artificial neural networks and perhaps others. The models will be evaluated for performance to identify, which if any, are effective at predicting suicides based on historic data. Ten-fold cross validation will be used. After identifying the best model, this model will be tested on a set of withheld data from the exploratory phase to measure the effectiveness of the model against data that played no role in the training or validation.

# Benefit to society

As noted above, suicides and attempts have a significant impact on society, both in terms of lives lost and economic impact. Attempts can result in debilitating lifelong injuries. However, if potential suicide attempts can be identified and the individual treated, these are preventable events. A system that can identify patients at risk and provide necessary treatment could reduce the adverse social and economic impact associated with these events. In addition, behavior health patients may be using ER’s at a higher rate for behavior health issues. Identifying these patients and referring them to behavioral health caretakers instead of ER doctors would be better for the patients’ health and reduce ER utilization.