DEALER: Data Enrichment Algorithms for Predicting Extremely Rare Multi-Drug-Resistant Healthcare Associated Infections

Objective/Motivation: Healthcare-associated infections (HAI) affects hundreds of thousands of patients every year in the U.S. and burden the healthcare systems with billions of dollars. Although the number of HAI cases has decreased due to the advancement of care and infection prevention guidelines, the rise of MultiDrug-Resistant Organisms (MDRO) continues to challenge HAI treatment. Timely and accurate prediction of patients' risk of developing MDRO-causing HAI during their hospital stay offers a potentially valuable vantage point for the care team to carry out proper preventative measures. We propose to leverage recent advances in machine learning to develop novel methods that accurately predict for each patient the risk of developing any of the extremely rare but hard to treat MDRO-caused infections using the longitudinal patients' data.

Our goal is to develop a new interpretable machine learning system, DEALER, to timely and accurately predict the risk of MDRO-caused infections and integrate it to the learning laboratory of the Ohio State University hospitals to continually use the flux of patients' data and provide the care team with actionable information to prevent HAIs. Desired prediction tasks include both classification, e.g., predicting if a patient develops an infection next week or not, and regression, e.g., time to infection. Many of the MDROs are extremely rare and therefore per infection prediction is very challenging and needs development of new tailored machine learning methods. We envision DEALER as a novel machine learning algorithm that leverages all of the pooled data from different MDROs infection cases but separately predicts the risk of each type of infection. In this manner, we enrich the dataset for each specific infection type by using all available information from other pathogens and learn both shared predictive risk factor between MDROs and individual per specific risk factors. DEALER will integrate heterogeneous sources of data like different hospitals and units therein while leveraging the available hierarchy of data source to improve learning outcomes.

Team, Challenges, and Intellectual Merit: We propose a four-year research initiative during which we will build a machine learning system for real-time MDRO-caused HAI prediction and integrate it to the Ohio State Hospital system. Our team brings together two PIs with expertise in machine learning and data mining (from the University of Minnesota and the Ohio State University) and two PIs with expertise in bioinformatics and infectious diseases (from the Ohio State University). Our project will address several fundamental challenges:

Prediction of rare events: We will develop novel machine learning algorithms that deal with extremely imbalance datasets by leveraging similarities and differences between rare events.

Structure of data: We will exploit the inherent structure of our heterogeneous data source to improve HAI prediction. Sources of data heterogeneity are hospitals and units inside each of them which makes the prediction of HAI for each patient interconnected not only to the patients in the same unit but also link the outcome to other units and patients therein.

Real-time Prediction: Timely prediction of HAI risk is a significant factor determining the utility of any preventative intervention. We will pursue a real-time prediction strategy that utilizes longitudinal patient data to notify the clinicians for proper, timely intervention.

Interpretability: We will incorporate recent advances in interpretable machine learning into our system in order to explain the prediction outcomes for the care team.

Missing data: We will investigate different data imputation strategy to address the common missing data phenomena in electronic health records.

Hidden factors: We will take into account the presence of many critical unknown factors like cleaning protocols and colonized carriers and ensure that DEALER outcome is robust to them.

Intellectual Merit

Broader Impacts

Project Description

1. Introduction

With the ever-growing availability of the *Electronic Health Records* (EHR), many data scientists and health-informaticians have been presented with an unprecedented opportunity to improve the healthcare system by providing evidence-based solutions. One of the costly, life-threatening but preventable problems of the healthcare system is *Healthcare-Associated Infections* (HAI) which are those infections that a patient get while receiving treatment for another disease. HAIs are categorize broadly into two categories of device-associated infections (DAI) or Surgical-Site Infections (SSI). The threat of HAIs has been amplified by the recent rise of *Multi-Drug Resistance Organisms* (MDRO) which makes the treatment options of the affected patients very limited. HAIs caused by MDROs are claiming tens of thousands lives every year and burden the U.S. healthcare system with around 10 billion dollars. Although there has been a lot of work for predicting the risk of each type of HAIs individually [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] or more prevalent pathogens like Clostridium difficile (CDI) bacteria [18, 19, 20, 21, 22, 23], there has been no effort in addressing the less common infections because of their scarcity. In this work, we propose to harness the recent advances in machine learning literature, specifically multi-task learning, multi-label prediction, and multi-target regression to holistically assess the patients' risk for many life-threatening MDRO-caused infections.

Another motivation behind our proposed work stems from the fact that the severity and extent of MDRO-caused HAIs are vastly depends on the facility and the population of patients in that facilities. For example, risk prediction models developed using a dataset from an intensive care units (ICU) can not be used directly to predict the risk of HAIs for patients in long-term care facilities. Also, since we are dealing with extremely rare events, it is not feasible to collect representative data from each facility or small care unit and develop their own predictive models from scratch. For example, although CDI cause 500,000 infections per year in U.S. hospitals, its prevalence is less than 10 percent, which leaves a 200 beds hospital with too few positive samples to build its own model for CDI risk prediction. We propose a multi-task learning framework with a tree structure task hierarchy to model the different facilities inside the hospital and to better predict the risk of various HAIs by leveraging the common and individual risk factors of patients hospitalized in different units or facilities. Furthermore, we will develop transfer learning methods for predicting risk of HAIs where patients features are not exactly equal in different facilities.

Creating risk scores and determining the risk factors for different HAIs is the limit of current risk prediction models. We want to move one step further and provide the clinicians with an ordered list of possible interventions which takes into account convenience of the patient and his/her family along with the financial cost of preventative interventions for the whole healthcare system. TODO: need to expand this based on John and Courtney's inputs.

Goal: Our long term goal is to systematically develop prediction models which exploit data collected from multiple (heterogeneous) sources to predict different (but related) extremely rare outcomes while taking into accounts the relation between data sources and also outcomes to improve prediction accuracy of all outcomes. In this project, we plan to investigate such models toward accurate risk prediction of various extremely rare MDRO-caused HAI, in order to provide clinicians with evidence-based list of possible intervention options. The work will focus on *enriching* risk prediction models with hierarchical side information about the relation of different tasks (multiple hospitals and units therein) as well as grouping of different types of HAIs (device-associated or SSI and sub-groups therein). We also build on top of the recent success of multi-label prediction and multi-target regression methods advanced by the machine learning community to simultaneously predict both the risk of developing multiple MDRO-caused HAI and also time to infections. Such predictive model will elevate our ability to better understand the risk factors involved in any of the HAIs, paving the way for accurate suggestion of preventative interventions to clinical practitioners.

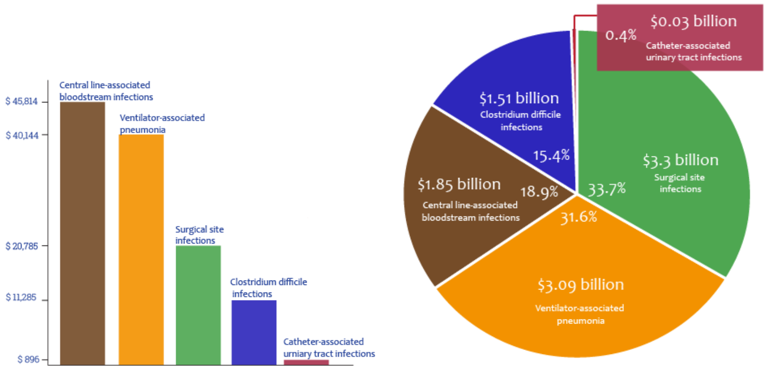
Specific Objectives: Working with a multi-disciplinary team including computer scientists, (machine learning, data mining), bioinformaticians, economists, and infectious diseases specialist, we will use the three years data of all patients admitted to the Ohio State University (OSU) hospitals which has been collected as a part of the infectious disease Learning Laboratory (LL) project (link to the website and; letter of support from Susan Moffatt-Bruce at the OSU hospital attached) to develop a Machine Learning (ML) system, called DEALER, that systematically uses the extremely rare HAI-causing MDROs incidents in the pooled data from multiple hospitals and units therein, and accurately predict the temporal risk of a patient developing any of the MDRO infections during their stay at a hospital. We hypothesize that enriching our samples by merging various heterogeneous source of data along with simultaneous prediction of multiple related outcomes will both improve the prediction accuracy of all adverse events and also provides more interpretable models. In particular, our long term goal is to integrate DEALER into the OSU's patient safety learning lab called IDEA4PS whose primary goal is to improve workflows and information transfers in the healthcare environment in order to enhance patients' outcomes. Advances in both computational and statistical aspects of data analysis to be pursued in the project will be key to translating the IDEA4PS dataset into a humna-in-the-loop system that reduces adverse health events.

The ML system that we envision is a framework for surveillance of "hot spots" of HAIs which exploits electronic medical records and provides real-time risk prediction and recognizes concerning trends sooner so that clinicians can implement timely and effective interventions. Our project is motivated by a strong desire to develop methods and systems to mitigate the threat of MDROs-caused HAIs in healthcare systems. To this end, we anticipate that this project will produce the following:

- It will enable accurate and timely identification of patients with high risk of developing multiple types of HAIs and provides practitioners with interpretable decisions.
- It will provide smaller healthcare institutes methods to tailor mathematical HAI's risk prediction models developed for larger institutes for their environment.
- It will assist clinicians in selecting proper preventative intervention which factors in patients preferences along with healthcare costs.

Significance: The United States has made significant progress toward the collective goal of eliminating HAIs, and as a result, healthcare in the U.S. is safer now than it was even 10 years ago. Building upon this success and continuing towards the elimination of HAIs is critical [24]. With the rise of MDROs in U.S. [25, 26, 27] and the CDC's goal of substantially reducing multiple types of HAIs by 2020 [28], addressing MDROs-causing HAIs is of extreme value.

There has been focused studies to predict patients' risk of developing infection due to the more frequent MDROs using EHRs. Clostridioides difficile (C. diff.), the most prevalent cause of HAI which sometimes become drug-resistant [29, 30, 31], has been extensively studied [20, 32, 18, 33]. Methicillin-resistant Staphylococcus aureus (MRSA) is the second most frequent pathogen whose risk prediction from EHR has been recently studied in isolation [34]. Due to the scarcity of HAIs associated with other MDROs, comprehensive analysis of risk factors and risk stratification of patients for them remained unexplored. Since the distribution of frequent pathogens and their resistance pattern are changing [35] and many of the extremely rare MDROs are more fatal than



(a) Per case cost.

(b) Contribution percentage.

Figure 1: Cost of the five most common HAIs in the U.S. reaches the total of $9.8 billion annually. Data source from [57] and figures from [58]

.

their rare counterparts [36], comprehensive investigation of all HAIs caused by MDROs is of extreme significance.

2. Scientific Background and Data Sources

One of the key initiatives of the US government to decrease cost and improve healthcare quality is the mandating of health care providers to implement Electronic Health Record (EHR) systems [37]. EHRs are real-time, patient-centered records that make information such as demographics, medications, laboratory test results, diagnosis codes, and procedures, available instantly and securely to authorized users. While the primary goal of an EHR system is electronic documentation of patients' care, the collected data is often serve as an input source for many clinical informatics applications with the goal of extracting actionable information to improve diagnostics and patients outcomes [38, 39, 40]. Some of the tasks that EHR has been successfully used for include cohort identification [41, 42], risk prediction [43, 33], biomarker discovery [44], and adverse event detection [45, 46]. Broadly, an adverse event is a detrimental effect of patient health as a result of medical care. The goal of our proposed work is to predict a specific adverse event, namely rare infections using EHR.

Hospital-Acquired Infection (HAI) or more broadly healthcare-associated infection is an infection a person get while he/she is receiving health care for another condition. Although often being preventable [47, 48, 49], HAIs have been the cause of many diseases and deaths among hospitalized patients [50, 51, 52, 53] and their elimination is a priority of the Department of Health and Human Services [54]. It is estimated that at any given time, about 1 in 25 inpatients have HAI [55, 54]. Other studies report the estimated number of HAI associated death to be around 99000 per year [53] while burdening the U.S. healthcare budget by $5 to $10 billion annually [56, 57], Figure 1.

Though the rate of many HAI types have dropped over the last decade [47] increasing trends in prevalence rates of HAIs caused by *Multi-Drug Resistant Organisms* (MDRO) has been observed [59, 24]. MDROs are pathogens, predominantly bacteria, that have the ability to defeat many of the known antibiotics, and therefore infections caused by MDROs are difficult or sometimes impossible to treat [60]. The prevention and control of MDROs is a national priority [61, 62]. MDROs may naturally exists in the population of bacteria like gut microbiom without any adverse health effect, but when antibiotics kill other types of organisms in the population, MDROs take over all of the

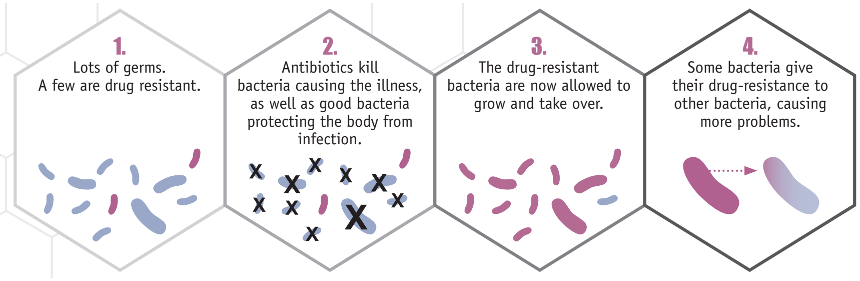

1.
Lots of germs.
A few are drug resistant.
2.
Antibiotics kill bacteria causing the illness, as well as good bacteria protecting the body from infection.
3.
The drug-resistant bacteria are now allowed to grow and take over.
4.
Some bacteria give their drug-resistance to other bacteria, causing more problems.

Figure 2: Antibiotic resistance as an outcome of antibiotic overuse and natural selection [36].

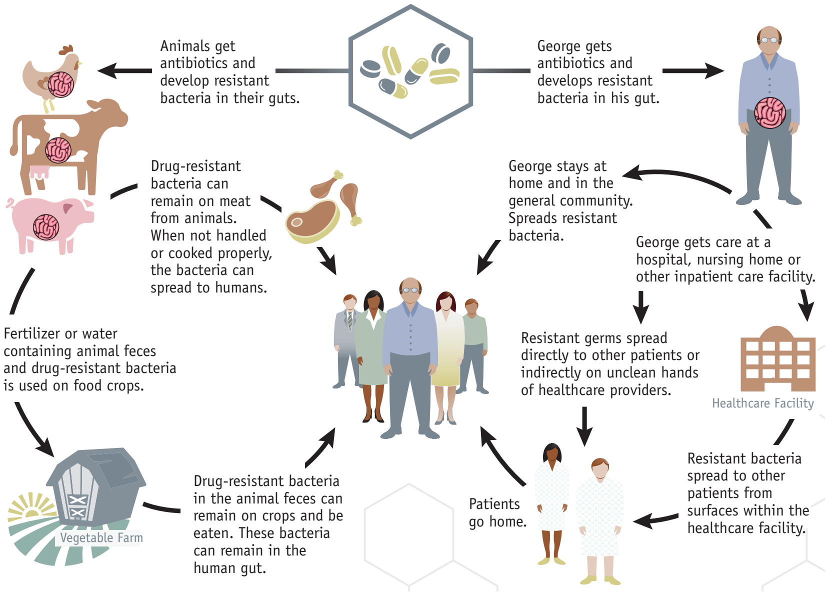

Animals get antibiotics and develop resistant bacteria in their guts.
George gets antibiotics and develops resistant bacteria in his gut.
Drug-resistant bacteria can remain on meat from animals. When not handled or cooked properly, the bacteria can spread to humans.
George stays at home and in the general community. Spreads resistant bacteria.
George gets care at a hospital, nursing home or other inpatient care facility.
Fertilizer or water containing animal feces and drug-resistant bacteria is used on food crops.
Resistant germs spread directly to other patients or indirectly on unclean hands of healthcare providers.
Healthcare Facility
Vegetable Farm
Drug-resistant bacteria in the animal feces can remain on crops and be eaten. These bacteria can remain in the human gut.
Patients go home.
Resistant bacteria spread to other patients from surfaces within the healthcare facility.

Figure 3: Animal food and healthcare system as sources of generating MDROs [36].

resources and produce offsprings that results in a generation of organisms all resistant to antibiotics. Finally, MDROs may amplify their presence in a population by pass their resistant-causing genetic materials to other bacteria [36]. The process of MDRO growth due to the evolutionary pressure is depicted in Figure 2.

Multi-drug resistance has been prevailed over the last decade due to overuse of antibiotics in different settings, e.g., food animal production [63]. One of the main places that hosts most of MDROs are healthcare facilities, Figure 3. Although transmission of MDROs is most frequently documented in acute care facilities, all healthcare settings are affected by the emergence and transmission of antimicrobial-resistant microbes. It should be noted that, the severity and extent of disease caused by MDROs varies by the population(s) affected and by the institution(s) in which they are found. For example, risk of patients in institutions with varying physical and functional characteristics, like long-term care facilities, intensive care units (ICU), burn units, neonatal ICUs are varying a lot [60]. Therefore one can not build a risk prediction model for an MDRO using data from one population and transfer the built model to another facility and use it for another population of patients without suffering from prediction loss [64]. Because of this, any risk prediction model of these pathogens need to be tailored to the specific needs of each population and individual institution [60, 64].

Table 1: Table of Acronyms and their Descriptions

Name	Description	Name	Description
General Terms		**Type of HAI**	
EHR	Electronic Health Record	CLABSI	Central Line-Associated Bloodstream Infection
CDC	Centers for Disease Control and Prevention	CAUTI	Catheter-Associated Urinary Tract Infections
NHSN	National Healthcare Safety Network	SSI	Surgical Site Infection
HAI	Healthcare-Associated Infection	VAP	Ventilator-Associated Pneumonia
MDRO	Multi-Drug Resistant Organism	**Type of Pathogens**	
LabID	Laboratory Identified Event	CDI	Clostridium difficile, C. diff.
LL	Learning Laboratory	VRE	Vancomycin-resistant Enterococci
TAP	Targeted Assessment for Prevention	MRSA	Methicillin-resistant Staphylococcus aureus
SIR	Standardized Infection Ratio	GNB	Gram-negative Bacteria

In most cases, antibiotic-resistant infections require extended hospital stays, additional follow-up doctor visits, and costly toxic alternatives [65, 36]. In 2013, the Centers for Disease Control and Prevention (CDC) published a threat report outlining the top eighteen drug-resistant threats to the U.S. eight of which are healthcare associated with more than 600,000 total occurrences and around 30000 death per year [36]. To dive into more scientific details of HAI and also elaborating our data source, in the following, we first introduce the required terminology and then present our dataset. For conveniance Table 1 summarizes all of the acronyms and their description.

Types of HAI: HAIs are categorize by CDC into two broad categories of device-associated infections (DAI) or Surgical-Site Infections (SSI) [1, 2, 3, 4]. Primary device-associated infections include Central Line-Associated Bloodstream Infection (CLABSI) [5, 6, 7, 8, 9], Catheter-Associated Urinary Tract Infection (CAUTI) [10, 11, 12], and Ventilator-Associated Pneumonia (VAP) [13, 14, 15, 16, 17]. Unfortunately, in all of the cited studies, only one type of HAI is studied in isolation. Based on these studies, many different risk scores have been developed, and preventative measures have been suggested.

To fully harness the power of data, we suggest to study *all* of different types of HAIs together while incorporating the relation of different prediction task in our model. We hypothesized that for a specific infection-causing pathogen, all of the HAIs may share similar risk factors and at the same time each one should have its dedicated important risk factor. We suggest to model the risk of an HAI as a multi-task learning where task are hierarchically related. For example, all HAIs are divided to DAIs and SSIs while DAIs are divided to subtasks based on the relevant device and SSIs are divided based on the place of surgery. Thus, we enrich our dataset by considering are HIAs together and we are capturing shared and individual per-task risk factors at the same time.

Types of Pathogens: A perpendicular categorization of HAIs is based on the pathogen causing the infection. Although bacteria, fungi, and viruses can cause HAI, four of the most threatening HAI causes are all bacteria and are mostly multi-drug resistant [36, 60]. In the followings, we briefly introduce each one of the MDROs studied in this proposal:

- Clostridium difficile (CDI): CDI also known by C. diff bacteria causes life-threatening diarrhea and colitis (an inflammation of the colon). Those most at risk are people, especially older adults, who take antibiotics and also get medical care. CDC classifies CDI as an urgent threat and estimates it causes 500,000 infections per year where 29,000 died within 30 days of the initial diagnosis and 15000 death cases were directly attributed to CDI [36]. Although CDI in general is not considered as MDROs but some variant of it shows resistant to antibiotic [29, 31] and it is monitored by CDC along with other MDROs [66].

- Methicillin-resistant Staphylococcus aureus (MRSA): Methicillin-resistant Staphylococcus aureus (MRSA) is a type of staph bacteria that is resistant to certain antibiotics called beta-lactams and has been classified as a serious threat by CDC [36]. These antibiotics include methicillin and other more common antibiotics such as oxacillin, penicillin, and amoxicillin. Severe or potentially life-threatening MRSA infections occur most frequently among patients in healthcare settings. By CDC estimate, every year 80400 MRSA infections are happening in the U.S. where more than 11200 of them lead to death [36].

- Vancomycin-resistant Enterococcus (VRE): Enterococci cause a range of illnesses, mostly among patients receiving healthcare. Vancomycin-resistant Enterococci are specific types of antimicrobial-resistant bacteria that are resistant to vancomycin, the drug often used to treat infections caused by enterococci. Enterococci are bacteria that are typically present in the human intestines and in the female genital tract and are often found in the environment. Vancomycin-resistant Enterococci infections is considered a serious threat with the death toll of 1300 per year [36].

- Gram-Negative Bacteria (GNB): GNB is a group of pathogens cause infections including pneumonia, bloodstream infections, wound or SSI, and meningitis in healthcare settings. GNB are resistant to multiple drugs and are increasingly resistant to most available antibiotics. These bacteria have built-in abilities to find new ways to be resistant and can pass along genetic materials that allow other bacteria to become drug-resistant as well [36]. Gram-negative infections include those caused by Klebsiella, Acinetobacter, Pseudomonas aeruginosa, and E. coli., as well as many other less common bacteria [67].

There have been many model developed ...

NHSN: The National Healthcare Safety Network, of the Centers for Disease Control and Prevention (CDC), is the nation's most widely used health care-associated infection tracking system. Since 2009, infection data has been reported to the NHSN to track the national progress of the reduction of HAIs.

Laboratory identified (LabID) Event: For reporting to the National Healthcare Safety Network; an infection is considered laboratory identified when a patient sample is tested and confirmed positive by laboratory test only (i.e., clinical evaluation of the patient is not required).

Hospital-onset HAI: For LabID events, an infection is considered hospital-onset if the positive specimen is collected on or after the fourth day of admission.

Targeted Assessment for Prevention (TAP) strategy: It is known that some hospitals and some units are more susceptible to MDROs and therefore have higher number of HAI cases. TAP is a method developed by the CDC to use data for action to prevent HAIs. The TAP strategy targets healthcare facilities and specific units within facilities with a disproportionate burden of HAIs to address infection prevention gaps.

Standardized Infection Ratio (SIR): SIR is a statistic used to track HAIs over time, at a national, state, or facility level. The SIR is the ratio of the actual number of HAIs at each hospital, to the predicted number of infections. The predicted number is an estimate based on national baseline data, and it is risk adjusted. Risk adjustment takes into account that some hospitals treat sicker patients than others.

IDEA4PS Dataset: In 2015, the OSU was awarded a four-year program project grant from the Agency of Healthcare Research and Quality to establish The Institute for the Design of Environments Aligned for Patient Safety (IDEA4PS). This grant is being used to identify and explore how feedback of information can be used to inform the development of robust practices that lead to improved patient safety. As a part of IDEA4PS temporal patient's data has been collected to conduct surveillance of healthcare-acquired infections in real time and to provide clinicians with actionable information.

We will leverage IDEA4PS data to train our ML system, DEALER and integrate it to the OSU hospital system. add data stat here.

Interventions: Although it seems that for the patients with higher risk of infection the care team should carry out the most effective interventions, in reality, this is not the best measure. Serious interventions are usually more difficult to tolerate by the patient, and they are of more significant cost for the healthcare system. Besides, for the specific intervention of antibiotic administration the rise of drug-resistant organisms urge us to use antibiotics only if they are absolutely necessary and by following strict "antibiotic stewardship" guidelines.

The epidemiology of emergent MDROs in health care settings must be monitored to allow for appropriate adaptations to current infection control interventions, including antimicrobial prophylaxis, isolation strategies, and screening strategies. Vaccines are also a powerful way to prevent thousands of infections and deaths that occur each year for diseases such as influenza and hepatitis. Currently, there are eight vaccines licensed in the U.S. that target pathogens that can be acquired in health care settings. The appropriate use of these lifesaving interventions needs to be defined.

A successful ML system should take into account the range of possible intervention and based on its confidence about the predicted outcome suggest a proper intervention.

3. Motivating Problems and Challenges

Problem 1: Given a set of MDROs where some of them are extremely rare, we want to leverage patients static (demographic, sex, etc.) and temporal features (vitals, lab results, etc.) to predict the desired outcomes. Examples of the desired outcome include, but not limited to, timely prediction of the occurrence of a specific MDRO incident (discrete outcome) or prediction of time to infection (continuous outcome).

Problem 2: Given a task of predicting an extremely rare HAI, incorporate the domain expert knowledge about the involved risk factors into the ML system and provide an interpretable answer that suggests which collection of risk factors are affecting the prediction outcome the most.

Problem 3: Given a heterogeneous source of data like hospitals and units therein where the whole environment that data has been gathered is different and also collected features are not exactly similar, tailor the prediction for each hospital or unit while leveraging the similarities between each environment.

We describe below some of the key technical challenges involved in addressing these problems and how we plan to solve them:

Prediction of rare events: As reported by many studies, HAIs affect less than 10% of the hospital admitted patients which makes any collected dataset highly imbalanced. Among all HAIs some of them like C. diff. are rare while others like VRE are extremely rare. Highly imbalanced data induce unique challenges for learning algorithms. We propose to ...

Structure of data: We propose to study admitted patients to the OSU hospital system which consists of x main hospitals each of which has between y to z units. A naive algorithm would put together all of the patients' data and analyze that as a single dataset. However, a more sophisticated approach will take into account the hierarchical structure that the hospitals induce on patients which is extremely important in the context of infectious diseases. Another important structure is the relation of different MDRO, exploiting similarities and differences between MDROs will potentially improve the outcome of prediction algorithms.

Timely Prediction: Data shows that the longer a patient stays in the hospital, the more it is probable to acquire an HAI. Therefore using patients longitudinal data is essential for any practical algorithm. Also, evaluation of any method should consider the length of time window between flagging a patient as high risk and the actual infection happening.

Interpretability: Since different outcomes of the risk prediction algorithm require the care team to carry out different preventative measures, e.g., isolation, or administration of prophylaxis, the algorithm should explain its decisions. Clinicians are hesitant to incorporate weakly evaluated black-box ML algorithms into their practice. Therefore, interpretability should be addressed in every level of the proposed ML system.

Missing data: The widespread prevalence of missing data in electronic health records presents a significant challenge for any ML algorithm. Different causes of missing data in the EHR data may introduce unintentional bias, and therefore any level of the proposed ML system should be robust to missing data.

Hidden factors: Many known risk factors for HAIs are hard to measure and therefore are unknown to the algorithm. Factors like cleaning protocols, quality of care, and the presence of colonized carrier are important hidden risk factors that make the prediction task more complicated. For example, it is known that the C. diff. spores can persist on environmental surfaces, and therefore the role of environmental cleaning is likely to be significant. For MRSA, it is widely held that the primary reservoir for transmission in the health care setting is infected or colonized patients and that patient-to-patient transmission occurs indirectly via transient carriage by health care personnel or through shared equipment that is contaminated.

4. Technical Approach

Notation:

Data sharing, data enrichment, multi task learning, sparse task differences, imbalanced classification.

Problem 1: A few questions/commentspotential directions:

- For patient specific prediction, consider/contrast with *Cox proportional hazards model* (CPHM), and other forms of *survival analysis*. Here, 'survival' refers to staying un-infected.
- For a group based model, consider/contrast with *multiple instance learning* (MIL). The goal here is to characterize if any one individual in a cohort will have MDR infection. The cohorts can be formed based on (static and temporal) patient covariates, say by clustering, and/orbased on physical location, e.g., a certain unit in a clinic. MIL have had mixed success in the literature, so we have to be careful about what we propose.

Problem 2: How do we plan to make things interpretable? Couple of possible approaches:

- Importance of individual features: One can assess individual feature importance in nonlinear models, including random forests, gradient boosted trees, and deep nets, by an ablation study, i.e., by studying predictive accuracy by leaving each feature out.
- Importance of sub-groups of features: Doing ablation study with sub-groups naively leads to an exponential blow-up in computation. One can consider doing PCA regression, sufficient dimensionality reduction, or Shapley value regression in such settings.

Few considerations in the current context:

- Can such feature importance be done efficiently?
- How does one take into account the fact that the target events are rare?

- How does one handle the small sample regime, so feature importance is assessed at the population level, not because it helps overfit on the training set?
- How do the proposed approaches relate to sparse methods? Shall we use sparse methods instead?
- How do the proposed approaches related to frequent pattern mining for prediction problems?

Problem 3: Since we will be combining data from multiple clinics, the samples are going to be disjoint (similar to our current data-sharing model), but the features will have some shared and some unique covariates (different from our current model).

5. System Evaluation

Evaluation of any method should consider the length of time window between flagging a patient as high risk and the actual infection happening.

6. Comparison of Our Approach to Related Work

7. Community Outreach and Education

8. Results from Prior NSF Support

9. Collaboration Plan

Name of PI: NSF-Program (Award Number) "Title of the Project" ($AMOUNT, PERIOD OF SUPPORT). **Publications:** List of publications resulting from the NSF award. A complete bibliographic citation for each publication must be provided either in this section or in the References Cited section of the proposal); if none, state: "No publications were produced under this award." **Research Products:** evidence of research products and their availability, including, but not limited to: data, publications, samples, physical collections, software, and models, as described in any Data Management Plan.

References Cited

[1] Y. Mu, J. R. Edwards, T. C. Horan, S. I. Berrios-Torres, and S. K. Fridkin, "Improving risk-adjusted measures of surgical site infection for the national healthcare safely network," *Infection Control & Hospital Epidemiology*, vol. 32, no. 10, pp. 970–986, 2011.

[2] A. C. de Oliveira, S. I. Ciosak, E. M. Ferraz, and R. S. Grinbaum, "Surgical site infection in patients submitted to digestive surgery: risk prediction and the nnis risk index," *American journal of infection control*, vol. 34, no. 4, pp. 201–207, 2006.

[3] N. D. Friedman, A. L. Bull, P. L. Russo, K. Leder, C. Reid, B. Billah, S. Marasco, E. McBryde, and M. J. Richards, "An alternative scoring system to predict risk for surgical site infection complicating coronary artery bypass graft surgery," *Infection Control & Hospital Epidemiology*, vol. 28, no. 10, pp. 1162–1168, 2007.

[4] H.-Y. Chiang, A. S. Kamath, J. M. Pottinger, J. D. Greenlee, M. A. Howard, J. E. Cavanaugh, and L. A. Herwaldt, "Risk factors and outcomes associated with surgical site infections after craniotomy or craniectomy," *Journal of neurosurgery*, vol. 120, no. 2, pp. 509–521, 2014.

[5] M. C. Wylie, D. A. Graham, G. Potter-Bynoe, M. E. Kleinman, A. G. Randolph, J. M. Costello, and T. J. Sandora, "Risk factors for central line–associated bloodstream infection in pediatric intensive care units," *Infection Control & Hospital Epidemiology*, vol. 31, no. 10, pp. 1049–1056, 2010.

[6] A. Y. Noaman, F. Nadeem, A. H. M. Ragab, A. Jamjoom, N. Al-Abdullah, M. Nasir, and A. G. Ali, "Improving prediction accuracy of "central line-associated blood stream infections" using data mining models," *BioMed research international*, vol. 2017, 2017.

[7] H. Schoonover, K. Kelley, and L. Thatcher, "Accurately predicting risk of central line-associated bloodstream infection—application of machine learning to predict and minimize incidence of central line-associated bloodstream infection," *American Journal of Infection Control*, vol. 45, no. 6, p. S46, 2017.

[8] E. Herc, P. Patel, L. L. Washer, A. Conlon, S. A. Flanders, and V. Chopra, "A model to predict central-line–associated bloodstream infection among patients with peripherally inserted central catheters: The mpc score," *Infection Control & Hospital Epidemiology*, vol. 38, no. 10, pp. 1155–1166, 2017.

[9] J. P. Parreco, A. E. Hidalgo, A. D. Badilla, O. Ilyas, and R. Rattan, "Predicting central line-associated bloodstream infections and mortality using supervised machine learning," *Journal of critical care*, vol. 45, pp. 156–162, 2018.

[10] M. Graña *et al.*, "Detection of healthcare-associated urinary tract infection in swedish electronic health records," *Innovation in Medicine and Healthcare 2014*, vol. 207, p. 330, 2015.

[11] P. A. Tambyah, "Catheter-associated urinary tract infections: diagnosis and prophylaxis," *International journal of antimicrobial agents*, vol. 24, pp. 44–48, 2004.

[12] D. M. Siddiq and R. O. Darouiche, "New strategies to prevent catheter-associated urinary tract infections," *Nature Reviews Urology*, vol. 9, no. 6, p. 305, 2012.

[13] A. H. Froon, M. J. Bonten, C. A. Gaillard, J. W. M. Greve, M. A. Dentener, P. W. de LEEUW, M. Drent, E. E. Stobberingh, and W. A. Buurman, "Prediction of clinical severity and outcome of ventilator-associated pneumonia: comparison of simplified acute physiology score with systemic inflammatory mediators," *American journal of respiratory and critical care medicine*, vol. 158, no. 4, pp. 1026–1031, 1998.

[14] J. Larsson, T. S. Itenov, and M. H. Bestle, "Risk prediction models for mortality in patients with ventilator-associated pneumonia: A systematic review and meta-analysis," *Journal of critical care*, vol. 37, pp. 112–118, 2017.

[15] T. Lisboa, E. Diaz, M. Sa-Borges, A. Socias, J. Sole-Violan, A. Rodríguez, and J. Rello, "The ventilator-associated pneumonia piro score: a tool for predicting icu mortality and health-care resources use in ventilator-associated pneumonia," *Chest*, vol. 134, no. 6, pp. 1208–1216, 2008.

[16] S. Krüger, D. Frechen, and S. Ewig, "Prognosis of ventilator-associated pneumonia: what lies beneath," 2011.

[17] M. Mirsaeidi, P. Peyrani, J. A. Ramirez, and I. M. through Pathway Assessment of Critical Therapy of Hospital-Acquired Pneumonia (IMPACT-HAP) Investigators, "Predicting mortality in patients with ventilator-associated pneumonia: The apache ii score versus the new ibmp-10 score," *Clinical infectious diseases*, vol. 49, no. 1, pp. 72–77, 2009.

[18] J. Wiens, W. N. Campbell, E. S. Franklin, J. V. Guttag, and E. Horvitz, "Learning data-driven patient risk str. jpegication models for clostridium difficile," in *Open forum infectious diseases*, vol. 1, no. 2. Oxford University Press, 2014.

[19] C. Sen, T. Hartvigsen, E. Rundensteiner, and K. Claypool, "Crest-risk prediction for clostridium difficile infection using multimodal data mining," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 52–63.

[20] J. Oh, M. Makar, C. Fusco, R. McCaffrey, K. Rao, E. E. Ryan, L. Washer, L. R. West, V. B. Young, J. Guttag *et al.*, "A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers," *infection control & hospital epidemiology*, vol. 39, no. 4, pp. 425–433, 2018.

[21] F. D. LaBarbera, I. Nikiforov, A. Parvathenani, V. Pramil, and S. Gorrepati, "A prediction model for clostridium difficile recurrence," *Journal of community hospital internal medicine perspectives*, vol. 5, no. 1, p. 26033, 2015.

[22] E. R. Dubberke, Y. Yan, K. A. Reske, A. M. Butler, J. Doherty, V. Pham, and V. J. Fraser, "Development and validation of a clostridium difficile infection risk prediction model," *Infection Control & Hospital Epidemiology*, vol. 32, no. 4, pp. 360–366, 2011.

[23] J. L. Kuntz, D. H. Smith, A. F. Petrik, X. Yang, M. L. Thorp, T. Barton, K. Barton, M. Labreche, S. J. Spindel, and E. S. Johnson, "Predicting the risk of clostridium difficile infection upon admission: a score to identify patients for antimicrobial stewardship efforts," *The Permanente Journal*, vol. 20, no. 1, p. 20, 2016.

[24] "National and state healthcare associated infectious: Progress report," https://health.gov/hcq/pdfs/hai-action-plan-acute-care-hospitals.PDF, accessed: 2018-11-17.

[25] A. E. Pop-Vicas and E. M. D'agata, "The rising influx of multidrug-resistant gram-negative bacilli into a tertiary care hospital," *Clinical infectious diseases*, vol. 40, no. 12, pp. 1792–1798, 2005.

[26] B. R. Edlin, J. I. Tokars, M. H. Grieco, J. T. Crawford, J. Williams, E. M. Sordillo, K. R. Ong, J. O. Kilburn, S. W. Dooley, K. G. Castro *et al.*, "An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome," *New England Journal of Medicine*, vol. 326, no. 23, pp. 1514–1521, 1992.

[27] C. G. Whitney, M. M. Farley, J. Hadler, L. H. Harrison, C. Lexau, A. Reingold, L. Lefkowitz, P. R. Cieslak, M. Cetron, E. R. Zell *et al.*, "Increasing prevalence of multidrug-resistant streptococcus pneumoniae in the united states," *New England Journal of Medicine*, vol. 343, no. 26, pp. 1917–1924, 2000.

[28] "National targets and metrics," https://health.gov/hcq/prevent-hai-measures.asp, accessed: 2018-11-28.

[29] F. C. Tenover, I. A. Tickler, and D. H. Persing, "Antimicrobial resistant strains of clostridium difficile from north america," *Antimicrobial agents and chemotherapy*, pp. AAC–00 220, 2012.

[30] Z. Peng, D. Jin, H. B. Kim, C. W. Stratton, B. Wu, Y.-W. Tang, and X. Sun, "An update on antimicrobial resistance in clostridium difficile: Resistance mechanisms and antimicrobial susceptibility testing," *Journal of clinical microbiology*, pp. JCM–02 250, 2017.

[31] P. Spigaglia, "Recent advances in the understanding of antibiotic resistance in clostridium difficile infection," *Therapeutic advances in infectious disease*, vol. 3, no. 1, pp. 23–42, 2016.

[32] J. Wiens, J. Guttag, and E. Horvitz, "Learning evolving patient risk processes for c. diff colonization," in *ICML Workshop on Machine Learning from Clinical Data*, 2012.

[33] J. Wiens, E. Horvitz, and J. V. Guttag, "Patient risk stratification for hospital-associated c. diff as a time-series classification task," in *Advances in Neural Information Processing Systems*, 2012, pp. 467–475.

[34] T. Hartvigsen, C. Sen, S. Brownell, E. Teeple, X. Kong, and E. A. Rundensteiner, "Early prediction of mrsa infections using electronic health records." in *HEALTHINF*, 2018, pp. 156–167.

[35] L. M. Weiner, A. K. Webb, B. Limbago, M. A. Dudeck, J. Patel, A. J. Kallen, J. R. Edwards, and D. M. Sievert, "Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the national healthcare safety network at the centers for disease control and prevention, 2011–2014," *infection control & hospital epidemiology*, vol. 37, no. 11, pp. 1288–1301, 2016.

[36] U. D. of Health, H. Services *et al.*, "Antibiotic resistance threats in the united states, 2013," *Centers for Disease Control and Prevention*, 2013.

[37] "Health it and health information exchange basics," https://www.healthit.gov/topic/health-it-basics/benefits-ehrs, accessed: 2018-11-24.

[38] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (ehrs): a survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 85, 2018.

[39] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[40] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[41] J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford, O. Gottesman, P. L. Peissig, J. A. Pacheco, G. Tromp, J. Pathak, D. S. Carrell *et al.*, “Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability,” *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1046–1052, 2016.

[42] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, “A review of approaches to identifying patient phenotype cohorts using electronic health records,” *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 221–230, 2013.

[43] K. Ng, A. Ghoting, S. R. Steinhubl, W. F. Stewart, B. Malin, and J. Sun, “Paramo: a parallel predictive modeling platform for healthcare analytic research using electronic health records,” *Journal of biomedical informatics*, vol. 48, pp. 160–170, 2014.

[44] A. Bitton and T. Gaziano, “The framingham heart study’s impact on global risk assessment,” *Progress in cardiovascular diseases*, vol. 53, no. 1, pp. 68–78, 2010.

[45] D. R. Levinson and I. General, “Adverse events in hospitals: national incidence among medicare beneficiaries,” *Department of Health and Human Services Office of the Inspector General*, 2010.

[46] C. M. Torio, A. Elixhauser, and R. M. Andrews, “Trends in potentially preventable hospital admissions among adults and children, 2005–2010: Statistical brief# 151,” 2006.

[47] “National and state healthcare associated infectious: Progress report,” https://www.cdc.gov/HAI/pdfs/progress-report/hai-progress-report.pdf, accessed: 2018-11-17.

[48] D. S. Yokoe, D. J. Anderson, S. M. Berenholtz, D. P. Calfee, E. R. Dubberke, K. D. Eilingson, D. N. Gerding, J. P. Haas, K. S. Kaye, M. Klompas *et al.*, “A compendium of strategies to prevent healthcare-associated infections in acute care hospitals: 2014 updates,” *Infection Control & Hospital Epidemiology*, vol. 35, no. S2, pp. S21–S31, 2014.

[49] C. A. Umscheid, M. D. Mitchell, J. A. Doshi, R. Agarwal, K. Williams, and P. J. Brennan, “Estimating the proportion of healthcare-associated infections that are reasonably preventable and the related mortality and costs,” *Infection Control & Hospital Epidemiology*, vol. 32, no. 2, pp. 101–114, 2011.

[50] B. A. Miller, L. F. Chen, D. J. Sexton, and D. J. Anderson, “Comparison of the burdens of hospital-onset, healthcare facility-associated clostridium difficile infection and of healthcare-associated infection due to methicillin-resistant staphylococcus aureus in community hospitals,” *Infection Control & Hospital Epidemiology*, vol. 32, no. 4, pp. 387–390, 2011.

[51] “Modeling infectious diseases in healthcare network (mind - healthcare), center for disease control and prevention,” https://www.cdc.gov/hai/research/MIND-Healthcare.html, accessed: 2018-11-17.

[52] R. D. Scott, "The direct medical costs of healthcare-associated infections in us hospitals and the benefits of prevention," 2009.

[53] R. M. Klevens, J. R. Edwards, C. L. Richards Jr, T. C. Horan, R. P. Gaynes, D. A. Pollock, and D. M. Cardo, "Estimating health care-associated infections and deaths in us hospitals, 2002," *Public health reports*, vol. 122, no. 2, pp. 160–166, 2007.

[54] "National action plan to prevent healthcare-associated infections: roadmap to elimination. washington, dc: Department of health and human services," https://health.gov/hcq/prevent-hai.asp, accessed: 2018-11-17.

[55] S. S. Magill, J. R. Edwards, W. Bamberg, Z. G. Beldavs, G. Dumyati, M. A. Kainer, R. Lynfield, M. Maloney, L. McAllister-Hollod, J. Nadle *et al.*, "Multistate point-prevalence survey of health care–associated infections," *New England Journal of Medicine*, vol. 370, no. 13, pp. 1198–1208, 2014.

[56] P. W. Stone, E. C. Hedblom, D. M. Murphy, and S. B. Miller, "The economic impact of infection control: making the business case for increased infection control resources," *American journal of infection control*, vol. 33, no. 9, pp. 542–547, 2005.

[57] E. Zimlichman, D. Henderson, O. Tamir, C. Franz, P. Song, C. K. Yamin, C. Keohane, C. R. Denham, and D. W. Bates, "Health care–associated infections: a meta-analysis of costs and financial impact on the us health care system," *JAMA internal medicine*, vol. 173, no. 22, pp. 2039–2046, 2013.

[58] "The center for disease dynamics, economics & policy," https://cddep.org/tool/overall_and_unit_costs_five_most_common_hospital_acquired_infections_hais_us/, accessed: 2018-11-28.

[59] A. Balkhair, Y. M. Al-Farsi, Z. Al-Muharrmi, R. Al-Rashdi, M. Al-Jabri, F. Neilson, S. S. Al-Adawi, M. El-Beeli, and S. Al-Adawi, "Epidemiology of multi-drug resistant organisms in a teaching hospital in oman: a one-year hospital-based study," *The Scientific World Journal*, vol. 2014, 2014.

[60] J. D. Siegel, E. Rhinehart, M. Jackson, and L. Chiarello, "Management of multidrug-resistant organisms in health care settings, 2006," *American journal of infection control*, vol. 35, no. 10, pp. S165–S193, 2007.

[61] J. Lederberg, P. F. Harrison *et al.*, *Antimicrobial resistance: issues and options.* National Academies Press, 1998.

[62] D. M. Shlaes, D. N. Gerding, J. F. John Jr, W. A. Craig, D. L. Bornstein, R. A. Duncan, M. R. Eckman, W. E. Farrer, W. H. Greene, V. Lorian *et al.*, "Society for healthcare epidemiology of america and infectious diseases society of america joint committee on the prevention of antimicrobial resistance: guidelines for the prevention of antimicrobial resistance in hospitals," *Clinical infectious diseases*, vol. 25, no. 3, pp. 584–599, 1997.

[63] T. F. Landers, B. Cohen, T. E. Wittum, and E. L. Larson, "A review of antibiotic use in food animals: perspective, policy, and potential," *Public health reports*, vol. 127, no. 1, pp. 4–22, 2012.

[64] J. Wiens, J. Guttag, and E. Horvitz, "A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions," *Journal of the American Medical Informatics Association*, vol. 21, no. 4, pp. 699–706, 2014.

[65] "Antibiotic/antimicrobial resistant," https://www.cdc.gov/drugresistance/about.html, accessed: 2018-11-24.

[66] "Surveillance for c. difficile, mrsa, and other drug-resistant infections," https://www.cdc.gov/nhsn/ltach/cdiff-mrsa/index.html, accessed: 2018-11-26.

[67] "Gram-negative bacteria infections in healthcare settings," https://www.cdc.gov/hai/organisms/gram-negative-bacteria.html, accessed: 2018-11-26.

Budget Justification

1. A. Senior Personnel

A1. Includes PI at 10% CY.

2. B. Other Personnel

B3. Includes stipend for one graduate student for each calendar year of the project.

3. C. Fringe Benefits

Fringe benefits are calculated at a rate of X% for faculty, Y% for graduate students.

4. E. Travel

1) all travel (both domestic and foreign) must now be justified. 2) temporary dependent care costs above and beyond regular dependent care that directly result from travel to conferences are allowable costs provided that the conditions established in 2 CFR § 200.474 are met.

5. G. Other Direct Costs

1) Includes coverage on costs of computing devices 2) The charging of computing devices as a direct cost is allowable for devices that are essential and allocable, but not solely dedicated, to the performance of the NSF award **G5.** Includes tuition for graduate students participating in the program.

6. H. Indirect Costs

Overhead at a rate of X% is charged on all direct salaries and wages, applicable fringe benefits, materials and supplies, services, travel and subawards up to the first $X of each subaward. Excluded are equipment and the portion of each subaward in excess of $X.

Current & Pending Support

Investigator:	
Project Title:	Put your Proposal title here
Project Location:	
Source of Support:	NSF
Total Award Amount:	
Total Award Period:	
Status:	Pending (this project)

Facilities, Equipments, & other Resources

This section of the proposal is used to assess the adequacy of the resources available to perform the effort proposed to satisfy both the Intellectual Merit and Broader Impacts review criteria. Proposers should describe only those resources that are directly applicable. Proposers should include an aggregated description of the internal and external resources (both physical and personnel) that the organization and its collaborators will provide to the project, should it be funded. Such information must be provided in this section, in lieu of other parts of the proposal (e.g., budget justification, project description). The description should be narrative in nature and must not include any quantifiable financial information. Reviewers will evaluate the information during the merit review process and the cognizant NSF Program Officer will review it for programmatic and technical sufficiency.

Data Management Plan

Proposals must include a supplementary document of no more than two pages labeled “Data Management Plan”. This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4)