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

Introduction

Hospital-Acquired Infections (HAI) or more broadly healthcare-associated infections are linked with high rates of disease and death among hospitalized patients [1] and their elimination is a priority of the Department of Health and Human Services [2]. HAIs are infections people get while they are receiving health care for another condition are often preventable [3]. It is estimated that at any given time, about 1 in 25 inpatients have HAI [4, 2]. Other studies report the estimated number of HAI associated death to be almost 99000 per year [5] while burdening the U.S. healthcare budget by $5 to $10 billion annually [6]. Approximately one-third of HAIs are preventable [7].

Though the rate of many device-associated HAIs has dropped over the last decade [3] increasing trends in prevalence rates of HAIs caused by *Multi-Drug Resistant Organisms* (MDRO) has been observed [8, 9]. A study found that 16% of all HAIs are caused by MDROs [10]. 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. four of which are healthcare associated with more than 600,000 total occurrences and 28000 death per year [11]. There has been focused studies to predict patients risk of developing Clostridioides difficile (C. diff.), the most prevalent cause of HAI which sometimes become drug-resistant [12, 13, 14, 15]. Due to the scarcity of HAIs associated with other MDROs, comprehensive analysis of risk factors and patient risk prediction for them remained unexplored.

We propose the development of a machine learning (ML) system, called DEALER, that systematically uses the pooled data from multiple extremely rare HAI-causing MDROs and accurately predict the temporal risk of a patient developing any of the MDROs during their stay at hospitals. We hypothesize that enriching our samples by merging various heterogeneous source of data will both improve the prediction accuracy of all adverse events and also provides more interpretable models. In particular, our goal is to integrate DEALER into the Ohio State University's (OSU) 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.

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. Beyond benefiting from enriched heterogeneous MDROs-caused incidents, DEALER also furthers its performance by taking into account hierarchy of hospitals, units, and wards along with their relationship with each other.

Contributions...

Background And Terminology

Types Of HAI

HAIs can be divided into two broad categories of device-associated infections or Surgical-Site Infections (SSI). Primary device-associated infections include Central Line-Associated Bloodstream Infection (CLABSI), Catheter-Associated Urinary Tract Infection (CAUTI), and Ventilator-Associated Pneumonia (VAP). Recent studies show reductions in CLABSI incidents due to the advancement in preventative activities while SSI and CAUTI have not experienced the same lessening.

Another way of grouping HAIs is based on the pathogen causing the infection. Although

bacteria, fungi, and viruses can cause HAI, three of the most threatening HAI causes are all bacteria. In the followings, we briefly introduce each one of the MDRO studied in this proposal:

Clostridium difficile (C. diff): C. diff bacteria causes life-threatening diarrhea and colitis (an inflammation of the colon). It was estimated to cause almost half a million infections in the United States in 2011, and 29,000 died within 30 days of the initial diagnosis. Those most at risk are people, especially older adults, who take antibiotics and also get medical care.

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. These antibiotics include methicillin and other more common antibiotics such as oxacillin, penicillin, and amoxicillin. In the community, most MRSA infections are skin infections. More severe or potentially life-threatening MRSA infections occur most frequently among patients in healthcare settings.

Vancomycin-resistant Enterococcus (VRE): Enterococci cause a range of illnesses, mostly among patients receiving healthcareVancomycin-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. These bacteria can sometimes cause infections. Most vancomycin-resistant Enterococci infections occur in hospitals.

Terminology

Antimicrobial resistance The result of bacteria changing in ways that reduce or eliminate the effectiveness of drugs, chemicals, or other agents used to cure or prevent infections. Antibiotic resistance is one type of antimicrobial resistance.

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. Methicillin-resistant Staphylococcus aureus (MRSA): A type of staph bacteria that is resistant to many antibiotics. In this report, the MRSA data include all laboratory identified hospital-onset MRSA bacteremia (bloodstream infections) reported to the National Healthcare Safety Network from all inpatient locations in the facility.

Multi-drug resistant organism (MDRO) infection: An infection caused by a germ that is resistant to multiple classes of antimicrobials. In some cases, the germs have become so resistant that no available antibiotics are effective against them.

Targeted Assessment for Prevention (TAP) strategy: A method developed by the Centers for Disease Control and Prevention (CDC) to use data for action to prevent healthcare-associated infections (HAIs). The TAP strategy targets healthcare facilities and specific units within facilities with a disproportionate burden of HAIs to address infection prevention gaps.

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.

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.

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 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.

Technical Approach

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

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.

Comparison Of Our Approach To Related Work

Community Outreach And Education

Results From Prior NSF Support

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] "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.

[2] "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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] A. I. Hidron, J. R. Edwards, J. Patel, T. C. Horan, D. M. Sievert, D. A. Pollock, S. K. Fridkin *et al.*, "Antimicrobial-resistant pathogens associated with healthcare-associated infections: annual summary of data reported to the national healthcare safety network at the centers for disease control and prevention, 2006–2007," *Infection Control & Hospital Epidemiology*, vol. 29, no. 11, pp. 996–1011, 2008.

[11] "Centers for disease control and prevention. antibiotic resistance threats in the united states, 2013. atlanta, ga:us department of health and human services, cdc," https://www.cdc.gov/drugresistance/biggest_threats.html, accessed: 2018-11-17.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

Budget Justification

A. Senior Personnel

A1. Includes PI at 10% CY.

B. Other Personnel

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

C. Fringe Benefits

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

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.

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.

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)