Project Narrative

Rationale: Animal influenza viruses emerging into humans have triggered devastating pandemics in the past. Yet, our ability to evaluate the pandemic potential of individual strains that do not yet circulate in humans, remains limited. Here we propose to develop an experimentally validated platform called the Emergenet (Enet), to predict in near-real-time where and when new variants of concern would emerge, using only observed sequences of key viral proteins, procured in ongoing global surveillance of Influenza A viruses. We bring together new machine learning algorithms customized to the problem at hand, key insights from information theory, evolutionary theory, epidemiology and precise statistical uncertainty quantification to develop a rigorous framework, to track evolutionary trajectories of pathogens through a complex, poorly characterized, and dynamically changing fitness landscape. Our deliverable is best described as the foundations for creating a platform akin to bio-NORAD, *identifying when and where an imminent zoonotic emergence event is likely, and if such novel strains are likely to achieve human-to-human transmission capability.*

Influenza viruses constantly evolve[1], sufficiently altering surface protein structures to evade the prevailing host immunity, and cause the recurring seasonal epidemic. These periodic infection peaks claim a quarter to half a million lives[2] globally. Additionally, Influenza A, partly on account of its segmented genome and its wide prevalence in animal hosts, can easily incorporate genes from multiple strains and (re)emerge as novel human pathogens[3], thus harboring a high pandemic potential. Strains spilling over into humans from animal reservoirs is thought to have triggered pandemics at least four times (1918 Spanish flu/H1N1, 1957 Asian flu/H2N2, 1968 Hong Kong flu/H3N2, 2009 swine flu/H1N1) in the past 100 years[4]. One approach to mitigating such risk is to identify animal strains that do not yet circulate in humans, but is likely to spill-over and quickly achieve human-to-human (HH) transmission capability. While global surveillance efforts collect wild specimens from diverse hosts and geo-locations annually, our ability to objectively, reliably and scalably risk-rank individual strains remains limited[5]. The Center for Disease Control's (CDC) current solution to this problem is the Influenza Risk Assessment Tool (IRAT)[6], which relies on time-consuming proteomics and transmission assays and potentially subjective evaluations by subject matter experts, taking weeks to months to compile for each strain of concern. With tens of thousands of strains being sequenced annually, this results in a scalability bottleneck.

Here we plan to develop a platform powered by novel pattern discovery and recognition algorithms to automatically parse out emergent evolutionary constraints operating on Influenza A viruses in the wild, to provide a less-heuristic theory-backed scalable solution to emergence prediction. We plan to show that this capability enables preempting strains which are expected to be in future human circulation, and approximate IRAT scores of non-human strains without experimental assays or SME scoring, in second as opposed to weeks or months. Our approach automatically takes into account the time-sensitive variations in selection pressures as the background strain circulation changes over time, and will potentially be able to rank-order strains adapatively. Additionally, we plan to validate our ability to predict future variations of viral proteins by showing that predicted variants of HA and NA fold correctly, and are functional, binding to the relevant human receptors in in-vivo laboratory experiments. Thus, bringing together rigorous data-driven modeling, and validation via tools from reverse genetics we plan to deliver an actionable and deployable platform that optimally exploits the current biosurvellance capacity.

The BioNORAD platform will enable proactive and actionable global surveillance for emerging pandemic threats from Influenza A. This importance of the ability to preempt pandemic risk to the national interest of the United States cannot be overstated, especially in the context of protecting DoD assest and personnel deployed in potentially high risk centers of emergence. Additionally, the BioNORAD will enable preemptive action including the inoculation of animal reservoirs before the first human infection, potentially eliminating the pandemic before it has a chance to trigger.

Hypotheses:

FY23 PRMRP Portfolio Category: Infectious Diseases | FY23 PRMRP Topic: proteomics | FY23 PRMRP Strategic Goal: Epidemiology: Identify strategies for surveillance or develop modeling tools and/or biomarkers to predict outbreaks or epidemics

1) Learning patterns of cross-dependency between mutations and genomic change reveals enough of the underlying rules of organization to meaningfully and actionably constrain the evolutionary trajectories of emerging pathogens, and in our context, that of Influenza A viruses in the wild.

2) Such patterns can be learned from biosurveillance data that is being collected now globally to ultimately develop a next-generation pro-active surveillance platform that acts as an early warning system for pandemic threats, and serves a similar function to the strategic goal of NORAD in the context of defending our airspace from adversarial intrusion.

Specific Aims:

Research Strategy and Feasibility:

Innovation: While numerous tools exist for ad hoc quantification of genomic similarity[7–12], higher similarity between strains in these frameworks is not sufficient to imply a high likelihood of a jump. To the best of our knowledge, the Emergenet algorithm is the first of its kind to learn an appropriate biologically meaningful comparison metric from data, without assuming any model of DNA or amino acid substitution, or a genealogical tree a priori. While the effect of the environment and selection cannot be inferred from a single sequence, an entire database of observed strains, processed through the right lens, can parse out useful predictive models of these complex interactions. Our results are aligned with recent studies demonstrating effective predictability of future mutations for different organisms[13,14].

Technical Abstract

Lay Abstract

Impact Statement

Relevance to Military Health Statement

Facilities, Existing Equipment, and Other Resources

The University of Chicago is a private non-profit institution located on the ethnically-diverse South Side of Chicago that has been a center of advanced learning and research since its inception in 1892. The University of Chicago is comprised of four graduate Divisions (Biological Sciences, Physical Sciences, Social Sciences, and Humanities), six professional schools (Chicago Booth School of Business, Divinity School, Harris School of Public Policy Studies, Law School, Pritzker School of Medicine, and School of Social Service Administration), the Graham School of General Studies, and the undergraduate College. The University has a unique history of organizing around research questions that cross disciplines rather than operating within disciplinary boundaries. The extent to which this strategy reflects University of Chicago is illustrated by its numerous interdisciplinary Committees, Centers and Institutes (described below). The University of Chicago maintains its commitment to scholarship, teaching, and research through its more than 2100 faculty members and a student population of approximately 15,600 with nearly 2/3 engaged in advanced research and professional study. Through the years, 86 Nobel Laureates (8 are current faculty), 44 members of the National Academy of Sciences, 169 members of the American Academy of Arts and Sciences, and 14 recipients of the National Medal of Science have been associated with the University as students, teachers or research investigators. The University of Chicago is ranked among the world's top universities by a number of criteria, including the amount of federal research funding received (despite a size much smaller than many of its academic peers). This spirit of discovery, innovation and public service provides a robust foundation for success.

South Shore Senior Health Center (SSSC): The SSSC is a 6800 sq. ft. university-owned geriatric facility 5 miles south of DCAM, with doorstep free parking in the heart of Chicago's South Side, specifically in the South Shore district. There are 13 patient exam rooms equipped with an examination bed, ophthalmoscope and otoscope, in addition to an on-site phlebotomist and capability for ECG. There is a large conference room for team meetings and support groups. The Memory Center team meets on Mondays at this site.

Center for Care and Discovery: Completed in 2013, the CCD is a ten-story adult hospital focused on cancer, advanced surgery, high tech imaging, the neurosciences and gastrointestinal procedures. The building is 1,200,000 GSF with floor plates of 102,000 GSF. The CCD includes 240 private patient rooms, 28 operating rooms (21 initially); an imaging department with 3 CT's, 2 MRI's, 1 fluoroscopy room, 2 general radiology rooms, 7 interventional radiology rooms; and a gastroenterology procedures suite with 11 GI procedure rooms, 2 fluoroscopy rooms and 2 bronchoscopy rooms. The CCD includes an inpatient kitchen, cafeteria, 7th floor sky lobby meeting rooms, ground floor retail space and clinical support services. Two floors are "shelled space" for future expansion of services. The building is centrally located between existing clinical facilities (DCAM, Comer) and new research facilities (KCBD, GCIS, Knapp). The facility is connected to both DCAM and Comer via above and below ground connections. There are procedure rooms, 2 fluoroscopy rooms and 2 bronchoscopy rooms. Neurology and neurosurgery inpatients as well as a Neuro- ICU are contiguously located on one floor of the CCD.

Computational Facilities

The principal investigators have access to extensive computational facilities available at the University of Chicago to carry out the tasks described.

Access to Clinical Data for AI-enabled Analytics: The ZeD lab (overseen by Professor Chattopadhyay) is housed within the Department of Medicine at the University of Chicago, and has access to the full range of high end computing resources offered by the University of Chicago. In addition, Prof. Chattppadyay's laboratory has access to the HIPAA compliant clinical data warehouse maintained by the Biological Sciences Division as detailed below:

The Clinical Research Data Warehouse: (CRDW) within the Biomedical Sciences Division of the University of Chicago is one of the deepest, richest, and most research-ready data repositories of its kind. Containing more than a decade of University of Chicago medical data, it seamlessly brings together multiple internal and external data sources to provide researchers with access to more than 12 million encounters for 2.3 million patients. The associated diagnoses, labs, medications, and procedures number in the tens of millions each. The CRDW is run on IBM Netezza Pure Data System for Analytics servers, a patented Asymmetric Massively Parallel Processing architecture designed to deliver exceptional query performance and modular scalability on highly complex mixed workloads.

TABLE 1
University of Chicago Research Computing Center Capabilities Summary

Cluster	Partition	Compute cores (CPUs)	Memory	Other configuration details
midway1	westmere	12 x Intel X5675 3.07 GHz	24 GB	
	sandyb	16 x Intel E5-2670 2.6GHz	32 GB	
	bigmem	16 x Intel E5-2670 2.6GHz	256 GB	
		32 x Intel E7-8837 2.67GHz	1 TB	
	gpu	16 x Intel E5-2670 2.6GHz	32 GB	2 x Nvidia M2090 or K20 GPU
		20 x Intel E5-2680v2 2.8GHz	64 GB	2 x Nvidia K40 GPU
	mic	16 x Intel E5-2670 2.6GHz	32 GB	2 x Intel Xeon Phi 5100 coprocessor
	amd	64 x AMD Opteron 6386 SE	256 GB	
	ivyb	20 x Intel E5-2680v2 2.8GHz	64 GB	
midway2	broadwl	28 x Intel E5-2680v4 2.4GHz	64 GB	
	bigmem2	28 x Intel E5-2680v4 @ 2.4 GHz	512 GB	
	gpu2	28 x Intel E5-2680v4 @ 2.4 GHz	64 GB	4 x Nvidia K80 GPU

In order to meet the acute need for data related to COVID-19, the CRDW team has constructed three data marts (de-identified, limited, and identified) to provide the most commonly requested data elements for this patient population. The initial instance of the COVID-19 data mart includes de-identified structured data on patient demographics, encounters, diagnoses, labs, medications, flow sheets, and procedures. Additional data will be added based on resource availability and urgency.

Cohort Discovery Tool: The purpose of this tool (SEE Cohorts) is to provide a secure web-based tool for the initial exploration of de-identified data. It allows researchers to search available data, build a cohort of patients, and view actual de-identified data within the interface. The data in SEE Cohorts is refreshed weekly.

Research Computing Center: The University of Chicago Research Computing Center (RCC) provides high-end research computing resources to researchers at the University of Chicago, which include high-performance computing and visualization resources; high-capacity storage and backup; software; high-speed networking; and hosted data sets. Resources are centrally managed by RCC staff who ensure the accessibility, reliability, and security of the compute and storage systems. A high-throughput network connects the Midway Compute Cluster to the UChicago campus network and the public internet through a number of high-bandwidth uplinks. To support data-driven research RCC hosts a number of large datasets to be accessed within the RCC compute environment.

RCC maintains three pools of servers for distributed high-performance computing. Ideal for tightly coupled parallel calculations, tightly-coupled nodes are linked by a fully non-blocking FDR-10 Infiniband interconnect. Loosely-coupled nodes are similar to the tightly-coupled nodes, but are connected with GigE rather than Infiniband and are best suited for high-throughput jobs. Finally, shared memory nodes contain much larger main memories (up to 1 TB) and are ideal for memory-bound computations. The types of CPU architectures RCC maintains are tabulated in Table 1.

RCC also maintains a number of specialty nodes:

- *Large shared memory nodes* - up to 1 TB of memory per node with either 16 or 32 Intel CPU cores. Midway is always expanding, but at time of writing RCC contains a total of 13,500 cores across 792 nodes, and 1.5 PB of storage.
- *Hadoop:* Originally developed at Google, Hadoop is a framework for large-scale data processing.
- *GPU Computing:* Scientific computing on graphics cards can unlock even greater amounts of parallelism from code. RCC GPU nodes each include two Nvidia Tesla-class accelerator cards and are integrated in the Infiniband network. RCC currently provides access to Fermi-generation M2090 GPU devices and Kepler-generation K20 and K40 devices.
- *Xeon Phi:* The Many Integrated-Core architecture (MIC) is Intel's newest approach to manycore computing. Researchers can experiment with these accelerators by using MIC nodes, each of which have two Xeon Phi cards, and are integrated into the Infiniband network.

Persistent and High-Capacity Storage. Storage is accessible from all compute nodes on Midway1 and Midway2 as well as outside of the RCC compute environment through various mechanisms, such as mounting directories as network drives on your personal computer or accessing data as a Globus Online endpoint (at the time of this writing, Globus Online is supported on Midway1). RCC takes snapshots of all home directories (users' private storage space) at regular intervals so that if any data is lost or corrupted, it can easily be recovered. RCC maintains GPFS Filesystem Snapshots for quick and easy data recovery. In the event of catastrophic storage failure, archival tape backups can be used to recover data from persistent storage locations on Midway. Automated snapshots of the home and project directories are available in case of accidental file deletion or other problems. Currently snapshots are available for these time periods: 1) 7 daily snapshots, 2) 4 weekly snapshots.

Tape Backups. Backups are performed on a nightly basis to a tape machine located in a different data center than the main storage system. These backups are meant to safeguard against events such as hardware failure or disasters that could result in the complete loss of RCC's primary data center.

Data Sharing. All data in RCC's storage environment is accessible through a wide range of tools and protocols. Because RCC provides centralized infrastructure, all resources are accessible by multiple users simultaneously, which makes RCC's storage system ideal for sharing data among your research group members. Additionally, data access and restriction levels can be put in place on an extremely granular level.

Data Security & Management. The HIPAA compliant security of the Research Computing Center's storage infrastructure, protected by two-factor authentication, gives users peace of mind that their data is stored, managed, and protected by HPC professionals. Midway's file management system allows researchers to control access to their data. RCC has the ability to develop data access portals for different labs and groups.

Publications and/or Patents

Letters of Organizational Support

Letters of Collaboration

Data and Research Resources Sharing Plan

References

[1] Dos Santos, G., Neumeier, E. & Bekkat-Berkani, R. Influenza: Can we cope better with the unpredictable? *Human vaccines & immunotherapeutics* **12**, 699–708 (2016).

[2] Huddleston, J. *et al.* Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza a/h3n2 evolution. *Elife* **9**, e60067 (2020).

[3] Reid, A. H. & Taubenberger, J. K. The origin of the 1918 pandemic influenza virus: a continuing enigma. *Journal of general virology* **84**, 2285–2292 (2003).

[4] Shao, W., Li, X., Goraya, M. U., Wang, S. & Chen, J.-L. Evolution of influenza a virus by mutation and re-assortment. *International journal of molecular sciences* **18**, 1650 (2017).

[5] Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS biology* **19**, e3001135 (2021).

[6] CDC. Influenza risk assessment tool (irat) — pandemic influenza (flu) — cdc. https://www.cdc.gov/flu/pandemic-resources/national-strategy/risk-assessment.htm. (Accessed on 07/02/2021).

[7] Posada, D. & Crandall, K. A. Modeltest: testing the model of dna substitution. *Bioinformatics (Oxford, England)* **14**, 817–818 (1998).

[8] Goldberger, A. L. & Peng, C.-K. Genomic classification using an information-based similarity index: application to the sars coronavirus. *Journal of Computational Biology* **12**, 1103–1116 (2005).

[9] Huelsenbeck, J. P. & Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and systematics* **28**, 437–466 (1997).

[10] Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting evolution from the shape of genealogical trees. *Elife* **3**, e03568 (2014).

[11] van der Meer, F. J. U. M., Orsel, K. & Barkema, H. W. The new influenza A H1N1 virus: balancing on the interface of humans and animals. *The Canadian veterinary journal = La revue veterinaire canadienne* **51**, 56–62 (2010).

[12] Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).

[13] Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS biology* **19**, e3001390 (2021).

[14] Maher, M. C. *et al.* Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science Translational Medicine* **14**, eabk3445 (2022).