A Framework for Integrating Multiple Data Sources for Modeling and Forecasting of Infectious Diseases

Children’s Hospital Corporation

PI: Elaine O. Nsoesie

Grant Number: 1-K01S25438-01

I am trained as a computational biologist and statistician, and I am currently a postdoctoral fellow at Boston Children's Hospital, Harvard Medical School. My main career goal is to become an independent researcher at a major research institution. I plan to continue my current research pursuits in global health and infectious diseases. Specifically, I aim to continue developing mathematical and computational approaches for modeling to understand disease transmission, forecasting future dynamics and evaluating interventions for public policy decisions. As a postdoctoral research fellow, I have had the wonderful opportunity of working with data from multiple sources. Although several of these data streams could be labeled as "Big Data", I typically work with the data after it is already processed, filtered and aggregated to a daily or weekly resolution. While I have developed the necessary skills for modeling these already processed data, there are three important areas where I require additional training, mentoring, and experience: (1) advanced computational skills especially in the use of high performance computing and informatics tools, (2) techniques in computational machine learning and data mining necessary for data acquisition and processing, and (3) biostatistical methodology needed for the statistical design of studies involving big data. These three training and mentoring aims would enable me to develop the skills necessary to become an independent investigator in Big Data Science for biomedical research. Boston Children’s School and Harvard Medical School are leading institutions in translational biomedical research, thereby making them the ideal environment to pursue the training and research aims in this proposal. The recent emergence of infectious diseases such as the avian influenza H7N9 in China, and re-emergence of diseases such as polio in Syria underscores the importance of strengthening immunization and emergency response programs for the prevention and control of infectious diseases. Researchers have developed computational and mathematical models to capture determinants of infectious disease dynamics and identify factors that support prediction of these dynamics, provide estimates of disease risk, and evaluate various intervention scenarios. While these studies have been extremely useful for the understanding of infectious disease transmission and control, most have been disease specific and solely used data from traditional disease surveillance systems. In contrast, there is a huge amount of internet-based data that have been extensively assessed and validated for public health surveillance in the last decade, but it has been scarcely used in conjunction with other data sources for modeling to predict disease spread. Using these novel digital event-based data sources in combination with climate and case data from traditional disease surveillance systems, we will establish a much needed framework for integrating these disparate data sources for modeling to estimate disease risk and forecasting temporal dynamics of infectious diseases. Our approach will be achieved through three aims. The first objective is to develop an automated process for acquiring, processing and filtering data for modeling (Aim 1). Once we gather this data, we will develop temporal models for the dynamical assessment of the relationship between the various data variables and infectious disease incidence (Aim 2). Finally, we will assess the utility of the modeling approaches developed under Aim 2 for forecasting temporal trends of infectious diseases (Aim 3). Through data acquisition, thorough processing, statistical and epidemiological modeling, and guided by advisers with expertise in biomedical informatics, computer science and statistics, we plan to achieve a comprehensive approach to integrating multiple data streams for modeling to forecast infectious diseases. PUBLIC HEALTH RELEVANCE: Although there have been significant medical and technological advances towards infectious disease prevention, surveillance and control, infectious diseases still account for an estimated 15 million deaths each year worldwide. Reliable forecasts of infectious disease dynamics can influence decisions regarding prioritization of limited resources during outbreaks, optimization of disease interventions and implementation of rigorous surveillance processes for quicker case identification and control of emerging disease outbreaks. Our goal is therefore to develop a data mining/informatics framework that leverages the huge amount of digital event-based data sources in combination with climate data, and data from traditional disease surveillance systems for modeling and forecasting infectious diseases.