Athens Journal of Technology and Engineering - Volume 8, Issue 4, December 2021 – Pages 283-298

A Web-Based Analytic for Pathogen Suggestion using Syndromic Data

By Nileena Velappan[*], Ashlynn R. Daughton[±], Yan Xu[‡], William Rosenberger[+], Geoffrey Fairchild[♦] & Alina Deshpande[°]

Historical disease outbreaks provide enhanced contextual information for an unfolding outbreak. Utilizing this concept, we have developed a visual analytic tool known as Analytics for Investigation of Disease Outbreaks (AIDO), a web accessible decision support tool available at aido.bsvgateway.org. AIDO currently contains more than 650 historical outbreaks for 40 human diseases. Recently, we have evaluated AIDO's ability to identify an outbreak pathogen using syndromic disease "families". The AIDO gastrointestinal family contains 122 outbreaks from five different pathogens and the mosquito-borne family has 151 outbreaks caused by eight pathogens. We identified epidemiological properties that are different by pathogen within families using Chi-squared tests. The similarity algorithms developed for each syndromic family, based on identified properties were tested using 90 different outbreaks spread across the thirteen pathogens. In our analyses, we are able to suggest the correct pathogen upwards of 75% of the time, using our algorithm that relies on differences in epidemiological properties. Our analyses with mosquito-borne outbreak family also showed that AIDO is capable of identifying outbreaks caused by emerging pathogens. This easy-to-use web-based analytic can be a useful tool in pandemic mitigation across the globe.

Keywords: *disease outbreaks, syndromic analyses, population epidemiology, web-based visual analytics*

Introduction

During an infectious disease outbreak, the number of people infected depends upon rapid identification of the pathogen, its transmission route and prompt implementation of effective control measures. Quick identification of the pathogen associated with an unfolding outbreak can save countless lives, especially if it is an emerging pathogen (AJMC 2021). Disease surveillance is used to identify clusters of related illnesses or outbreaks. Three general methods are used to identify these disease clusters: pathogen specific surveillance, syndromic surveillance, and complaint system (CIFOR 2014). Pathogen specific surveillance system detects

[*]Research Technologist, Biosciences Division, Los Alamos National Laboratory, USA.
[±]Scientist II, Analytics, Intelligence and Technology Division, Los Alamos National Laboratory, USA.
[‡]Software Developer, Biosciences Division, Los Alamos National Laboratory, USA.
[+]Scientist I, Analytics, Intelligence and Technology Division, Los Alamos National Laboratory, USA.
[♦]Scientist III, Biosciences Division, Los Alamos National Laboratory, USA.
[°]R&D Manager, Biosciences Division, Los Alamos National Laboratory, USA.

https://doi.org/10.30958/ajte.8-4-1 doi=10.30958/ajte.8-4-1

clusters of a specific pathogen that were identified by health-care providers and laboratorians. This method is highly sensitive and specific for a given disease, however it is time consuming and expensive. Syndromic surveillance is a highly complex and technology driven automated tool used in North America and Europe (May et al. 2009). This type of surveillance often involves extraction of health information such as school and work absenteeism, nurse help-lines, sales of certain over-the-counter drugs, complaints to regulatory authorities (e.g., Water Company). The data for syndromic surveillance can be collected during pre-diagnostic and post diagnostic periods. These data are analyzed by agencies such as state health departments to identify possible disease clusters. Syndromic surveillance can identify outbreaks earlier and faster than traditional pathogen detection. However, syndromic surveillance may not make the associated pathogen immediately apparent. The complaint based surveillance system is the simplest surveillance method where similar complaints from multiple individuals are used to identify disease outbreaks. For example, "Pneumonia of unknown cause" was the first report that the world received at the start of COVID-19 pandemic (AJMC 2021). However, the disadvantage associated with this fast and cheap method is again the inability to identify specific pathogen (CIFOR 2014). In most parts of the world, pathogen based surveillance system and syndromic surveillance are not implemented due to prohibitory costs associated with these systems. The complaint based system is often the best (only) surveillance system that is used to identify disease outbreaks (CIFOR 2014). Identification of a pathogen associated with a given disease symptom is essential in designing control measures for an outbreak. However, the laboratory confirmation of a pathogen takes both time and money.

Laboratory tests often employ different techniques for specific identification of a pathogen. Culturing of the organism on differential plates is a first step in bacterial pathogen detection. These tests can be followed up with biochemical testing and serological confirmation (Thermo Scientific Inc. 2021). Nucleic acid based assays (e.g. pulse field gel electrophoresis, sequencing) are further employed to identify the strain causing the outbreak. Multiple parallel tests are conducted to identify the causative organism during the initial stages of an outbreak (Foddai and Grant 2020). Antibody based serological testing and nucleic acid based analyses, such as those developed for SARS CoV2 are employed for viral pathogen detection (Wang et al. 2020). Pathogens that belong to the same viral family sharing similarity in envelop protein structures (e.g., Chikungunya virus, Dengue, virus and Zika virus) often produces cross reactive antibodies that make the identification of a specific virus challenging (Paixão et al. 2018). These tests are also sensitive to sample collection time, as the antibody response time in patients will vary based on disease progression. The expensive detection methods described above are performed by specialized laboratories (Paixão et al. 2018). Different parts of the world use these laboratory confirmation tests differently. In India, only the initial tests that cost $10-$20 are performed for each pathogen (MediFee 2021). Therefore, multiple tests are ordered by the physicians during early stages of an outbreak to identify the pathogen. US Centers for Disease Control (CDC) uses a specialized algorithm to distinguish Zika virus patients from other closely related mosquito borne pathogens (CDC 2021) and the tests cost about USD 1240 if all

comprehensive tests listed in the algorithm are performed (Findlabtests.com 2021). Similarly, in US the average cost of stool tests to identify enteric pathogen ranges from $150-$200. While the cost of negative sample is trivial, identification of positive sample using multiple test costs about $400 and 72-96hrs (Labcorp 2021). Different laboratories that perform these diagnostic tests balance their expenses by charging the same price for all samples. While, this is the situation in countries with well-established diagnostic laboratories, such facilities are not readily available in many parts of the world.

A web-based analytic that can facilitate pathogen suggestion during an unfolding outbreak using syndromic data will have many advantages. The information provided can be used to narrow down number of tests required to identify a specific pathogen. This will help to cut down expenditure and time required. In parts of the world where extensive laboratory networks are not available, these analyses can be used prioritize expensive laboratory tests. Moreover, these analyses can be used to plan possible mitigation efforts, so that their speedy implementation is possible upon pathogen confirmation (e.g., control strategy for viral vs. bacterial pathogen). Data aggregation sites such as Healthmap[1], ProMED-mail[2], and Flu Trackers[3] are sites that collect information from various sources. Google search query analyses tools such as Data Collaboratives (2021) and Google Dengue Trends (Strauss et al. 2017) collect search queries and correlate to disease data to identify trends and build models. These data can serve as early warning of disease outbreaks. Government resources such as the US CDC and World health Organization (WHO) also maintain and display surveillance reports and weekly summary reports. Increasingly social media and e-mail groups (Granowitz et al. 2004) are also being used by individuals and small organization to gather information on possible disease outbreaks. Websites such as the Global Early Warning System[4], and Global Infectious diseases and Epidemiology networks[5] compile reported outbreak information. Web-based tools such as Premier Biosoft (2021),Virus finder (Ding et al. 2004), pathosphere.org (Kilianski et al. 2015) are available for rapid identification of pathogen based on nucleic acid sequence and/or biochemical test results. These tools accelerate the pathogen identification only after the laboratory sequencing test results are available. The web-based technologies described above are not intended for identification of pathogen in the early stages of an unfolding outbreak, when limited information and data is available.

Perhaps the most similar work to ours is Bogich et al. (2013), who describe a method using network theory to identify the pathogen of disease outbreaks. In this method, they used properties such as disease symptoms, seasonality, and case-fatality ratio to link an ongoing outbreak to outbreaks of known etiology. This method was used to identify outbreaks from 10 different diseases with 76%

[1]Healthmap (2021) Retrieved from: https://healthmap.org/en/. [Accessed 16 March 2021]
[2]ProMED-mail (2021) Retrieved from: http://www.promedmail.org/. [Accessed 16 March 2021]
[3]Flu Trackers (2021) Retrieved from: https://flutrackers.com/forum/. [Accessed 16 March 2021]
[4]The Global Early Warning System – GLEWS (2021) Retrieved from: http://www.glews.net/. [Accessed 18 March 2021]
[5]Global Infectious Diseases and Epidemiology Network – GIDEON (2021) Retrieved from: https://www.gideononline.com/. [Accessed 18 March 2021]

sensitivity and 88% specificity. This study utilized the method to study ten diseases which can cause encephalitis, which is a rare symptom compared to fever or stomach ache. Uniqueness of the symptom was found to be important for this method. However, the source code used in their analyses is not publicly available currently.

Analytics for the Investigation of Disease Outbreaks (AIDO) (Velappan et al. 2019) is a web-based tool that contains a library of more than 650 historical outbreaks for 40 different diseases that represent the diversity of outbreak presentation for each disease. This tool currently can be used to identify the closest matching historical outbreak for an unfolding infectious disease epidemic to develop a better understanding of how the unfolding epidemic may progress and understand possible mitigation strategies based on what has been used in previous outbreaks. In AIDO, we grouped together different pathogens that produce similar symptoms to create "syndromic disease families". A set of pathogens that cause stomach ailments were grouped together to form the Gastrointestinal (GI) family. Similarly eight different pathogens that cause febrile illness transmitted by mosquitoes were grouped together and analyzed as mosquito-borne (MB) family. AIDO's historical outbreak library family of gastrointestinal pathogens includes outbreaks caused by five pathogens: *Campylobacter sp., Escherichia coli, Salmonella sp, Shigella sp* and norovirus. The pathogens included in AIDO mosquito-borne disease family includes: chikungunya virus, dengue virus, Japanese encephalitis virus, Rift Valley fever virus, Yellow Fever virus, Zika virus, West Nile virus and *Plasmodium spp*. (the parasite causing malaria). We evaluated the ability of AIDO to suggest the causative agent of an unfolding gastrointestinal or mosquito-borne disease outbreak using the similarity algorithm within particular disease families. Our initial analyses with AIDO showed that the biology of the pathogen and its transmission patterns contribute to different epidemiological features during an outbreak. For example, norovirus outbreaks have high case numbers with rapid peak time compared to *Salmonella* outbreaks that have months-long durations. In contrast, *Campylobacter* (not a hardy organism) causes low case number and outbreaks of short duration and rapid peak time. Our analyses also showed similar difference in epidemiological features for mosquito-borne diseases. The difference in case numbers, geographical area, and duration were characteristics of the mosquito species that carried the specific pathogen. We postulated that these differences in epidemiological features could allow AIDO based pathogen suggestions from syndromic disease families.

We utilized gastrointestinal and mosquito-borne disease families to develop a pathogen "suggestion" algorithm based on similarity of a user's input to historical disease outbreaks in the AIDO library. Here, we report the statistical methods used to identify epidemiological features (properties) that can distinguish between outbreaks caused by different pathogens within a syndromic disease family. We discuss development of our user interphase and display of results on AIDO. We analyzed the pathogen suggestion algorithm of both families using three types of test outbreaks: outbreaks that are part of AIDO, outbreaks currently not included in AIDO, and blinded analyses performed by our colleagues to simulate analyses performed by public health officials around the world. Results of these analyses

are discussed below. The world nowadays is acutely aware of dangers posed by emerging pathogens; therefore, we also evaluated AIDO's ability to distinguish between emerging and non-emerging outbreaks using mosquito-borne family analyses. Our results indicate the strong potential for AIDO to be used to identify a pathogen for a syndrome based outbreak, in the early stages when limited data are available.

Materials and Methods

Historical Outbreak Data Collection and Disease Specific AIDO Libraries

AIDO outbreak data is collected from publicly available data sources such as ProMED-mail, CDC, WHO[6], Eurosurveillance[7], government Ministry of Health databases as well as other scholarly journals. If data are only available as bar graphs or plots in a pdf, the plots are digitized using PlotDigitizer (WebPlotDigitizer 2021). AIDO uses statistical analyses to identify disease specific properties from a list of 27 different properties as described in Velappan et al. (2019). These disease specific properties were further analyzed for their ability to distinguish between outbreaks caused by different pathogens within a syndromic disease family.

Chi-Squared Test Analyses to Identify Properties that can distinguish between Outbreaks by Different Pathogens

The GI family has a total of 122 outbreaks in AIDO. Distribution of outbreaks for each of the disease specific properties was collected for the GI family. For example, there were 21 outbreaks caused by *Salmonella* contaminated product and 10 outbreaks of *Salmonella* that were associated with a specific location or event. Similar data was collected for the other four pathogens. Chi-test or Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories, epidemiological properties in our analyses (Wikipedia 2021). Chi-test analysis was conducted using Microsoft Excel and p-value for the property was noted. The properties chosen for chi-test has been previously shown to be significant for each of the individual diseases that make up each of the families (Velappan et al. 2019). Properties with p-values <0.05 were considered significant for inter-outbreak pathogen discrimination. Similar analyses were performed for physician density, contaminating food source, population, human development index, case definition, season, and outbreak curve shape to identify properties suitable for GI syndromic family analyses. Data collection and chi-test analyses was performed on the following properties for mosquito-borne disease family: human development index (HDI), population, precipitation category,

[6]World Health Organization – WHO (2021) Retrieved from: https://www.who.int/. [Accessed 18 March 2021]

[7]Eurosurveillance (2021) Retrieved from: https://www.eurosurveillance.org/. [Accessed 18 March 2021]

disease endemicity status, physician density (PD), case definition, climate category, presence or absence of a natural disaster, general population vs. special group, rural/urban/both (proxy for population density), population movement, case fatality rate, outbreak curve shape, ecosystem (coastal/river vs. other) and WHO region. Data for each of the statistically significant properties were entered into excel sheets for each of the outbreaks and uploaded. AIDO's automated weight calculation algorithm was used to determine the weights for each of the properties and these were used to calculate the similarity score using a weighted sum as previously described (Velappan et al. 2019).

Table 1. *Chi-Test Based Statistical Analyses to Determine AIDO Properties for Syndromic Family Analyses*

	Properties analyzed	**p-value for chi-test**	**p-value<0.05**
Gastrointestinal family			
1	contamination source	4.81E-08	TRUE
2	population	2.75E-12	TRUE
3	human development index (HDI)	0.028298777	TRUE
4	case definition	3.06E-19	TRUE
5	physician density	0.008022595	TRUE
6	outbreak curve	2.92E-12	TRUE
7	season	0.197586372	FALSE
Mosquito-borne family			
1	HDI	5.81E-10	TRUE
2	population	5.40E-09	TRUE
3	precipitation	1.36E-15	TRUE
4	disease status	1.51E-10	TRUE
5	physician density	0	TRUE
6	case definition	3.62E-11	TRUE
7	climate	0	TRUE
8	natural disaster	1.41E-20	TRUE
9	general vs. special population group	0.00094069	TRUE
10	rural vs. urban	2.32E-05	TRUE
11	population movement	0.000401464	TRUE
12	outbreak curve	0.435360193	FALSE
13	Case fatality rate (CFR)	too many unknowns	FALSE

Properties selected for analyses, p-value for chi-test and information on whether they met the criteria for inclusion are given.

Implementation of AIDO Family Web User Interphase and Development of AIDO Mobile App Mock Ups

AIDO functionalities are written in Python, using the Django, web framework and PostgreSQL, for the backend. Bootstrap, jQuery, and Plotly are used on the frontend for overall user interface design/functionality and graphs, respectively. These methods are described in Velappan et al. (2019).

Evaluation of the AIDO Pathogen Suggestion Algorithm

The syndromic family algorithms were initially tested using four outbreaks for each disease, i.e., 20 GI outbreaks and 32 mosquito-borne outbreaks. These outbreaks were already in the AIDO library and we evaluated the ability of the algorithm to display the specific pathogen as one of top five outbreaks with highest similarity score. Data were entered into user-interphase (UI) of the given family and the name of the pathogen, outbreak ID, and percentage similarity information was collected. This information was used to determine whether the specific pathogen was present in the top five matches and if yes, at which position with how much similarity. Number of positive identification was then used to calculate percent accuracy of pathogen identification. The analyses were repeated with test outbreaks that are not part of AIDO currently. We used 22 test outbreaks for the GI family and 16 test outbreaks for the mosquito-borne family. The third test on syndromic family simulated real life outbreak analyses with minimal data available during early stages of an outbreak. Peers who did not have prior knowledge outbreak performed analyses on AIDO as part of a blind study and accuracy values from tests were calculated.

Results

Properties for Inter-Pathogen Outbreak Analyses

We analyzed 122 outbreaks of the gastrointestinal family using seven different properties as shown in Table 1. Six different properties had p-value less than 0.05 and was considered statistically significant for inter pathogen outbreak analyses. These properties were contamination source, population, HDI, case definition, physician density and outbreak curve shape. Similarly chi-test analyses identified 13 statistically significant properties for distinguishing outbreaks caused by eight different pathogens in the mosquito-borne disease family. These properties are also listed in Table 1. Seasonality property of GI family and outbreak curve shape and case fatality rate (CFR) properties for mosquito-borne family did not meet the criteria for inclusion.

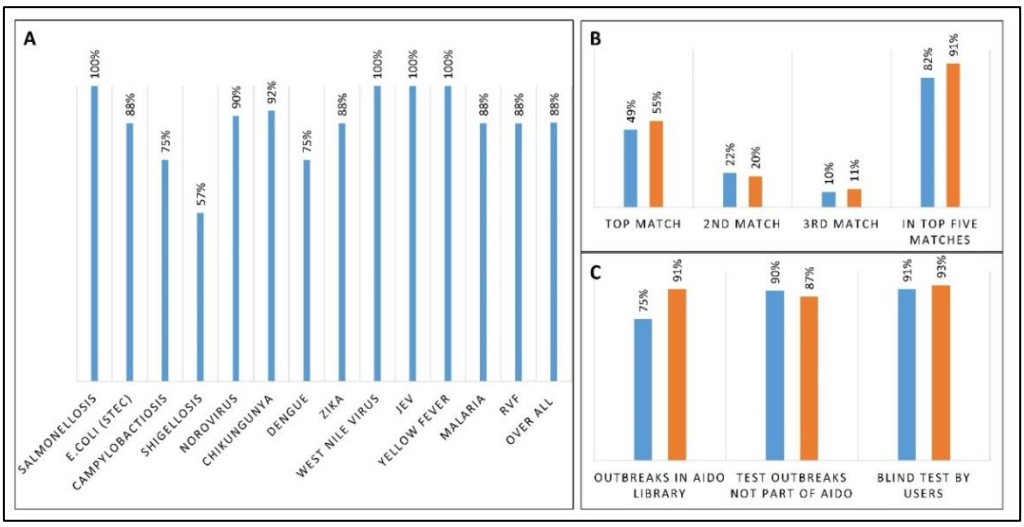
Analyses of the Syndromic Families Using Test Outbreaks

Pathogen Suggestion

We used 20 GI outbreaks and 32 mosquito-borne outbreaks for initial evaluation of the syndromic family algorithm. The results showed that AIDO algorithm brought forth a historical outbreak caused by the specific pathogen as one of the top five matches in our tests 88% of the test case scenarios. We further analyzed the syndromic disease families with 38 (22 GI and 16 MB) outbreaks not currently included in AIDO library. Our results showed that overall 88% specific pathogen suggestion was achieved with our algorithm. The success rate for individual pathogens varied (57-100%) and the data are shown in Figure 1A.

We further analyzed the specific pathogen suggestion pattern among the top five matches. Our results showed (Figure 1B) that on an average 75% of the time the specific pathogen was identified as the top matching outbreak or the second scoring historical outbreak. In ~10% of the time the third matching historical outbreak suggested the correct pathogen. Over all, success rate of pathogen identification using top five similar historical outbreaks in AIDO was 82% for gastrointestinal syndromic family and 91% for mosquito-borne disease family. Success rate of pathogen identification is similar when the tests were performed with outbreaks in AIDO or not included AIDO as well as blind test analyses (Figure 1C).

Figure 1. *Accuracy of Pathogen Suggestion AIDO Algorithm for Gastrointestinal and Mosquito-Borne Disease Families*



Data from ninety outbreaks from all parts of the world was used as test case scenarios. Panel A gives values for number of times the correct pathogen was part of the top five matches in AIDO disease family analyses. The analyses were performed with 6-10 outbreaks for each pathogen and average values are shown. Data for individual outbreak pathogen and overall value for all 13 pathogens are presented. Panel B and C show pathogen match pattern for gastrointestinal family (blue) and mosquito-borne disease family (orange). Panel B shows the average values for specific pathogen as top 1-3 matches and overall value for top five matches are given. Panel C shows the pathogen identification pattern for three types of test conducted in our evaluation and the average values are calculated based on top five matches.

Emerging Disease (Non-Endemic Outbreaks) Identification

Zika virus caused several mosquito-borne outbreaks in the Americas during 2015-2016 in non-endemic areas and AIDO contains fourteen of these emerging outbreaks of Zika. Similarly, non-endemic outbreaks were also included in AIDO library for other mosquito-borne diseases. We utilized these outbreaks occurring in non-endemic areas to assess the utility of AIDO for emerging pathogen detection. We evaluated our emerging disease detection algorithm with 11 non-endemic outbreaks (emerging in a new geographical area) and 21 outbreaks in endemic regions. In our analyses when emerging outbreak data was used as input, AIDO algorithm matched to other emerging disease outbreaks. For example, when 2016

Zika outbreak in Aruba was used as test outbreak (12 cases in 1 month), the top matching outbreaks were 2009 Dengue outbreak in Florida, USA (78%), 2014 malaria outbreak in Aswan Egypt (77%), 2016 Zika outbreak in British Virgin Island (76%), and 2001 dengue outbreak in Hawaii (71%). All four of these outbreaks were outbreaks caused by emergence of pathogen in a new location among immunologically naïve populations. Similarly, when 2004 yellow fever outbreak data from Bolivia was used as input, the AIDO algorithm matched to other outbreaks of yellow fever in Bolivia, indicating another seasonal outbreak, not an emerging pathogen. In our analyses, we were able to differentiate emerging and endemic outbreaks in 100% of the test case scenarios. Since gastrointestinal diseases are endemic in all parts of the world, these analyses were not performed for the GI family.

Case Study - Rift Valley Fever Outbreak in Mayotte, France in 2018-2019

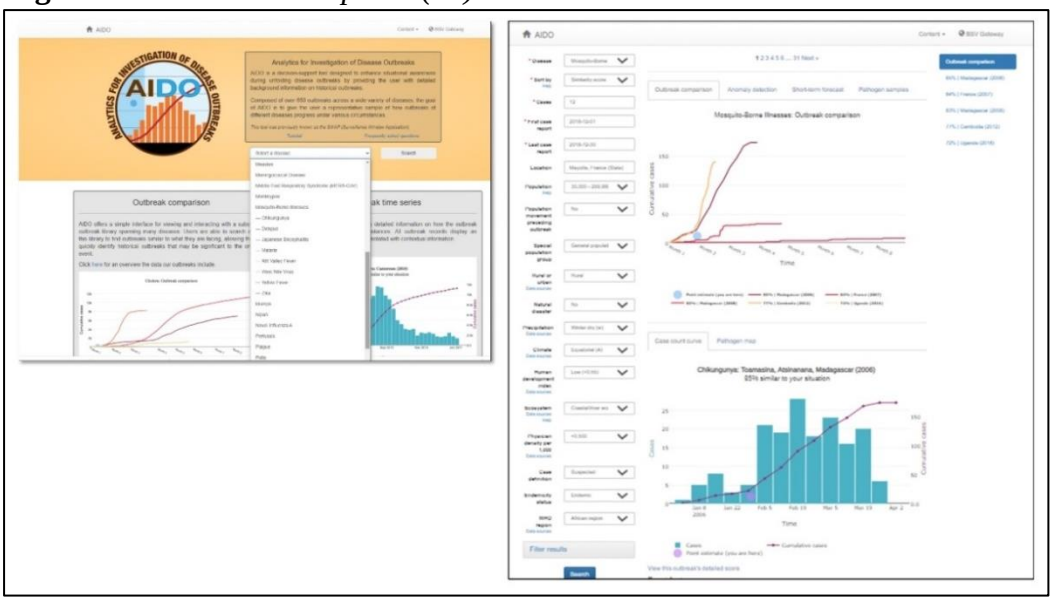
On January 4, 2019 the WHO was notified of possible Rift Valley Fever (RVF) outbreak in the Indian Ocean island of Mayotte, about 12 cases have occurred in the previous month (WHO 2019). A public health official in France is interested in understanding more about this unfolding situation. The analyst would like to perform AIDO analyses prior to sending samples for laboratory tests. The analyst accesses the AIDO family website (beta version at http://swap-play.bsvstaging.lanl.gov/)[8] and enters in 12 cases in 1-month period. Mayotte is a small island in the Indian Ocean with 30-300,000 inhabitants. The outbreak is occurring among general population in rural area, there have not been any severe natural disaster events in recent weeks. The climate and precipitation category for Mayotte are equatorial (A) and winter dry (w). Even though this island is part of France, the HDI and PD values are low and <0.55 was entered for both. As an island they have coastal/river ecosystem and this part of Africa is endemic for mosquito-borne diseases. After entering all input values, AIDO library is searched for the closest matching historical outbreaks. The UI and results are shown in Figure 2. The top matching outbreaks are 1) 2006 Chikungunya outbreak from Taomasina, Madagascar (85%) 2) 2007 RVF outbreak in Mayotte (84%), 3) 2008 RVF outbreak in Madagascar (83%) 4) Chikungunya in Cambodia (2012) (77%) and 5) 2016 RVF outbreak in Uganda (72%).

The results indicate that most likely the outbreak pathogen is RVF or chikungunya virus and laboratory tests for these two pathogens should be prioritized. Data also indicate that while the RVF historical outbreaks had 10-30 cases in 1-4 months, the chikungunya outbreaks had about 150 cases in couple of months. These suggest a possible larger outbreak of RVF or a small outbreak of chikungunya as a possibility. Reading through the historical outbreaks suggests, patients may test positive for both pathogens and testing of animal population may allow confirmation of RVF. AIDO also provides details of outbreak control measures that were successful in this region previously e.g., educational and awareness campaign regarding possible dangers of sick animal milk/meat consumption, and vector control programs. AIDO analyses also matched the

[8]AIDO 2021. Retrieved from: http://swap-play.bsvstaging.lanl.gov/. [Accessed 22 March 2021]

outbreak to endemic outbreaks, therefore it is unlikely to be an emerging zoonotic pathogen in this case. The outbreak news from WHO reported 129 cases of RVF in Mayotte, during November 2018 - May 2019, proving the AIDO analyses during early stages of the outbreak accurate (WHO 2019).

Figure 2. *AIDO User Interphase (UI)*



The left hand panel shows the AIDO home page. The right-hand panel shows the UI for mosquito-borne family and display of results.

Discussion

Traditional surveillance systems using clinical diagnosis, laboratory confirmation, and communication with public health official have been an effective strategy. However, pathogen identification and outbreak declaration tends to be quite slow and may result in loss of lives during the early stages of an outbreak. Our world has changed dramatically in the last 20 years, the availability of the internet and its usage for novel applications is increasing at a dramatic pace. Nowadays the globe is confronted with threats of bioterrorism, possible pandemics, massive population movement, and emerging infectious diseases. In fact, the global pandemic caused by SARS CoV2 virus has highlighted the extreme need for surveillance systems that provide adequate lead-time for optimal public health response (Thaker et al. 2012, Morgan et al. 2021). Syndromic surveillance was the first tool used by epidemiologists to identify occurrence and spread of COVID 19 in various localities. For example, the Illinois Department of Health collected information on patients reporting pneumonia and shortness of breath in patients coming in for emergency hospital visits. These syndromic data mapped to the time scale of increased case count for COVID-19 (IDPH 2021). Internet based syndromic disease surveillance systems offers a unique opportunity to bridge the gap between outbreak declaration and pathogen identification.

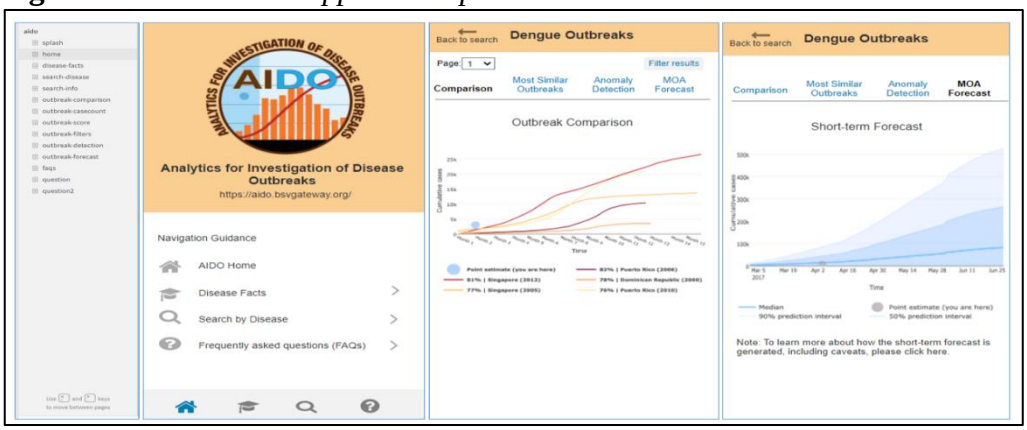
Web-based syndromic analyses using AIDO allows pathogen suggestion at no cost to users around world during early stages of an outbreak when minimal data are available. Data from 90 different outbreaks were used for these analyses involving gastrointestinal family and mosquito-borne family. We have 88% overall accuracy in suggesting the correct pathogen when considering the top five most similar historical outbreak. This high success rate can be attributed to wide array of epidemiological features used for AIDO analyses. We had originally identified 27 different properties for our intra-pathogen outbreak analysis (Velappan et al. 2019). The case count and duration are the top weighted properties in AIDO, they allow differentiation based on the outbreak trajectory, which in turn reflects the biology and transmission pattern of the pathogen. AIDO's health infrastructure properties (HDI, physician density) allow inclusion of historical pattern of effective of control measures in different nations in our analyses. Location specific properties such as climate and precipitation allows pathogen and vector species distinction in AIDO analysis. Here, we included two additional properties; ecosystem (coastal/river vs. other) and WHO region for mosquito-borne family analyses. The ecosystem property allowed us to distinguish the habitats preferred by different mosquito species (Discover Life 2021). For example, *Culex* and *Anopheles* are permanent water mosquitos, their eggs require water for survival. *Aedes spp*. is floodwater mosquitos, their eggs can dry out and then hatch once the water is present (American Mosquito Control Association 2021). Including ecosystem as a property thus allowed us to discriminate between West Nile virus, malaria and dengue/chikungunya diseases transmitted by Culex, Anopheles and Aedes mosquitoes, respectively. Different parts of the world are endemic to different mosquito-borne diseases (Dahmana and Mediannikov 2020) and the WHO region property allowed us to perform the chi-test to analyze significance and capture this valuable distinction for the mosquito-borne family. Pathogen specific properties included in GI family AIDO analyses are product vs. site/event and contamination source (e.g., cooked food, uncooked food, water, person-to-person) (Lee and Greig 2010). These properties allowed differentiation of GI pathogen biology and transmission pattern. Population based properties such as population movement, occurrence of natural disaster, and outbreak among general vs. special group population were also found to statistically significant properties for mosquito-borne family. In addition, as shown in the case study AIDO analyses can be used to glean information on possible case count and duration of the outbreak as well as effective control measures taken in historical outbreaks. Taken together our methodology that uses comprehensive list of epidemiological properties, statistical analyses, and the AIDO algorithm showed that this is valuable tool for public health officials around the world.

AIDO analyses will greatly benefit from increased outbreak library size, continued addition of newer outbreak would capture more nuance about different evolving situations around the globe. The analyses presented here can be further improved by combining the population based pathogen identification data with individual symptom based pathogen identification algorithm (Robertson et al. 2010, Grantz et al. 2020, Ni et al. 2015, Koch 2016). These two algorithms can be used to complement and increase confidence in pathogen suggestion during early

stages of an outbreak. Our analyses also showed that the AIDO family algorithm may be effective for identifying emerging disease outbreaks that represent the occurrence of a known pathogen in a new location. This is based on our analyses of outbreaks caused by Zika virus in 2015-2016 as well as outbreaks caused by other mosquito-borne pathogens in new geographic areas. We showed 100% success in distinguishing between emerging and non-emerging outbreaks. However, these analyses are cumbersome since the user has to read the outbreak factors for the top matching outbreaks or look at similarity score spider chart to determine if the top matching outbreaks are emerging or endemic. Development of new visual analytics (e.g., color coding emerging outbreaks) will enhance the emerging pathogen detection using AIDO. These visual analytics combined with an enhanced anomaly detection algorithm that we also offer in AIDO will be an effective surveillance mechanism for emerging pathogen and bioterrorism detection. The analyses shown here can be further enhanced by developing machine learning algorithm to suggest outbreak pathogen. We can explore two approaches: 1) Suggesting outbreak pathogen(s) based on top n similar outbreaks identified by AIDO similarity score instead of using a fixed number for top n (5 as of now) outbreak. 2) Suggesting outbreak pathogen(s) based on a threshold on similarity score. AIDO outbreak library will be used as a training data to evaluate our approaches. We will use nested cross-validation to choose optimal parameters and evaluate our approaches on the held out dataset (Parikh et al. 2021).

Another dramatic change in our life style in the past decade is the ubiquitous nature of mobile phones around the world. Availability of outbreak detection and analyses algorithms on mobile phones will be revolutionary (Robertson et al. 2010, Grantz et al. 2020). In an attempt to facilitate mobile app development for AIDO, we have developed 14 mobile UI mockups with Axure tool using the iPhone 8 screen size. The mocks ups are available at this link. (https://l9429o.axshare.com) and a few examples are given in Figure 3.

Figure 3. *AIDO Mobile App Mock Ups*



AIDO mobile app user interphase and examples of screen shots are given.

The data presented here and in Velappan et al. (2019) that details development of AIDO, involve analyses of disease outbreaks in humans. However, outbreaks

also occur in animal and plant populations. Future work can extend AIDO's library and framework to outbreaks among crops and livestock.

Conclusions

Here, we present a web-based visual analytic tool that can be used in the early stages of a disease outbreak anywhere in the world at no cost to the user. AIDO uses disease family-based properties to suggest comparable historical outbreaks and possible pathogens using a similarity algorithm. AIDO disease family analysis suggests plausible pathogens within specific a disease families, which can ultimately reduce the costs associated with excess laboratory testing, and quicker onset of mitigation measures. Easy-to-use, easy-to-interpret outbreak analysis tools such as AIDO are important tools for containing and preventing global pandemics.

Acknowledgments

AIDO family analyses were funded by Biosciences Division royalty fund grant sponsored by Feynman Center for Innovation, Los Alamos National laboratory. The original AIDO was funded by the Defense Threat Reduction Agency, and the Department of Homeland Security Science and Technology Directorate. Dr. Dave Osthus provided invaluable advice on the appropriate statistical tests to use in property analysis. All data for this project was collected from previously published manuscripts/reports, and was determined not to be human subjects research by the LANL Institutional Review Board.

References

AIDO (2021) Retrieved from: https://aido.bsvgateway.org/ (internal beta version http:// swap -play.bsvstaging.lanl.gov/) [Accessed 22 March 2021]

AJMC (2021) *A timeline of COVID-19 developments in 2020*. Retrieved from: https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020. [Accessed 16 March 2021]

American Mosquito Control Association (2021) *Mosquito-borne diseases*. Retrieved from: https://www.mosquito.org/page/diseases. [Accessed 25 October 2021]

Bogich TL, Funk S, Malcolm TR, Chhun N, Epstein JH, Chmura AA, et al. (2013) Using network theory to identify the causes of disease outbreaks of unknown origin. *Journal of the Royal Society Interface* 10(81): 20120904.

Centers for Disease Control and Prevention – CDC (2021) *NEW Zika and dengue testing guidance*. Retrieved from: https://www.cdc.gov/zika/hc-providers/testing-guidance.html. [Accessed 16 March 2021]

Council to Improve Foodborne Outbreak Response – CIFOR (2014) *Foodborne disease surveillance and outbreak detection*. Retrieved from: http://cifor.us/uploads/resources/CIFOR14_Chapter4_FINAL.pdf. [Accessed 16 March 2021]

Dahmana H, Mediannikov O (2020) Mosquito-borne diseases emergence/resurgence and how to effectively control it biologically. *Pathogens* 9(4): 310.

Data Collaboratives (2021) *Google flu trends.* Retrieved from: https://datacollaboratives.org/cases/google-flu-trends.html. [Accessed 16 March 2021]

Ding Y, Cao H, Wen S (2004) Virusfinder: a web-based virus identification system using viral nucleotide signatures. In Proceedings of the *2004 IEEE Computational Systems Bioinformatics Conference, 2004*, 620–621. Stanford, CA, USA: IEEE.

Discover Life (2021) *Global mapper*. Retrieved from: https://www.discoverlife.org/mp/20m?act=make_map. [Accessed 20 March 2021]

Findlabtests.com (2021) *Zika Test Cost in online lab tests stores*. Retrieved from: https://www.findlabtest.com/lab-test/infectious-disease-testing/zika-test-cost-quest-93870. [Accessed 16 March 2021]

Foddai ACG, Grant IR (2020) Methods for detection of viable foodborne pathogens: current state-of-art and future prospects. *Applied Microbiology and Biotechnology* 104(10): 4281–4288.

Granowitz EV, Srinivasan A, Clynes ND (2004) Using the Internet to identify infectious-disease outbreaks. *New England Journal of Medicine* 351(24): 2558–2559.

Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, et al. (2020) The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications* 11(1): 4961.

Illinois Department of Public Health – IDPH (2021) *COVID-19 syndromic surveillance*. Retrieved from: http://www.dph.illinois.gov/covid19/syndromic-surveillance. [Accessed 19 March 2021]

Kilianski A, Carcel P, Yao S, Roth P, Schulte J, Donarum GB, et al. (2015) Pathosphere.org: pathogen detection and characterization through a web-based, open source informatics platform. *BMC Bioinformatics* 16(1): 416.

Koch L (2016) Pathogen diagnostics for the masses. *Nature Reviews Genetics* 17(7): 378–378.

Labcorp (2021) *008144: stool culture*. Retrieved from: https://www.labcorp.com/tests/008144/stool-culture. [Accessed 18 March 2021]

Lee MB, Greig JD (2010) A review of gastrointestinal outbreaks in schools: effective infection control interventions. *Journal of School Health* 80(12): 588–598.

May L, Chretien JP, Pavlin JA (2009) Beyond traditional surveillance: applying syndromic surveillance to developing settings – Opportunities and challenges. *BMC Public Health* 9(1): 242.

MediFee (2021) *Chikungunya test cost*. Retrieved from: https://www.medifee.com/tests/chikungunya-test-cost/. [Accessed 18 March 2021]

Morgan OW, Aguilera X, Ammon A, Amuasi J, Fall IS, Frieden T, et al. (2021) Disease surveillance for the COVID-19 era: time for bold changes. *The Lancet* 397(10292): 2317–2319.

Ni P-X, Ding X, Zhang Y-X, Yao X, Sun R-X, Wang P, et al. (2015) Rapid detection and identification of infectious pathogens based on high-throughput sequencing. *Chinese Medical Journal* 128(7): 877–883.

Paixão ES, Teixeira MG, Rodrigues LC (2018) Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases. *BMJ Global Health* 3(Suppl 1): e000530.

Parikh N, Daughton AR, Rosenberger WE, Aberle DJ, Chitanvis ME, Altherr FM, et al. (2021) Improving detection of disease re-emergence using a web-based tool (red alert): design and case analysis study. *JMIR Public Health and Surveillance* 7(1): e24132.

Premier Biosoft (2021) *Pathogen detection*. Premier Biosoft Inc. Retrieved from: http://www.premierbiosoft.com/tech_notes/pathogen-detection.html. [Accessed 16 March 2021]

Robertson C, Sawford K, Daniel SLA, Nelson TA, Stephen C (2010) Mobile phone-based infectious disease surveillance system, Sri Lanka. *Emerging Infectious Diseases* 16(10): 1524–1531.

Strauss RA, Castro JS, Reintjes R, Torres JR (2017) Google dengue trends: an indicator of epidemic behavior. The Venezuelan case. *International Journal of Medical Informatics* 104(Aug): 26–30.

Thaker SB, Qualters J, Lee L (2012) Public health surveillance in the United States: evolution and challenges. *Morbidity and Mortality Weekly Report (MMWR)* 61(03): 3–9.

Thermo Scientific Inc. (2021) *Thermo scientific food testing solutions*. Retrieved from: http://www.remel.com/pdf/Food%20Testing%20Solutions%20Guide%20-%20FINAL.PDF. [Accessed 16 March 2021]

Velappan N, Daughton AR, Fairchild G, Rosenberger WE, Generous N, Chitanvis ME, et al. (2019) Analytics for investigation of disease outbreaks: web-based analytics facilitating situational awareness in unfolding disease outbreaks. *JMIR Public Health and Surveillance* 5(1): e12032.

Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, et al. (2020) Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA* 323(18): 1843–1844.

WebPlotDigitizer (2021) *Extract data from plots, images, and maps*. Retrieved from: https://automeris.io/WebPlotDigitizer/. [Accessed 18 March 2021]

World Health Organization – WHO (2019) *Rift valley fever – Mayotte (France)*. Retrieved from: https://www.who.int/csr/don/13-may-2019-rift-valley-fever-mayotte-france/en/. [Accessed 19 March 2021]

Wikipedia (2021) *Chi-squared test*. Retrieved from: https://en.wikipedia.org/w/index.php?title=Chi-squared_test&oldid=1008237521 [Accessed: 21 March 2021]