A Web-based Analytic for Pathogen Identification Using Syndromic Data

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.bsgateway.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 of five different pathogens and the mosquito-borne family has 151 outbreaks caused by eight pathogens. Chi-squared test based statistical analysis was used to identify epidemiological properties that can distinguish between inter-family outbreaks. 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, the top first and second matching historical outbreak correctly identified the outbreak pathogen in 75% of test case scenarios. The success rate was 88% when top five matches were used in pathogen identification. Our analyses with mosquito-borne outbreak family also showed that AIDO is capable of identifying outbreaks caused by emerging pathogen. This easy-to-use web-based analytic will be a useful tool in pandemic preparation 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 rapid implementation of effective control measures. Rapid identification of pathogen of an unfolding outbreak can save countless lives, especially if it is an emerging pathogen. Disease surveillance is used 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. The pathogen specific surveillance system detects 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 involves automatic 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 independent agencies such as state health departments to identify possible disease clusters. The syndromic surveillance has the ability identify outbreaks earlier and faster than pathogen detection. However, the pathogen associated with an outbreak may not be apparent immediately. 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\(^1\). However, the disadvantage associated with this fast and cheap method is again the inability to identify specific pathogen\(^2\). 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. 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\(^3\). 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\(^4\). Antibody based serological testing and nucleic acid based analyses, such as those developed for SARS CoV2 are employed for viral pathogen detection\(^5\). 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\(^6\). 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\(^6\). 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\(^7\). 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\(^8\) and the tests cost about USD 1240 if all comprehensive tests listed in the algorithm are performed\(^9\). 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\(^10\). 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 identification of 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 as first line of defense against an ongoing outbreak. 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, Promed mail, and FluTracker.com are sites that collect information from various sources. Google Search analyses tools such as Google flu trends and google dengue trends collect search data from google.com searches and correlate data to perform trend analyses. 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 are also being used by individuals and small organization to gather information on possible disease outbreaks. Websites such as the Global Early Warning system (GLEWS), and Global Infectious diseases and Epidemiology networks (GIDEON) compile reported outbreak information. Web based tools such as Premier Biosoft, Virus finder, pathosphere.org 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 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. Bogich et al. described 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% sensitivity and 88% specificity. This study utilized the method to study ten diseases causing 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 publically available currently.

Analytics for Disease outbreak Investigation (AIDO) 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 outbreak to develop a better understanding of how the unfolding outbreak may progress and obtain possible mitigation strategies. AIDO’s historical outbreak library includes representative outbreaks caused by five different gastrointestinal (GI) pathogens: Campylobacter sp., Escherichia coli, Salmonella sp, Shigella sp and norovirus. The pathogens included in AIDO mosquito-borne (MB) disease family are: chikungunya virus, dengue virus, Japanese encephalitis virus, Rift Valley fever virus, Yellow Fever virus, Zika virus, West Nile virus and Plasmodium spp. that causes malaria. We evaluated the ability of AIDO to use these two disease families and identify the causative agent of an unfolding gastrointestinal outbreak and mosquito borne disease outbreak. 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 often lasts into months.
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 difference in epidemiological features would allow AIDO based pathogen identification from syndromic disease families.

We utilized gastrointestinal and mosquito-borne disease families to develop a pathogen “suggestion/identification” 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 identification 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 is available.

Materials and Methods

Historical outbreak data collection and disease specific AIDO libraries

AIDO outbreak data is collected from publically available data sources such as ProMED, CDC, WHO, Eurosurveillance, 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. 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 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 were 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. 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. 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, 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.

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/identification algorithm

The syndromic family algorithms were initially tested using four outbreaks
for each diseases, 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 specific identification was then used to
calculate percent accuracy of pathogen identification. The analyses was 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 identification/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 36 (22 GI and 16 MB) outbreaks not
currently included in AIDO library. Our results showed that overall 88% specific
pathogen identification 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 identification 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).
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 distinguish 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 2019-2020

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. 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/) 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 had 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. This suggests 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 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.

Discussion

Traditional surveillance systems using clinical diagnosis, laboratory confirmation, and communication to 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 early stages of an outbreak. Our world has changed dramatically in the last 20 years, the availability of 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. 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. Internet based syndromic disease surveillance system offers a unique opportunity to bridge this gap.

Web-based syndromic analyses using AIDO allows pathogen identification/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 success rate in identification of specific pathogen in 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. The case count and duration are the top weighted property in AIDO, they allow distinction 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. 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. Including ecosystem as a property thus allowed us to distinguish between West Nile virus, malaria and dengue/chikungunya diseases transmitted by *Culex*, *Anopheles* and *Aedes* mosquitos, respectively. Different parts of the world are endemic to different mosquito-borne diseases 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). These properties allowed distinction between 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 possible case count and duration of the outbreak as well as effective control measures taken in past 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 will capture the 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. These two algorithm can be used to complement and confirm pathogen 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 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.

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. 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 4.

The data presented here and in the 2019 publication that details development
of AIDO involves analyses of disease outbreaks in humans. Unfortunately,
outbreaks occur among animal and plant populations as well. Disasters of famine
as a result of outbreaks among crops are part of human history, for example the
Irish famine of 1800s caused by potato blight and the ensuing mass migration. In
todays interconnected world both pathogens and vectors spread across the globe at
jet speed and therefore rapid identification of an outbreak pathogen with easy to
use tools are essential for health and economic security of all nations and their
citizens. The concepts behind AIDO, using historical outbreaks provide contextual
information during an unfolding outbreak can be extended to outbreaks among
crops and livestock. In conclusion, we present a web-based visual analytic tool
that can be used in syndromic surveillance at population level during early stages
of a disease outbreak anywhere in the world at no cost to the user.

Acknowledgement

AIDO family analyses was 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
Ostthus 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


### Table 1. Chi-test based statistical analyses to determine AIDO properties for syndromic family analyses. Properties selected for analyses, p-value for chi-test and information on whether they met the criteria for inclusion are given

<table>
<thead>
<tr>
<th>properties analyzed</th>
<th>p-value for chi-test</th>
<th>p-value&lt;0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gastrointestinal family</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 contamination source</td>
<td>4.8072E-08</td>
<td>TRUE</td>
</tr>
<tr>
<td>2 population</td>
<td>2.7507E-12</td>
<td>TRUE</td>
</tr>
<tr>
<td>3 human development index (HDI)</td>
<td>0.028298777</td>
<td>TRUE</td>
</tr>
<tr>
<td>4 case definition</td>
<td>3.05648E-19</td>
<td>TRUE</td>
</tr>
<tr>
<td>5 physician density</td>
<td>0.008022595</td>
<td>TRUE</td>
</tr>
<tr>
<td>6 outbreak curve</td>
<td>2.92329E-12</td>
<td>TRUE</td>
</tr>
<tr>
<td>7 season</td>
<td>0.197586372</td>
<td>FALSE</td>
</tr>
<tr>
<td><strong>Mosquito-borne family</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 HDI</td>
<td>5.81442E-10</td>
<td>TRUE</td>
</tr>
<tr>
<td>2 population</td>
<td>5.40146E-09</td>
<td>TRUE</td>
</tr>
<tr>
<td>3 precipitation</td>
<td>1.36209E-15</td>
<td>TRUE</td>
</tr>
<tr>
<td>4 disease status</td>
<td>1.51466E-10</td>
<td>TRUE</td>
</tr>
<tr>
<td>5 physician density</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>6 case definition</td>
<td>3.62336E-11</td>
<td>TRUE</td>
</tr>
<tr>
<td>7 climate</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>8 natural disaster</td>
<td>1.40579E-20</td>
<td>TRUE</td>
</tr>
<tr>
<td>9 general vs. special population group</td>
<td>0.00094069</td>
<td>TRUE</td>
</tr>
<tr>
<td>10 rural vs. urban</td>
<td>2.32303E-05</td>
<td>TRUE</td>
</tr>
<tr>
<td>11 population movement</td>
<td>0.000401464</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
Figure 1: Accuracy of pathogen identification/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 was 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 shows the specific pathogen match pattern for the 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.
Figure 2: AIDO user interface (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.

Figure 3: AIDO mobile app mock ups. AIDO mobile app user interface and examples of screen shots are given.