Distribution Agreement

Artificial Intelligence to Modernize Public Health Surveillance

by

Oluwafunbi Idowu Awoniyi

Master of Public Health

Hubert Department of Global Health

Scott JN McNabb, PhD, MS

Thesis Advisor/ Committee Chair

Artificial Intelligence to Modernize Public Health Surveillance

by

Oluwafunbi Idowu Awoniyi, M.D

B.A., Georgia State University, 2010

M.Sc., Georgia State University, 2012

M.D., St. George's University School of Medicine, 2018

Thesis Committee Chair: Scott JN McNabb, PhD. MS

An abstract of

A thesis submitted to the faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of Master

of Public Health in Hubert Department of Global Health,

2022

Abstract

Artificial Intelligence to Modernize Public Health Surveillance

by

Oluwafunbi Idowu Awoniyi

Introduction

During COVID-19, countries like the United States – with well-financed public health surveillance (PHS) – realized their systems had become overwhelmed and could not keep up with public health demands. With U.S. PHS siloed (i.e., based on disease and data abstracted from proprietary healthcare systems) it was impossible to address (re)emerging disease issues and healthcare complications. Awareness of these gaps led the global public health community to consider and reframe the modernization of PHS.

Objective

Show the important role of artificial intelligence (AI) and machine learning (ML) in modernizing PHS data sources, collection, case detection, prediction, analyses, reporting, and forecasting.

Methods

I performed a systematic literature review of published articles using PubMed™, Science Direct™, SCOPUS™, Web of Science™, Academic Search Complete and Institute of Electrical and Electronics Engineers (IEEE)™ from Jan 2000 – Jan 2022. Articles were imported, screened, and assessed for eligibility using Covidence™ systematic review software. Selection based on eligibility criteria (e.g., published in English, peer-reviewed, related to PHS, AI, ML, and their subsets) allowed 1,983 articles imported into Covidence™. After removing duplicates, 1,443 were screened in the first stage based on title and abstract, and a full-text review of 274 articles was conducted where 259 were eliminated; fifteen articles remained and were analyzed.

Results

Specific details such as author's name, title, aim/objectives, data sources, AI methods used, key findings, publication year, and country of study were extracted. Most articles analyzed were from the United States (n = 6), South Korea (n=3) China (n=2), Canada (n = 2); France, Pakistan, United Kingdom, Italy, Singapore, and Philippines published one study each. Most of the articles were published in 2021(n=7). The synthesis of these articles resulted in the formulation of three themes: data sources and collection, prediction and forecasting, and detection.

Discussion

AI and ML should be applied in modernizing PHS for data sourcing, collection, case detection, prediction, analyses, reporting, and forecasting. Modernizing PHS will take funding plus scientific and political commitment; AI and ML can make it successful.

Artificial Intelligence to Modernize Public Health Surveillance

by

Oluwafunbi Idowu Awoniyi

B.A. Georgia State University, 2010

M.Sc. Georgia State University, 2012

M.D. St. George's University School of Medicine, 2018

Thesis Committee Chair: Scott JN McNabb, PhD. MS

A thesis submitted to the faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Hubert Department of Global Health,

2022

Acknowledgements

First and foremost, I am extremely grateful to my Advisor, Dr. Scott McNabb, for his invaluable advice, continuous support, and patience in the past year. His immense knowledge, steady guidance and enthusiasm for the topic made a positive impression on me throughout my entire academic research and in my daily life.

I would also like to thank the Research librarian, Ms. Shenita Peterson, for her technical assistance while I was working on this project. Her dedicated support through all the stages of writing my project is deeply appreciated. I would like to thank my ADAPs, Ms. Flavia Traven and Ms. Catherine Strate for their supportive involvement with my class enrollments and assistance in resolving the challenges faced during my academic program.

Completing this thesis required more than academic support and for that, I would like to express my gratitude and appreciation to my amazing family- Olubukola, Adesoji, Adegbemisola, Olufunmilola, Kome, Olubusola, Olayinka, Tayelolu, Joe, my nieces and nephew. Without their tremendous understanding, support, and encouragement in the past two years, it would have been impossible for me to complete my studies.

Finally, I would like to thank God, for guiding me and keeping me healthy throughout my academic research especially during an unprecedented pandemic period. I have experienced your guidance daily and I will keep on trusting you for a bright future in my career.

Table of Contents

Chapter 1. Introduction..1-4

Chapter 2. Methods...5-6

2.1 Information sources and search strategy……………………………….…………5

2.2 Study selection………………………………………………………...…...…5

2.3 Eligibility criteria…………………………………………………….........…5

2.4 Data collection, and extraction……………………………….................6

Chapter 3. Results 7-15

3.1 Distribution of studies by publication year and country…………………... 7

3.2 Data synthesis and charting Process………………………...…………...8

Chapter 4. Discussion..16-20

4.1 Conclusion 16

4.1.1 Data sources and collection 16

4.1.2 Use of AI to Predict and Forecast infectious disease events 17

4.1.3 The utility of AI in infectious disease Detection……………………..18

4.2. Limitation……………………………………………………......………19

4.3. Recommendation…………………………………………….....…..……19

References...20-24

Chapter 1. Introduction

The COVID-19 pandemic has shaken the world and put global Public Health Surveillance (PHS) to the test. While the World Health Organization (WHO) and member states (MS) responded to the initial outbreak and the new reality, many could not keep up with overwhelming healthcare demands. To respond to an outbreak or pandemic, there must be effective and efficient PHS. During the COVID-19 outbreak, countries like the United States with well-funded PHS quickly realized their systems had become overwhelmed and could not keep up with increasing demands. For example, the Global Health Security (GHS) index which measured a country's preparedness for an epidemic or a pandemic, ranked the United States No. 1 for being prepared.[1] But, when COVID-19 hit, the United States was not prepared and fell dramatically to the bottom of the list.

In the United States, the National Notifiable Disease Surveillance system (NNDSS), managed by the U.S. Center for Disease Control and Prevention (CDC) monitors, controls, and prevents reportable infectious diseases, bio-terrorism agents, and non-infectious conditions. However, all data are managed through >120 siloed systems.[2] With each PHS silo based on the disease *de jure*, it has been impossible to address (re)emerging diseases or new health complications because vital information was not available.

Beginning with the COVID-19 pandemic, it was difficult to obtain information about the virus and its symptoms because researchers could not obtain the right data using PHS. Awareness of these gaps led CDC and the global public health community to consider modernizing PHS. CDC laid out the following three goals they would like to achieve their objectives.

- *Enhance the accountability, resource use, workforce, and innovation for surveillance at CDC and in support of federal and state, territorial, local, and tribal agencies;*
- *Accelerate the use of emerging tools and approaches to improve the availability of quality and timely surveillance data;*

- *Demonstrate early success through four crosscutting surveillance system initiatives to improve public health surveillance outcomes.*[3]

One approach to modernize PHS is the use of Artificial Intelligence (AI). AI improves how services are provided, specifically in healthcare. The advancement of AI has been accelerated by the growing accessibility to large datasets such as social media and keyword searches.[4]

To understand how to modernize PHS, there must be an understanding of PHS. PHS is *defined as the systematic collection of health information for the purposes of monitoring, preventing, or controlling the spread of diseases in a population.*[5] Its purpose is to provide information used to implement and guide public health policies and programs by decision-makers (e.g., local, state, and national public health officials).[3] PHS involves various fields of public health (e.g., environmental health, community health, epidemiology, infectious diseases, global health, and health communication and policy).[6] Elements of PHS include data collection, analyses, interpretation, dissemination of findings, and response.[7]

Information collected from PHS should characterize disease patterns, detect disease outbreaks, suggest hypothesis for further investigation, identify cases for further research, and guide and evaluate public health and disease control programs.[8] There are several forms of PHS, but the two major ones are passive and active. Passive PHS involves standard reporting of notifiable conditions by healthcare workers or others involved in patient care.[9] Examples of passive surveillance include Adverse Event Reporting System (AERS) by the Food and Drug Administration (FDA) and Vaccine Adverse Reporting System (VARS), by CDC.[10]

Active PHS involves local and state health officials actively seeking information from healthcare workers, and anyone involved in patient care. This type of PHS is used to investigate diseases or exposures that pose a health risk to the public.[9] Examples of active PHS include Polio and *E. coli* outbreak PHS.[11] Other types of PHS are Sentinel, Syndromic, population-based, laboratory-based, community-based, and digital PHS.[5]

In 2004, John McCarthy, one of the founders of Artificial Intelligence (AI), defined AI as the *science and engineering of making intelligent machines, especially intelligent* *computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable*.[12]

Over the decades, AI has evolved but the concept remains the same. AI has six major branches: expert systems; robotics; machine learning (ML); neural network; fuzzy logic; and natural language processing.[13] There are two types of AI are supervised and unsupervised learning. Supervised learning requires human input to accurately label data used for training and development of a model, while unsupervised learning has no human input and is left to learn on its own.[14]

AI, especially ML, uses data collected from various credible sources; the data are analyzed and can be used to develop a high-level algorithm or model showing patterns in the data to predict future events.[15] In the health sector, these data can be used to modernize detection, confirmation, analyses, reporting, mitigation, and elimination. The analyses of data used can support public health research and PHS to guide policy and program initiation and evaluation.[16]

AI, specifically ML, has being used in various aspects of PHS (e.g., disease prediction, forecasting disease future event values, and identifying disease cases). ML uses past collected data to develop a prediction model that eventually predicts future circumstances of importance.[17] ML is now being used in medicine, finance, global security, and other fields.[18] ML in medicine is used to identify diseases by interpreting, analyzing large datasets and predicting their outcome. This is used to treat diseases such as pneumonia, diabetes, and cancer. It has also been used to decide the type of treatment a patient might need. ML is used for forecasting and predicting an infectious disease pattern, past data trains the machine and develops an algorithm producing a predictive and forecasting model. To evaluate the accuracy of the newly developed model, new data are entered. During COVID-19, ML identified patients infected with the SARS-CoV-2 virus based on their symptoms; it was also used in India to forecast future confirmed COVID-19 cases.[17]

The objective of this systematic review is to highlight the application of AI for modernizing PHS, especially the use of AI processes such as ML and deep learning (DL) to improving data collection, case prediction, forecast and detection. To understand this process of modernizing PHS, three research questions have guided this research.

- How is AI currently being used in the health services sector?
- What is the role of AI in improving data collection sources?
- How are case detection and prediction created using AI?

Chapter 2. Methods

2.1. Information Sources and Search Strategy

A systematic literature review was conducted on articles published between Jan 2000 and Jan 2022. Six databases (five health science databases and one computer science database) including PubMed™, Science Direct™, SCOPUS™, Web of Science™, Academic Search Complete and Institute of Electrical and Electronics Engineers IEEE™ were thoroughly searched. Search themes and keywords were synthesized using previously described classifications of AI (e.g., machine and deep learning) including a broad category of a few types of PHS. Relevant articles were identified using search terms (e.g., “public health surveillance", "passive surveillance", "syndromic surveillance,” “sentinel surveillance", "disease surveillance,” “active surveillance", “artificial intelligence,” “augmented intelligence,” and “machine learning”).

2.2. Study Selection

Articles were exported from the various databases mentioned and imported into the Covidence™ (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia). Covidence™ restructures the production of systematic reviews through citation importing, title, and abstract screening, full text review, extraction of study characteristics, and export of data and references. Duplicate articles were eliminated before meticulously screening in two stages. First, the source articles were screened by one reviewer for significance based on the information provided in the title and abstract. These criteria included “Public Health Surveillance” and its subsets, “Artificial Intelligence” and its subsets, “Infectious diseases” and then evaluated for inclusion based on the full text.

2.3. Eligibility Criteria

- Articles related to PHS, AI, and their subsets (ML, deep learning); articles published in English; published in peer reviews; articles with available full text, and articles that focused on infectious disease (ID) were included.
- Articles unrelated to ID, “public health surveillance” OR “Artificial Intelligence”

- Articles that were unrelated to AI prediction, forecasting, data sources, collection, reporting and diagnosis abstracts, not peer-reviewed, and books not published in English were excluded.

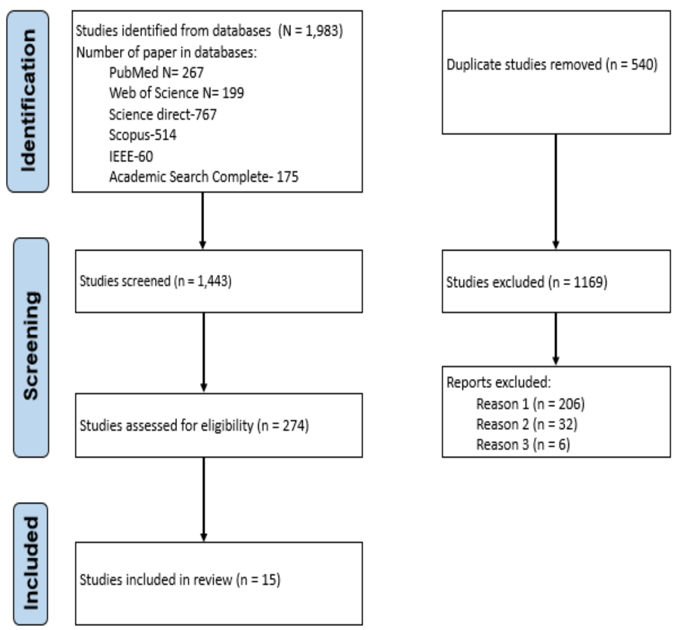
2.4. Data Collection and Extraction

Full text articles discovered to met the inclusion criteria were moved to the extraction phase. Both quantitative and qualitative data (e.g., author, study title, study aim/objectives, data sources, AI method used, key findings, publication year, and country of study) were extracted using a Covidence™ data extraction template. The extracted data were then categorized based on the three main themes observed in the results (data sources and collection, prediction and forecasting, and detection). These data were then organized into a table and inserted into Microsoft Word™. Institutional Review Boards (IRB) approval was not required, as this research did not involve human subjects.

Chapter 3. Results

This systematic review was conducted using six databases by applying a combination of the keywords PHS and AI. A total of 1,983 articles were imported into Covidence™; duplicates were eliminated producing 1,443 articles that were carefully screened, in the first stage, based on title and abstract. A full-text review of the remaining 274 articles was completed resulting in the exclusion of 259 articles because they did not meet the specified inclusion criteria. This systematic review was conducted using the Preferred Reporting Items of Systematic Reviews and Meta-Analysis (PRISMA) guidelines. At the end of the exercise, fifteen articles were extracted. (Figure 1)

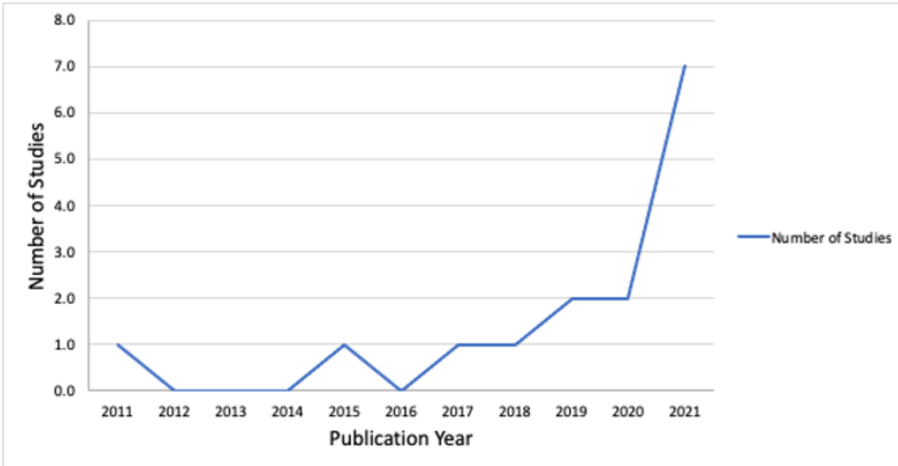
Figure 1. PRISMA flow diagram for Public Health Surveillance and Artificial Intelligence



3.1. Distribution of Studies, by Publication Year and Country

The first study was published in 2011 while most were published in 2021. (Figure 2)

Figure 2. Distribution of Studies, by Publication Year, 2011 – 2021



Most studies were conducted in the United States (n = 6), South Korea (n=3) China (n=2), Canada (n = 2), France, Pakistan, United Kingdom, Italy, Singapore, and Philippines had published one study each (some studies were produced by more than 1 country).

3.2. Data Synthesis and Charting Process

Fifteen studies were extracted and specific details such as author, study title, study aim and objectives, data sources for AI development, AI methods used, key findings, publication year, and country of study were extracted. The synthesis of these studies resulted in the formulation of three themes: data sources and collection, prediction and forecasting, and detection. (Table 1)

Table 1. Summary of PHS Artificial Intelligence Approaches from 2000-2022

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Data Sources and Collection					
Mackey, *et* al.[19]	Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study	Detect and characterize user-generated conversations associated with COVID-19-related symptoms, testing, and recovery using an unsupervised ML	4,492,954 tweets containing terms that could be related to COVID-19 symptoms from March 2020	Unsupervised machine learning approach called BTM	BTM was used to analyze the filtered dataset to identify relevant topic clusters. BTM was able to extract 35,786 tweets for the topic clusters
Chae, *et* al.[20]	Predicting Infectious Disease Using Deep Learning and Big Data	Design a model that uses the infectious disease occurrence data provided by the Korea Center for Disease Control (KCDC), search query data from search engines in South Korea, social media, and weather data.	1. Naver search engine data for data collection, three infectious diseases in south Korean. 2. Daily Social media data from twitter and average daily weather data were collected. 3. Daily number of confirmed infectious disease diagnoses from 2016-2017	Prediction models- DNN, LSTM, ARIMA, and OLS	DNN model can be used as a prediction tool when predicting the minimum value \ for disease occurrence and using the LSTM model when predicting the maximum value

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Poirier, *et* al.[21]	Influenza forecasting for French regions combining EHR, web and climatic data sources with a machine learning ensemble approach	Present a machine-learning modeling approach that produces real-time estimates and short-term forecasts of influenza activity for the twelve continental regions in France by using multiple data sources.	1. Weekly Influenza-like-illness (ILI) incidence rates from the French Sentinelles network from 2004-2017 2. Electronic health record data (EHR) from the clinical data warehouse	The ARGO, Net, AGROnet, autoregressive, baseline models	The proposed ensemble approach, named ARGONet, produced forecasts with the lowest errors, highest correlation, and most reliable longer-term forecasts.
Kim, *et* al.[22]	Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches.	Provide a quantitative evaluation of AI in the automated identification of information related to ID occurrences using online sources.	Data containing the names of 100 major infectious diseases were collected from WHO-DON, WHO-IHR, WHO-AFRO, NCDC, and SAMOH.	Two machine learning algorithms - ConvNet and BiLSTM and Two classification methods- DocClass and SenClass	The performance of BiLSTM with SenClass yielded an overall accuracy of 92.9% in classifying infectious disease occurrences.
Edo-Osagie, *et* al.[23]	Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance	Establish the utility of social media data, specifically, Twitter data for syndromic surveillance	Tweets from twitter from different seasons for 2015-2017	Compared supervised classification algorithms (Naive Bayes, DTs, LR, SVM and MLP to semi-supervised approach	Reliable performance was obtained in classifying symptomatic tweets with supervised and semi-supervised and the proposed semi supervised algorithms

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Prediction and Forecasting					
Ajayi *et* al.[24]	Forecasting herd-level Porcine Epidemic Diarrhea (PED) frequency trends in Ontario (Canada)	Determine the most accurate machine-learning approach for forecasting future Porcine Epidemic Diarrhea Virus (PEDV) trends in Ontario swine herds.	Weekly incidence and prevalence measures from the Ontario Area Regional Control and Elimination (ARC&E) project in Canada.	The RF, artificial neural nets, and classification tree algorithms were selected for forecasting PEDV trends	The results show that the RF classification model with 30 explanatory variables was the best model for forecasting future PEDV trends for this target population.
Guo *et* al.[25]	Developing a dengue forecast model using machine learning: A case study in China	construct an accurate forecast model to track the epidemic trajectory of dengue by comparing different prediction algorithms	1. Weekly dengue case data of Guangdong, from 2011 to 2014, were obtained from the Guangdong Provincial CDC. 2. Meteorological Data obtained from the China Meteorological Data Sharing Service System	ML algorithms - SVR, step-down linear regression, GBM, NBM, LASSO, linear regression model and GAM	The SVR model had the consistently smallest prediction error rates for tracking the dynamics of dengue and forecasting the outbreaks in other areas in China

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Necesito *et* al.[26]	Combination of univariate long-short term memory network and wavelet transform for predicting dengue case density in the Philippines	Develop a modeling technique that can be used to help guide control and prevention measures during an early stage of a dengue surge	Monthly dengue case density data from the National Capital Region (NCR) the Philippines from 1994 to 2018	Univariate LSTM network	LSTM combined with DWT could be used to model disease data containing surges or noises associated with an oscillating system or a seasonal pattern, such as dengue
Bellocchio *et* al.[27]	Enhanced Sentinel Surveillance System for COVID-19 Outbreak Prediction in a Large European Dialysis Clinics Network	Develop an advanced sentinel surveillance system supported by a machine learning prediction model, where the occurrence of COVID-19 cases in a clinic propagates distance-weighted risk estimates to adjacent dialysis units	1. SARS-CoV-2 infections aggregated data from April 2020 from Treatment Incident Report (TIR) module in EuCLiD 2. Aggregated data on biochemical assays prescriptions from dialysis clinics 3. Open-source detailing epidemic dynamics in European countries	ML sentinel surveillance system	The ML sentinel surveillance system can provide a robust strategy to assess the level of community transmission of COVID-19 and to guide the selection and mitigation measures

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Althouse *et* al.[28]	Prediction of Dengue Incidence Using Search Query Surveillance	Accurately predict an increase in incidence to generate a series of clinical interventions, and public health interventions to reduce the transmission of dengue fever	1. Google searches of signs and symptoms related to dengue and chikungunya, in English, Chinese Malay, and Tamil in Singapore and Bangkok from 2004-2011 2. Weekly dengue incidence data from Singapore and Monthly dengue incidence data from Bangkok from 2004 - 2011	Step-down linear regression, GBR, and negative binomial regression were used to predict incidence cases LR and SVM models were used to predict periods of high incidence	1. The AIC step-down model outperformed the GBR and negative binomial model for predicting numbers of incident cases and was chosen as optimal in both Singapore and Bangkok 2. For both Singapore and Bangkok, logistic regressions and SVM models were fit to predict the binary outcome of incidence above or below a threshold
Detection					
Jingxin *et* al.[29]	COVID-19 lesion detection and segmentation-A deep learning method	Supply rapid and precise assistance for disease surveillance on the medical imaging aspect using deep learning methods	CT images of 50 patients with pulmonary lesions and 30 other healthy people from the China-Japan Union Hospital, Jilin University	Deep Learning	Results showed that spatial-only model (Ours-SP) and Complete one (Ours) both had performed better compared to the Mask R-CNN and U-net across the four metrics

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Amin *et* al.[30]	Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches	Propose a machine-learning-based approach to detect dengue and flu outbreaks in social media platform Twitter, using four machine learning algorithms	Tweets extracted from twitter on dengue and flu	4 Machine learning algorithms: RF, KNN, SVM, and DT	Results showed that the RF classifier had outperformed SVM, DT, and KNN in terms of accuracy, precision, recall, and F1 measure
Peterson *et* al.[31]	Automated Travel History Extraction from Clinical Notes for Informing the Detection of Emergent Infectious Disease Events: Algorithm Development and Validation	Assess the feasibility of annotating and automatically extracting travel history mentions from unstructured clinical notes in the Department of Veterans Affairs across disparate health care facilities	Clinical notes for 57 patients from 2015-2018 were obtained from the Veterans Affair (VA) Corporate Data Warehouse.	ML and neural language models	The proposed model was the best performing model, with precision 88.0%, recall 83.3%, and F1 measure 85.6% which showed that automated extraction is feasible from patient travel history clinical notes for passive PHS.

Author	Title	Aim/Objective	Data sources	AI Method used	Key Results
Pineda *et* al.[32]	Comparison of machine learning classifiers for influenza detection from emergency department free-text reports	Systematically evaluate the performance of ML classifiers based on three different missing data category configurations applied to both training and test datasets with Brier score	31,268 Emergency Department (ED) reports from four EDs at the University of Pittsburgh Medical Center Health System were collected between 2008-2011	7 Machine Learning Classifiers-Bayesian Network (Naive Bayes, K2, EBMC), Logistics Regression, ANN, SVM, RF	1. Using the Brier Skill Score (BSS, all ML classifiers can perform better compared to the expert constructed classifiers given a particular NLP extraction system
Parikh *et* al.[33]	Improving Detection of Disease Re-emergence Using a Web-Based Tool (RED Alert): Design and Case Analysis Study	Compose disease-related data needed to understand disease detection and disease re-emergence	Case prevalence, vaccination rates, and indicators related to disease transmission from WHO, Pan American Health Organization (PAHO), World Bank, and Gideon	A machine learning approach was used, using RED Alert (web-based tool) and random forest classifier model	Supervised learning models were able to classify 82%-90% of the local re-emergence cases, however, with 19% to 31% (except 46% for dengue) false positives

* Algorithm- Artificial Neural Network (ANN), Autoregressive Integrated Moving Average (ARIMA), Bidirectional Long Short-Term Memory (BiLSTM), Biterm Topic Model (BTM), Deep Neural Network (DNN), Decision Tree (DT), Generalized Additive Model (GAM), Gradient Boosted Regression (GBM), Generalized Boosted Regression (GBR), K-Nearest Neighbor (KNN), Least Absolute Shrinkage and Selection Operator (LASSO),Logistic Regression (LR),Long-Short Term Memory (LSTM), Multilayer Perceptron (MLP), Negative Binomial Regression (NBM), Ordinary Least Squares (OLS), Region-based Convolutional Neural Networks (R-CNN), Random Forest (RF), Support Vector Machine (SVM), and Support Vector Regression (SVR).

Five of the fifteen studies selected for extraction had a common theme of the various sources of data used to collect data. Other others for extraction also had similar themes on prediction, forecasting, and detection of ID using AI methods for PHS.

Chapter 4. Discussion

4.1. Conclusion

With AI going mainstream in the early 2010s, there has been a nonstop advancement and expansion in its development to improve PHS, especially in data collection, prediction, and disease detection. Several ML and DL algorithms and models were observed in this review. They were ANN- Artificial Neural Network, SVM- Support Vector Machine, Random Forest (RF) and a few others. Random forest (RF) is a type of supervised ML classification algorithm, that is *constructed from decision tree algorithms,* it is commonly used because of its simplicity and flexibility for most users. [34] Support Vector Machine (SVM) is also a supervised ML algorithm that evaluates data for classification and regression analysis, this is achieved by taking data and categorizing it into one or two classes.[35] Artificial Neural Network (ANN) is an information processing method that contains a large number of connected processing units that are intended to process information like neurons in the human brain. [36]

4.1.1. Data Sources and Collection

Two distinct types of datasets were observed in the data sources used by the extracted studies: structured and unstructured. Structured data adheres to a pre-defined data model and is therefore straightforward to analyze (e.g., excel files, demographic data, spatial data, physiological data).[37] The two most used structured data in the review were epidemiology and meteorological. *Epidemiology is the method used to find the causes of health outcomes and diseases in populations.*[38] Epidemiology data observed in the review were weekly number of disease incidence and prevalence and meteorological data observed were humidity and temperature measurement.

Unstructured data are information that either does not have a predefined data model or is not organized in a pre-defined manner. Examples are video, audio, images, emails, and text files.[37] In this review, three types of unstructured data were observe: social media data (tweets from twitter) and search keywords (from google and other search engines); and patient clinical notes. Social media data (e.g., users post, content, and reviews gathered from social media platforms) extracted could include posted text, date, location, and time of the post. Social media data can be used to increase awareness of trends and issues.[39] Social media data used in PHS can be used to gather Information for syndromic surveillance. If social media data were obtained from a specific community and there were trends of users posting about similar signs and symptoms of a disease such as diarrhea or vomiting, public health offices can be alerted of the trends and investigate the causes in the community. Keyword research is another way for researchers to know what people search for, especially when it comes to health-related issues or infectious diseases. Keyword searches for coronavirus worldwide first peaked Jan 31, 2020, then again Mar 12[t], 2020 after the WHO declared the coronavirus outbreak a pandemic Mar 11, 2020.[40] Keyword searches can be used to predict Infectious outbreak worldwide. This type of dataset can monitor disease control of other health issues such as HIV and cancer.[40]

4.1.2. Use of AI to Predict and Forecast Infectious Disease Events

There ML models forecast events; forecasting an infectious disease outbreak is especially important because it assists healthcare facilities, local, and state public health leadership prepare resources needed. In the initial stages of the COVID-19 pandemic, healthcare facilities had a shortage of masks and PPE equipment needed to handle the surge of patients. With this dilemma, prediction and forecasting ML tools became popularized, as healthcare facilities and public health officials tried to predict and forecast new COVID-19 cases, determine how resources (e.g.,N-95 masks, gloves, and PPE equipment) should be distributed to manage patient surges.[41]

In this review, ML was used as sentinel surveillance in dialysis clinic networks in Europe. The ML model was developed as a risk assessment tool to forecast the possibility of COVID-19 outbreaks two weeks beforehand at the dialysis clinics. Having a risk assessment tool such as this is essential especially for the fact that most dialysis patients would be immunocompromised with a higher risk of severe complications if they were infected with SARS-CoV 2 virus.[27]

Apart from its use in healthcare centers, ML prediction models are used to limit transmission of ID such as Dengue fever. Researcher combined dengue epidemiological, and meteorological data to develop an ML model that could predict and track the epidemic trajectory of dengue in Guangdong, China.[25] The advantage of this type of ML model is that it can be used to predict and forecast seasonal infectious diseases, especially in countries that have seasonal infectious diseases.

4.1.3. The Utilization of AI in Infectious Disease Detection

The use of ML or DL for detecting and diagnosing an infection disease is a relatively new development. In 2014, only one AI model had been approved by the FDA for clinical use, but today more than 75 AI models have been approved for clinical use in radiology, cardiology, and pathology.[42] The emergency of the COVID-19 pandemic made information such as patient lung imaging available.

Researchers around the world shared information to analyze and research how to diagnosis, treat, and prevent infection of SARS-CoV 2.[43] In the initial stages, one of the hallmarks of COVID-19 was the *non-specific ground-glass opacities consolidation* on Chest x-ray or CT scan.[29] As efforts to quickly diagnosis COVID-19 in patients increased, researchers developed DL models called Convolutional Neural Network (CNN) to accurately detect/ diagnosis the infection. CNN is mostly used to analyze visual imagery; it uses a sequence of layers that *transforms one volume of activations to another through a differentiable function* and trained on the several types of CT scan. The model can identify regions of interest (ROI) and eventually detect the infection based on lesions on the CT scan.[42]

With the digitalization of hospitals, patient clinical notes have now been updated from handwritten records to electronic health records (EHRs). EHRs include structured data such as gender, race, age, and labs. EHRs also contains unstructured data such as free text clinical notes. Retrieving information from free text clinical notes is usually done manually, which can be challenging and time consuming.[44] With the increase in international travel and the emergence of infectious disease such as Zika, Ebola, and COVID-19, tracking patient travel history was one of the essential information obtained from patients, as this information helped generate a chronological timeline so as to assess if the patient fits the criteria for the infectious disease diagnosis.[45] To make the process of retrieving essential information, especially travel history, from clinical notes easier researchers have developed a ML and neural language model that automatically extracts travel history from clinical notes.[31] Application of these models would greatly improve infectious disease detection.

4.2. Limitations

This review had limitations that must be considered. First, only articles published in English, peer reviewed, and were published between 2000-2022 were imported into Covidence™. Second, although AI can be applied to various aspects of PHS, only articles that were related to infectious diseases were considered. Third, because the objective of the review was to show the role of AI and ML in PHS data sources and collection, case detection, prediction, analyses, and forecasting, quality of the articles were not assessed.

4.3. Recommendations

The future of AI in PHS is broad and rapidly growing as public and private organizations are developing innovative technology and methods to modernize PHS. These include the discovery of new data sources such as data from wearable personal devices and internet of things (IoT) using AI to predict the number of cases of an infectious disease a hospital can expect and extracting pertinent information from a clinical note so that healthcare workers and public health officials can rapidly get the much-needed information to developing early warning alert systems. Currently, AI in PHS has programs such as Epidemic Intelligence from Open Sources (EIOS) that pulls data and articles in real-time on public health events. EIOS is an open-source tool by the World Health Organization (WHO), that creates *a unified all-hazards, One Health approach to early detection, verification, assessment, and communication of public health threats using publicly available information.*[46] EIOS is a platform where the global public health community can share information and work together towards a common goal. With this technology, the future of AI in PHS will include real-time data sharing of patient records between hospitals, healthcare facilities and public health agencies.

References

1. Bell JA, Nuzzo. JB. *Global Health Security Index: Advancing Collective Action and Accountability Amid Global Crisis, 2021*. Accessed April 12, 2022. https://www.ghsindex.org/country/united-states/

2. CDC. National Notifiable Diseases. *Waterborne Dis Outbreak Report*. Published online 2019. https://www.cdc.gov/healthywater/surveillance/nndss.html

3. Services O of PHS. *Public Health Surveillance: Preparing for the Future.* Centers for Disease Control and Prevention (CDC); 2018. Accessed February 1, 2022. https://www.cdc.gov/surveillance/improving-surveillance/Public-health-surveillance.html

4. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. 2020;395(10236):1579-1586. doi:10.1016/S0140-6736(20)30226-9

5. McNabb S, Conde JM, Ferland L et al. Transforming Public Health Surveillance. Proactive Measures for Prevention, Detection, and Response. In: Elsevier Health Sciences; :4. https://www.elsevier.com/books/transforming-public-health-surveillance/jn-mcnabb/978-0-7020-6337-4

6. Rivier University. Public Health Fields and Specialties. Rivier University. Published 2022. Accessed February 5, 2022. https://www.rivier.edu/academics/online/resources/program-resources/public-health-fields-and-specialties/

7. Institute of Medicine (US) Forum on Emerging Infection, Davis JR, Lederberg J. Public Health Systems and Emerging Infections: Assessing the Capabilities of the Public and Private Sectors: Workshop Summary. In: National Academies Press (US); 2000. Accessed February 1, 2022. https://www.ncbi.nlm.nih.gov/books/NBK100249/

8. Center for Disease Control and prevention (CDC). Introduction to Public Health 101 Series. Department of Health and Human Services, CDC. Published 2014. https://www.cdc.gov/training/publichealth101/public-health.html

9. Nsubuga P, White ME, Thacker SB, et al. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. In: Jamison DT, Breman JG, Measham AR, et al., eds. *Disease Control Priorities in Developing Countries*. The International Bank for Reconstruction and Development / The World Bank; 2006:997-1015. Accessed February 1, 2022. https://www.ncbi.nlm.nih.gov/books/NBK11770/

10. Institute of Medicine (US) Committee on a National Surveillance System Cardiovascular and select Chronic Diseases. Existing Surveillance Data Sources and Systems. In: *A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Disease*. National Academies Press (US); 2011. Accessed February 1, 2022. https://www.ncbi.nlm.nih.gov/books/NBK83157/

11. London School of Hygiene and Tropical Medicine. Types of surveillance. London School of Hygiene and Tropical Medicine. Published 2009. Accessed April 18, 2022. http://conflict.lshtm.ac.uk/page_75.htm

12. Mccarthy J. What is Artificial Intelligence? Published online 2004. Accessed January 27, 2022. http://www-formal.stanford.edu/jmc/

13. Tyagi N. 6 Major Branches of Artificial Intelligence (AI). AnalyticSteps. Published 2020. Accessed January 27, 2022. https://www.analyticssteps.com/blogs/6-major-branches-artificial-intelligence-ai

14. Delua J. Supervised vs. Unsupervised Learning: What's the Difference? IBM. Published 2021. Accessed April 12, 2022. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

15. Burns E, Laskowski N, Tucci L. What is Artificial Intelligence (AI)? TechTarget. Published 2022. Accessed February 5, 2022. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence

16. Center for Disease Control and prevention (CDC). Artificial Intelligence, Public Trust, and Public Health. Center for Disease Control and Prevention (CDC). Published 2021. Accessed February 5, 2022. https://blogs.cdc.gov/genomics/2020/09/17/artificial-intelligence/

17. Naseem M, Akhund R, Arshad H, Ibrahim MT. Exploring the Potential of Artificial Intelligence and Machine Learning to Combat COVID-19 and Existing Opportunities for LMIC: A Scoping Review. *J Prim Care Community Heal*. 2020;11. doi:10.1177/2150132720963634

18. Kaur M. Top 10 real-life examples of Machine Learning. Big Data Made Simple. Published 2019. Accessed March 24, 2022. https://bigdata-madesimple.com/top-10-real-life-examples-of-machine-learning/

19. Mackey T, Purushothaman V, Li J, et al. Machine Learning to Detect Self-Reporting of

Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. *JMIR Public Heal Surveill*. 2020;6(2). doi:10.2196/19509

20. Chae S, Kwon S, Lee D. Predicting infectious disease using deep learning and big data. *Int J Environ Res Public Health*. 2018;15(8). doi:10.3390/ijerph15081596

21. Poirier C, Hswen Y, Bouzille G, et al. Influenza forecasting for French regions combining EHR, web and climatic data sources with a machine learning ensemble approach. *PLoS One*. 2021;16(5). doi:10.1371/JOURNAL.PONE.0250890

22. Kim M, Chae K, Lee S, Jang HJ, Kim S. Automated classification of online sources for infectious disease occurrences using machine-learning-based natural language processing approaches. *Int J Environ Res Public Health*. 2020;17(24):1-13. doi:10.3390/ijerph17249467

23. Edo-Osagie O, Smith G, Lake I, Edeghere O, De La Iglesia B. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PLoS One*. 2019;14(7). doi:10.1371/journal.pone.0210689

24. Ajayi T, Dara R, Poljak Z. Forecasting herd-level porcine epidemic diarrhea (PED) frequency trends in Ontario (Canada). *Prev Vet Med*. 2019;164:15. doi:10.1016/J.PREVETMED.2019.01.005

25. Guo P, Liu T, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. *PLoS Negl Trop Dis*. 2017;11(10). https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005973

26. Necesito I V., Velasco JM, Kwak J, et al. Combination of univariate long-short term memory network and wavelet transform for predicting dengue case density in the national capital region, the philippines. *Southeast Asian J Trop Med Public Health*. 2021;52(4):479-494.

27. Bellocchio F, Carioni P, Lonati C, et al. Enhanced sentinel surveillance system for covid-19 outbreak prediction in a large european dialysis clinics network. *Int J Environ Res Public Health*. 2021;18(18). doi:10.3390/ijerph18189739

28. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis*. 2011;5(8). doi:10.1371/journal.pntd.0001258

29. Jingxin L, Mengchao Z, Yuchen L, et al. COVID-19 lesion detection and segmentation- A

deep learning method. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8256684/

30. Amin S, Uddin MI, Alsaeed DH, Khan A, Adnan M. Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches. *Complexity*. 2021;2021. doi:10.1155/2021/5520366

31. Peterson KS, Lewis J, Patterson O V., et al. Automated Travel History Extraction From Clinical Notes for Informing the Detection of Emergent Infectious Disease Events: Algorithm Development and Validation. *JMIR public Heal Surveill*. 2021;7(3). doi:10.2196/26719

32. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Rich Tsui F. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *J Biomed Inform*. 2015;58:60-69. doi:10.1016/j.jbi.2015.08.019

33. Parikh N, Daughton AR, Rosenberger WE, et al. Improving Detection of Disease Re-emergence Using a Web-Based Tool (RED Alert): Design and Case Analysis Study. *JMIR public Heal Surveill*. 2021;7(1). doi:10.2196/24132

34. Mbaabu O. Introduction to Random Forest in Machine Learning. Section. Published 2020. Accessed April 12, 2022. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

35. Pisner DA, Schnyer DM. Support vector machine. *Mach Learn Methods Appl to Brain Disord*. Published online January 1, 2020:101-121. doi:10.1016/B978-0-12-815739-8.00006-7

36. Park YS, Lek S. Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. *Dev Environ Model*. 2016;28:123-140. doi:10.1016/B978-0-444-63623-2.00007-4

37. Enterprise Big Data Framework. Data Types: Structured vs. Unstructured Data. Enterprise Big Data Framework. Published 2019. Accessed April 10, 2022. https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/

38. Center for Disease Control and prevention (CDC). What is Epidemiology? Center for Disease Control and Prevention (CDC). Published 2016. Accessed April 10, 2022. https://www.cdc.gov/careerpaths/k12teacherroadmap/epidemiology.html

39. Stieglitz S, Mirbabaie M, Ross B, Neuberger C. Social media analytics – Challenges in topic discovery, data collection, and data preparation. *Int J Inf Manage*. 2018;39:156-168.

doi:10.1016/J.IJINFOMGT.2017.12.002

40. Effenberger M, Kronbichler A, Shin J Il, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with Internet Search Volumes: A Google TrendsTM Analysis. *Int J Infect Dis*. 2020;95:192. doi:10.1016/J.IJID.2020.04.033

41. Painuli D, Mishra D, Bhardwaj S, Mayank A. Forecast and prediction of COVID-19 using machine learning. *Data Sci COVID-19*. 2021;(January):381-397. doi:10.1016/B978-0-12-824536-1.00027-7

42. Khemasuwan D, Sorensen JS, Colt HG. Artificial intelligence in pulmonary medicine: computer vision, predictive model and COVID-19. *Eur Respir Rev*. 2020;29(157):1-16. doi:10.1183/16000617.0181-2020

43. Khemasuwan D, Colt HG. Applications and challenges of AI-based algorithms in the COVID-19 pandemic. *BMJ Innov*. 2021;7(2):387-398. doi:10.1136/BMJINNOV-2020-000648

44. Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns*. 2021;2(7). doi:10.1016/J.PATTER.2021.100289

45. Duong TN, Waldman SE. Importance of a Travel History in Evaluation of Respiratory Infections. *Curr Emerg Hosp Med Rep*. 2016;4(3):141-152. doi:10.1007/S40138-016-0109-Y

46. World Health Organization (WHO). Epidemic Intelligence from Open Sources (EIOS): Zero Impact from Health Threats. World Health Organization (WHO). Published 2021. Accessed April 12, 2022. https://www.who.int/initiatives/eios