A systematic literature review of the public health informatics predictive models that utilize data from EHR systems as a data source

Jim Clavin[1], Miles Norton[1], Anup Haridas[1], Linda Le[1], Jack Shan[1], and Raad Mustafa[1]

[1]University of Maryland, Baltimore County

May 10, 2018

Abstract

Background: A systematic literature review was executed to identify data sources used in place of, side by side with, or in conjunction with, electronic health record (EHR) data in predictive models for influenza like illness (ILI) outbreaks.

Objectives: To determine how predictive models for ILI outbreak use EHR data and how often EHR data is used in ILI surveillance and forecasts.

Methods: Articles were sourced from Pubmed and the Journal of Medical Internet Research (JMIR). Results from these online databases were filtered down to a corpus of 48 studies. From these studies, 10 dummy and 10 categorical variables were identified and placed into a Google sheet; data visualizations were built from the Google sheet using Tableau public; and descriptive analytics reviewed.

Results: From the articles, eighty-four data sources were identified, of which 14 (or 17%) were data from EHRs. EHR data was utilized in 5% of those studies that also leveraged either governmental or syndromic surveillance data. Likewise, EHR data was used in 5% of studies that incorporated Google search and trend data. Most studies' models used autoregression (15%), with machine learning algorithms referenced second most often (13%). The utilization of EHR data was found only in the United States (9 studies) and Europe (4 studies).

Conclusion: EHR data used in tandem with other data sets in an ensemble approach, or in isolation, can be used by predictive models to signal alert levels earlier than existing government-provided models in those regions where such data is available but its adoption remains limited.

Corresponding author: James Clavin, University of Maryland Baltimore County, 1000 Hilltop Circle, Sondheim Hall 306-C, Baltimore, MD 21250, jclavin@umbc.edu

Keywords: EHR, flu, data, predictive modeling, machine learning, syndromic surveillance

Introduction

The incorporation of EHR data into syndromic surveillance alert systems has the potential to significantly improve the prediction models used to alert the public of disease outbreak, but faces technical, regulatory, data standard, and quality control challenges. [1].

Data Sources for Modeling Flu

Many government health agencies, oftentimes a country's Center for Disease Control, collect regional ILI data for reporting and forecasting [2]; [3]; [4]; [5]. The United States CDC categorizes surveillance data into five bins: virologic surveillance, outpatient illness surveillance, mortality surveillance, hospitalization surveillance, and summarized geographic state reports on the spread of flu and provides the data to the public via web application [2]; [6].

Relatively newer data sources have become available due to the growth of the Internet, including self-reporting (volunteer) websites, Google Search and Twitter data; models built using such data have been found to detect trends earlier than governmental methods alone [7]; [8]; [9]. Google failed at building a flu model that could reliably predict outbreaks based upon search alone [10].

EHR data has also been used as data input into predictive models and used in isolation or in combination with historical data, has been found to alert as early as Google and Twitter based models and be as reliable as CDC based models [11]; [12].

This Study

From the 48 studies included, 55 national data sets from 6 continents with 24 distinct countries were identified ; 15 from the United States and 8 from China ranking first and second respectively.

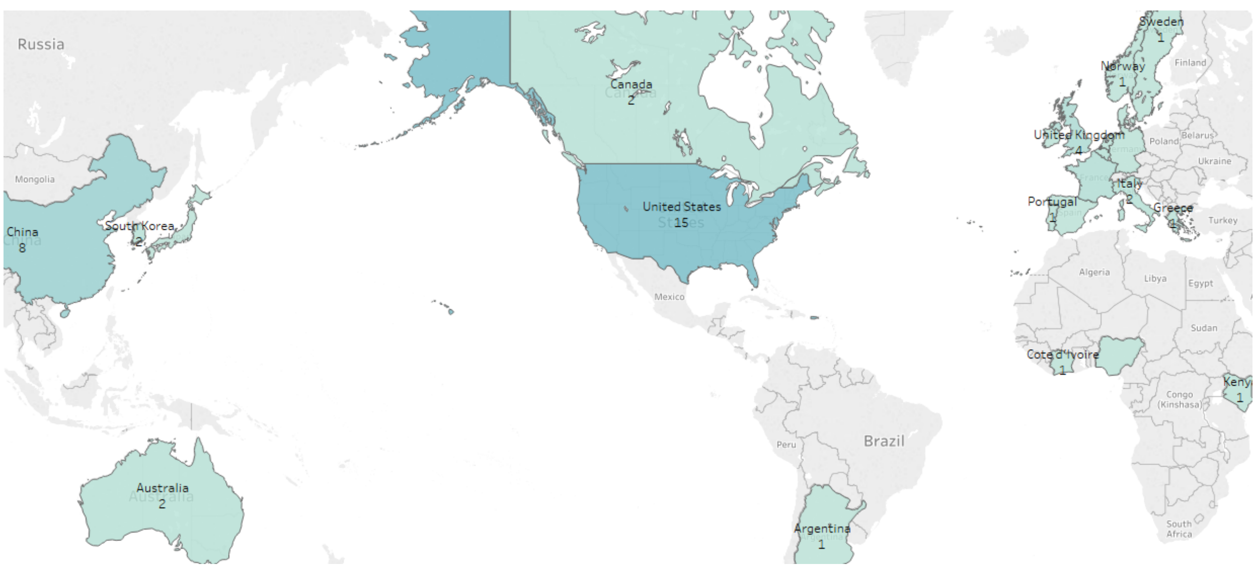


Figure 1: Countries and Number of Articles

Methods

A systematic process was developed and applied for this study, and included the following components: search methodology, variable selection, data presentation, and analysis.

Search Methodology

The reports researched were from the Journal of Medical Internet Research (JMIR) and Pubmed. The JMIR website offered peer reviewed and open access medical journals that available for reference on any publication related to the health domain [13]. Pubmed was used as an online searchable database of citations from biomedical literature [14] .

When combing through search results to identify articles to include, the following five step method was executed:

1. From the search results, quickly read a resource that seems promising, then rule it in or out.
2. If it was to be included, the article was assigned a number, and was summarized. If not included, then return to step 1.
3. Documented the keywords from the summary.
4. Reviewed article footnotes for additional journal articles to review. If so, proceeded to step 2.
5. Go to step 1.

48 studies were identified for inclusion using these process.

Variable Selection

From the summaries, we took the keywords and used them to create a possible set of dummy and categorical variables that would enable aggregation of attributes to identify patterns and compare results across articles. Twenty variables were identified: 10 dummy variables and 10 categorical variables.

Dummy Variables

A value of 1 was used to indicate the study included that data set, and a value of 0 to indicate the data set was not mentioned. Dummy variables enabled us to create subgroups for comparison, and is a method used in regression analysis [15].

- EHR Data: EHR data was used in a model
- Pharmacy Data: prescriptions for specific drugs used for ILI symptoms or influenza
- General Practitioner Data: data from providers not located in hospitals or emergency departments.
- Governmental Data: data sourced from regional or national databases. E.g, Centers for Disease Control (CDC) data.
- Surveillance Sites Data: providers that monitor ILI levels and often report
- Google: search data, or Google Flu Trends
- Facebook: a stand in for social media that is non-Twitter
- Twitter: twitter data
- Self reported: volunteer data, or online we application data for users who self select, or mobile apps
- Meteorological Data: weather data

Categorical Variables

Where possible, the number of possible values to include in the following categories was limited; some of these categories held NULL (or no value) if the study did not provide the data point.

- Model: modeling technique used
- Country: one ore more countries per article could happen
- Continent: the continent in which the country is located.
- Data Source: the government entity or provider name. E.g., "CDC"
- Flu: the flu strain studied. E.g., H1N1
- Regions: geospatial areas that do not necessarily align with political borders.
- Healthcare System Type: Universal or public-private mix.
- Date Range: Begin and end years of when the flu outbreak was studied
- Study Year of Publication: the year of publication
- Coding Scheme: any reference to a coding standard. E.g., ICD-9.

Google Sheet

A google sheet was created with the variables identified, plus two columns for this article:

- Citation: for this paper, the citation to include in the references
- Study #: from the search methodology; an arbitrary number
- EHR
- Pharm
- Gen Pract
- Govtl
- Surveil
- Goog
- Fbook
- Twiiter
- Self
- Meteor
- Country
- Continent
- Region
- Healthcare System
- Model
- Data Source
- Flu
- Date Range
- Study Year of Pub.
- Coding Scheme

Results

Google Sheet

Citation	Study #	EHR	Pharm	Gen Pract	Govtl	Surveil	Goog	Fbook	Twitter	Self	Meteor
[16]	40	0	0	0	1	0	0	0	0	0	0
[17]	37	0	0	0	1	0	0	0	0	0	1
[18]	27	0	0	0	0	0	0	0	0	0	1
[19]	26	0	0	0	0	1	0	0	0	0	0
[20]	32	0	0	1	1	1	1	0	0	0	1
[21]	38	0	0	0	0	1	0	0	0	0	1
[22]	28	0	0	0	0	1	0	0	0	0	0
[23]	29	0	0	0	1	0	0	0	0	0	1
[24]	7	1	0	1	0	1	0	0	0	0	0
[25]	25	0	0	1	1	1	0	0	0	0	0
[26]	10	0	1	0	0	0	0	0	0	0	0
[27]	11	0	0	0	0	0	0	0	0	1	0
[28]	12	0	0	0	0	0	0	0	0	1	0
[11]	13	1	0	1	1	1	0	0	0	0	0
[29]	14	0	0	1	1	1	1	0	1	1	0
[30]	15	1	0	0	1	0	1	0	0	0	0
[31]	16	1	0	0	0	0	0	0	0	0	0
[32]	17	0	0	0	0	0	0	0	0	1	0
[33]	18	0	0	0	0	0	0	0	1	0	0
[34]	33	0	0	0	1	1	0	0	1	0	0
[35]	20	1	0	0	0	1	1	0	0	0	0
[36]	22	1	0	0	0	0	0	0	0	0	0
[37]	23	1	0	0	0	0	0	0	0	0	0
[38]	30	0	0	0	1	0	0	0	0	0	0
[39]	34	0	0	0	0	0	0	0	0	0	0
[40]	35	0	0	0	0	1	0	0	0	0	0
[41]	36	0	0	0	1	0	0	0	0	0	1
[42]	39	0	0	0	0	0	0	0	0	0	0
[43]	44	0	0	0	0	1	0	0	0	1	0
[44]	43	0	0	0	0	1	0	0	0	0	0
[45]	45	0	0	0	0	0	0	0	0	0	0
[46]	1	1	0	0	0	0	0	0	0	0	0
[47]	19	1	0	0	1	1	0	0	0	0	0
[48]	2	1	0	1	0	0	0	0	0	0	0
[49]	3	1	0	0	1	0	1	0	0	0	0
[50]	5	1	0	0	0	0	0	0	0	0	0
[51]	31										
[52]	41										
[53]	6	1	0	0	0	0	0	0	0	0	0
[54]	9	1	0	0	0	0	1	0	1	1	0
[55]	8	0	0	0	0	0	0	0	0	1	0
[56]	21	0	0	0	1	1	1	1	1	0	0
[57]	42	0	0	1	1	1	0	0	0	0	0
[58]	24	0	0	0	1	1	1	0	0	0	0

Study #	Country	Continent	Regions	Healthcare System Typ
40	China	Asia	Provinces	Public-private mix

	United States	North America	States	Public-private mix
37	China	Asia	Provinces	Public-private mix
27	Cote d'Ivoire	Africa	Districts	Public-private mix
26	Kenya	Africa	Provinces	Public-private mix
32	China	Asia	Provinces	Public-private mix
38	China	Asia	Provinces	Public-private mix
28	Nigeria	Africa	States	
29	Eygpt	Africa		
7	France	Europe	Departments	Universal
25	United Kingdom	Europe	Countries	Universal
10	Japan	Asia	Prefectures	Universal
11	South Korea	Asia	Provinces	Universal
12	Netherlands	Europe	Countries	Public-private mix
	Belgium			Public-private mix
	Portugal			Universal
	Italy			Universal
	United Kingdom			Universal
	Sweden			Public-private mix
	France			Universal
	Spain			Universal
	Ireland			Universal
	Denmark			Universal
13	Belgium	Europe	Provinces	Public-private mix
14	United States	North America	States	Public-private mix
15	United States	North America	HHS Regions	Public-private mix
	United States	North America		Public-private mix
16	United States	North America	New York City	Public-private mix
17	Norway	Europe		
18	United States	North America		Public-private mix
33	United States	North America	11 U.S. Cities	Public-private mix
20	France	Europe	Brittany	Universal
22	United States	North America	Pennsylvania	Public-private mix
23	United States	North America		
30	Argentina	South America	Provinces	
34	China	Asia		Public-private mix
35	China	Asia	Provinces	Public-private mix
36	China	Asia	Provinces	Public-private mix
39	China	Asia	Beijing	Public-private mix
44	Canada	North America	Hutterite colonies in Canadian provinces	Universal
43	Germany	Europe	140 regions of Southern Germany	Public-private mix
45	Canada	North America		Universal
1				
19	United States	North America		Public-private mix
	Australia	Australia		Universal
	United Kingdom	Europe		Universal
2	United Kingdom	Europe	England	Universal
			Wales	

			Scotland	
3	United States	North America		
Three part study				
5	United States	North America	New York City	Public-private mix
31				
41				
6	United States	North America	Massachusetts	
9	United States	North America	Boston	
8	United States	North America	Multiple	
21	South Korea	Asia	Provinces	Universal
42	Australia	Australia	Melbourne	Universal
24	Italy			
	Greece			

Study #	Model
40	D-R algo built by study authors, using the Grey prediction model
37	Zero truncated Poisson regression model
27	Auto regressive
26	Sensitivity and Specificity
32	GLM
	LASSO
	Deep Learning with Neural Network
	Statistical model fusion with Bayesian model averaging
	Auto regressive
38	RS-SVM
28	None
29	Multi-variate risk model
7	periodic regression
	robust periodic regression
	Markov model
25	Moving Epidemic Model
	Percentile approach
10	multiple regression model
11	Ordinary Least Squares Linear Regression
12	Multi-variate
13	Auto regressive
14	Machine learning
	Stacked linear regression
	Support Vector Machine regression
	AdaBoost regression with decision trees
15	Machine learning
	Auto regressive
16	None
17	None
18	Machine Learning
	Classifier
	Support Vector Machine (SVM)
33	Machine Learning
	Classifier
20	Serfling regression model

22	Regression models
23	Machine Learning
	Unsupervised
30	Monte Carlo
	Regression models
34	Positive Predictive Value
	Negative Predictive Value
35	Positive Predictive Value
	Negative Predictive Value
36	Logistic Regression
	Spatial Autocorrelation Analysis
	Temporal Cluster Analysis
39	Multivariate logistic regression analysis
44	Positive Predictive Value
	Negative Predictive Value
	Pearson chi-square
	Uni-variate logistic regression
43	Akaike's information criterion (AIC)
	Bayesian information criterion (BIC)
	One-step-ahead
	Ranked probability score (RPS)
45	Logistic regression
1	Logistic regression
19	
2	Logistic regression
	Positive Predictive Value
	Negative Predictive Value
3	Auto regressive
Three part study	None
5	
31	
41	
6	None
9	Auto regressive
8	Simulation
	Ensemble
	Pearson Correlation
	Machine Learning
21	LASSO
	SVR
42	
24	Auto regressive

Study #	Data Source
40	China Weather Network
	Local Health Bureau Website and Database
	WHO
	CDC
37	Ministry of Heatlh
27	Database of the influenza surveillance network of the National Institute of Public Hygiene (INHP)

	Airport Operating Development Aviation and Meteorology Company
26	Access or SQL Databases at the CDC offices in Nairobi
32	Google search data
	meteorological data
38	China CDC
	National Meterological Information Center
28	
29	WHO
	U.N. data
7	Computerized medical records from participating surveillance hospitals
	ED
	General Practitioner data
25	General Practitioner data
10	Prescription drug data
11	Mobile App called Fever Coach
12	Websites that volunteers go to
13	General Practictioner Data
	Surveillance Data from the national Influeza Center (WIV-ISP data)
14	CDC
	athenahealth
	Google Trends
	Twitter
	FluNearYou
15	athenahealth
	CDC
16	NY Presbyterian Hospital EHR
17	Blood glucose monitor
18	Twitter
33	Twitter
	County and city websites
20	
22	Lancaster General health system in Lancaster County, PA
23	IMS Health
30	National Institute of Microbiology
	Argentina Health Ministry
	National Institute of Geography
34	Xiamen International Airport
35	Gansu Province
36	WHO
	State Forestry Administration
	National Fundament GIS
	NDVI data from Geospatial Data Cloud
	Data Sharing Infrastructure of Earth System Science
39	Clinical data: Peking University Health Science Center
44	
43	https://survstat.rki.de/
45	Six acute care hospitals in Ontario
1	EPIC rheumatology patients from a large pediatric hospital
19	US Indian Health Service
	Australian National Influenza Surveillance Systems
	UK Heatlh Protection Agency

2	30 General Practitioners
3	athenahealth
Three part study	
5	
31	
41	
6	
9	
8	WISDM
	Influenzanet
	FluNearYou
21	Korean CDC
42	
24	

Study #	Flu	Date Range	Study Year of Pub.	Coding Scheme
40	H1N1	2009	2014	
	H1N1			
37	H7N9 (avian)	2013 - 2014	2010	
27	H1N1	2007 - 2012	2016	
26		2013 - 2014	2017	
32				
38		2005 - 2009	2012	
28	H5N1 (avian)	2006 - 2009	2014	
29	H5N1 (avian)			
7	H3N2	2010 - 2016	2017	
25			2015	
10			2012	
11			2017	
12			2017	
13		2003 - 2015	2017	ICPC-2
14		2009 - 2015		
15		2014 - 2015	2016	
16	H1N1		2010	ICD-9
				Internal hospital codes
17			2005	
18		2012 - 2013	2017	
33		2013 - 2014	2014	
20		2010 - 2015	2018	ICD-10
				Unstructured data
22	H1N1	2009	2010	ICD-9
23	H1N1	2009 - 2010	2015	ICD-9
30	H1N1	2009		
34	H3N2	2015 - 2016	2018	
35		2014 - 2015	2016	
36	H7N9 (avian)	2013 - 2014	2015	
39	H1N1	2009 - 2010	2012	

	H3N2			
44		2008 - 2009	2011	
43		2001 - 2008	2012	
45	H3N2			
	H1N1			
1		2007 - 2009		
19	H1N1	2009	2010	
2		2013 - 2015	2015	
3			2017	
Three part study				
5				CVX
31				CPT
41				ICD-9
6			2016	
9			2018	
8			2017	
21		2011 - 2014	2016	
42			2017	
24		2011 - 2012	2017	

Using a free data visualization building application, called “Tableau Public,” we created graphs and tables using the Google sheet as the data source [59].

Analysis

Geospatial

6 continents were represented in the studies, with North America accounting for 40% of the data sources. The health care systems of the 26 distinct countries were assigned a value of either “Public-private mix” or “Universal.” Public-private mix was more common, occurring 57% of the time.

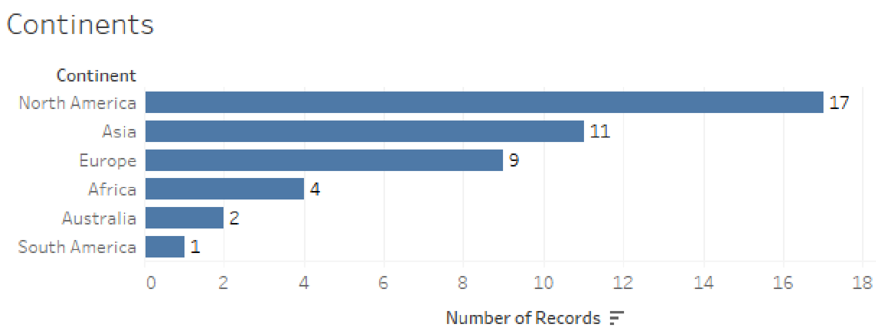
The H1N1 flu strain was studied 11 times, with all others tallying 8 references. The 2013-14 flu season was most studied; most articles were published in 2017 (29%).

EHR Data Source Utilization

In four studies EHR data was used in combination with, or in comparison to, surveillance data [24]; [11]; [35]; [60]. When used in combination, some studies referred to this as an “ensemble” method[29]; [12].

Modeling Techniques

Based on our summary from the 48 papers, we classified the modeling techniques adopted in the papers into three main categories: Machine Learning (ML)Models, Statistical Models (except ML models) and Other models.



Countries

Argentina	1
Australia	2
Belgium	2
Canada	2
China	8
Cote d'Ivoire	1
Denmark	1
Eygpt	1
France	3
Germany	1
Greece	1
Ireland	1
Italy	2
Japan	1
Kenya	1
Netherlands	1
Nigeria	1
Norway	1
Portugal	1
South Korea	2
Spain	1
Sweden	1
United Kingdom	4
United States	15
Grand Total	55

Healthcare System Types

Public-private mix	25
Universal	19

Flu Strains

H1N1	11
H3N2	4
H5N1 (avian)	2
H7N9 (avian)	2

Figure 2: Categorical Variable Counts

Machine Learning Models

Machine Learning Methods					
Regression Model		Others			
Logisitic Regression	Linear regression	GLM	RS-SVM	decision tress	Classiers
5	10	1	1	1	8

Table 6: Summary of Machine Learning Models

Machine learning (ML) included supervised, unsupervised and self-supervised algorithms, which could help

EHR/Google

Google	EHR Data 0	EHR Data 1
0	24	10
1	4	4

EHR/Twitter

Twitter	EHR Data 0	EHR Data 1
0	24	13
1	4	1

EHR/Surveillance

Surveillance Sites Data	EHR Data 0	EHR Data 1
0	14	10
1	14	4

EHR/Government

Governmental Data	EHR Data 0	EHR Data 1
0	15	10
1	13	4

Figure 3: EHR data compared with Google, Surveillance Data, Twitter data, and Government-sourced data

construct prediction models based on past/previous data [61]. In all of the 48 papers, ML models were used 21 times. In all of the ML models adopted, the linear regression method was utilized most. This model included multiple regression model, ordinary least square linear regression, periodic regression, and robust periodic regression. In six studies, all from locations in the United States, ML was used in three ways. First, ML techniques were used three times as part of an ensemble data approach, in which two or more data sets related to flu surveillance were combined[12]; [29]; [55]. ML was used with twitter data twice [34]; [33] and with EHR data once [37]. All studies found that ML techniques performed as good as or better than existing surveillance techniques in detecting influenza outbreaks [62].

Statistical Modeling Techniques

Statistical Modeling Techniques are mathematical models that embody a set of statistical assumptions

Statistical Models (total =22)									
Models				Evaluation Metricx				Pearson	
Auto Regressive Model	Grey Model	Markov Model	spatial model	AIC	BIC	NPV	PPV	Pearson Chi-square	Pearson Corr
7	1	1	1	1	1	4	4	1	1

Table 7: Summary of Statistical Models

concerning the generation of some sample data and similar data from a larger population [63]. In the 48 studies, statistical models were adopted 22 times. Among the application of statistical models, 7 were auto regressive models, 10 were used for evaluating the flu outbreak predictive model accuracy and 2 were about adopting Pearson methods.

Other Models (total = 10)					
Algorithm			Deep learning		Other
Monte Carlo	One-step	simulation	Neural network	Ensemble	Self-defined
1	1	1	1	1	5

Table 8: Summary of other Models

Other Models

Aside from adopting models, such as machine learning and statistical models, some researchers applied some algorithms, such as Monte Carlo simulation, deep learning methods, and other self-defined models in their investigation of the flu outbreak probability. Based on our reading on the 48 studies, we summarized there were 10 other model types used or referenced. Of these , the self-defined methods, such as ranked probability score and percentile approach were used most.

Discussion

Based upon our research public health informatics flu prediction models on a global basis continue to primarily use syndromic surveillance data from sentinel providers (18 data sources) which is often data provided and maintained by the government (17 data sources). Google search trends have been used, as well as Twitter data, in some models around the globe (China, France, Italy, U.S. South Korea). EHR data was found to be used as in input to influenza predictive models only in the United States and Europe (France, England, Belgium). China continues to use primarily surveillance data, maintained at the provincial and national level; sometimes meteorological data was also use as input. Japan was an outlier, in the sense that it was the only nation in our study cohort that used pharmacy prescription data. Australia models used surveillance data. Countries in Africa were found to use surveillance data and meteorological data as inputs to predictive models.

Modeling techniques that used EHR data were most often Auto Regression (3 studies) and Machine Learning (2 studies). EHR-powered models in the studies most often used data provided by the cloud EHR service provide athenahealth (4 studies). [54]; [29]; [12]; [49]. ML techniques and statistical modeling techniques were evenly used (21 times and 22, respectively). ML techniques were capable of utilizing one or more data sets and building prediction models. This ensemble approach was found to have success in one study, and to have challenges with controlling for known non-demographic population variables that are specific to the strain and dispersion of the influenza strain present in the data[29]; [57].

Limitations

Modeling limitations noted were selection and sample bias, especially for those studies that used volunteer, self-reported data as in [28] and [28], respectively. Some models that use social data selection bias of regions with larger populations with access to the Internet may influence the predictive power of the models and the likelihood of EHRs being present.

Limitations of this study included the translation of article attributes into dummy variables, which required interpretation and synthesis and therefore left room for ambiguity and challenge. E.g., one study utilized data from an EHR which was populated into a claims level data warehouse [35]. In such instances, a determination was made of either 1 or 0, and a note added into the Google sheet (referenced above) to record the reasoning behind the decision.

Conclusions

The choice of a surveillance system affects the prediction model used to predict influenza epidemics [57]. EHR data was found to be used as in input to influenza predictive models only in the United States and Europe (France, England, Belgium). A cloud service provider, athenahealth, has provided de-identified EHR data for use in research studies and may hold promise for a future model in which "nowcast" models outperform in terms of robustness and timeliness those existing surveillance systems which are monitored by governmental entities.

Acknowledgments, Competing Interests, Funding and all other required statements.

Competing Interests: none

References

1 Shortliffe EH, Cimino JJ, editors. *Biomedical Informatics*. Springer London 2014. doi:10.1007/978-1-4471-4474-8

2 Overview of Influenza Surveillance in the United States — Seasonal Influenza (Flu) — CDC. https://www.cdc.gov/flu/weekly/overview.htm

3 Epidemic Prediction Initiative. https://predict.phiresearchlab.org

4 French GPs Sentinelles network ¿ France ¿ Home. https://websenti.u707.jussieu.fr/sentiweb/?lang=en

5 Sources of UK flu data: influenza surveillance in the UK. https://www.gov.uk/guidance/sources-of-uk-flu-data-influenza-surveillance-in-the-uk

6 FluView Interactive — Seasonal Influenza (Flu) — CDC. https://www.cdc.gov/flu/weekly/fluviewinteractive.htm

7 Google Flu Trends. https://www.google.org/flutrends/about/

8 Alessa A, Faezipour M. A review of influenza detection and prediction through social networking sites. *Theoretical Biology and Medical Modelling* 2018;**15**. doi:10.1186/s12976-017-0074-5

9 Influenzanet. https://www.influenzanet.eu/influenzanet/

10 What We Can Learn From the Epic Failure of Google Flu Trends. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

11 Michiels B, Nguyen VK, Coenen S, *et al.*. Influenza epidemic surveillance and prediction based on electronic health record data from an out-of-hours general practitioner cooperative: model development and validation on 2003–2015 data. *BMC Infectious Diseases* 2017;**17**. doi:10.1186/s12879-016-2175-x

12 Santillana M, Nguyen AT, Louie T, *et al.*. Cloud-based Electronic Health Records for Real-time Region-specific Influenza Surveillance. *Scientific Reports* 2016;**6**. doi:10.1038/srep25732

13 JMIR-Journal of Medical Internet Research. https://www.jmir.org

14 Home - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/

15 Social Research Methods - Knowledge Base - Dummy Variables. https://socialresearchmethods.net/kb/dummyvar.php

16 Ding F, Zarlenga DS, Qin C, *et al.*. A novel algorithm to define infection tendencies in H1N1 cases in Mainland China. *Infection Genetics and Evolution* 2011;**11**:222–6. doi:10.1016/j.meegid.2010.09.015

17 Zhang Y, Feng C, Ma C, *et al.*. The impact of temperature and humidity measures on influenza A (H7N9) outbreaks—evidence from China. *International Journal of Infectious Diseases* 2015;**30**:122–4. doi:10.1016/j.ijid.2014.11.010

18 N'gattia AK, Coulibaly D, Nzussouo NT, *et al.*. Effects of climatological parameters in modeling and forecasting seasonal influenza transmission in Abidjan Cote d'Ivoire. *BMC Public Health* 2016;**16**. doi:10.1186/s12889-016-3503-1

19 Ndegwa LK, Emukule G, Uyeki TM, *et al.*. Evaluation of the point-of-care Becton Dickinson Veritor™ Rapid influenza diagnostic test in Kenya 2013–2014. *BMC Infectious Diseases* 2017;**17**. doi:10.1186/s12879-016-2131-9

20 Xu Q, Gel YR, Ramirez LLR, *et al.*. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PLOS ONE* 2017;**12**:e0176690. doi:10.1371/journal.pone.0176690

21 Hu X-qin, Quirchmayr G, Winiwarter W, *et al.*. Influenza early warning model based on Yunqi theory. *Chinese Journal of Integrative Medicine* 2012;**18**:192–6. doi:10.1007/s11655-012-1003-4

22 Evaluating a surveillance system: live-bird market surveillance for highly pathogenic avian influenza, a case study. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/25328630

23 Rabinowitz PM, Galusha D, Vegso S, *et al.*. Comparison of Human and Animal Surveillance Data for H5N1 Influenza A in Egypt 2006–2011. *PLoS ONE* 2012;**7**:e43851. doi:10.1371/journal.pone.0043851

24 Pelat C, Bonmarin I, Ruello M, *et al.*. Improving regional influenza surveillance through a combination of automated outbreak detection methods: the 2015/16 season in France. *Eurosurveillance* 2017;**22**. doi:10.2807/1560-7917.es.2017.22.32.30593

25 GREEN HK, CHARLETT A, MORAN-GILAD J, *et al.*. Harmonizing influenza primary-care surveillance in the United Kingdom: piloting two methods to assess the timing and intensity of the seasonal epidemic across several general practice-based surveillance schemes. *Epidemiology and Infection* 2014;**143**:1–12. doi:10.1017/s0950268814001757

26 Sugawara T, Ohkusa Y, Ibuka Y, *et al.*. Real-time Prescription Surveillance and its Application to Monitoring Seasonal Influenza Activity in Japan. *Journal of Medical Internet Research* 2012;**14**:e14. doi:10.2196/jmir.1881

27 Kim M, Yune S, Han HW. Detecting Influenza Epidemics Using Self-reported Data Through Mobile App (FeverCoach). *Iproceedings* 2017;**3**:e56. doi:10.2196/iproc.8686

28 Koppeschaar CE, Colizza V, Guerrisi C, *et al.*. Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe. *JMIR Public Health and Surveillance* 2017;**3**:e66. doi:10.2196/publichealth.7429

29 Santillana M, Nguyen AT, Dredze M, *et al.*. Combining Search Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology* 2015;**11**:e1004513. doi:10.1371/journal.pcbi.1004513

30 Santillana M, Nguyen AT, Louie T, *et al.*. Cloud-based Electronic Health Records for Real-time Region-specific Influenza Surveillance. *Scientific Reports* 2016;**6**. doi:10.1038/srep25732

31 Khiabanian H, Holmes AB, Kelly BJ, *et al.*. Signs of the 2009 Influenza Pandemic in the New York-Presbyterian Hospital Electronic Health Records. *PLoS ONE* 2010;**5**:e12658. doi:10.1371/journal.pone.0012658

32 IOS Press Ebooks - Using Blood Glucose Data as an Indicator for Epidemic Disease Outbreaks. http://ebooks.iospress.nl/publication/10303

33 Kagashe I, Yan Z, Suheryani I. Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. *Journal of Medical Internet Research* 2017;**19**:e315. doi:10.2196/jmir.7393

34 Aslam AA, Tsou M-H, Spitzberg BH, *et al.*. The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance. *Journal of Medical Internet Research* 2014;**16**:e250. doi:10.2196/jmir.3532

35 Bouzillé G, Poirier C, Campillo-Gimenez B, *et al.*. Leveraging hospital big data to monitor flu epidemics. *Computer Methods and Programs in Biomedicine* 2018;**154**:153–60. doi:10.1016/j.cmpb.2017.11.012

36 Journal of Lancaster General Health - The Spread of the H1N1 Outbreak in the Lancaster Community. http://www.jlgh.org/Past-Issues/Volume-5---Issue-3/The-Spread-of-the-H1N1-Outbreak-in-the-Lancaster-C.aspx

37 Ramanathan A, Pullum LL, Hobson TC, *et al.*. Discovering Multi-Scale Co-Occurrence Patterns of Asthma and Influenza with Oak Ridge Bio-Surveillance Toolkit. *Frontiers in Public Health* 2015;**3**.

doi:10.3389/fpubh.2015.00182

38 Predictibilidad de la propagación espacial y temporal de la epidemia de influenza A H1N1 en la Argentina por el método de percolación. http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0325-75412011000300005&lng=en&nrm=iso&tlng=en

39 Chen J, Yang K, Zhang M, *et al.*. Rapid identification of imported influenza viruses at Xiamen International Airport via an active surveillance program. *Clinical Microbiology and Infection* 2018;**24**:289–94. doi:10.1016/j.cmi.2017.05.027

40 Probe into the Precautionary Measures of Influenza in Gansu Province. http://zhlxbxzz.yiigle.com/CN112338201603/868760.htm?locale=zh_CN

41 Dong W, Yang K, Xu Q-L, *et al.*. A Predictive Risk Model for A(H7N9) Human Infections Based on Spatial-Temporal Autocorrelation and Risk Factors: China 2013–2014. *International Journal of Environmental Research and Public Health* 2015;**12**:15204–21. doi:10.3390/ijerph121214981

42 Clinical Predictors for Diagnosing Pandemic (H1N1) 2009 and Seasonal Influenza (H3N2) in Fever Clinics in Beijing, China. http://www.besjournal.com/Articles/Archive/archive/No1/201202/t20120228_57844.html

43 A comparison of self-report and health care provider data to assess surveillance definitions of influenza-like illness in outpatients. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/22338332

44 Held L, Paul M. Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* 2012;**54**:824–43. doi:10.1002/bimj.201200037

45 Lam P-P, Coleman BL, Green K, *et al.*. Predictors of influenza among older adults in the emergency department. *BMC Infectious Diseases* 2016;**16**. doi:10.1186/s12879-016-1966-4

46 Patwardhan A, Kelleher K, Cunningham D, *et al.*. Improving the influenza vaccination rate in patients visiting pediatric rheumatology clinics using automatic best practice alert in electronic patient records. *Pediatric Rheumatology* 2012;**10**:A106. doi:10.1186/1546-0096-10-s1-a106

47 IOS Press Ebooks - Design and Assessment of a Common, Multi-National Public Health Informatics Infrastructure to Enable H1N1 Influenza Surveillance. http://ebooks.iospress.nl/publication/13476

48 Venkatesan S, Myles PR, McCann G, *et al.*. Development of processes allowing near real-time refinement and validation of triage tools during the early stage of an outbreak in readiness for surge: the FLU-CATs Study. *Health Technology Assessment* 2015;**19**:1–132. doi:10.3310/hta19890

49 Yang S, Santillana M, Brownstein JS, *et al.*. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infectious Diseases* 2017;**17**. doi:10.1186/s12879-017-2424-7

50 Newton-Dame R, McVeigh KH, Schreibstein L, *et al.*. Design of the New York City Macroscope: Innovations in Population Health Surveillance Using Electronic Health Records. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2016;**4**:26. doi:10.13063/2327-9214.1265

51 McVeigh KH, Newton-Dame R, Chan PY, *et al.*. Can Electronic Health Records Be Used for Population Health Surveillance? Validating Population Health Metrics Against Established Survey Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2016;**4**:27. doi:10.13063/2327-9214.1267

52 Thorpe LE, McVeigh KH, Perlman S, *et al.*. Monitoring Prevalence, Treatment, and Control Of Metabolic Conditions In New York City Adults Using 2013 Primary Care Electronic Health Records: A Surveillance Validation Study. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2016;**4**:28. doi:10.13063/2327-9214.1266

53 Cutrona SL, Sreedhara M, Goff SL, *et al.*. Improving Rates of Influenza Vaccination Through Electronic Health Record Portal Messages Interactive Voice Recognition Calls and Patient-Enabled Electronic

Health Record Updates: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols* 2016;**5**:e56. doi:10.2196/resprot.5478

54 Lu FS, Hou S, Baltrusaitis K, *et al.*. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. *JMIR Public Health and Surveillance* 2018;**4**:e4. doi:10.2196/publichealth.8950

55 Brownstein JS, Chu S, Marathe A, *et al.*. Combining Participatory Influenza Surveillance with Modeling and Forecasting: Three Alternative Approaches. *JMIR Public Health and Surveillance* 2017;**3**:e83. doi:10.2196/publichealth.7344

56 Woo H, Cho Y, Shim E, *et al.*. Estimating Influenza Outbreaks Using Both Search Engine Query Data and Social Media Data in South Korea. *Journal of Medical Internet Research* 2016;**18**:e177. doi:10.2196/jmir.4955

57 MOSS R, ZAREBSKI A, DAWSON P, *et al.*. Retrospective forecasting of the 2010–2014 Melbourne influenza seasons using multiple surveillance systems. *Epidemiology and Infection* 2016;**145**:156–69. doi:10.1017/s0950268816002053

58 Samaras L, García-Barriocanal E, Sicilia M-A. Syndromic Surveillance Models Using Web Data: The Case of Influenza in Greece and Italy Using Google Trends. *JMIR Public Health and Surveillance* 2017;**3**:e90. doi:10.2196/publichealth.8015

59 Tableau Public. https://public.tableau.com/en-us/s/

60 Advani A. Design and Assessment of a Common, Multi-National Public Health Informatics Infrastructure to Enable H1N1 Influenza Surveillance. *Studies in Health Technology and Informatics* 2010;**160**.

61 David E. Goldberg JHH. *Genetic Algorithms and Machine Learning.* https://link.springer.com/article/10.1023%2FA%3A1022602019183?LI=true 1988. https://link.springer.com/article/10.1023/A:1022602019183

62 Comparison of rapid detection methods for influenza A virus and their value in health-care management of institutionalized geriatric patients.. *Journal of clinical microbiology* 1994;**32**.http://jcm.asm.org/content/32/1/70.short

63 Wikipedia. Statistical model - Wikipedia. https://en.wikipedia.org/wiki/Statistical_model