b
UNIVERSITÄT
BERN

INPUT Research Meeting

Utilizing electronic health records for epidemiological surveillance

PD Dr. Christian L. Althaus

20 January 2022, Institute of Social and Preventive Medicine, University of Bern

Electronic health records

Definition

- Electronic health records (EHRs) are systematized collection of patient and population electronically stored health information in a digital format.

- EHRs can include demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information.

- EHRs can be used anonymously for statistical reporting in matters such as quality improvement, resource management, and public health surveillance.

Electronic health records

Use for public health surveillance

- EHRs for infectious disease ("communicable disease") surveillance
  - Many examples, e.g., tuberculosis, acute viral hepatitis, sexually transmitted infections, influenza/ influenza-like-illness
- EHRs for chronic disease surveillance
  - Fewer examples, e.g., asthma

Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health

Guthrie S. Birkhead,[1,2] Michael Klompas,[3,4] and Nirav R. Shah[5]

[1]New York State Department of Health, Albany, New York 12237; email: guthrie.birkhead@health.ny.gov

[2]School of Public Health, University at Albany, Rensselaer, New York 12144

[3]Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts 02215; email: mklompas@partners.org

[4]Brigham and Women's Hospital, Boston, Massachusetts 02115

[5]Office of the Chief Operating Officer, Kaiser Foundation Hospitals, Kaiser Permanente, Pasadena, California 91188; email: nirav.r.shah@kp.org

Annu. Rev. Public Health 2015. 36:345–59

First published online as a Review in Advance on January 2, 2015

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

This article's doi: 10.1146/annurev-publhealth-031914-122747

Copyright © 2015 by Annual Reviews. All rights reserved

Keywords

public health surveillance, electronic health records, meaningful user

Abstract

Public health surveillance conducted by health departments in the Unite States has improved in completeness and timeliness owing to electronic lal oratory reporting. However, the collection of detailed clinical informatic about reported cases, which is necessary to confirm the diagnosis, to unde stand transmission, or to determine disease-related risk factors, is still heavi dependent on manual processes. The increasing prevalence and functiona ity of electronic health record (EHR) systems in the United States prese important opportunities to advance public health surveillance. EHR da have the potential to further increase the breadth, detail, timeliness, ar completeness of public health surveillance and thereby provide better da to guide public health interventions. EHRs also provide a unique opport nity to expand the role and vision of current surveillance efforts and to he bridge the gap between public health practice and clinical medicine.

Assessment of electronic health records for infectious disease surveillance

Technical report

19 Nov 2021

Cite:

This is the final report for the mapping study 'Assessment of electronic health records (EHRs) for infectious disease surveillance, prevention and control'. The study was commissioned by ECDC and was delivered by RAND Europe. The objective of the project is to investigate the current status of EHR systems in the European Union and European Economic Area (EU/EEA) and the potential capacity for the use of these data for surveillance of infectious diseases within ECDC's remit.

Download

Assessment of electronic health records for infectious disease surveillance - EN - [PDF-4.76 MB]

Example

Infectious disease surveillance

- **Aim:** Forecasting seasonal influenza
- **Data:** Electronic health records, traditional surveillance data, internet search traffic, and social media activity.
- **Method:** Hierarchical framework using multiple regression, greedy optimization to choose the most predictive combinations of data sources.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Optimal multi-source forecasting of seasonal influenza

Zeynep Ertem[1]*, Dorrie Raymond[2], Lauren Ancel Meyers[3,4]

1 Department of Statistics and Data Science, The University of Texas at Austin, Austin, Texas, United States of America, **2** athenaResearch, Watertown, Massachusetts, United States of America, **3** Departments of Integrative Biology and Statistics and Data Science, The University of Texas at Austin, Austin, Texas, United States of America, **4** The Santa Fe Institute, Santa Fe, New Mexico, United States of America

* zeynepsertem@gmail.com

Check for updates

OPEN ACCESS

Citation: Ertem Z, Raymond D, Meyers LA (2018) Optimal multi-source forecasting of seasonal influenza. PLoS Comput Biol 14(9): e1006236. https://doi.org/10.1371/journal.pcbi.1006236

Editor: Mark M. Tanaka, University of New South Wales, AUSTRALIA

Received: January 30, 2018

Accepted: May 28, 2018

Published: September 4, 2018

Copyright: © 2018 Ertem et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available from the AthenaHealth for researchers who meet the criteria for access to confidential data. Data is restricted to ensure patient confidentiality and prevent disclosure of PHI. Data is de-deidentified, but it is athenahealth's policy to require a legal data use agreement between athena and the researcher prior to providing the data. Please reach out to Josh Gray, Vice President of athenaResearch at athenahealth: jogray@athenahealth.com.

Funding: We would like to acknowledge funding from Defense Threat Reduction Agency (US) HDTRA1-14-C-0114 and NIH/NIGMS MIDAS grant

Abstract

Forecasting the emergence and spread of influenza viruses is an important public health challenge. Timely and accurate estimates of influenza prevalence, particularly of severe cases requiring hospitalization, can improve control measures to reduce transmission and mortality. Here, we extend a previously published machine learning method for influenza forecasting to integrate multiple diverse data sources, including traditional surveillance data, electronic health records, internet search traffic, and social media activity. Our hierarchical framework uses multi-linear regression to combine forecasts from multiple data sources and greedy optimization with forward selection to sequentially choose the most predictive combinations of data sources. We show that the systematic integration of complementary data sources can substantially improve forecast accuracy over single data sources. When forecasting the Center for Disease Control and Prevention (CDC) influenza-like-illness reports (ILINet) from week 48 through week 20, the optimal combination of predictors includes public health surveillance data and commercially available electronic medical records, but neither search engine nor social media data.

Author summary

In the United States, seasonal influenza causes thousands of deaths and hundreds of thousands of hospitalizations. The annual timing and burden of the flu season vary considerably with the severity of the circulating viruses. Epidemic forecasting can inform early and effective countermeasures to limit the human toll of severe seasonal and pandemic influenza. With a growing toolkit of sophisticated statistical methods and the recent explosion of influenza-related data, we can now systematically match models to data to achieve timely and accurate warning as flu epidemics emerge, peak and subside. Here, we introduce a framework for identifying optimal combinations of data sources, and show that public health surveillance data and electronic health records collectively forecast seasonal influenza better than any single data source alone and better than influenza-related search engine and social media data.

Example

Infectious disease surveillance

- **Target data:** Aggregate flu data from ILINet, the CDC national sentinel surveillance system (delayed)

- **Predictor data:**

  - Flu-related electronic health records data (Athena): flu vaccination, flu diagnosis, ILI diagnosis, flu test, positive flu test, flu-related prescription, patients seeking medical attention for ILI.

  - Lab-confirmed cases, Wiki/blog/Twitter activity

Example

Infectious disease surveillance

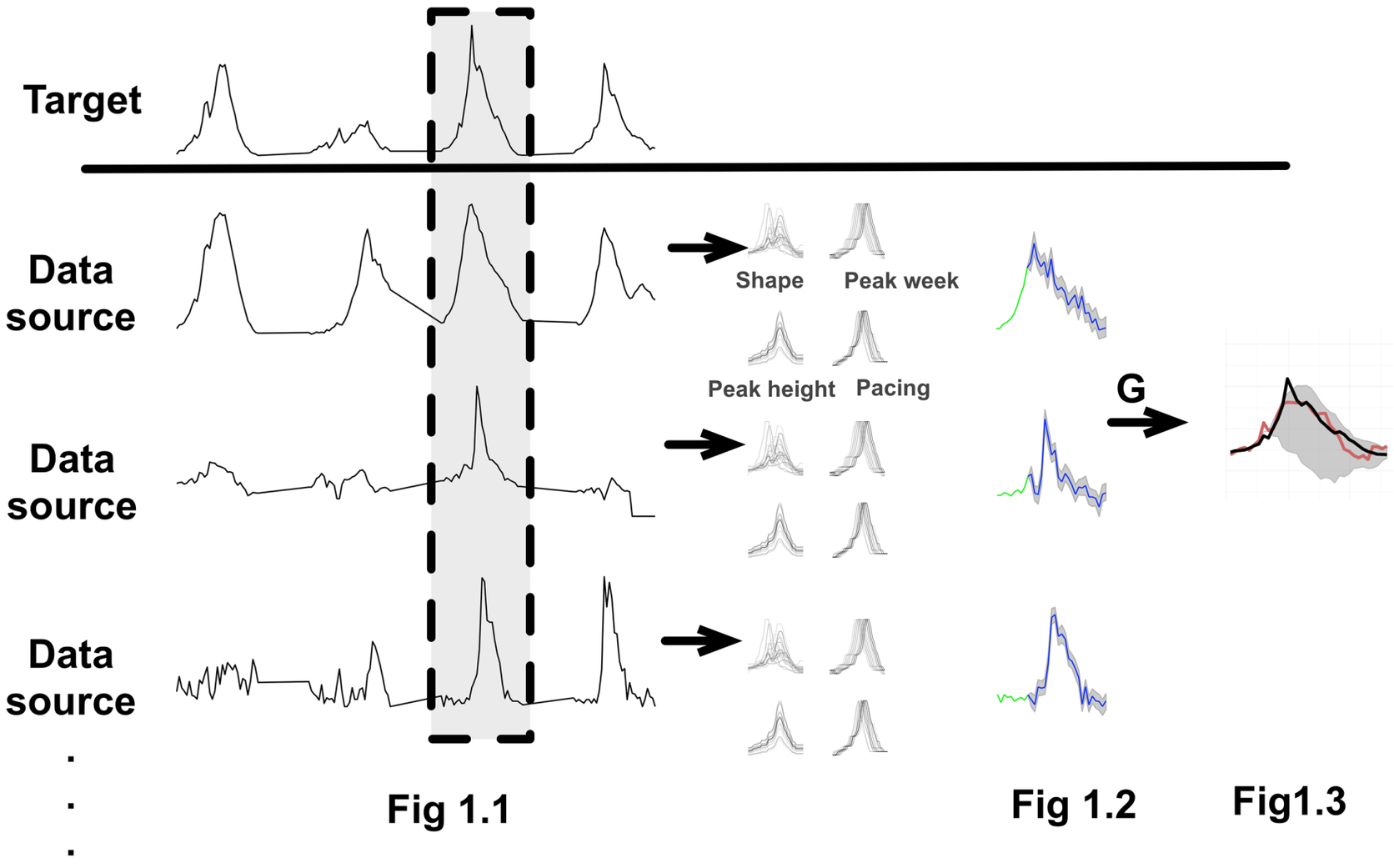


Fig 1. Multi-linear forecast of a historical influenza season. When evaluating a candidate data source, we combine it with previously selected data sources and perform a series of leave-one-out forecasts. Each forecast involves three steps. (1) Align data and remove the *focal* season from all time series (gray band). (2) Make separate Bayes forecasts for each predictor, using the method introduced in [6] (green curves indicate observed weeks and blue curves indicate forecasts). The forecasts are derived from distributions of prototypical curves generated by perturbing and combining characteristics of historic seasons for each candidate data source (shape, pace, peak timing and peak height). (3) Integrate the predictor forecasts into a target forecast (red curve) using the multi-linear model *g* fit to the historical predictor and target data. We evaluate this approach by comparing our target forecasts with the true values of the *focal* season (black curve).

Example

Infectious disease surveillance

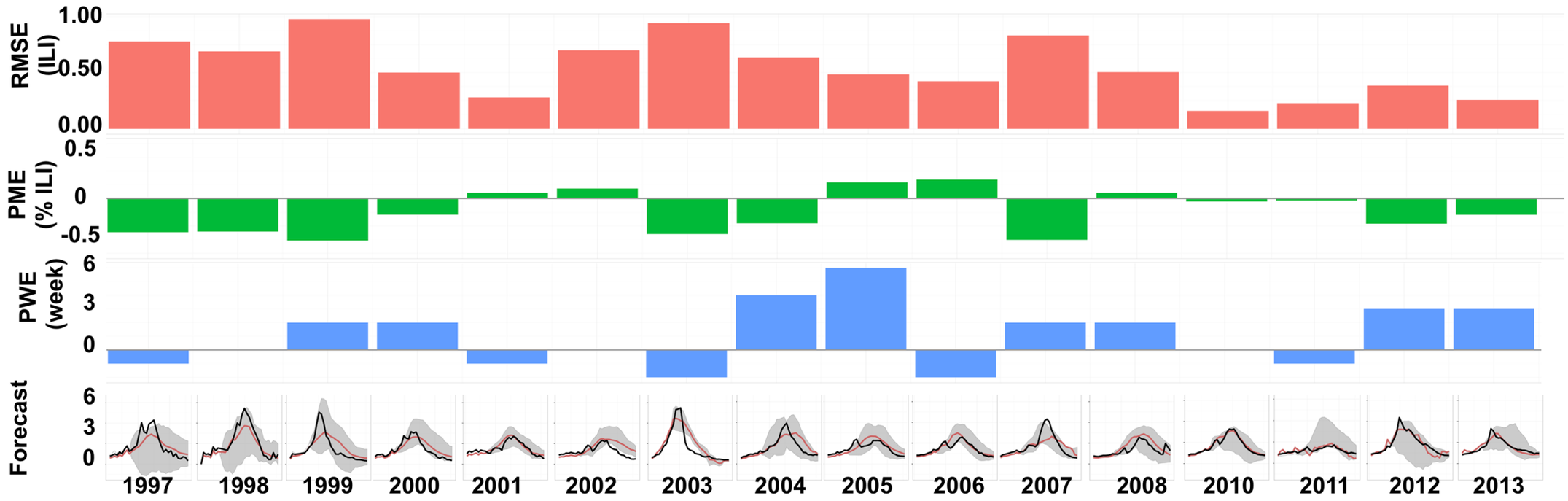


Fig 2. Forecasts of historical flu seasons from 1997-1998 through 2013-2014 (excluding 2009-2010) by the optimized five-source surveillance system. The system includes ILINet, WHO, and three Athena data sources. Forecast performance is summarized in top rows of graphs, by RMSE (red), PWE (green), and PME (blue). The bottom row compares the forecasted (red) and actual (black) times series with 95% credible intervals (gray). Vertical dashed lines indicate the last week of the observational periods, after which all predictor and target data are forecasted.

Example

Infectious disease surveillance

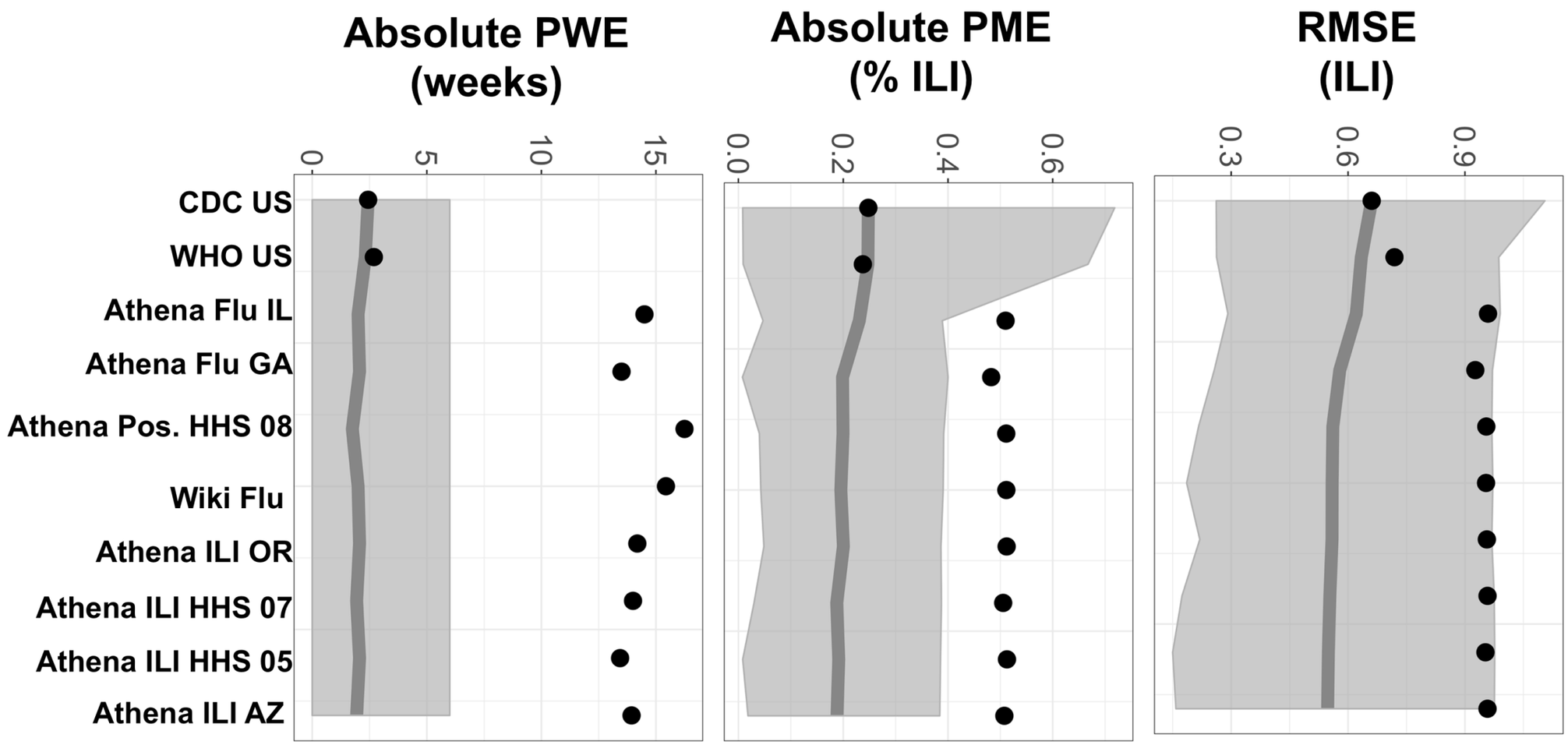


Fig 3. Performance curves for the first ten selected data sources. The system was built through the sequential addition of data sources to minimize RMSE, as listed from left to right along the x-axis. Graphs show the changing performance of the growing system, where points indicate the quality (mean RMSE, PWE, or PME) of forecasts made using all data sources to the left of and including the given x-axis label. Circles indicate individual performance of selected data sources; shading indicates performance range across the 16 seasons tested.

Example

Syndromic surveillance during heat waves

- **Data:** Following the heat wave in 2003, the French National Institute for Public Health Surveillance set up a syndromic surveillance system including data from **emergency departments (ED)** on a daily basis:

  Age, gender, zip-code, reason for emergency admission, main medical diagnosis (ICD-10), admission to hospital.

- **Syndromes:** Hyponatremia, dehydration, malaise, hyperthermia.

- **Analysis:**

  - True positive: Number of above-threshold days in terms of the number of visits during “On Alert Periods” (ONAP).

  - Sensitivity: Proportion of days that exhibited elevated heat-related disease counts detected by the surveillance system during ONAP.

  - Specificity: Proportion of days with normal numbers of heat-related diseases during “Off Altert Periods” (OFAP).

  - Positive predictive value: Number of days with a significant count of heat-related visits during ONAP among the total number of days with a significant count of heat-related visits.

OPEN ACCESS Freely available online

PLoS one

Assessment of a Syndromic Surveillance System Based on Morbidity Data: Results from the Oscour® Network during a Heat Wave

Loïc Josseran[1]*, Anne Fouillet[1], Nadège Caillère[1], Dominique Brun-Ney[2], Danièle Ilef[1], Gilles Brucker[3], Helena Medeiros[1], Pascal Astagneau[4]

1 Department of Alert Coordination and Regions, French Institute for Public Health Surveillance, Saint Maurice, France, **2** Assistance Publique des Hôpitaux de Paris, Paris, France, **3** Groupement d'Intérêt Public (GIP) Esther, Paris, France, **4** Department of Public Health, Pierre and Marie Curie University School of Medicine, Paris, France

Abstract

Background: Syndromic surveillance systems have been developed in recent years and are now increasingly used by stakeholders to quickly answer questions and make important decisions. It is therefore essential to evaluate the quality and utility of such systems. This study was designed to assess a syndromic surveillance system based on emergency departments' (ED) morbidity rates related to the health effects of heat waves. This study uses data collected during the 2006 heat wave in France.

Methods: Data recorded from 15 EDs in the Ile-de-France (Paris and surrounding area) from June to August, 2006, were transmitted daily via the Internet to the French Institute for Public Health Surveillance. Items collected included diagnosis (ICD10), outcome, and age. Several aspects of the system have been evaluated (data quality, cost, flexibility, stability, and performance). Periods of heat wave are considered the most suitable time to evaluate the system.

Results: Data quality did not vary significantly during the period. Age, gender and outcome were completed in a comprehensive manner. Diagnoses were missing or uninformative for 37.5% of patients. Stability was recorded as being 99.49% for the period overall. The average cost per day over the study period was estimated to be €287. Diagnoses of hyperthermia, malaise, dehydration, hyponatremia were correlated with increased temperatures. Malaise was most sensitive in younger and elderly adults but also the less specific. However, overall syndrome groups were more sensitive with comparable specificity than individual diagnoses.

Conclusion: This system satisfactorily detected the health impact of hot days (observed values were higher than expected on more than 90% of days on which a heat alert was issued). Our findings should reassure stakeholders about the reliability of health impact assessments during or following such an event. These evaluations are essential to establish the validity of the results of syndromic surveillance systems.

Citation: Josseran L, Fouillet A, Caillère N, Brun-Ney D, Ilef D, et al. (2010) Assessment of a Syndromic Surveillance System Based on Morbidity Data: Results from the Oscour® Network during a Heat Wave. PLoS ONE 5(8): e11984. doi:10.1371/journal.pone.0011984

Editor: Landon Myer, University of Cape Town, South Africa

Received September 25, 2009; **Accepted** June 22, 2010; **Published** August 9, 2010

Copyright: © 2010 Josseran et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was funded by the French National Institut for Public Health Surveillance (www.invs.sante.fr). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: l.josseran@invs.sante.fr

Introduction

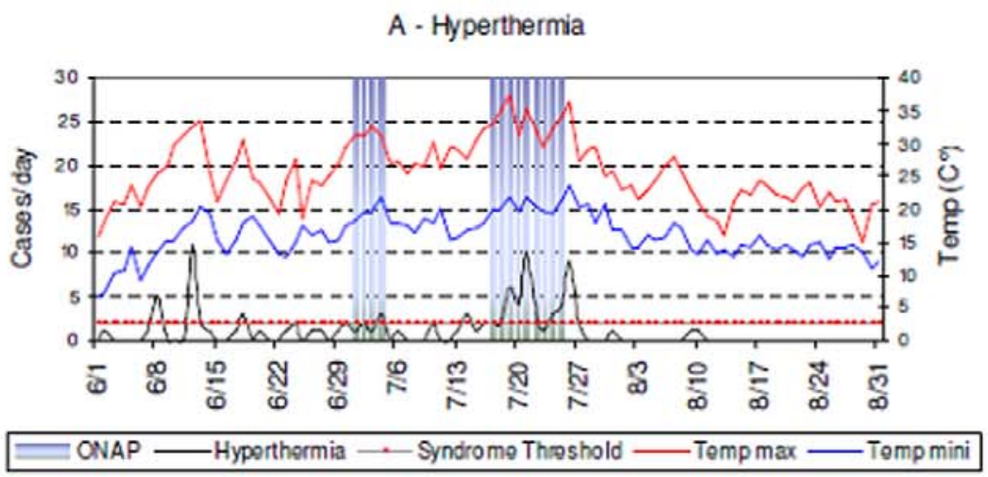
From the time John Graunt published the first epidemiological analysis in 1662 until recently, data recording was limited to paper-based modalities [1]. The Réseau Sentinelles® in 1984, set up in France using the Minitel® (French electronic network), first demonstrated the utility of electronic data recording for routine infectious disease system alerts and feedback transmission for general practitioners [2]. With improvements in electronic technologies, the concept of syndromic surveillance, based on non-specific disease data recorded routinely by healthcare professionals [3–5] and transmitted automatically via the Internet [6], has emerged as a valuable resource.

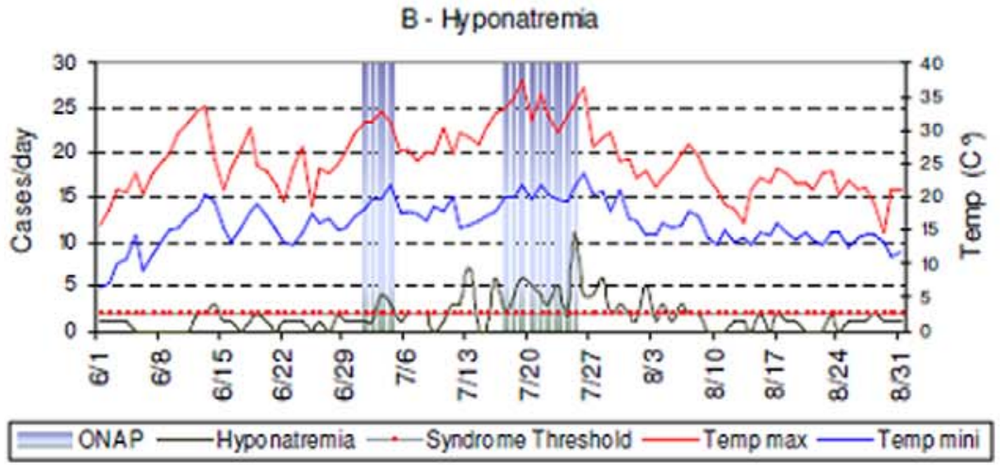
Several syndromic surveillance systems have been developed and deployed worldwide in response to bioterrorist threats [4,7,8]. However, syndromic surveillance systems geared towards a public health approach (not limited to bioterrorism) are of growing interest [7,9,10]. This method has potential for a number of applications, such as monitoring environmental health effects (heat waves, cold spells, and carbon monoxide poisoning) and infectious diseases (influenza, gastroenteritis, and viral meningitis) [6,10–12]. Evaluation of new public health tools is infrequently reported, however, and should be accorded greater importance [3,13–16]. Although gold standards are lacking, it is important to assess the quality of data, which are increasingly used by decision makers to evaluate public health threats [17].

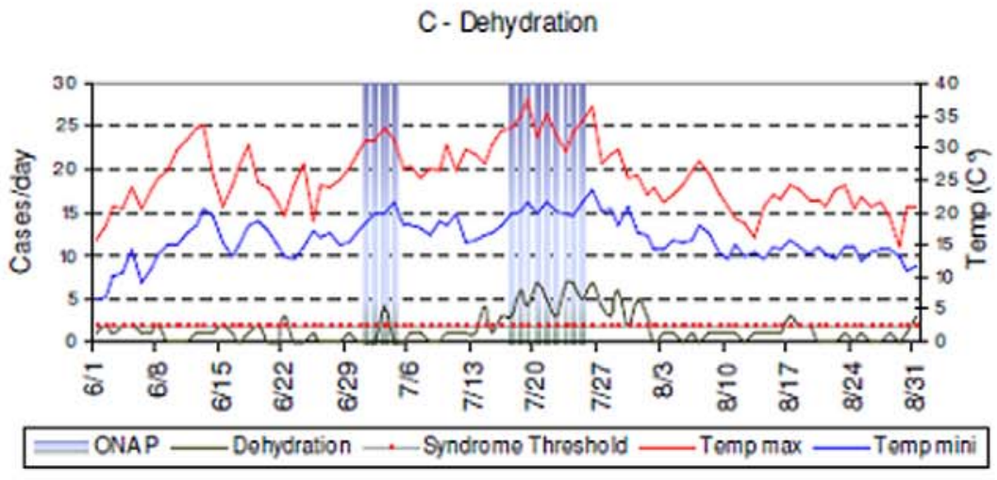
In July 2004, the French National Institute for Public Health Surveillance (Institut de Veille Sanitaire - InVS) set up a syndromic surveillance system based on three data sources: emergency departments (ED), emergency General Practitioners

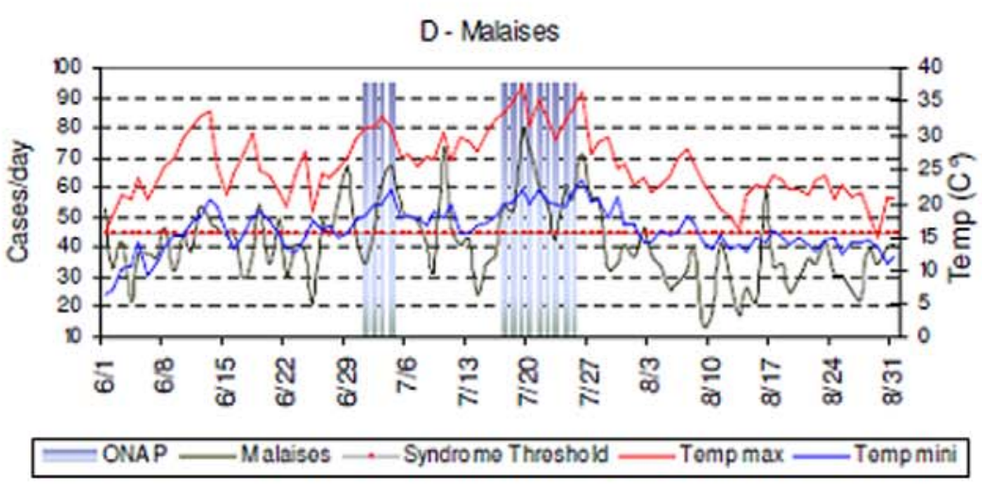
Example

Syndromic surveillance during heat waves









Temp (C°): Temperatures in Celsius degree
ONAP : On Alert Period

Example

Syndromic surveillance during heat waves

Table 3. Sensitivity, specificity, positive predictive value and correlation coefficient of syndromes and ED visits according to age group, compared with ONAP.

All adults	A/D	Sensitivity (CI 95%)	Specificity (CI 95%)	PPV (CI 95%)	Corr Coeff
ED Visits N = 139,433	1/77	0.08 (0.02–0.13)	0.97 (0.94–1.00)	0.33 (0.23–0.43)	0.44
Dehydration N = 133	10/63	0.77 (0.68–0.86)	0.80 (0.72–0.88)	0.38 (0.28–0.48)	0.47
Hyperthermia N = 53	5/72	0.38 (0.28–0.48)	0.91 (0.85–0.97)	0.42 (0.32–0.52)	0.52
Malaise N = 3,711	11/58	0.85 (0.78–0.92)	0.73 (0.64–0.82)	0.34 (0.25–0.43)	0.58
Hyponatremia N = 157	9/68	0.69 (0.60–0.78)	0.86 (0.79–0.93)	0.45 (0.35–0.55)	0.53
15–74 yrs					
ED Visits N = 124,717	2/73	0.15 (0.08–0.22)	0.92 (0.87–0.97)	0.25 (0.16–0.34)	0.43
Dehydration N = 31	0/76	0.00 (0.00–0.00)	0.96 (0.92–1.00)	0.00 (0.00–0.00)	0.25
Hyperthermia N = 44	4/72	0.31 (0.21–0.41)	0.91 (0.85–0.97)	0.36 (0.26–0.46)	0.51
Malaise N = 2,872	9/58	0.69 (0.60–0.79)	0.73 (0.64–0.82)	0.30 (0.21–0.39)	0.57
Hyponatremia N = 52	3/73	0.23 (0.14–0.32)	0.92 (0.87–0.98)	0.33 (0.24–0.42)	0.32
75 and above					
ED Visits N = 14,716	5/64	0.38 (0.28–0.48)	0.81 (0.73–0.89)	0.25 (0.16–0.34)	0.22
Dehydration N = 102	10/67	0.77 (0.68–0.86)	0.85 (0.78– 0.92)	0.45 (0.35–0.55)	0.46
Hyperthermia N = 9	2/74	0.15 (0.07–0.23)	0.94 (0.89–0.99)	0.29 (0.19–0.39)	0.30
Malaise N = 839	11/59	0.85 (0.78–0.92)	0.75 (0.66–0.84)	0.35 (0.26–0.44)	0.31
Hyponatremia N = 105	10/63	0.77 (0.68–0.86)	0.80 (0.72–0.88)	0.38 (0.28–0.48)	0.49

Example

Syndromic surveillance during heat waves

Table 2. Syndromes or situations monitored using Oscour ® Network. July 2004 to April 2009– France.

Syndromes or situations	Monitored period
Infectious diseases	
Influenza	Winter
Bronchiolitis	Fall and Winter
Viral meningititis	All year
Gastro-enterititis	Fall and Winter
Measles	All year
Dengue	Winter ED located in French over seas departments
Environmental health	
Asthma	Spring, Summer, Fall
Cold weather impact	Winter
Hot weather impact	Summer
Carbon monoxide poisoning	All year
Extreme weather event (hurricane, floods, heat)	All year
Others	
Industrial accident impact	All year
Stakeholders reassurance	All year
Mass gathering (health Monitoring)	All year

Natural language processing

From unstructured to structured text

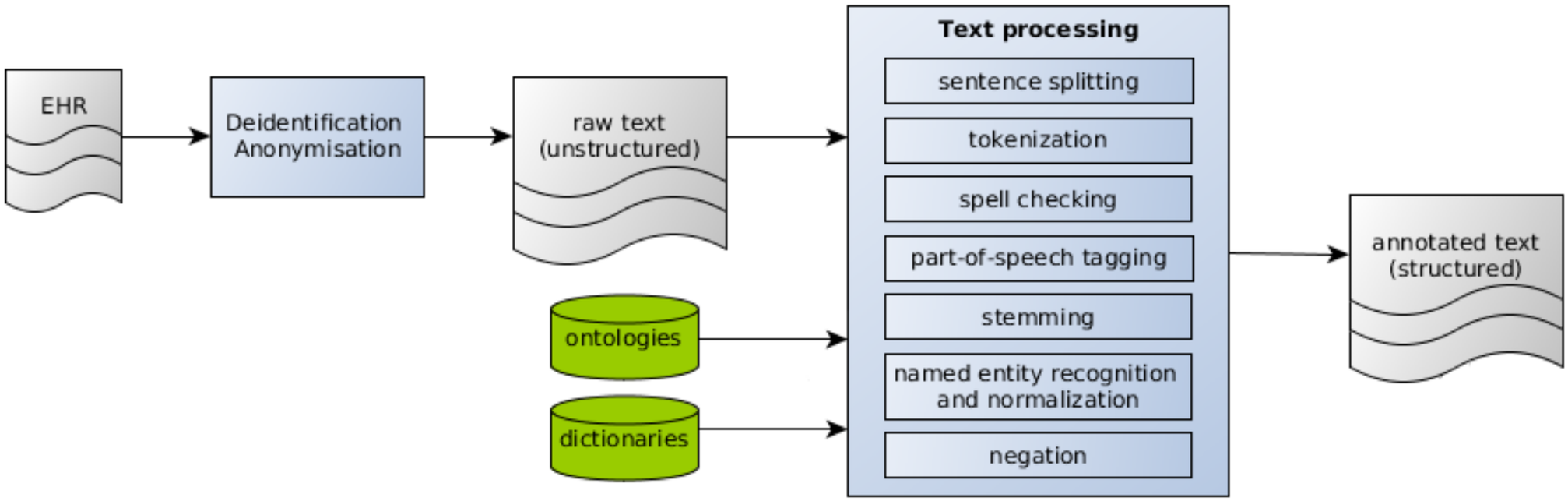
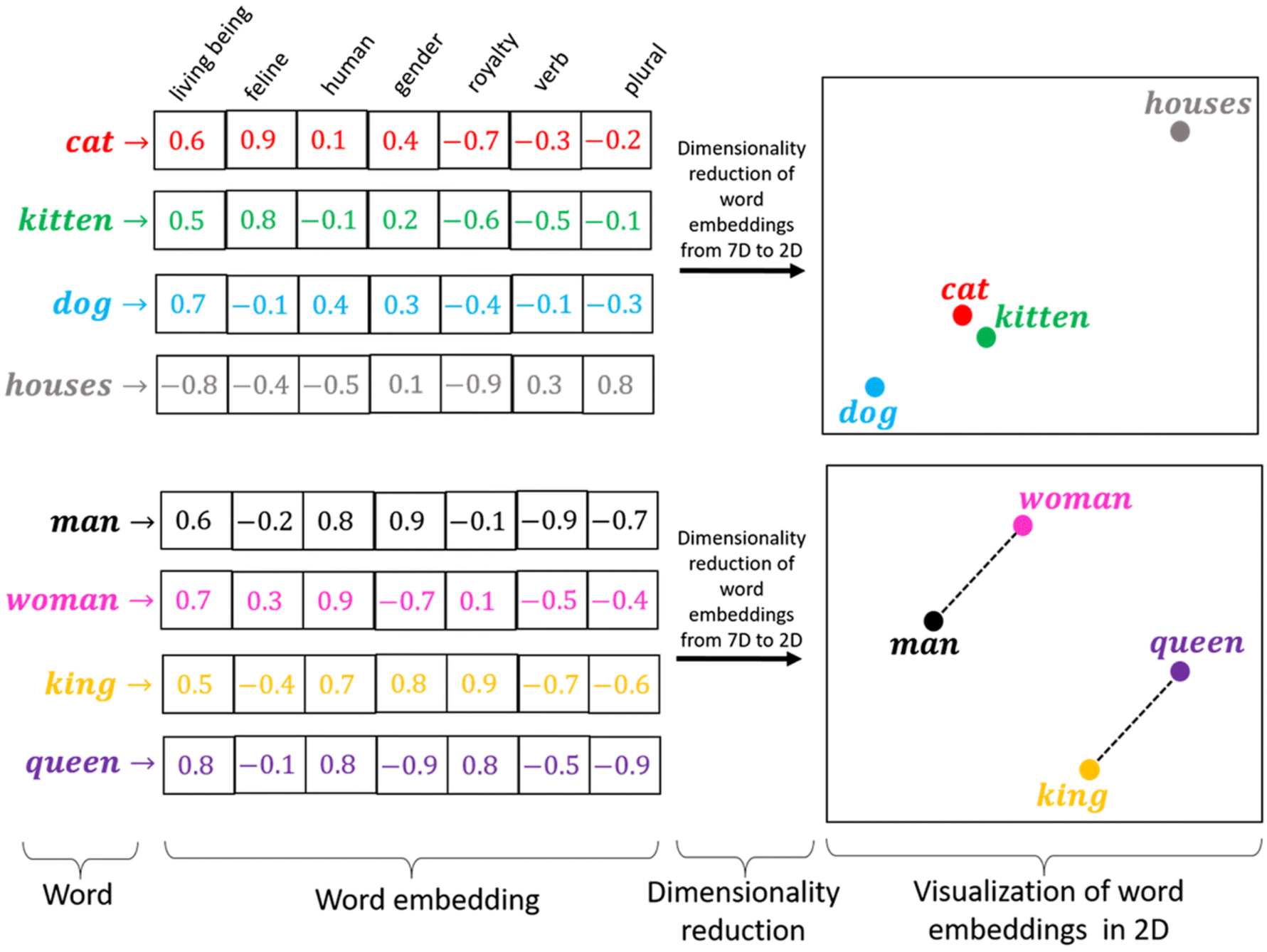


Figure 2: Common setup of NLP pipeline: Electronic health records (EHR) are de-identified and input to the text processing module performing several subtasks including linguistic preprocessing, named entity recognition and normalization, and negation detection to produce annotated text. Ontologies and dictionaries are used as resources for entity recognition and normalization.

Reference: *Starlinger et al. (2016, Information Technology)*

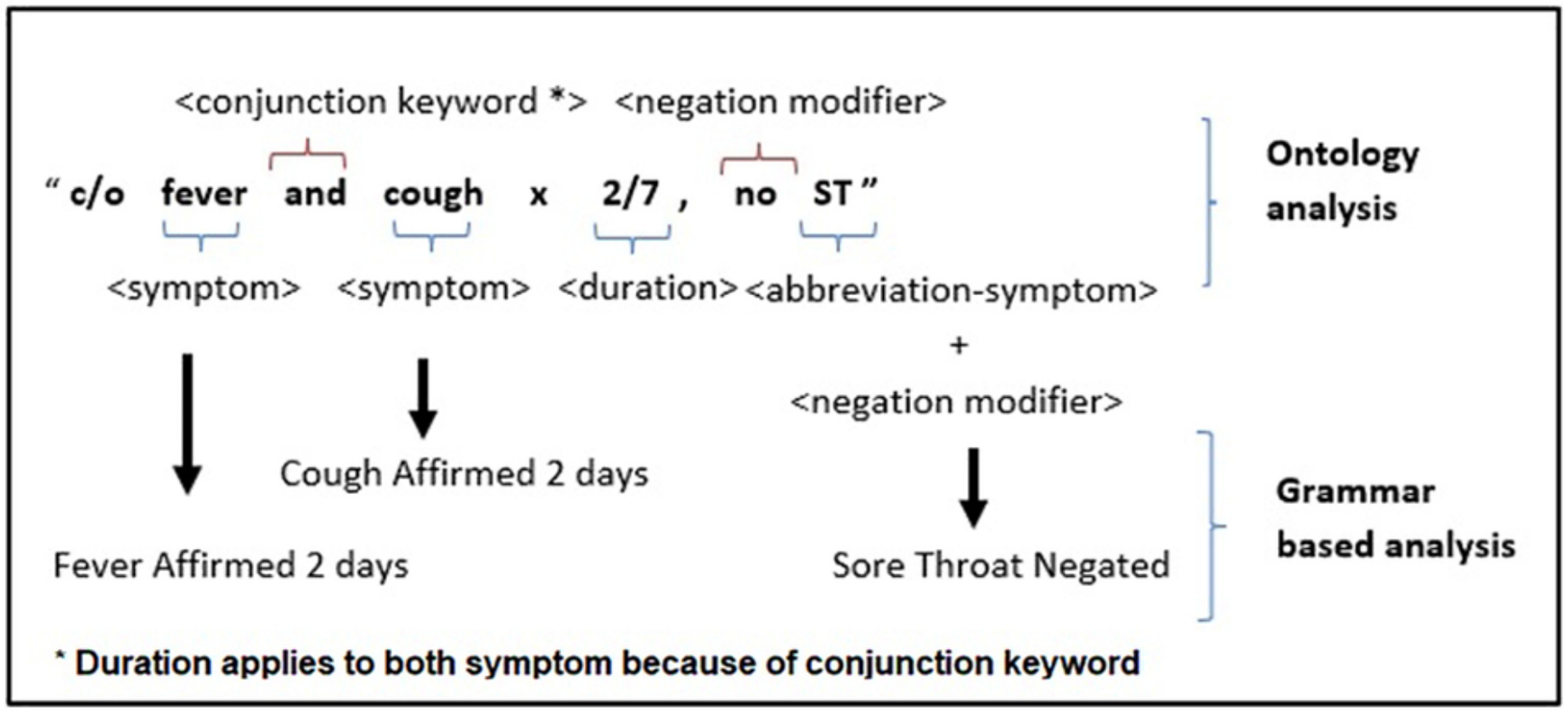
Natural language processing

Word embedding



Natural language processing

Ontology and grammar-based analysis

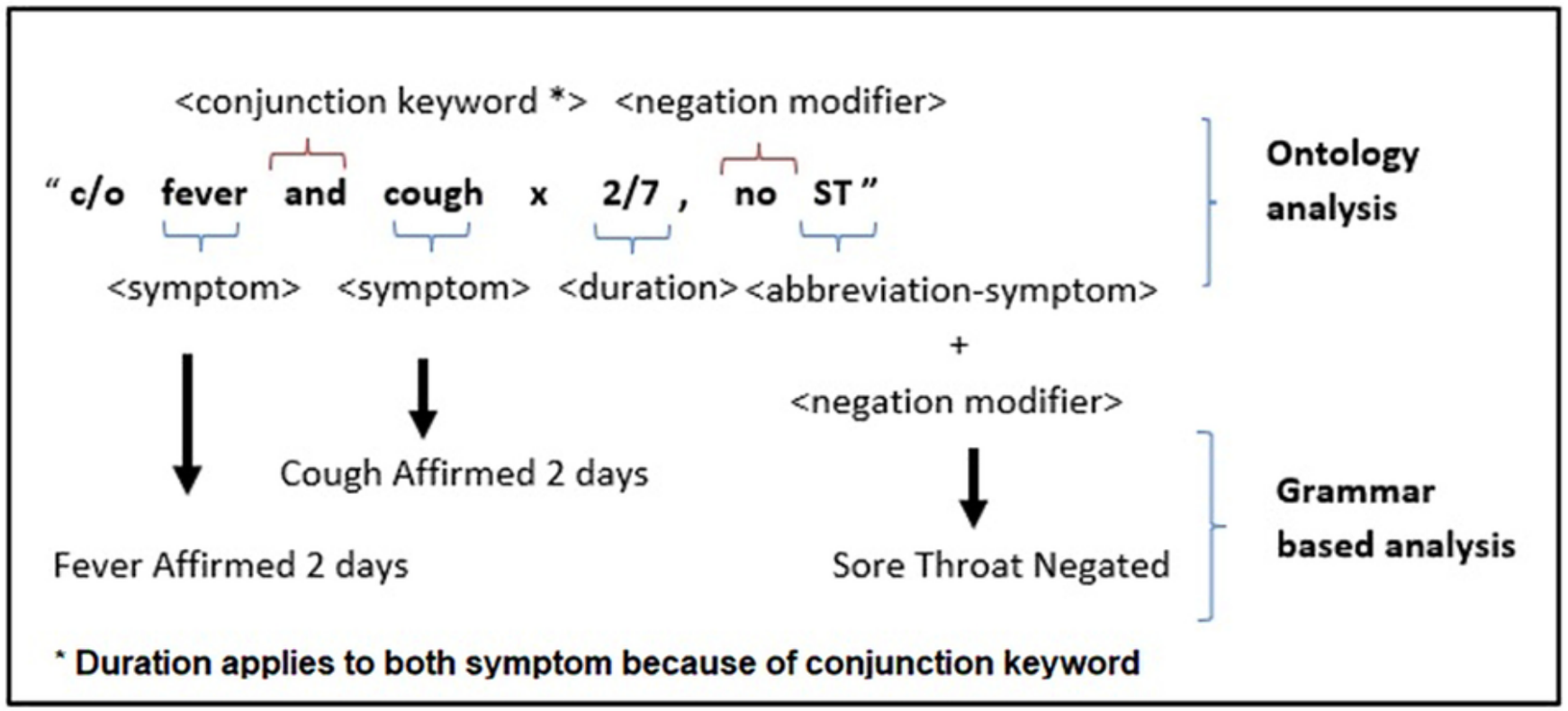


Ontology and grammar-based analysis of the rule-based natural language processing (NLP) algorithm. Signs and symptoms and information on assertion status and duration are captured and tokenized in the ontology analysis. Relationships between tokens are built up in the grammar-based analysis. C/o: complain of; ST; sore throat.

Reference: *Hardjojo et al. (2018, JMIR Med Inform)*

Natural language processing

Extract clinical information from unstructured, free text

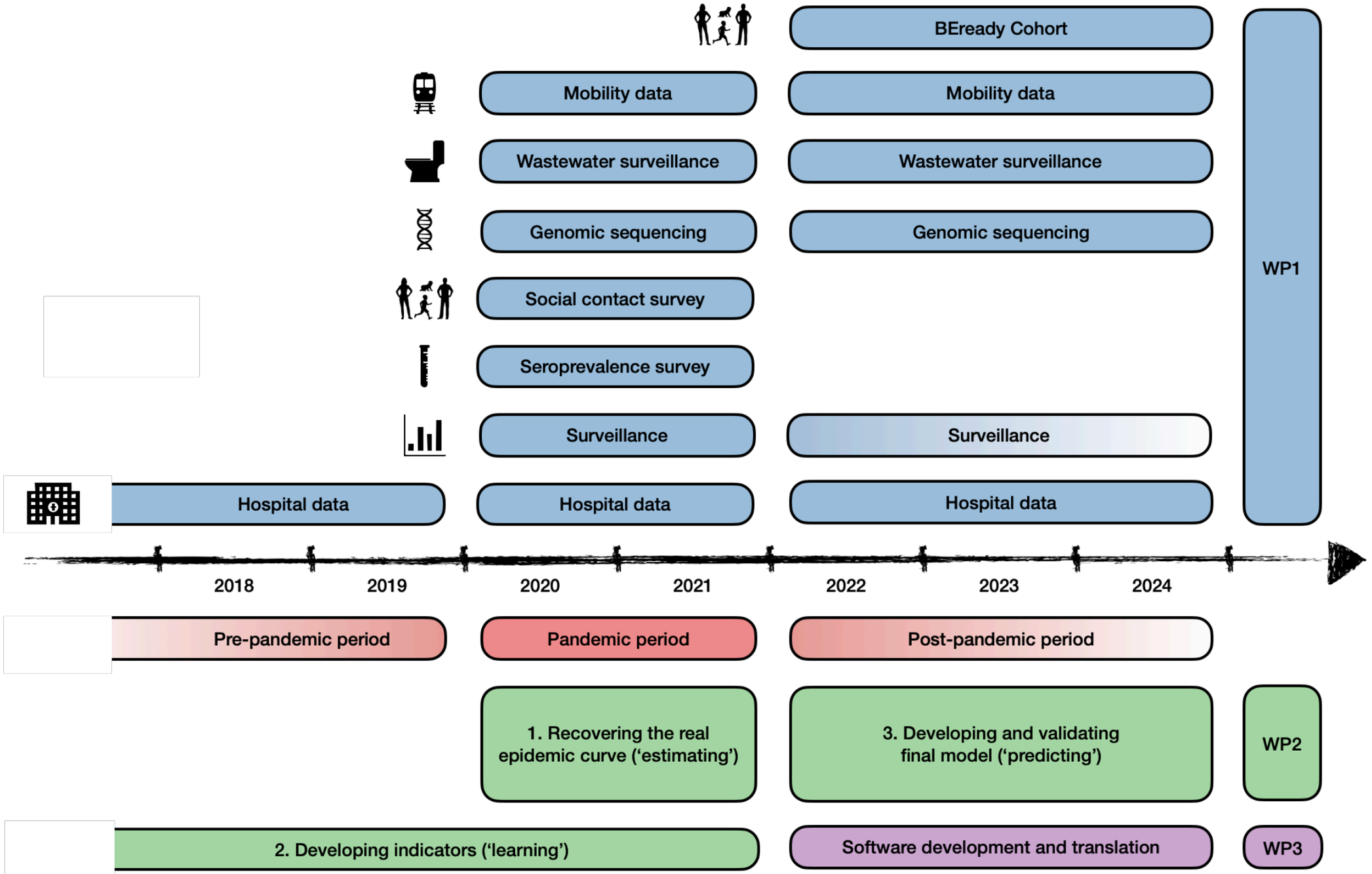


Ontology and grammar-based analysis of the rule-based natural language processing (NLP) algorithm. Signs and symptoms and information on assertion status and duration are captured and tokenized in the ontology analysis. Relationships between tokens are built up in the grammar-based analysis. C/o: complain of; ST; sore throat.

Reference: *Hardjojo et al. (2018, JMIR Med Inform)*

MCID project

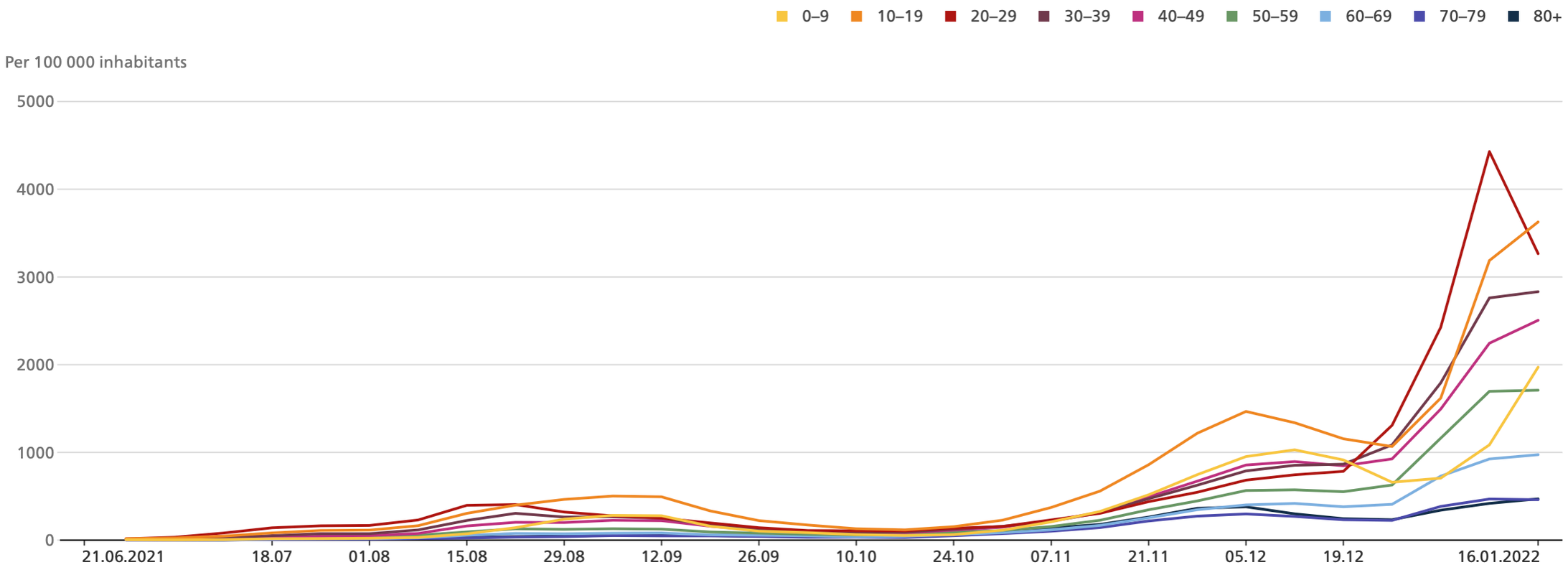
Integrating multiple data sources for monitoring SARS-CoV-2



MCID project

Integrating multiple data sources for monitoring SARS-CoV-2

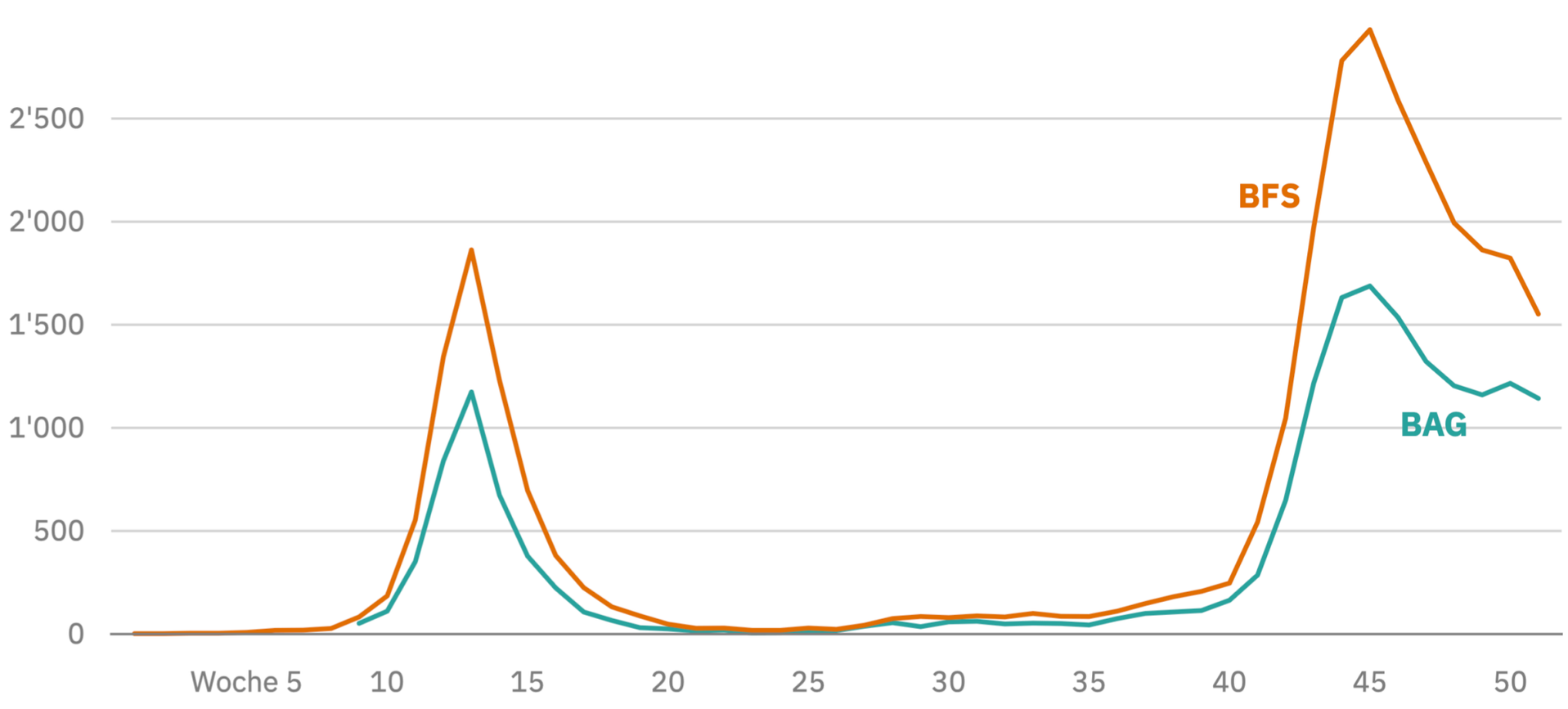
Problem: Current monitoring tools typically rely on routine surveillance data (e.g., confirmed cases) that are heavily biased due to heterogenous testing strategies.



MCID project

Integrating multiple data sources for monitoring SARS-CoV-2

Same for hospitalizations…



MCID project

Integrating multiple data sources for monitoring SARS-CoV-2

- **Hypothesis:** EHRs can be used to develop indicators that predict the epidemic dynamics.
- **Aim:** To identify correlates between the **multidimensional** data from EHRs and the epidemic dynamics of SARS-CoV-2 across different population strata.
- **Data:** EHRs of hospital admissions
  - demographics (e.g., age, sex)
  - patient administrative data (e.g., emergency care, ICU stay)
  - temperature, laboratory results (e.g., CRP levels, leukocytosis, blood pressure, pulse)
  - ICD codes (e.g., comorbidities), medication data (e.g., ATC codes)
  - specific SARS-CoV-2-related data
  - unstructured text files

Discussion

Utilizing electronic health records for epidemiological surveillance

- **Potential use of EHRs for other research projects at ISPM**
  - Non-communicable and chronic diseases
  - Association of disease with climate (heat-related morbidity and mortality)
- **Statistical and machine learning: What are the most appropriate tools to analyze EHRs?**
  - Linear and logistic regression
  - Generalized additive models
  - Random forests
  - Deep networks?
- **Related methods from physics/natural sciences**
- **Natural Language Processing (NLP)**
  - Other data sets consisting of unstructured text data
  - Demand at ISPM and other institutes?
  - Applications in epidemiology/public health outside EHRs?
  - Experts at the University of Bern?
- **Potential for collaborations across faculties?**