HHS Public Access

Author manuscript

J Travel Med. Author manuscript; available in PMC 2025 June 03.

Published in final edited form as:

J Travel Med. 2024 June 03; 31(4): . doi:10.1093/jtm/taae013.

From GeoSentinel data to epidemiological insights: a multidisciplinary effort towards artificial intelligence-supported detection of infectious disease outbreaks

Stan Heidema, Msc[1,*], **Ivo V. Stoepker, Msc**[1], **Gerard Flaherty, MD, PhD**[2,3], **Kristina M. Angelo, DO, MPH&TM**[4], **Richard A. J. Post, PhD**[1], **Charles Miller, MSOR**[4], **Michael Libman, MD**[5], **Davidson H. Hamer, MD**[6,7,8], **Edwin R. van den Heuvel, PhD**[1,9], **Ralph Huits, MD, PhD**[10]

[1]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[2]School of Medicine, University of Galway, University Road, Galway, H91 TK33, Ireland

[3]School of Medicine, International Medical University, Bukit Jalil, 57000, Kuala Lumpur, Malaysia

[4]Division of Global Migration and Health, National Center for Emerging and Zoonotic Infectious Disease, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA 30329, USA

[5]J.D. MacLean Centre for Tropical Diseases, McGill University, Room E05.1830, 1001 Boulevard Décarie, Montréal, Québec H4A 3J1, Canada

[6]Department of Global Health, Boston University School of Public Health, Crosstown 308, 801 Massachusetts Avenue, Boston, MA 02118, USA

[7]Section of Infectious Diseases, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Crosstown 308, 801 Massachusetts Avenue, Boston, MA 02118, USA

[8]Center for Emerging Infectious Diseases Policy and Research, Boston University, 111 Cummington Mall, #104, Boston, MA 02215, USA

[9]Department of Preventive Medicine and Epidemiology, School of Medicine, Boston University, 72 East Concord Street, Floor L-5, Boston, MA 02218, USA

[*]To whom correspondence should be addressed. s.g.a.m.heidema@tue.nl.

Author Contributions

Stan Heidema (Conceptualization-Equal, Writing—original draft-Lead, Writing—review & editing-Equal), Ivo Stoepker (Conceptualization-Equal, Supervision-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Gerard Flaherty (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Kristina Angelo (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Richard Post (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Charles Miller (Conceptualization-Equal, Writing—review & editing-Equal), Michael Libman (Conceptualization-Equal, Writing—review & editing-Equal), Davidson Hamer (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Edwin van den Heuvel (Conceptualization-Equal, Supervision-Equal, Writing—review & editing-Equal), Ralph Huits (Conceptualization-Equal, Supervision-Equal, Writing—original draft-Equal, Writing—review & editing-Equal).

Conflict of interests: ML, DHH, RH receive salary support via the cooperative agreement between ISTM and the CDC for GeoSentinel (1 U01CK000632-01-00). All remaining authors have declared no conflicts of interest.

[10]Department of Infectious Tropical Diseases and Microbiology, IRCCS Sacro Cuore Don Calabria Hospital, 37024, Via Don A Sempreboni 5, Negrar di Valpolicella, Verona, Italy

Keywords

Sentinel event; disease surveillance; travel medicine; outbreak detection; artificial intelligence; data science

Sentinel surveillance of international travellers has enabled GeoSentinel, a global surveillance and research network collaboration between the International Society of Travel Medicine (ISTM) and the US Centers for Disease Control and Prevention (CDC), to help identify multiple unrecognized outbreaks of public health importance (e.g. dengue in Angola 2013, Zika in Costa Rica 2016 and yellow fever in Brazil 2018[1]) mostly using manual analysis techniques. Since its inception in 1995, the number of participating international GeoSentinel clinical sites has increased to 71 across 29 countries located on six continents.

Standardized data (e.g. demographic, clinical and travel information) of ill travellers seen during and after travel are collected and entered in the GeoSentinel database by expert clinicians at travel and tropical medicine clinical sites. The database holds records from over 400 000 international travellers, however, the evolving system of data collection (e.g. addition or removal of variables) and the dynamic changes in both the number and geographic coverage of reporting sites have led to increased complexity in detecting sentinel cases or outbreaks. Although the application of standard statistical methodologies has led to successful detection of travel-associated illness trends and clusters,[2] data from the GeoSentinel Network present further opportunities to enhance our understanding of travel-related diseases through development of more sophisticated outbreak detection methodologies.

Important progress in developing early-warning systems for disease surveillance has been made by incorporating artificial intelligence (AI) algorithms that can extract insights from complex datasets for signals of infectious disease events with high accuracy.[3] Between 1900 and 1935, modelling techniques were developed in which populations were assigned to compartments (e.g. Susceptible, Infected, Recovered) to describe the characteristics of the spread of infectious diseases. Such models, now considered foundational to mathematical epidemiology, were not developed by statisticians but by public health physicians.[4] In a similar spirit, we argue that the development of modern outbreak detection methodologies is accelerated by multidisciplinary collaboration between data scientists and epidemiologists.[5] While AI has increasingly replaced human tasks in other industries, given the necessary global collaboration in combating disease outbreaks, an outbreak detection methodology should complement rather than replace human decision-making.[6]

In this perspective, we identify challenges associated with applying novel data science methods to GeoSentinel surveillance data for outbreak detection. Subsequently, we demonstrate how effective multidisciplinary collaboration can overcome these challenges. Finally, we highlight the advantages of analysing the GeoSentinel data using such methods.

Collaborative efforts to overcome inherent challenges in outbreak detection

Multiple statistical methods have been developed for early detection of infectious disease outbreaks such as control charts, scan statistics and regression-based techniques.[7] As these methods have become more sophisticated with the integration of AI,[3] the following inherent challenges persist in automated outbreak detection: determining and modelling background behaviour (e.g. endemic transmission rates), evaluating model performance, handling outbreak signals, the nature of outbreaks and their identification and evaluating overall system performance.[8]

In addressing the challenge of determining and modelling background behaviour, baseline prevalence data are essential. Since GeoSentinel data are limited to travellers seeking healthcare at GeoSentinel member sites, calculating prevalence, incidence and risk is challenging. However, given a diagnosis, timeframe and geographical range, approximate baseline limits for non-outbreak-like frequency patterns can still be established.[9] Outbreak detection models signal epidemiologists when observed case numbers exceed these baseline limits. It is thus crucial for epidemiologists to label past non-outbreak periods accurately to avoid erroneously using data on epidemic behaviour to establish such baselines. For instance, between March and June 2022, five European travellers across three GeoSentinel sites were diagnosed with Zika.[9] These cases were recognized as part of a 2022 outbreak because there were no reported cases during the baseline period that began in early 2020. If March to May 2022 was used as the baseline period, four cases would have been identified, and the single case in June would not have been recognized as part of an outbreak.

The key to evaluating the performance of an outbreak detection methodology is to have access to labelled datasets that clearly identify prior outbreaks. Accurate labels can be used to show how well a method detects past outbreaks, therefore helping to evaluate potential performance in detecting future outbreaks. Given predefined threshold parameters designed to control the false signal rate, receiver operating characteristic curves are an effective tool for evaluating performance by illustrating the relationship between true and false positives. Labeling historical data also allows estimation of the likelihood a signal corresponds to an outbreak through calculation of a positive predictive value. Another benefit of labelled datasets is that they are amenable to the application of powerful supervised machine learning methods. An example is the successful implementation of a method, based on a hidden Markov model which used endemic state data to reflect the expected endemic baseline, for the detection of *Salmonella* and *Campylobacter* outbreaks.[10]

Users of large surveillance systems, such as GeoSentinel, may experience difficulties handling outbreak signals because new alerts appear on a daily basis.[8] In contrast to fields such as engineering, where conservative adjustments for multiple testing are routine, in outbreak detection higher false signal rates are generally preferred over under-detection of significant events.[8] Determining the optimal thresholds, however, is challenging for data scientists and requires input from epidemiologists and other public health professionals. User-customizable threshold parameters are common,[6] but a methodology that is self-adaptive to ongoing feedback from epidemiologists is highly desirable. For example, suppose an initial method proves overly sensitive for identifying clusters of diagnoses

such as influenza-like-illnesses. In that case, the user can correct the method to alert more conservatively. Conversely, positive feedback can preserve sensitivity in detecting outbreaks with high mortality rates (e.g. yellow fever). Incorporating such feedback is easier in explainable methods (e.g. Bayesian networks) than in complex, overparametrized machine learning models, often deemed impossible to interpret and commonly referred to as black box models.

Due to the nature of outbreaks, the heterogeneity in magnitude, shape and expected lengths between outbreaks is important to recognize.[8] Outbreaks identified by GeoSentinel have varied from local (chikungunya in Bali, 2022[11]), to country wide (Zika in Cuba, 2017[12]), to global (mpox, 2022[13]). Modelling tools, such as spatial control charts and the spatial–temporal scan statistics available in SaTScan™ (https://www.satscan.org/), allow disease surveillance at varying levels of spatial and temporal aggregation.[7] Increasing the number of aggregation levels can explain more complex relations, but it comes at a cost. To prevent the unnecessary inflation of the number of statistical tests, epidemiologists must choose practically relevant aggregation levels.

Evaluating overall system performance in comprehensive disease surveillance, which constitutes a complex and interdependent network of methodologies and data sources, presents challenges beyond assessing outbreak detection methodology among international travellers alone. One notable challenge in this evaluation is the dependence on local reporting systems, as sentinel surveillance of travellers can supplement surveillance activities in source countries. For example, many cases of Zika among travellers to Cuba were reported to GeoSentinel in 2017,[12] although the reported numbers were not higher than previous years. However, when the GeoSentinel case numbers were compared to those reported by domestic surveillance systems, there was a discrepancy—many more cases were reported among international travellers than reported domestically during that year. Furthermore, outbreaks may not always manifest as increased case frequency but as increased clinical severity, e.g. through pathogens gaining virulence. Although monitoring multivariate time series is mathematically challenging,[8] tools such as HealthMap (https://www.healthmap.org/) and ESSENCE[6] are the results of multidisciplinary efforts and enable us to comprehensively understand disease dynamics through various data sources, such as over-the-counter pharmaceutical sales, web searches, school/work absences, wastewater and climatological[14] data.

GeoSentinel's data advantage for outbreak detection

It is important to recognize the impact of data quality on the performance of data science methods—analysing vast amounts of data is beneficial only if the data are valid. The expertise of GeoSentinel site members in diagnosing travel-related diseases is key in ensuring data validity and accuracy. In addition, regular summary reports[1] of outbreaks detected by the network will make labeling of the dataset a feasible process. When training models to automate outbreak detection, the use of GeoSentinel data entered by clinical experts has obvious advantages (cf. Table 1) over the analysis of raw data from a variety of sources on the internet where misinformation may be present.[15] Additionally, although GeoSentinel data are not representative of all travellers, the international catchment of

travel-related illnesses among various types of travellers (e.g. tourists, migrants) to diverse destinations is important for the identification of outbreaks via AI methods.

Conclusion

GeoSentinel is advancing to develop and deploy AI-supported outbreak detection methods to further impact clinical medicine, patient care and public health. The early signals generated by outbreak detection methods using GeoSentinel data may influence policymaking, shape public health responses and contribute to global disease control strategies. Timely communication and collaboration among epidemiology, clinical and data science partners, including GeoSentinel sites, affiliate members, ISTM members, CDC, European Centre for Disease Prevention and Control, Public Health Agency of Canada, World Health Organization, ProMED, TropNet, EpiCore and HealthMap is crucial for combating global health threats.

Funding

This project was funded through a Cooperative Agreement between the Centers for Disease Control and Prevention and the International Society of Travel Medicine (Federal Award Number: 1 U01CK000632-01-00). Public Health Agency of Canada also provides a grant to the International Society of Travel Medicine. This work is part of the research project ‘GLOBAL OUTBREAK DETECTION’ at Eindhoven University of Technology co-funded by GeoSentinel.

References

1. Hamer DH, Rizwan A, Freedman DO, Kozarsky P, Libman M. GeoSentinel: past, present and future. J Travel Med 2020; 27:1–8.
2. Leder K, Torresi J, Brownstein JS et al. Travel-associated illness trends and clusters, 2000–2010. Emerg Infect Dis 2013; 19:1049–73. [PubMed: 23763775]
3. Brownstein JS, Rader B, Astley CM, Tian H. Advances in artificial intelligence for infectious-disease surveillance. New England Journal of Medicine 2023; 388:1597–607. [PubMed: 37099342]
4. Brauer F Mathematical epidemiology: past, present, and future. Infect Dis Model 2017; 2:113–27. [PubMed: 29928732]
5. Flaherty GT, Piyaphanee W. Predicting the natural history of artificial intelligence in travel medicine. J Travel Med 2023; 30:1–3.
6. Burkom H, Loschen W, Wojcik R et al. Electronic surveillance system for the early notification of community-based epidemics (ESSENCE): overview, components, and public health applications. JMIR Public Health Surveill 2021; 7:e26303. [PubMed: 34152271]
7. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. J R Statist Soc A 2012; 175:49–82.
8. Shmueli G, Burkom H. Statistical challenges facing early outbreak detection in biosurveillance. Dent Tech 2010; 52:39–51.
9. Seers T, Rothe C, Hamer DH et al. Zika virus infection in European travellers returning from Thailand in 2022: a GeoSentinel case series. Tropical Medicine and International Health 2023; 28:576–9. [PubMed: 37269191]
10. Zacher B, Czogiel I. Supervised learning using routine surveillance data improves outbreak detection of salmonella and campylobacter infections in Germany. PloS One 2022; 17:e0267510. [PubMed: 35511793]
11. Mayer AB, Consigny PH, Grobusch MP, Camprubí-Ferrer D, Huits R, Rothe C. Chikungunya in returning travellers from Bali - a GeoSentinel case series. Travel Med Infect Dis 2023; 52:102543. [PubMed: 36682574]

12. Grubaugh ND, Saraf S, Gangavarapu K et al. Travel surveillance and genomics uncover a hidden Zika outbreak during the waning epidemic. Cell 2019; 178:1057–1071.e11. [PubMed: 31442400]
13. Angelo KM, Smith T, Camprubí-Ferrer D et al. Epidemiological and clinical characteristics of patients with monkeypox in the GeoSentinel network: a cross-sectional study. Lancet Infect Dis 2023; 23:196–206. [PubMed: 36216018]
14. Pramanik M, Singh P, Kumar G, Ojha VP, Dhiman RC. El Niño southern oscillation as an early warning tool for dengue outbreak in India. BMC Public Health 2020; 20:1498. [PubMed: 33008350]
15. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021; 23:e17187. [PubMed: 33470931]
16. Ruff L, Kauffmann JR, Vandermeulen RA et al. A unifying review of deep and shallow anomaly detection. Proc IEEE 2021; 109:756–95.

Table 1

Key features of the GeoSentinel database and their modelling benefits for outbreak detection

Feature	Description	Modelling benefits
Reliability	Use of validated diagnostic testing leads to high-quality data.	Reliable output depends on valid input. Misinformation on social media,[15] increasing the need for trusted sources.
Scalability	Large amount of global, historical data (since 1995) and a growing network.	Labelled data allow for supervised learning.[10] Sustained growth allows for powerful deep anomaly detection methods.[16]
Dimensionality	Variety of data collected (e.g. patient demographic information, travel history, reason for travel).	Potential for multivariate monitoring methods,[7] spatial-temporal monitoring[7] and high-risk subgroup identification.[1]
Timeliness	Sites incentivized to promptly enter records, leading to rapid alerts.[11,12]	Automated monitoring of travel-related health data in real-time or at high frequency enables efficient insights and alerts.