Research Output Journal of Engineering and Scientific Research 4(1): 43-49, 2025

ROJESR Publications Online ISSN: 1115-9790

https://rojournals.org/roj-engineering-and-scientific-research/ Print ISSN: 1115-6155

https://doi.org/10.59298/ROJESR/2025/4.1.4349

Predictive Analytics in Public Health: Anticipating Disease Outbreaks

Ivan Mutebi

Department of Pharmacognosy Kampala International University Uganda
Email: ivan.mutebi@studwc.kiu.ac.ug

ABSTRACT

Predictive analytics is an emerging approach in public health that leverages data-driven methodologies, statistical modeling, and machine learning to anticipate and mitigate disease outbreaks. By analyzing historical and real-time data, predictive analytics enables decision-makers to implement proactive measures, allocate resources efficiently, and enhance public health responses. This paper examines the scope of predictive analytics, key data sources, and the various statistical and machine learning models used in outbreak prediction. Additionally, it presents case studies showcasing successful applications and discusses the ethical challenges and limitations of predictive analytics in public health. The study emphasizes the importance of integrating predictive models into public health decision-making while addressing data quality, privacy, and equity concerns.

Keywords: Predictive Analytics, Public Health, Disease Outbreak Prediction, Machine Learning, Statistical Modeling, Health Data Sources.

INTRODUCTION

Data drives decision-making across numerous public health sectors and is supported by growing technological advances in various domains. Public health gains are more likely to occur when using data-driven approaches to anticipate relevant events. Predictive analytics involves anticipating future events based on current and historical data. In public health, much of the interest lies in anticipating when the next person will experience an avoidable negative health event. Once the occurrence can be anticipated, a series of monitoring, prevention, and therapeutic actions can be taken by decision-makers. The outcome of applied predictive analytics within public health can be leveraged for resource allocation to the jurisdictions anticipated to have the maximum public health needs; used in smart public health intelligence systems; supplemented with additional modeling to help in policy development; and leveraged for general public health decision support [1, 2]. Predictive analytics is broadly defined as computationally intensive modeling methodologies. Within predictive analytics, there is a wide variety of advanced analytics and tools that contribute to predicting health-related outcomes. Tools range from out-of-the-box algorithms and normalizing health data patterns to analyzing patterns within spatial-temporal trends. Both sides of the predictive analytics continuum contribute to developing the tools that can be used for public health decision analysis in various types of public health big data. It is a growing and dynamic field: the techniques are shifting and gaining complexity in tandem with the increasing volumes of electronic health data. The new tools developed within a predictive analytics framework are particularly vital as there is a unique and timely need for public health analysis to inform the current ongoing and progressively increasing demands on regional public health services around the world. In

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

times of emerging public health crises causing population movement, we need analysis now in the current rapidly warming globe [3, 4].

Definition and Scope of Predictive Analytics

Predictive analytics leverages statistical techniques, algorithms, and machine learning to forecast future events or behaviors. In predictive analytics, past and current data are input into models that then generate and test multiple guesses or hypotheses to produce risk determinations. The discussion of predictive analytics focuses on its application in public health, facilitating the anticipation of emerging health threats. In public health, predictive analytics is multidisciplinary, drawing on expertise from statistics, computer science, and epidemiology. Public health benefits greatly from the timely and accurate analysis of data for making quick evidence-based decisions rather than anecdotal or intuition-based assumptions [5, 6]. Predictive analytics forms the highest layer of analysis with the potential of proactively anticipating adverse health outcomes and engaging in activities that may make a difference. The three general layers of analytics visualization informatics used in data analysis are: 1) Descriptive Analytics: answering the question—what happened? 2) Diagnostic Analytics: answering the question—why did it happen? 3) Predictive Analytics: answering the question—what is likely to happen? Descriptive analytics is an interpretation of data that highlights important characteristics, patterns, trends, and associations in data; diagnostic analytics are research methods used for the determination of potential causal relationships between identified patterns and outcomes, events, or associations in data. Predictive analytics is the analysis and interpretation of data using statistical or machine learning techniques to generate models and predict the likelihood of future events or outcomes. In healthcare, it involves using patients' electronic health records of the past and present to predict hospital readmission, adverse events, nosocomial infections, mortality rates, disease tendencies, individualized healthcare costs, and the efficacy of surgical management [7, 8].

Data Sources and Collection in Public Health

Starting with an understanding of the basics of public health and survival, data about demography, morbidity, mortality, and other disease indicators is quintessential. During primary data collection, surveys and interviews provide live data related to the demographic profile of the country's population. A wide range of data is collected live during interviews with women between 15 and 49 years of age. Secondary data are data that have already been collected by others. These can include hospital records, clinic records, morbidity records, police records, and death records. There is often a delay in the availability of data, due to which the deductions made are not "live" for predictive study [9, 10]. Data quality is important for analysis. Data source quality may range from clinical trials to insurance claims. In a predictive application, data your model didn't include may be lost forever. Make data reliable and ensure adequate collection checking. Time is essential in the availability of data. Hospital data is immediately available. Public health data are reliable but have delays. Emerging sources of data come from casual informants or professionals, and rapid screening tests using emerging technologies. Traditional sources of data in public health include telephone surveys, written surveys, personal interviews, and medical reports when possible. These models furnish an "average" for the complete data set and exclude the live diseased persons who are unable to or do not attend clinics. The rapidity of death syndromic surveillance commenced after a significant event. Emergent sources of data for predictive studies are now available from social media and mobile health applications. Although these data cannot just be evaluated without the patient's informed consent, a risk policy on data sharing between agencies has been made. It is preferred to integrate various types of databases for your studies. You should also try to evaluate as many new data mining tools and models. Data should be as fresh as possible to evaluate a service or an epidemic. Key Messages: Reliance on routine data and classically collected data for predictive analytics can lead to inaccurate predictions. Integrating varied data sets can enhance your analytics [11, 12].

Types of Data Sources

Different types of data sources can be used in the analysis. In general, for public health purposes, they are forensic or surveillance data. In this paper, we focus mostly on non-forensic data, which include clinical and epidemiological data and health survey data. Clinical data are collected in medical facilities and are mostly related to acute diseases. Longitudinal clinical cohorts are established for this type of data, and the European patients' data generate a substantial mass of such information. Health survey data are often interlinked with electronic health records or claims data rich in socioeconomic and environmental factors. Population-based patient registries or clinically mapped patient registries are other examples of clinical data. Non-forensic data that are usually not shared for public health currently are derived from wellness

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

programs, phone applications, sensors, etc. With some exceptions, all of the data sources are heterogeneous. Most of them are managed by different governmental institutions or agencies in charge of public health in individual countries. Some are managed by international platforms, which mostly cluster data collated by international or supranational regional or national public health agencies. A long list of governmental and non-governmental data repositories can be found [13, 14]. In public health, epidemiological or clinical information generated from real-time data sources that directly reflect the status of the event can be used for early effect detection, selection of test populations, or response measure monitoring. Historical data sources have been used as long as public health has existed. The oldest registries include death registries and communicable disease registries. From the very beginning, the value of the geographic presentation of the information on communicable disease registries was recognized in the form of cholera and plague geographic mapping. Geospatial data nowadays must be gathered from very diverse sources. Studying the Internet for disease tracking using a broad information-monitoring platform based on a range of themes is a relatively new option. Collecting and processing data manually from the interviews was very time-consuming. Nowadays, we can use web-based interviews or questionnaires that save time and also ensure a standardized approach, yet the challenges in that are the multiplicity of data sources, data integration, real-time monitoring, and provision of feedback. Data aggregated from diverse sources differ in terms of representativeness of the population, data quality, and available data sources. Public health stakeholders interested in data to be analyzed should collaborate with institutions, organizations, communities, etc., having a reporting system based primarily on their social mission [15, 16].

Statistical and Machine Learning Models for Disease Outbreak Prediction

Forecasting disease outbreaks can involve analyzing past patterns, developing predictive models within a given context, and validating predictions. In contrast to traditional statistical models that require data conforming to certain characteristics, predictive disease models tend to involve large and complex multivariate datasets and have shifted towards machine learning methods. For public health, machine learning is typically used to generate indicators or scores that can be interpreted in the context of the public health problem at hand. Most machine learning models can obtain checks and metrics, especially in cases of changing datasets obtained from real-time feeds and databases. These models can be categorized into three groups. The first group contains statistical and machine learning techniques that aim to describe the frequency distribution and the accumulation of disease counts over time. This includes time series analysis and regression models for infectious disease surveillance. The second group tries to predict the severity of the outcome of an event in the future. Hence, they are not used for making real-time forecasts. The third group consists of classification algorithms, which have an outcome in the form of classes labeled as either one event or a non-event. This includes machine learning algorithms that lend themselves to detecting either rare or emerging patterns from a data stream. In public health, this mainly seeks the development of methodologies that help in real-time early detection of outbreaks of infectious conditions and forecasting their impacts [17, 18]. Models that detect outbreaks provide an assessment of past outbreaks rather than a prediction. Similarly, models that report "sentinels" also do not forecast. They report the status or condition of a predetermined criterion. The primary advantage of using such detection models lies in their ability to detect outbreaks of unknown conditions or modes of attack promptly. Until a few years ago, the use of statistical tools for forecasting the significance of outbreaks or their trends was a mere dream for both developing and developed countries. Advances in statistical and machine learning techniques have led to the development of efficient forecasting tools. The main advantage of such techniques is that these models are dynamic and adaptive to the data, provided the nature of the model does not change. An accurate understanding of the problem and available resources can help us decide and develop the best model using the most appropriate method. Time series analyses have been applied to monitor and forecast the spread of respiratory and gastrointestinal diseases such as whooping cough and influenza from various countries around the globe. These models are not restricted to indicators from human illness and can also predict the onset of epidemics for infectious and non-infectious diseases, ranging from animal warts outbreaks to chicken-related illnesses. Some of the following models have handled more than 1,000 outbreaks per day and thus are proficient in segregating true alarms from this range of outbreaks. Some of these models also provide measures of uncertainty for these alarms and have been applied to some biological warfare scenarios where resources were stretched and hence provided never-reported cases of dead bodies being used to check for outbreaks. A few of the mechanisms used by these modeling tools include the use of Granger causality to find the linkages and

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

trigger epidemic models. Some of these tools also use adaptive models that learn from previous thresholds and hence can be used in varied conditions [19, 20].

Commonly Used Models

A variety of machine learning and statistical models have been applied to the problem of capturing heterogeneities in the data associated with disease outbreaks, with many of these models combining data types or forecast methods.

SIR: This process-based model includes susceptible, infectious, and recovered states and can be expanded to include exposed or severe compartments of disease. Application: The SIR model has been used to predict outbreaks of diseases. Special Characteristics: This model operates at the individual level and can simulate the course of an outbreak when relevant data are unavailable. Applications: SIR-like models have been used to forecast infectious disease outbreaks, including influenza and cholera. Special Characteristics: This method may be more versatile for non-herd diseases. Case Example: A forest distal extreme learning machine, which combined a random forest with neural network hidden layer computations, concluded that their model outperforms the current state-of-the-art hindcasting model [21, 22].

Random Forests: A statistical model designed to reduce overfitting to a random subset of the original data.

Application: This model is commonly used in disease ecology and forecasting. Special Characteristics: Recent research suggests that the random forest does not improve forecasts over autoregression for known mechanisms of disease systems. Applications: In bioterrorism literature, a rolling self-organizing max TSS neural network was used to predict smallpox outbreaks. Special Characteristics: The network size was increased optimally to include 36 agents, but the method was not extensively validated. Case Example: Logistic regression was used to forecast an anthrax outbreak. Special Applications: Research used a hybrid cost of illness approach to estimate the marginal cost impact of a smallpox outbreak [23, 24].

Case Studies and Applications in Disease Outbreak Prediction

To demonstrate how data science approaches have been or could be used operationally to predict disease outbreaks, we present key case studies. This review focuses on: (1) how different data sources have been leveraged to forecast outbreaks of different diseases and in different settings; (2) the upstream and downstream implications arising from predictive approaches. The review also identifies challenges to translating effective predictions into operational public health responses [25, 26]. This case study reflects an application of data science predictive methods within the public health sphere to "predict the unpredictable" — emerging infectious diseases. In contrast to the preceding examples, this study adopted a more formal modeling approach that utilized multiple data sources and integrated models. Likely due to the novelty and uncertainty of influenza during the time of this study, no formal evaluation was conducted. The last part of this study is a review of the potential impacts of using such predictions with public health practice. In some cases, the methods were not (or could not be) formally evaluated, and as such, no performance results are presented. In others, successful predictions were obtained to various degrees of success and for more advanced time pairs. Overall, several themes emerged from this review for predicting emerging infectious diseases: (1) predictions can affect the preparedness and response measures implemented in health; (2) timely prediction is imperative; (3) collaboration between decision-makers and data scientists is critical to making predictions actionable [27, 28].

Challenges and Ethical Considerations in Predictive Analytics for Public Health

Despite the potential of predictive analytics in public health, its use presents several challenges as well as ethical considerations. One of the main technical problems is the quality of the data used by predictive models, as well as the limited accuracy of their predictions. Added to that, prediction models are typically developed on single data sources, making it difficult to integrate predictions obtained from different sources of heterogeneous data. Moreover, there is a risk of misinterpretation of the results: predictive analytics can identify associations but do not decode mechanisms of causation, possibly leading to potentially inappropriate public health decisions. There are also technical challenges including data quality, inadequate performance of predictive models, the dynamic nature of populations and pathogens, limited generalizability of findings, difficulties in the integration of predictive scores in health practice, and the complex relationship between theory and data applied in predictive modeling [29, 30]. In addition to these, many ethical issues have also been raised, such as privacy legislation or recommendations, concerns about informed consent, and the quality and integrity of data for which individuals or societies are responsible. Concerning health domains, crucial are possible biases and the

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

opaqueness of algorithms. Critics often point to the risk that this mechanism will operate to the disadvantage of groups already burdened by a social and economic burden. From a public health perspective, transparency is crucial in the use of data in order to establish and maintain public trust. Especially for, but not limited to, problems of infectious diseases, different social, economic, and structural disparities should always be considered. Finally, the use of data mining has some considerations that make it difficult both conceptually and operationally in practice. Standardizing predictive analytics in the public domain can ensure both high-quality data and a high degree of validity of such knowledge. It is crucial to recognize the role of predictive analysis for public health as an area of promoting stakeholders' competence and an evidence-based public domain. It provides important knowledge to guide or offer policy and decision-making. Moreover, best practices and ethical guidelines can also offer support for researchers, funders, policymakers, journal editors, and referees in public health to balance technological developments with ethical practices. It can highlight the need to integrate predictive data analysis into interventions. There is a need to include predictive models in data systems to increase the overall understanding of health and risks. It may also help prepare future health professionals. In all the toolkits mentioned above, policymakers, health practitioners, and professionals use predictive models for knowledge exchange and decision-making. Setting out an ethical framework aims to ensure greater protection and commitment to the rights and public interest of individuals.

CONCLUSION

Predictive analytics plays an important role in modern public health by enabling the early detection and prevention of disease outbreaks. By leveraging diverse data sources, statistical methods, and machine learning models, public health officials can make informed decisions to mitigate health crises. However, the field faces challenges, including data quality, ethical concerns, and model reliability. Addressing these challenges through standardized frameworks, interdisciplinary collaboration, and ethical guidelines will enhance the effectiveness and credibility of predictive analytics. As technology advances, integrating predictive analytics into public health systems will be vital for improving global health security and preparedness against future outbreaks.

REFERENCES

1. Sheng J, Amankwah-Amoah J, Khan Z, Wang X. COVID-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions. British Journal of Management. 2021 Oct;32(4):1164-83. wiley.com
2. Sarker IH. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science. 2021 Sep;2(5):377.
3. Alsunaidi SJ, Almuhaideb AM, Ibrahim NM, Shaikh FS, Alqudaihi KS, Alhaidari FA, Khan IU, Aslam N, Alshahrani MS. Applications of big data analytics to control COVID-19 pandemic. Sensors. 2021 Mar 24;21(7):2282. mdpi.com
4. Zeng D, Cao Z, Neill DB. Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control. InArtificial intelligence in medicine 2021 Jan 1 (pp. 437-453). Academic Press.
5. Nwosu NT, Babatunde SO, Ijomah T. Enhancing customer experience and market penetration through advanced data analytics in the health industry. World Journal of Advanced Research and Reviews. 2024;22(3):1157-70. wjarr.co.in
6. Rahmanti AR, Ningrum DN, Lazuardi L, Yang HC, Li YC. Social media data analytics for outbreak risk communication: public attention on the "New Normal" during the COVID-19 pandemic in Indonesia. Computer Methods and Programs in Biomedicine. 2021 Jun 1;205:106083. nih.gov
7. Nancy AA, Ravindran D, Raj Vincent PD, Srinivasan K, Gutierrez Reina D. Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. Electronics. 2022 Jul 22;11(15):2292. mdpi.com
8. Cao Q, Zanni-Merk C, Samet A, Reich C, De Beuvron FD, Beckmann A, Giannetti C. KSPMI: a knowledge-based system for predictive maintenance in industry 4.0. Robotics and Computer-Integrated Manufacturing. 2022 Apr 1;74:102281. google.com
9. Zajacova A, Margolis R. Trends in disability and limitations among US adults age 18-44, 2000-2018. American Journal of Epidemiology. 2024 Aug 12:kwae262.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

10. Amezcua-Castillo E, González-Pacheco H, Sáenz-San Martín A, Méndez-Ocampo P, Gutierrez-Moctezuma I, Massó F, Sierra-Lara D, Springall R, Rodríguez E, Arias-Mendoza A, Amezcua-Guerra LM. C-Reactive Protein: The Quintessential Marker of Systemic Inflammation in Coronary Artery Disease—Advancing toward Precision Medicine. Biomedicines. 2023 Sep 2;11(9):2444. mdpi.com
11. Strieth-Kalthoff F, Sandfort F, Kühnemund M, Schäfer FR, Kuchen H, Glorius F. Machine learning for chemical reactivity: the importance of failed experiments. Angewandte Chemie International Edition. 2022 Jul 18;61(29):e202204647. [HTML]
12. Nam D, Chapiro J, Paradis V, Seraphin TP, Kather JN. Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. Jhep Reports. 2022 Apr 1;4(4):100443.
13. Christian H, Suhartono D, Chowanda A, Zamli KZ. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data. 2021 May 17;8(1):68.
14. Cvetkov-Iliev A, Allauzen A, Varoquaux G. Analytics on non-normalized data sources: More learning, rather than more cleaning. IEEE Access. 2022 Apr 18;10:42420-31.
15. Hoban S, Archer FI, Bertola LD, Bragg JG, Breed MF, Bruford MW, Coleman MA, Ekblom R, Funk WC, Grueber CE, Hand BK. Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. Biological Reviews. 2022 Aug;97(4):1511-38. wiley.com
16. Amdaoud M, Arcuri G, Levratto N. Are regions equal in adversity? A spatial analysis of spread and dynamics of COVID-19 in Europe. The European Journal of Health Economics. 2021 Jun;22(4):629-42. springer.com
17. Udegbe FC, Nwankwo EI, Igwama GT, Olaboye JA. Real-time data integration in diagnostic devices for predictive modeling of infectious disease outbreaks. Computer Science & IT Research Journal. 2023 Dec;4(3). researchgate.net
18. Peng Y, Li C, Rong Y, Pang CP, Chen X, Chen H. Real-time prediction of the daily incidence of COVID-19 in 215 countries and territories using machine learning: model development and validation. Journal of Medical Internet Research. 2021 Jun 14;23(6):e24285. jmir.org
19. Barrera-Animas AY, Oyedele LO, Bilal M, Akinosho TD, Delgado JM, Akanbi LA. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications. 2022 Mar 15;7:100204. sciencedirect.com
20. Spiliotis E, Makridakis S, Semenoglou AA, Assimakopoulos V. Comparison of statistical and machine learning methods for daily SKU demand forecasting. Operational Research. 2022 Jul;22(3):3037-61. academia.edu
21. Turkyilmazoglu M. Explicit formulae for the peak time of an epidemic from the SIR model. Physica D: Nonlinear Phenomena. 2021 Aug 1;422:132902.
22. Lazebnik T. Computational applications of extended SIR models: A review focused on airborne pandemics. Ecological Modelling. 2023 Sep 1;483:110422.
23. Yoon J. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. Computational Economics. 2021 Jan;57(1):247-65.
24. Gupta R, Pierdzioch C, Salisu AA. Oil-price uncertainty and the UK unemployment rate: A forecasting experiment with random forests using 150 years of data. Resources Policy. 2022 Aug 1;77:102662.
25. Khodadadi E, Towfek SK. Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System. Journal of Intelligent Systems & Internet of Things. 2023 Nov 1;10(1). [HTML]
26. Stockdale JE, Liu P, Colijn C. The potential of genomics for infectious disease forecasting. Nature Microbiology. 2022 Nov;7(11):1736-43.
27. Knight GM, Glover RE, McQuaid CF, Olaru ID, Gallandat K, Leclerc QJ, Fuller NM, Willcocks SJ, Hasan R, Van Kleef E, Chandler CI. Antimicrobial resistance and COVID-19: Intersections and implications. Elife. 2021 Feb 16;10:e64139. elifesciences.org
28. Shmueli L. Predicting intention to receive COVID-19 vaccine among the general population using the health belief model and the theory of planned behavior model. BMC public health. 2021 Dec;21:1-3.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

29. Nguyen QH, Ly HB, Ho LS, Al-Ansari N, Le HV, Tran VQ, Prakash I, Pham BT. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering. 2021;2021(1):4832864. wiley.com
30. Murugan Bhagavathi S, Thavasimuthu A, Murugesan A, George Rajendran CP, Raja L, Thavasimuthu R. Retracted: Weather forecasting and prediction using hybrid C5. 0 machine learning algorithm. International Journal of Communication Systems. 2021 Jul 10;34(10):e4805. [HTML]

CITE AS: Ivan Mutebi (2025). Predictive Analytics in Public Health: Anticipating Disease Outbreaks. Research Output Journal of Engineering and Scientific Research 4(1): 43-49. https://doi.org/10.59298/ROJESR/2025/4.1.4349

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.