Spatiotemporal Disease Spread Modelling and Intervention Efficacy Estimation: AI-Based Predictive Analytics for Epidemiological Modelling and Public Health Planning

Dr. Amira El-Shafei, Associate Professor of Computer Science, Ain Shams University, Egypt

1. Introduction to Epidemiological Modeling

Epidemiological modeling, the development of software models of infectious disease spread and their application to specific diseases, assists in the control and prevention of infectious disease in multiple ways. It seeks to facilitate basic scientific understanding of disease, estimate and track the epidemiological burden of disease, and investigate the likely impacts of potential interventions. It also provides a relative assessment of the quality of the evidence regarding diseases. Moreover, by simulating alternative potential interventions, it may contribute to the more rational allocation of resources for empirical research. Through these processes of data simulation and validation, epidemiological modeling is increasingly being used to inform real-time public health decisions and advance the development of data-informed early action systems for health. Epidemiological modeling is at the crossroads of several disciplines, combining biological understanding of infection and disease pathways with mathematical and statistical approaches to model building. As computational power has increased, models have emerged as powerful frameworks for weaving together biological, statistical, and behavioral data to form coherent descriptions of epidemics. Methods from network science are also increasingly underpinning epidemic modeling work. Despite these advances, many challenges remain for predictive modeling in epidemic settings. Disease dynamics are exquisitely sensitive to particular local social and environmental conditions. Few disease transmissibility parameters can be usefully parameterized in the generic models used, despite great efforts at large-scale simulation. Moreover, much of the critical disease intelligence required for effective early action can only be obtained with epidemic-scale involvement of affected communities. Meta-analyses of RCTs and

natural experiments focus on a particular outcome measured as part of disease symptomatology rather than the quality of the evidence in our understanding of disease dynamics. There is, thus, a pressing need for faster, more adaptive system approaches to disease early warning and time-sensitive systems for policy-sensitive disease intelligence.

1.1. Overview of Epidemiology and its Importance

1.1. Overview of Epidemiology

This section provides an overview of epidemiology, a field dedicated to studying the patterns, causes, and impacts of health and disease states in human populations. By using data collected at population levels, we can assess trends and patterns and use this information to determine effective interventions and make meaningful predictions. One objective of epidemiology is to describe patterns of disease. Another objective is to understand the determinants of a particular health outcome, including the risk factors associated with it. When we collect data to understand causes, epidemiologists identify the risk factors for diseases, the groups that are most affected, and attempt to determine which factors are influencing disease risk.

Epidemiology is also the foundation of public health, as it helps formulate policies and planning strategies for public health. For example, if cases of influenza are on the rise, public health agencies might increase their purchase orders for vaccines to ensure an adequate supply of vaccine is available when it is needed, which can prevent illnesses and save lives. If a particular population is shown to have high rates of diabetes, public health may target the population with programs designed to help prevent diabetes, such as nutrition or exercise programs. Conditions and diseases do not occur haphazardly. In order to truly address ill health, we also need to understand why and where disease patterns are occurring. These determinants are shaped by equity, human rights, and ensuring everyone has the ability to live a healthy life. If epidemiologists are to truly understand the causes of health and disease, we also need to assess social, community, and physical environments, as these factors are often determinants of the likelihood of disease, rather than biological factors.

It is not possible to accurately and meaningfully predict when the next pandemic will occur, what it will be, or how severe it will be. The 1918 influenza pandemic initially

appeared to be a more severe form of annual seasonal influenza. However, we now know that it resulted in the highest number of known influenza deaths in a single year; this projection was informed by surveillance and epidemiological studies. It was not foreseen by nor could it have been predicted by lawmakers and citizens, yet this pandemic has shaped public health, as the field grew out of the response to the 1918 influenza pandemic. Epidemiology is an early and essential public health science that informs policy, programs, and practices. Yet in our modern era, what some have labeled as an 'information revolution,' our capability to track, store, analyze, and generate datasets using computational means has been transformed. Our data generation has long outstripped our ability to analyze data using traditional epidemiological statistical methods, necessitating novel analytical techniques including machine learning strategies. Much of this modern era has seen an unprecedented capacity for data collection, generation, storage, and analysis, catapulting us into a world where we often do not want for data to analyze.

Epidemiology has traditionally been defined as the study of populations and the occurrence of health states. In the past decade, there has been a shift in this approach to elicit the relevance of epidemiology in studying social determinants of disease. For over three centuries, epidemiology has been an essential and evolving public health science. Epidemiologists traditionally utilize statistical tools to describe the patterns of disease and to identify causes or risk factors of epidemic events. While accurate at the time, these definitions do not characterize the development of epidemiology nor its role in public health. Epidemiology has always been a measurement tool and methodological framework for population health studies, but increasingly, with the advent of modern data science, epidemiology is required to adapt and integrate new technologies and methodologies into mainstream practice.

2. Machine Learning in Epidemiology

Machine learning has transformed the way in which we conduct epidemiological research. It has paved the way to depict a novel way of disease modeling that captivates the complex interrelationships between variables. Epidemiologists traditionally rely on stochastic and mathematical models, where events unfold according to statistical probabilities. Machine learning algorithms rely on data-driven models that develop over time by learning about the available data and developing predictive models based on

these datasets. Several researchers have employed machine learning algorithms to analyze large volumes of data and to predict the occurrence of events, including disease spread, hospital visits, and spikes in emergency traffic during extreme events. Moreover, AI-based algorithms can be used for the assessment of different intervention strategies through real-time evaluation of interventions in terms of speed and efficacy in curbing disease spread in the affected population.

Researchers have already integrated machine learning for the prediction of several diseases such as diabetes, cancer, mental disorders, and others. Two broad types of machine learning techniques, supervised learning and unsupervised learning, have made it possible to develop latent variable models like Gaussian mixture models and hidden Markov models for disaggregating the population based on the observations in the available data. The convolutional neural network is another AI technique that is used as a forecasting model that requires input corresponding to the disease spread in one or many areas along with their corresponding outcomes. Although such algorithms are not widely employed in epidemiological forecasting studies, their integration could prove to be effective in improving prediction accuracies. Indeed, the main advantage of using AI-based techniques in epidemiological forecasting is that their integration into existing models can considerably reduce the computational time of forecasting.

2.1. Applications of Machine Learning in Disease Prediction

Machine learning has been utilized for various applications, such as predicting outbreaks, the onset of a disease in a person, the number of cases, the intensity of the epidemic, the fatality rate of the disease, patient outcomes, and responses to existing treatments or drugs. It also includes predicting helpful responses to newly launched drugs, predicting toxicities, visualizing quantified data, and classifying known and unknown data. Machine learning techniques can:

▪ Predict the emergence of a novel influenza strain in a timely manner.

▪ Develop a simple diagnostic protocol for the early stages of measles, mumps, rubella, and varicella in children.

▪ Tailor surveillance programs to maximize information while minimizing resources to meet public health needs effectively.

▪ Predict the need for administration of the vaccine to reduce infection transmission in pedestrian populations during a bioterror event, smallpox outbreak, or pandemic.

▪ Identify candidate genes involved in complex disease processes in diabetes.

▪ Use syndromic surveillance data to analyze public health trends.

Key machine learning tools and frameworks have been used for automating the processes of prediction and increasing the efficiency of prevention. For the development of predictive modeling in the public health field, near real-time data processing has been shown to have the potential for providing better results. Machine learning approaches have made it feasible to predict the occurrence and significance of imminent epidemics, which is critical for determining disease control strategies. In particular, they are expected to have considerable relevance in predicting zoonoses in public health systems. The predictive information provided by epidemiological intelligence may be utilized for early response and prioritization if real-time models of infectious disease transmission are developed, generated, and verified. However, a number of ethical concerns may arise, especially in the utilization of machine learning predictions to propose interventions to prevent future infections. Accordingly, all perturbing efforts aimed at changing the flow of a public health crisis may need to be treated cautiously using ethical principles and include further research that adopts a wide viewpoint, including hierarchical analyses of effectiveness before implementing them as a public directive.

3. Data Sources and Preprocessing

A cornerstone of any epidemiological modeling activity is the pertinent available data regarding the disease and the population of interest. These datasets range from local to national-level time series of epidemiological indicators to environmental or demographic data useful to highlight relations with the disease of interest. From a surveillance standpoint, epidemiological datasets can be collected either at a nationwide level or through integration of local, ongoing surveys and information systems aimed at financial aid distribution. From a methodological point of view, spatially detailed epidemiological information in principle allows richer and more robust predictions because it includes additional signals. A recent review of modeling pandemic influenza places emphasis on the critical role of data in building both process- and data-driven predictive models. In summary, data issues play a major role in leading to improved

disease supervision and better detection of low-level epidemics in early stages, when spatial distribution of cases looks more like a random phenomenon.

The preprocessing of the data itself is concerned with any pre-analysis work that aims to optimize the querying and mining of the data into the data models that are part of the modeling pipeline. Specifically, preprocessing entails cleaning, transformation, and the development of necessary derived variables. The data is inspected for systematic inconsistencies arising from different surveillance systems and data collection procedures inherent to diverse healthcare practices in order to remove known biases from the analysis of the data. As an ongoing process, these data preprocessing plans also make use of the emergence of new technologies. Examples include the utility of remote sensing for high-frequency spatial information about certain zoonotically active organisms. In remote sensing particularly, as the spatial resolution of commercial satellites is increasing, this can now provide a robust and frequent enough data stream in high-resolution images or current and relevant data. Similarly, another new source of models to use is those from digitized health data. They can provide a large range of attractive features of patients at all population levels, which can significantly enrich reports for digital health data prediction. Despite the many advances, the usual problem in comparing quantities of data from regions is one of accessibility, so data from developing countries is subject to the data from disease-free zones. Feature reduction and feature selection are used practically to reduce the number of possibly useful features into models to help dimensional problems. In some cases, where levels of recorded features are at the detail of something like hourly measurements, these may be aggregated into more meaningful daily variables for modeling. In the epidemiological applications, features are curated from ten time series datasets into a single feature before adding spatial information to form an input to the neural network models.

3.1. Types of Data Used in Epidemiological Modeling

Epidemiological models can be created to focus on different determinants of disease, including predictive models of the incidence of disease, causal models to test hypotheses about health determinants, and decision models to evaluate alternative policy interventions or prevention strategies. We generally categorize the data required to parameterize these relationships into three sets: epidemiological/surveillance data, social and environmental data, and qualitative data from which we can learn more

about the spread of the disease, like who is the index case, who else is vulnerable, and who is the most likely to become infected. Quantitative variables contain observations that can be measured, counted, or somehow treated as numerical data. Categorical variables contain labels that place each observation into a particular group. Epidemiological/surveillance data, such as infection rates and the number of cases, are often reported by national or international surveillance networks. Notifiable or reportable diseases have been specified based on public health importance and the ability to prevent and control them. Disease case count data are supplemented by targeting integrated networks of public health, environmental, agriculture, and food monitoring systems to develop a single coordinated global surveillance network. Volunteered itinerary and activities data may be harvested by firms using global positioning system enabled smartphone data and/or by police and governments using video that records license plate numbers of passing cars. Health, social, and environmental variables come from many sources, including international, national, or regional agencies, or academic research. Social data include variables that influence health because they describe the relationships among people and societies. We can define these as social determinants of health or causes of the causes. Some of these variables are usually included in predictive models of infectious disease incidence. Others, such as mobility, contact patterns, quality of relationships in general, and trust in or suspicion of government, require more complex environment-response models to estimate their contributions to the spread of disease. One of the models used in the pandemic required census data, chain of transmission data, commuting data, and an estimate of the relationship between commuter flow and contact.

4. Modeling Techniques

Modeling techniques can be divided into traditional and modern machine learning models. The most traditional model in epidemiology is the susceptible-infected-recovered (SIR) model, a set of ordinary differential equations that consider reinfections or births of new individuals susceptible to the disease. These can be expanded to consider immunization and deaths, and can also be used for compartmental (susceptible-infected-susceptible, SI; susceptible-exposed-infected-susceptible, SEIS), individual (agent-based), or meta-population (spatial) levels of granularity. Very elaborate techniques are used, specifically the message-passing inference for estimating compact communities and expectation maximization for the efficient re-training of the

model regarding the identity of the disease carrier and semantics. They are relatively interpretable, so public health decisions can be based on them. However, the theories behind these traditional models are not grounded in data analysis, and the exploration of such models can be costly.

Studying disease outbreaks and their dynamics at local and global levels has been one of the main cornerstones that guide epidemiological predictions and measures to control outbreaks. Despite advantages, traditional methods are facing limitations when dealing with real-world problems, which are governed by complex nonlinearities, higher dimensions, and very large populations. Although data-driven approaches have been around for some time in general, putting data at the center of identifying and forecasting epidemics has been gaining popularity among data science and AI researchers. Ease of use, better ability to capture changes in the data, reduced strong assumptions about the form of the data, and better predictive ability are just some of the attractiveness and benefits of machine learning modeling. Machine learning models like recurrent neural networks, long short-term memory, and convolutional neural networks can capture the temporal elements many diseases follow, especially where diseases could incubate and even go into remission before full onset. In addition, as machine learning does not rely on differential equations, it is more flexible and can be applied to understanding and predicting phenomena at many different levels; in particular, considering disease dynamics at the individual level, rather than assuming "homogeneous mixing" traditionally documented in SIR-type models. Non-differential equation models are potentially more accurate for certain diseases and demographics. Hybrid models combine traditional epidemiological models with machine learning. These "cross-compartmental" approaches could indeed help to overcome some of the limitations of models that belong to a single category and to exploit the benefits of each. These models can provide not only more accurate results but, in some cases, additional insights.

4.1. Comparison of Traditional vs. Machine Learning Models

When comparing traditional epidemiological models to machine learning models, the strengths and weaknesses of each modeling approach become apparent. Traditional models, established over fifty years ago, are built on a strong epidemiological theory and are widely used in practice. These traditional models are straightforward to interpret, require a minimal amount of computer resources, and are easily utilized by

healthcare professionals. However, traditional models are poorly suited to the representation of complex non-linear relationships between disease outcomes and risk factors, which are commonly observed in large epidemiological datasets. Given the advent of big data in public health and the plethora of latent variables that could affect the contraction and propagation of pathogens, we establish that more complex machine learning models have the potential to discern these complexities. Regression models are limited to linearity and interaction specifications, while decision trees are limited by the values of predictors in the available data, making predictive outcomes questionable when making predictions.

In reality, both traditional epidemiological models and machine learning models have their strengths and weaknesses, and the choice between these two approaches greatly depends upon the scope and format of the epidemiological question being addressed. For instance, modeling with a regression model that includes interaction in known or hypothesized relationships would be the most pragmatic approach for testing the results of an existing epidemiologic theory. Similarly, machine learning could be reasonably employed for the discovery of interesting patterns that could later be confirmed, refuted, or revised in future theory-driven work. Hybrid modeling, including statistical and machine learning approaches, has also been adopted to analyze disease studies, where some factors are causally known, but the remaining variables are considered as predictors of prognosis at the same time and no mechanism of action is provided. Hybrid models where machine learning algorithms investigate all variables and provide a set of risk factors controlled for the primary purpose of the study can be particularly useful.

5. Case Studies and Real-World Applications

During the COVID-19 pandemic, numerous case studies appeared showcasing the important application of AI-based predictive analytics in epidemiology. Infectious Disease Modeling achieves more accurate forecasts when using data-driven approaches. They adopted a data-driven approach employing a Markov Chain Monte Carlo estimator to forecast seasonal and pandemic influenza in thousands of US cities. Standard statistical methods typically use historical and current disease data to forecast, but Seasonal Flu Mobility uses the spatiotemporal dynamics of infected individuals' cell phone data to predict the next week of influenza. Evaluating their model on real-world

data, their approach increased forecast accuracy over the typically employed model. Another case study of cell phone data and modeling in a communicable disease—tuberculosis (TB) shows great promise in improving future predictions of TB prevalence and can be used to examine qualitatively how societal changes affect disease prevalence. This is important in disease control and implies our model has value in public health planning.

The case study of these works, along with similar ongoing efforts, exemplifies data-driven approaches that increase forecast accuracy. The success of these case studies suggests that these types of results can be used in outbreaks to better allocate resources and model outbreak spread to better protect populations. Most of these case studies associate one or more experts in intrinsic epidemiological modeling with experts in predictive modeling, with a goal that the predictive tool is to be tested in the field to better forecast real-world disease outbreaks. Similar to these implemented models in forecasting epidemic peaks and resource allocation, they used cell phone data and advanced modeling to better model the spatiotemporal spread of disease. While useful tools, modeling a single outbreak can be affected by many limitations that modelers face in data, timing, and other constraints. A useful next study would be a strategic model for using models in calls for increased vaccination in response to outbreak indicators and early predictions. Further operational tests also need to be conducted for bettering and policymaking widespread use in robust forecasting.

5.1. Successful Implementations in Disease Forecasting

5.1 Successful Implementations in Disease Forecasting

Many of the projects listed above, despite their varying degrees of success, leverage machine learning algorithms or other state-of-the-art analytic approaches for improved disease outbreak prediction over traditional methodologies. In a meta-analysis of outbreak prediction models, machine learning-based models were 10%–35% more likely to predict the correct outbreak start date relative to classical time series-based methods. The added value of these predictive modeling approaches has not been limited to improved accuracy, though, with successful forecasts leading to effective and informative adaptive response strategies in practice. Implemented in the Philippines since 2012, the Dengue Sentinel Surveillance, which employed both epidemiological and remote sensing data to accurately forecast dengue outbreaks four months in advance,

led to further studies on employing similar approaches in other select countries. Also cited in a meta-analysis is a study from New York City, which, with a random forest model, made accurate predictions of some diseases about one week before the Department of Health in the region was able to make such predictions.

To improve forecast accuracy and timeliness during these implementations, many practitioners have chosen to integrate diverse data sources into their models. In particular, the use of non-traditional data has demonstrated utility in improving model forecast accuracy. In Sweden, the European Centre for Disease Prevention and Control makes vaccine efficacy estimates using media and population surveillance data, as this unpublished data was found to be more frequently updated and had better geographic granularity compared to traditional reporting for seasonal flu vaccine strain prediction. Similarly, an experimental highly pathogenic avian influenza forecast conducted in Vietnam tested the inclusion of wastewater sampling results into their predictive model. The wastewater forecast was found to offer a moderate improvement in forecast accuracy when used in combination with reporting metrics alone, with the greatest impact being realized when other data sources reported moderate to high risk of pathogen spillover from humans to wildlife. Both studies demonstrate that, within the scope of the underlying application, these atypical data streams can lead to improvements in model performance on an evolving public health surveillance problem. Many of the professionals who contributed to the outbreak prediction models cited throughout this paper cite several implementation challenges and lessons learned that can help inform new teams and projects in developing predictive models. One resounding lesson is the importance of model validation across diverse datasets, as models with consistent measures of predictive capacity across diverse sample populations of varying size, outbreak patterns, and biases are more likely to have general predictive value and less likely to perform well only under the conditions represented by the training data.

6. Future Direction

The past two decades have seen continued advancement and innovation in AI and epidemiological models. Emerging AI technologies such as self-supervised learning, transformers, and graph neural networks enable learning from general time series trends and iteration with few labeled data, as well as the discovery of complex

interactions within crowdsourced data and the identification of historical patterns or rules for secondary transmission. Nevertheless, challenges always abound in the development of AI-enabled surveillance and prediction, especially in the following aspects: 1) Increase in computational power and data availability. These expand the use of agent-based models and explicit mechanistic models, which cost intractable parameters and real-time infectivity pathways until recently. 2) Real-time integration of further data. Expanded computing ability and crowdsourced platforms provide more geospatial and temporal real-time behavioral and disease data for more detailed and faster updating of the AIs in the simulation models to improve responsiveness and fine-tuning of regional alerts and interventions. 3) Developing a new set of baseline and critical peak algorithms. All of the AI models, which were designed for straightforward and quick infection rate projections or case surveillance, clearly require substantial finessing to become part of an integrated computational predictive algorithm useful for predicting complex alerts such as high-density school opening and closing days and for the management of regional herd immunity allowing temporary increases in the infectious population as a strategy to reduce permutations of potential transmission impact zones. 4) Further development and data mining of tag-based population simulation. Such approaches will surely expand our understanding of epidemiology into the next sphere of interactive agent-based population disease models. 5) Ethical considerations and addressing legal and long-term community concerns with respect to crowdsourced AI integrated into the public health system. AI has the capability to create population panic with the consideration for the social and economic impacts if real-time crowdsourced data and alert suggestions are constantly misrepresented to the general public.

In the near future, we anticipate the joining and divergence of these predictive endeavors by infectious disease modelers and gravitational epidemiologists to tackle the outstanding epidemiological world phenomena in a structured collaborative way. In fact, such methodologies widening across the non-medical modeling community are already in progress, in an attempt to address complex environments and social behavior in public health. This will start to address challenges such as integrating on-the-ground action disease transmission data into models and advocacy for the development of society-wide agreed-upon crowdsourced project-tagged educational international initiatives and research. It will also fulfill more high-throughput criteria and create

assessment checks and balances to ultimately provide epidemiology ethical acceptance and recommendations for positive health social good use. It will bring together the correct people from around the world from diverse modeling backgrounds. It will also gather an interdisciplinary panel that is free of funding and social biases, combining elite scientific and long-term environmental risk assessment and mitigation from leading experts in climate change, nuclear medicine, and pollution source modeling risk assessment fields in addition to immunologists, virologists, computational modelers, and epidemiologists.

7. Conclusion

In this article, we illustrate how epidemiological modeling has evolved over the past few decades, from using purely deterministic policy-agnostic models to embracing increasingly complex models with the aim of integrating larger amounts of information. Our models are getting better and better, capable of incorporating the incredibly complex behavior of individuals and their interactions and, thus, more and more complex and data-demanding. As a result of this increasing complexity, at each step of this revolution, our needs for large-scale data analytics and computational resources have increased and will continue to do so. As we are increasing the complexity of our models and the level of detail they incorporate, the potential to apply policymaking strategies over these models so that they produce robust behaviors, that are as good as possible while not overfitted to the patterns present in the input data, is diminished, and the potential that these models are not general in time and space is increased. In practice, particularly since the pandemic, AI-based machine learning and predictive analytics techniques are becoming more and more personal in many areas, including the epidemiological domain.

In summary, we argue that leveraging AI-sensitive predictive analytics for the purposes of epidemiological modeling and parameter forecasting offers the public health community the ability to gain agility in responsiveness to outbreaks or pandemics. This toolkit, combining different AI algorithms, is key for monitoring disease and evaluating non-pharmaceutical interventions. We also discuss the opportunities and challenges these methods bring. In doing so, it is important to remember that both modelers and data scientists are continuous learners and adapters, and they do so best in concert: public health and infectious disease experts provide transversal knowledge while data

scientists leverage technological breakthroughs. Ethically, the use of AI, as it brings benefits, also brings responsibilities. Research directions from here are the following: innovation in terms of alternatives; integration of additional information sources; and assurance of the accuracy of these tools.