Journal of Advances in Medicine and Medical Research

33(5): 8-21, 2021; Article no.JAMMR.65476
ISSN: 2456-8899
(Past name: British Journal of Medicine and Medical Research, Past ISSN: 2231-0614, NLM ID: 101570965)

Literature Review on Epidemiological Modelling, Spatial Modelling and Artificial Intelligence for COVID-19

Danial Saraee[1*] and Charith Silva[2]

[1]School of Medicine, UHW Main Building, Heath Park, University of Cardiff, England.
[2]School of Science, Engineering and Environment, University of Salford, England.

Authors' contributions

This work was carried out in collaboration among both authors. Author DS designed the study, performed the data collection, analysis, interpretation and wrote the first draft of the manuscript. Author CS contributed on data analysis and interpretation. Author DS managed the literature searches. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMMR/2021/v33i530841
Editor(s):
(1) Dr. Rameshwari Thakur, Muzaffarnagar Medical College, India.
(2) Dr. Emin Umit Bagriacik, Gazi University, Turkey.
(3) Dr. Syed Faisal Zaidi , King Saud bin Abdulaziz University for Health Sciences, Kingdom of Saudi Arabia.
Reviewers:
(1) Zakir Hussain, FELTP,Pakistan.
(2) Jean Pierre NAMAHORO, China University of Geosciences, China.
(3) R Kesavan, University of Jaffna, Sri Lanka.
Complete Peer review History: http://www.sdiarticle4.com/review-history/65476

Review Article

Received 25 January 2021
Accepted 07 March 2021
Published 16 March 2021

ABSTRACT

Introduction: Following the outbreak of Coronavirus (COVID-19) in Wuhan, China in December 2019, the World Health Organisation (WHO) has declared this infectious disease as a pandemic. Unlike previous infectious outbreaks such as Severe Acute Respiratory Syndrome (SARS) and Middle Eastern Respiratory syndrome (MERS), the high transmission rate of COVID-19 has resulted in worldwide spread. The countries with the highest recorded incidence and mortality rates are the US and UK.

Rationale/Objective: This review will compare studies that have used epidemiological models for disease forecasting and other models that have identified sociodemographic factors associated with COVID-19. We will evaluate several models, from basic equation-based mathematical models to more advanced machine-learning ones. Our expectation is that by identifying high impact models used by policy makers and discussing their limitations, we can identify possible areas for future research.

**Corresponding author: E-mail: SaraeeD@cardiff.ac.uk;*

Evidence Review: The bibliographic database google scholar was used to search keywords such as 'COVID-19', 'epidemiological modelling' and 'machine learning'. We examined data review articles, research studies and government-released articles.
Results: We identified that the current SEIR model used by the UK government lacked the spatial modelling to enable an accurate prediction of disease spread. We discussed that machine-learning systems which can identify high-risk groups can be used to establish the disparities in COVID-19 death in BAME groups. We found that most of the data hungry AI models used were limited by the lack of datasets available.
Conclusion: In conclusion, advances in AI methods for infectious disease have overcome challenges presented in mathematical models. Whilst limitations do exist, when optimised, these highly advanced models have a great potential in public health surveillance, particularly infectious disease transmission.

Keywords: COVID-19; machine-learning; artificial intelligence; spatial modeling; epidemiological modeling.

1. BACKGROUND

Coronavirus disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) and originated in the Wuhan, Hubei Province, China [1]. Since the first cases traced back to the Huanan wholesale seafood market, the deadly virus has spread worldwide with just under 110 million cases globally and approximately 2,434,048 COVID-19 related death as of 18th February 2021 [2]. Despite the development of several vaccines, challenges with widespread distribution and extensive herd immunity evince that efforts to minimise disease spread are still a global priority. As a result, data modelling and big data have become the forefront in the battle against COVID-19. By predicting disease spread, forecasting the effects of social distancing measures and identifying high-risk regions, these tools have the potential to protect the vulnerable population in both the current crisis and future disease outbreaks.

The aim of this literature review is to identify the benefits and limitations of epidemiological models currently used in disease forecasting. In addition, we will discuss the current and future role of AI and machine learning in COVID-19 epidemiology and future public health interventions.

2. CURRENT EQUATION-BASED EPIDEMIOLOGICAL MODEL FOR COVID-19

The most frequently used epidemiological model worldwide is the susceptible-exposed-infected-removed (SEIR) epidemiological model. This compartmental model has been used to quantify transmission dynamics in order to derive epidemic curves and to observe the impact of control measures placed by the government. The SEIR model in short places a population in four different states: susceptible (S), exposed (E), infected (I), and recovered (R) at a given time (Fig. 1) [3]. Assuming the population is constant based on equal birth rates (α) and death rates (γ), the model predicts the number of individuals in each compartment by integrating differential equations formed [4]. The predictions assume that the whole population is considered susceptible (s). The rate at which a susceptible (S) individual enters the exposed (E) compartment is β, which is the number of contacts a susceptible individual has with infected individuals per unit time. These are dynamic numbers due to the variance of population patterns in the susceptible and infectious group. Social distancing intends to reduce the β value by limiting contact between susceptible and infectious individuals. The rate at which an exposed individual becomes infectious is the average disease incubation time (ε^{-1}), which is constant for a specific disease. Finally, the rate at which an infectious (I) individual enters the removed (R) compartment is the average duration of infection (γ) until the person has recovered or died. This number is intrinsic to the specific disease and remains constant.

Using these equations from the SEIR model, the reproductive ratio (R_0) can be produced (Fig. 1).

$$R_0 = \frac{\beta\epsilon}{(\epsilon+\mu)(\gamma+\alpha+\mu)}$$

The reproductive ratio indicates the average secondary infections arising from one infected

person in an unvaccinated population initially free of disease [5]. This estimated R_0 is frequently and ubiquitously used by governments and WHO to measure the spread of disease and allow early strategy planning. In the UK, the R_0 is referred to in the daily Coronavirus press conference updates. It is used as a quantitative measurement to determine when the virus transmissibility is low enough to ease lockdown restrictions [6]. Undoubtedly, the SEIR epidemiological model has been incredibly useful for government and policy makers worldwide to measure the effectiveness of control measures.

A review by The London School of Hygiene & Tropical Medicine (LSHTM) [7] discusses the use of the SEIR model by the UK government. The team analysed data from Wuhan and simulated control measures into the model. A limitation of the SEIR model is that contact between individuals is not considered, which is relevant in healthcare settings where workers are in close contact with confirmed cases. This can be overcome by an agent-based model, where household and healthcare infrastructure domains can be incorporated into the model to provide explicit detail of disease transmission in these different settings.

Another challenge faced is that the SEIR model assumes that the population recovered from the disease cannot become reinfected due to lifelong immunity. Whilst evidence from the Flu Watch cohort study shows reinfection from the same strain of coronavirus within the same season is highly unlikely, there is no robust data to support this [8]. In fact, diseases such as malaria and cholera show a waning of immunity after recovery from infection [9]. In consideration of the somewhat limited evidence for lifelong immunity in recovered patients, the SEIR model must acknowledge the possibility of reinfection and thus all recovered individuals should return to the susceptible compartment after a given time (ξ).

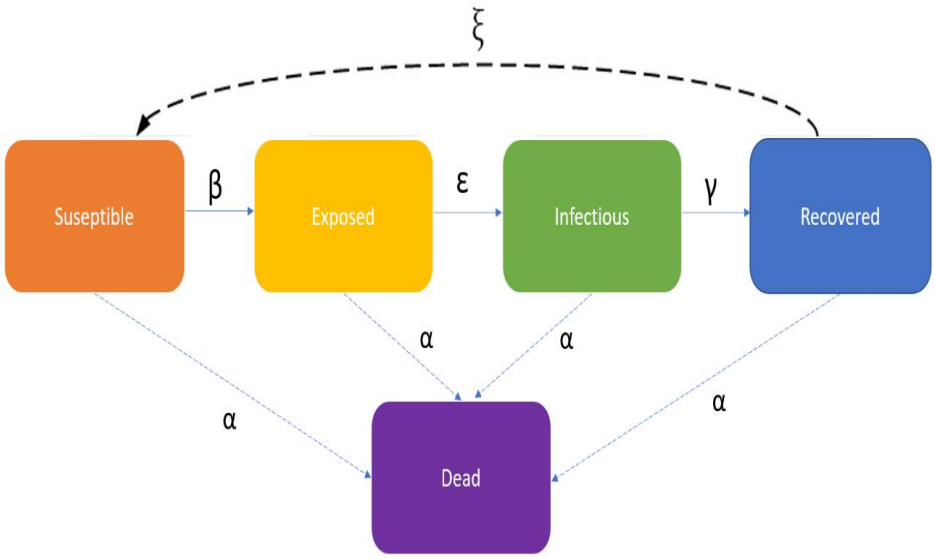

Fig. 1. SEIR epidemiological model of disease spread: This epidemiological model places a constant population in four categories. The model assumes that everyone in the population is susceptible from birth, therefore does not consider congenital immunity. The rate at which a susceptible person is exposed is termed β, this depends on social distancing, population density and social behaviour. Those in the exposed category have the virus however are not able to transmit the virus. The time it takes an exposed person to become infectious is termed ε and is an incubation period that remains constant for a specific virus. The average time for an infectious person to have recovered or died is termed γ. At each stage, individuals can enter the Dead phase (α) from natural causes. If lifelong immunity is not assumed with disease, a recovered individual can return to susceptible group at a rate of (ξ)

A benefit of SEIR models is their adaptability. As such, many models have included additional features. Lin et al. [10] for instance, included public risk perception and number of cumulative cases in the SEIR model [10]. Additional compartments can be added, for example, Giordano et al. [11] used a mean-field epidemiological model, known as SIDARTHE, encompassing susceptible (S), infected (I), diagnosed (D), ailing (A), recognized (R), threatened (T), healed (H) and extinct (E) categories [11]. This model detects the difference between undiagnosed and diagnosed infections whilst also distinguishing the severity of disease into life threatening and non-life threatening. This allows the model to determine the variation between the actual and perceived case fatality rate in order to identify discrepancies between the actual infection dynamics and the disease forecast projections. This model has been highly effective in presenting the epidemic curves of mortality and morbidity rates of COVID-19 in both symptomatic and asymptomatic individuals in different scenarios. These scenarios compared more lenient measures to stricter lockdown rules, as well as the effects of population-wide testing and contact tracing versus low testing and no contact tracing over 350 days. The study emphasized a limitation of using aggregated data; it cannot be supplemented with spatial modelling. The lack of incorporation of spatial modelling is also seen in SIR models and its predecessors such as the SEIR and SEIRS models. This reduces the accuracy in tracking the spread of disease, especially in populated cities. Incorporating spatio- temporal domains is incredibly important in highly contagious diseases such as COVID-19.

The spread of contagious diseases depends highly on the interactions of individuals in the population [12]. These interactions are dependent on geographic and social agents. The transmission of disease can be accurately forecasted if factors such as spatial distribution and mobility of individuals through geographical areas are incorporated. Implementing georeferenced GIS data layers and assimilating real landscapes with geospatial data provides more detail of the discrete interactions of individuals in realistic networks. Thus, providing a clearer picture of disease transmission [13]. An example of one such complex model is the previously described agent-based models, which track the progression of disease through each individual [12]. The model tracks individual movements in different social and geographic environments such as the workplace and incorporates daily activities such as the use of public transport in order to track contacts within geographic and social networks. This very model was used by Neil Ferguson, a mathematical epidemiologist at Imperial College London [14]. The paper published on 16th March 2020 used the same agent-based model used by the team in 2006 to reduce the impact of a potential flu-pandemic [15]. The model assumes transmissions between susceptible and infectious individuals occur in the household, workplace, school and outer communities as these are places that have the highest contact time for disease transmission. Therefore, the individual simulation model uses travel data, population density data, census data containing household size, workplace densities and data on average school class size. The model predicted that a lack of government action would lead to 510,000 deaths (R_0 of 2.4) in the UK and 2.2 million in the US. The model incorporated case isolation, social distancing and home quarantine scenarios which predicted significant reduction in the number of COVID-19 cases. These projections published by Ferguson and his team prompted the UK and US government to enforce lockdown rules [16]. Whilst the agent-based model reflects reality more than the equation-based models, the agent-based model requires a high volume of social contact data, such as how people travel to work or where they go during their free time.

A SEIR model was also used by the same Imperial university team which produced the similar results, with a US death rate of 2.18 million compared to 2.2 million predicted with the agent-based model [17]. Both aforementioned models have benefits and limitations. The equation-based SEIR model is a quick and easy solution to disease modelling for a grouped population, whilst the agent-based model predicts individual contacts in a population which is complex and data hungry.

2.1 Sociodemographic Factors Associated with Infectious Disease

As observed throughout history and the current pandemic, infectious disease does not discriminate and can affect anyone. However, sociodemographic factors, such as race, gender, age and poverty can result in disproportionate morbidity in certain groups. Recognising risk factors for increased mortality is important for protection of both high risk individuals and the wider community.

Table 1. Summary of studies reviewed

Title	Author(s)	Advantages	Disadvantages
A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model	Carcione J et al. (2020)	The model allows for inclusion of spatial diffusion of the virus for more accurate 2 dimensional estimation of disease spread	Lacks extra compartments such as asymptomatic (A) and dead-infected (D).
Dynamic Analysis of an SEIR Model with Distinct Incidence for Exposed and Infectives	Prem K et al. (2020)	Created synthetic contact matrices in order to incorporate location-specific physical distancing measures	Does not consider increase individual contact in healthcare settings and school/work environment Does not assume population immunity in model
A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action.	Lin Q et al. (2019)	Used parameter estimates from 1918 flu pandemic which showed similar infection-fatality rate. Using these parameters allowed for time-varying report rate- this provides high fitting performance.	Transmission from asymptomatic groups not incorporated into the model.
Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy	Giordano G et al. (2020)	Model implements diagnosed vs non-diagnosed infected cases which highlights the difference between actual infection dynamic versus predicted.	The model overestimated the number of patients with symptoms or life-threatening symptoms because the average infected individual was younger and may not show symptoms
An agent-based approach for modelling dynamics of contagious disease spread	Perez L, et al. (2009)	The model shows the progression of disease through each individual, taking into account the interaction between individuals in different environmente via GIS data layers. This allows for more accurate disease dynamic predictions	Not all the population is considered in the model due to limited memory space. This may underestimate individual contacts. Model validation in geographical data is limited
Spatiotemporal Infectious Disease Modelling	Angulo J et al. (2009)	Infectious disease data is highly sparse. This model benefits by a state-space model with disease data of different certainties in order to produce real-time disease estimates. More importantly, these estimates are in real-time as new observations are made	As the proportion of susceptible individuals reduces over time, the transmission rate estimation accuracy also declines.
Strategies for containing an	Ferguson et al.	Rather than using past estimates of transmission	A very basic model observing changes to

Title	Author(s)	Advantages	Disadvantages
emerging influenza pandemic in Southeast Asia	(2006)	from previous pandemics, they re-analysed incubation periods and household transmission data to provide a transmission estimate more consistent with the disease they are modelling.	reproductive number (R_0) following different scenarios such as social distancing. A major disadvantage is that very basic assumptions are incorporated into the model. For example, they implemented social distancing measures by reducing contacts at workplace by 50%.
Use of artificial intelligence in infectious diseases. Artificial Intelligence in Precision Health	Agrebi S et al. (2020)	Models in this review such as the ARIMA model can filter out noisy data and use linear dependence to find local trends	Lack of datasets
Machine Learning Techniques for Cognitive Decision Making	Chandiok A and Chaturvedi D (2015)	Artificial neural networks had the highest accuracy rate and performs excellent for learning non-linear (separable) problems	The neural networks take a long time to train the model
Artificial intelligence in radiology	Hosny A et al. (2018)	Deep learning methods in imaging analysis mean that algorithms can learn from data without initial input by human experts.	The AI model was unable to carry out more than one task at a time in order to detect multiple abnormalities
Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile	Wiens J et al. (2014)	Electronic medical record-based methods used in this study have a 10% improvement in the AUROC over that of the Curated model. The EMR model also reduced the number of false positives.	The model is a basic linear model. Not time-varying parameters
Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach	Wiens J et al. (2016)	Includes time-varying parameters	Decision threshold remained the same every day. Daily variable decision thresholds could be helpful.
Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making	Pourhomayoun M and Shakibi M. (2020)	A confusion matrix showed that the neural network was able to accurately predict mortality risk in patients	The random Forest algorithm only contained 20 trees.
A machine learning-based model for survival prediction in patients with severe COVID-19 infection	Yan L et al. (2020)	XGBoost recursive based tree system allows for clearer interpretability Any blood sample can be used in the model	Of the 3000 patients used in this study, none of the outcomes had been released A single-centred study with limited dataset

Title	Author(s)	Advantages	Disadvantages
Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan	Yan L et al. (2020)	Of the 300 clinical features tested, 3 clinical features were identified via the model that give poor prognosis. Scientific research backed the features the model identified.	Improved interpretability means loss of performance. The performance can be improved with a black box model. Only 3 biomarkers used for mortality prediction. However, risk of reduced capacity of prediction when more biomarkers are used
Spatially explicit models for exploring COVID-19 lockdown strategies	O'Sullivan D, Gahegan M, Exeter D and Adams B (2020)	The branching process variant of the model allows for the detailed generation of tracking connected cases (contact tracing)	Uncertainty around the COVID-19 parameters – more data configuration is needed.
Spatial modelling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran	Pourghasemi H et al (2020)	Anthropogenic factors such as person-person contact and person-surface contact was incorporated into the model via data obtained from Open Street Map and human footprint map. The random forest model used minimised "preconception and inconsistency owing to the assimilation of outcomes of each tree"	The accuracy of the training dataset was only 78.86%
Geographical information systems and tropical medicine	Khan O et al. (2010)	Disease surveillance can be carried out to predict disease spread via contact by satellite imagery and image interpretation. This model presented data in an interpretable way for the general public and healthcare professionals to understand.	The spatial analytic tool is limited by the lack of a temporal component- this means information is presented as a static snapshot.
Towards a Web GIS-based approach for mapping a dengue outbreak	Butt M et al. (2019)	Satellite data of Landsat V was overlaid the GIS model to determine hotspots of disease. The PRISM model used provided synchronous sharing of geospatial information to government.	

As COVID-19 spread from China to more ethnically diverse regions of the world such as the UK, where 13% of the population are from ethic minority backgrounds, it became more obvious that these groups are more severely affected [18]. For example, 11 doctors who died from COVID-19 were from black, Asian, and minority ethnic (BAME) communities [19]. In addition, a recent National intensive care audit reported a third of patients in intensive care units being of ethnic minority backgrounds [20]. Further studies showed that up to 22nd April 2020, 63% of COVID-19 related deaths in health workers were from BAME groups. Evidently, the data shows BAME individuals are at increased risk of death from COVID-19 [21].

The first and largest cross-sectional analysis was carried out by Lusignan et al. [22]. Using R software, they analysed data by multivariable logistic regression models with multiple imputation in order to identify variables associated with COVID-19. They found that 15.5% who tested positive were white and 62.1% were ethnically black. Overall, black people were disproportionately affected, even when adjustments were made for confounding variables such as hypertension and diabetes. Other clinical and demographic risk factors of COVID-19 included chronic kidney disease, obesity, being aged between aged 40 to 64 years and living in deprived areas. Challenges with this study, as seen in many other studies using NHS data, is that a large proportion of data is missing. For example, of the 3802 patients, 1014 (26·7%) had no data on ethnicity due to death certificates not including this information. This was overcome by randomly assigning an ethnic group based on proportion ethnicity in the area from Census data. Other incomplete data was addressed via multivariate Imputation by chained equations (MICE) and sensitivity analysis using complete case analysis [23].

Recent data was published by the UK National Statistics Office looking at COVID-19 related deaths by ethnic group during the 2 March and 10 April 2020 period [24,25]. Using a logistic regression model, they found BAME groups had a higher risk of dying from COVID-19. In order to quantify the risk of ethnicity without cofounding variables, adjustments were made for sociodemographic factors. Level of deprivation, rural urban classification, age, household composition and socioeconomical status were all adjusted for separately and presented as odd ratios.

Initially, an explanation for these findings is that BAME groups have a higher incidence of pre-existing chronic conditions such as high blood pressure, diabetes, asthma and obesity which increases their risk of COVID-19 [19]. Despite this, racial disparities in those pre-existing conditions do not reflect the huge COVID-19 death disparities. In reality, ethnicity influences culture and behaviour which can increase susceptibility to disease. For example, BAME groups tend to work in lower paid jobs, live in higher household size and live in more densely populated areas, all of which increases risk of exposure [26]. Many of the studies in this literature review did not consider socioeconomic factors such as employment in high-risk positions, income, literacy rates, access to healthcare and education. Aggregating groups of population to mortalities is an oversimplification and cannot explain the excessive deaths in BAME groups. Dissecting these groups by accounting for characteristics such as place of residence, leads to a better understanding of which factors cause these disproportionalities. However, due to the complex nature of ethnicity which is comprised of genetic, behavioural and social factors that interplay, further exploration of these observations is needed with associated robust analysis. Describing these findings is important in understanding more about this complex disease. Recent correspondence published by the Institute for Global Health, University College London in the Lancet expressed research into ethnicity as “an urgent public health research priority” [18]. However, one of the future challenges with identifying high risk ethnic groups is that the UK mortality reporting does not legally need to include ethnicity. Furthermore, the published data on ethnicity is aggregated, meaning associations identified may lead to the formation of inaccurate conclusions. Nevertheless, finding the definitive cause for any ethnic disparities is important. Artificial intelligence (AI) and machine learning technology in this area could be useful in understanding this disparity and ultimately protecting those at higher risk.

3. ADVANCED MODELLING: AI AND MACHINE LEARNING

Artificial intelligence is a branch of computer science where computers are able to mimic human intelligence in order to carry out complex tasks [27]. One type of AI is machine learning which identifies patterns and nuances from data

whilst the system learns and adapts from experience in order to improve overtime [28].

3.1 Use of AI and Machine Learning in Medicine

Infectious disease is a complex problem which requires innovative solutions. AI could be the modern technology with these solutions. As data storage capabilities have improved over the years in healthcare and more valuable data has been gathered, machine learning (ML) tools have been implemented to identify data patterns and predict future outcomes [29,30]. AI has already made its mark in medicine with its role in predicting phenotypes from genotypes, radiology and pathology diagnosis [31]. Particularly, advances in AI research became apparent with AI success in the field of radiology. ML algorithms used to interpret complex patterns of imaging data in image-based tasks have surpassed humans in certain task-specific applications [32]. The ability of an ML method to learn and improve as it extracts patterns and features from the high volume of input datasets found in healthcare records improves its predictive features and decision-making ability.

3.2 Use of AI And Machine Learning in Infectious Disease

AI goes beyond the normal equation-based models previously described. The aforementioned AI algorithms learn from the data, allowing other algorithms to make decisions using experience stored in the knowledge base [29]. A decision-making method called Gaussian process regression was used by a team at Oxford University to find the optimal malarial policy the government should use based on the spread of disease in areas of the population [33]. Implementing the correct policies at the right time was important as funding cuts in disease prevention necessitated budgeting changes.

Using ML based applications, efforts are already underway for predicting individual health risk factors that contribute to the overall risk of disease. Wiens et al. [34,35] developed an application that learns to map data such as patient history, lab results and demographics to predict patient risk to Nosocomial Clostridium difficile; a common hospital acquired bacterial infection [35]. The data-driven risk-stratification model can also identify hotspot areas of a hospital that may require more thorough disinfection. A limitation the model posed was that it assumes a constant risk of Clostridium difficile infection during a patient's time in hospital, whilst this is a dynamic variable. Daily risk estimates were produced by a multitask machine, allowing time-dependent variables to be considered. These estimates allow healthcare staff to rapidly isolate a recently identified high-risk patients from other patients hence reducing the spread of disease [35].

Following the review of many AI applications in the medical field, the main challenge posed is the vast volume of medical data stored in the electronic health records. Whilst this high volume of data is crucial for data hungry machine learning models, much of the data is described as ‘noisy’ [36]. Noisy data contains a high proportion of missing or irrelevant data. Unfortunately, these fully automated ML systems begin to learn new concepts from the noisy data resulting in inaccurate predictions; a concept known as overfitting. More commonly, this issue is overcome by using training regularization methods. Unfortunately, AI faces more challenges with infectious disease analytics as every disease has its own unique natural characteristics, such as the incubation period and transmissibility [33]. These characteristics are difficult to predict until the disease has emerged, which makes early forecasting challenging.

3.3 Use of AI and Machine Learning for COVID-19

As COVID-19 incidence increased exponentially, healthcare systems have been overwhelmed and resources strained. In order to reduce the burden, proposed prediction models from rule-based models to highly advanced ML models have been implemented. This enables identification of high-risk patients, diagnosis and prediction of disease outcomes and prognosis [37].

An AI model proposed by Pourhomayoun et al. [38] calculated mortality risks for COVID-19 positive patients, to help triage and prioritise at-risk patients more efficiently and accurately in the current overwhelmed healthcare system [40]. The ML algorithms used 117,000 COVID-19 positive patients to train a model that would predict mortality risks based on 42 features. These consisted of physiological conditions and demographic features extracted from medical records. Support Vector Machine (SVM), Neural

Networks and Random Forest were the algorithms used to create the predictive model. Performance measurements were made using AUC - ROC curves and the accuracy of the predictions was calculated using 10-fold cross-validation. The results showed that neural network algorithms were the most accurate at 93.75%.

A systematic review critically appraising 66 models from 51 studies found that the majority of the studies that used ML models were used for image analysis of CT scans and X-rays for COVID-19 analysis [37]. The remaining ML models were used for mortality prediction using demographics such as age and biomarkers from blood tests [36,40]. A group in China used the XGBoost model which is a high-performance ML algorithm [40]. Using 909 blood samples from 485 patients, an operable decision tree was developed from three selected blood biomarkers (LDL, hsCRP, lymphocytes). Across all literature these three biomarkers are frequently used for COVID-19 prognosis prediction. The predictive model can quantify the risk of death at 90%. More importantly, this model can be used to prioritise patients that require specialised care which is incredibly helpful where resources such as ventilation support are limited. The accuracy of the model could be further improved with more datasets. Despite currently using only three biomarkers to avoid overfitting, advances in our understanding of COVID-19 pathophysiology will enable identification more useful biomarkers that influence disease severity, such as Interleukin-6 [41]. Implementing these into a predictive model could further improve the model performance.

Limitations across all models reviewed included the high-risk of bias due to low sample size and poor reporting [37]. In addition, the models showed variable predicted outcomes, which can result in miscalibration. The authors also lacked adequate identification of target populations which is important for a contextualised approach to model appraisal. Currently, none of these models can be applied in practice and due to severe time constraints to publish findings, the majority of the papers in the systematic review are preprint and have not been peer-reviewed. Nevertheless, these models provide a platform for further development to ultimately produce a robust predictive model that can be used in practice. As more clinical datasets are gathered, stronger predictive models can be developed. Following validation by an independent external validator, clinicians can implement these models in practice.

There are limited papers that have used ML models to identify sociodemographic factors associated with COVID-19. This is an area of future research, where these dynamic ML models can identify people with COVID-19 related risk-factors in the population. Isolating these individuals and preventing exposure to the disease could reduce the number of patients admitted to hospital with severe COVID-19 symptoms that require a transfer to an ICU, which is essential in a time where critical care capacity is limited.

4. Spatial Modelling and Epidemiology

Spatial modelling in disease epidemiology involves the input of spatial data with geographic information systems into models in order to find and visualise the geographical distribution of disease based on demographic, environmental and sociodemographic factors [42]. Most, if not all infectious diseases are heavily influenced by environmental factors. Therefore, incorporating geospatial data in complex spatial models in order understand more about the transmission dynamics has been incredibly useful in disease epidemiology. GIS and spatial modelling go beyond simply visualizing data for the public. The spatial models can be vital in identifying sub-scale areas with emerging COVID-19 cases [43]. As national lockdown measures begin to be lifted, applying small spatial scales using spatial modelling allows sub-national areas such as cities or neighbourhoods that are at higher risk to be identified. This should help to significantly prevent the spread of COVID-19 in the wider community and thus prevent further peaks.

A team in Iran carried out spatial modelling and risk mapping based on a random forest machine learning technique (RF-MLT) using the 'random forest package' in R software [44]. Random forest is a type of supervised machine learning algorithm that combines multiple decision trees via Bootstrap aggregation. The advantage of combining these decision trees is the reduction of both bias and the likelihood of overfitting, resulting in high forecast accuracy. Sixteen variables were tested in the model including climatic factors, such as temperature, which were taken from WorldClim datasets and geographical factors such as distance from roads and city densities. These variables were mapped to predict disease transmission. Socially dense places such as bus stations and places of worship were accounted for using ArcGIS spatial tools. Heat maps were created which showed

the distribution of higher incidences of infection in different regions of Iran. The spatial analysis showed a correlation of high infection rate between two regions. For example, the cities of Alborz and Qazvin both had high levels of infection as a result of busy motorway connections. Whilst validation of risk maps via ROC-AUC showed a score of 0.886 which is classed as 'very good', GIS- based methods have been a more commonly used method.

4.1 GIS Modelling

Geoinformatics has played an important role in predicting future disease outbreaks and tracking the spread of disease. Using epidemiology mapping technology with location-based alert systems allows visualisation of trends and patterns which are a helpful resource for public health decision makers [45]. For over 20 years the World Health Organization's (WHO) greatest tool in disease mapping has been geoinformatics, specifically geographical information systems (GIS). Previously, WHO used GIS in order to track vector borne disease outbreaks like dengue and to analyse the vulnerable areas which required urgent vaccination programs [46]. Mapping of climate and population density for the Dengue virus showed that population density is the most significant factor in the spread of Dengue [47].

During the early pandemic, Johns Hopkins University created a widely popular interactive web-based dashboard using real-time data from US CDC (Centres for Disease Control and Prevention) and ECDC (European Centre for Disease Prevention and Control) [48]. The dashboard was updated regularly via Esri's ArcGIS Living Atlas team, streaming real-life data which detailed and visualised virus progression and confirmed cases worldwide [49]. However, the lack of real-time travel data prevents analysis of mobility patterns which is important in mapping disease spread. GIS is particularly helpful as it allows the identification of high-risk populations areas. This enables local authorities to provide timely information to both the public and channel healthcare resources for early healthcare intervention. For example, in Canada using data from Canada's Community Health Survey, areas with a high density of the population with pre-existing conditions and an aging population are identified via the dashboard, thereby allowing local government to redirect resources to prepare for a surge in hospital admissions [50].

4.2 Using GIS to Identify Sociodemographic Risk Factors of COVID-19

The majority of GIS has been used to predict new outbreaks, track disease spread and analyse the impact of social distancing. However, limited publications have used GIS to identify those risk-factors associated with COVID-19 in relation to various geographic regions. A useful tool in ArcGIS software is adding data layers of demographic characteristics from the "Living atlas of the world" onto the dashboard in order to identify sociodemographic features associated with COVID-19 hotspots. Identifying these risk factors that influence severity and prognosis of disease is crucial in understanding disease spread, facilitating public health planning and protecting vulnerable people.

5. CONCLUSION

As efforts to distribute vaccines continue, improving our understanding of the spread of COVID-19 and high-risk groups has become of great importance in order to contain the virus and reduce mortality rates. Whilst current epidemiological models used by the government have been incredibly useful, more innovative solutions such as AI and machine-based learning are now being embraced. The vast volume of health data available and advances in AI based applications, provide a great tool in forecasting disease spread and identifying high-risk patients.

A review of the current literature has shown the great potential of AI and machine-based learning technology in the field of medicine. Therefore, with the potential to solve the biggest crisis facing the world, the authors encourage more use of AI-based applications together with the equation-based models commonly used. This has the potential to make a major contribution to halting the current pandemic and reducing the burden on healthcare systems and damaged economies.

More than ever, epidemiological modelling is vital in understanding the spread of disease outbreaks and allowing for advanced intervention in order to prevent further disease spread, reduce mortality rates and ultimately reduce economical damage.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Parasher A. COVID-19: Current understanding of its pathophysiology, clinical presentation and treatment. Postgraduate Medical Journal. 2020;138577.
2. COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Johns Hopkins Coronavirus Resource Center; 2020. Available: https://coronavirus.jhu.edu/map.html
3. Carcione J, Santos J, Bagaini C, Ba J. A simulation of a COVID-19 epidemic based on a deterministic SEIR model. Frontiers in Public Health. 2020;8.
4. Li J, Cui N. Dynamic analysis of an SEIR model with distinct incidence for exposed and infectives. The Scientific World Journal. 2013;1-5.
5. Ridenhour B, Kowalik J, Shay D. Unraveling R_0: Considerations for public health applications. American Journal of Public Health. 2014;104(2):32-41.
6. Covid-19 Reproduction Number – R. Health; 2020 Available:https://www.health-ni.gov.uk/news/covid-19-reproduction-number-r
7. Prem K, Liu Y, Russell T, Kucharski A, Eggo R, Davies N et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. The Lancet Public Health. 2020;5(5):261-270.
8. Fragaszy E, Warren Gash C, Wang L, Copas A, Dukes O, Edmunds W et al. Cohort Profile: The Flu Watch Study. International Journal of Epidemiology. 2016;370.
9. Nelson K, Williams C. Infectious disease epidemiology. 3rd ed. Massachusetts: Burlington, MA; 2014.
10. Lin Q, Zhao S, Gao D, Lou Y, Yang S, Musa S et al. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. International Journal of Infectious Diseases. 2020;93:211-216
11. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine; 2020.
12. Perez L, Dragicevic S. An agent-based approach for modelling dynamics of contagious disease spread. International Journal of Health Geographics. 2009;8(1):50
13. Angulo J, Yu H, Langousis A, Kolovos A, Wang J, Madrid A et al. Spatiotemporal infectious disease modeling: A BME-SIR Approach. PLoS ONE. 2013;8(9):72168.
14. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand; 2020. Available:https://doi.org/10.25561/77482
15. Ferguson N, Cummings D, Fraser C, Cajka J, Cooley P, Burke D. Strategies for mitigating an influenza pandemic. Nature. 2006;442(7101):448-452.
16. Ferguson NM et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 2005;437:209–214.
17. Walker P, Whittaker C, Watson P, et al. The Global impact of COVID-19 and strategies for mitigation and suppression. Imperial College London; 2020. DOI: https://doi.org/10.25561/77735
18. Pareek M, Bangash M, Pareek N, Pan D, Sze S, Minhas J et al. Ethnicity and COVID-19: An urgent public health research priority. The Lancet. 2020;395(10234):1421-1422
19. Kirby T. Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities. The Lancet Respiratory Medicine. 2020;8(6):547-548.
20. Office of National Statistics, 2011 census; 2011.
21. UK government urged to investigate coronavirus deaths of BAME doctors the Guardian; 2020. Available:https://www.theguardian.com/society/2020/apr/10/uk-coronavirus-deaths-bame-doctors-bma
22. De Lusignan S, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, et al. Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and surveillance centre primary care network:

A cross-sectional study. The Lancet Infectious Diseases; 2020.
23. Buuren S, Groothuis Oudshoorn K. mice: Multivariate imputation by chained equations inR. Journal of Statistical Software. 2011;45(3).
24. Coronavirus (COVID-19) related deaths by ethnic group, England and Wales - Office for National Statistics. Ons.gov.uk; 2020. Available:https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronavirusrelateddeathsbyethnicgroupenglandandwales/2march2020to10april2020
25. Coronavirus-related deaths by ethnic group, England and Wales methodology - Office for National Statistics. Ons.gov.uk; 2020. Available:https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/coronavirusrelateddeathsbyethnicgroupenglandandwalesmethodolog
26. COVID-19: Review of disparities in risks and outcomes. GOV.UK; 2020. Available from: Available:https://www.gov.uk/government/publications/covid-19-review-of-disparities-in-risks-and-outcomes
27. Kersting K. Machine Learning and Artificial Intelligence: Two fellow travelers on the Quest for intelligent behavior in machines. Frontiers in Big Data. 2018;1.
28. Das S, Dey A, Pal A, Roy N. Applications of artificial intelligence in machine learning: Review and prospect. International Journal of Computer Applications. 2015;115(9):31-41.
29. Wong Z, Zhou J, Zhang Q. Artificial intelligence for infectious disease big data analytics. Infection, Disease & Health. 2019;24(1):44-48.
30. Agrebi S, Larbi A. Use of artificial intelligence in infectious diseases. Artificial Intelligence in Precision Health. 2020;415-438.
31. Chandiok A, Chaturvedi DK. Machine learning techniques for cognitive decision making. Computational Intelligence: Theories, Applications and Future Directions (WCI). 2015;1-6. DOI:10.1109/WCI.2015.7495529
32. Hosny A, Parmar C, Quackenbush J, Schwartz L, Aerts H. Artificial intelligence in radiology. Nature Reviews Cancer. 2018;18(8):500-510. DOI: 10.1038/s41568-018-0016-5
33. Bent O, Remy S, Roberts S, Walcott Bryant A. Novel Exploration Techniques (NETs) for Malaria Policy Interventions; 2017;.
34. Wiens J, Campbell W, Franklin E, Guttag J, Horvitz E. Learning data-driven patient risk stratification models for clostridium difficile. Open Forum Infectious Diseases. 2014;1(2).
35. Wiens J, Guttag J, Horvitz E. patient risk stratification with time-varying parameters: A multitask learning approach. Journal of Machine Learning Research. 2016;1(23).
36. Sarkar J, Chakrabarti P. A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19; 2020.
37. Wynants L, Van Calster B, Collins G, Riley R, Heinze G, Schuit E et al. Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. BMJ. 2020;1328.
38. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. Med Rxiv; 2020. DOI: https://doi.org/10.1101/2020.03.30.2004730
39. Yan L, Zhang H, Goncalves J, Xiao Y, Wang M, Guo Y et al. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. Med Rxiv; 2020.
40. Yan L, Zhang HT, Xiao Y, et al. Prediction of criticality in patients with severe COVID-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan. Med Rxiv; 2020. DOI:10.1101/2020.02.27.20028027
41. Messner C, Demichev V, Wendisch D, Michalick L, White M, Freiwald A et al. Clinical classifiers of COVID-19 infection from novel ultra-high-throughput proteomics. Med Rxiv; 2020.
42. Elliott P, Wartenberg D. Spatial epidemiology: Current approaches and future challenges. Environmental Health Perspectives. 2004;112(9):998-1006.
43. O'Sullivan D, Gahegan M, Exeter D, Adams B. Spatially explicit models for exploring COVID-19 lockdown strategies. Transactions in GIS. 2020;24(4):967-1000.
44. Pourghasemi H, Pouyan S, Heidari B, Farajzadeh Z, Fallah Shamsi S, Babaei S et al. Spatial modeling, risk mapping, change detection, and outbreak trend

analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). International Journal of Infectious Diseases. 2020;98:90-108.

45. Franch Pardo I, Napoletano B, Rosete Verges F, Billa L. Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment. 2020;739:140033
46. Khan O, Davenhall W, Ali M, Castillo Salgado C, Vazquez-Prokopec G, Kitron U et al. Geographical information systems and tropical medicine. Annals of Tropical Medicine & Parasitology. 2010;104(4):303-318.
47. Butt M, Khalid A, Ali A, Mahmood S, Sami J, Qureshi J et al. Towards a Web GIS-based approach for mapping a dengue outbreak. Applied Geomatics. 2019;12(2): 121-131.
48. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020;20(5):533-534.
49. Kamel Boulos M, Geraghty E. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. International Journal of Health Geographics. 2020;19(1).
50. COVID-19 Response: GIS best practices in Local Government. Data-Smart City Solutions; 2021. Available:https://datasmart.ash.harvard.edu/news/article/covid-19-response-gis-best-practices-local-government

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/65476