Animal Health Research Reviews

cambridge.org/ahr

Review

Cite this article: Yousefi Naghani S, Dara R, Poljak Z, Sharif S (2019). A review of knowledge discovery process in control and mitigation of avian influenza. *Animal Health Research Reviews* **20**, 61–71. https://doi.org/10.1017/S1466252319000033

Received: 31 March 2018
Revised: 25 September 2018
Accepted: 24 January 2019
First published online: 16 September 2019

Key words:
Avian influenza; data mining; knowledge discovery process; machine learning; modeling

Author for correspondence:
Rozita Dara, School of Computer Science, University of Guelph, Guelph, Ontario, Canada.
E-mail: drozita@uoguelph.ca

A review of knowledge discovery process in control and mitigation of avian influenza

Samira Yousefi Naghani[1], Rozita Dara[1], Zvonimir Poljak[2] and Shayan Sharif[3]

[1]School of Computer Science, University of Guelph, Guelph, Ontario, Canada; [2]Department of Population Medicine, University of Guelph, Guelph, Ontario, Canada and [3]Department of Pathobiology, University of Guelph, Guelph, Ontario, Canada

Abstract

In the last several decades, avian influenza virus has caused numerous outbreaks around the world. These outbreaks pose a significant threat to the poultry industry and also to public health. When an avian influenza (AI) outbreak occurs, it is critical to make informed decisions about the potential risks, impact, and control measures. To this end, many modeling approaches have been proposed to acquire knowledge from different sources of data and perspectives to enhance decision making. Although some of these approaches have shown to be effective, they do not follow the process of knowledge discovery in databases (KDD). KDD is an iterative process, consisting of five steps, that aims at extracting unknown and useful information from the data. The present review attempts to survey AI modeling methods in the context of KDD process. We first divide the modeling techniques used in AI into two main categories: data-intensive modeling and small-data modeling. We then investigate the existing gaps in the literature and suggest several potential directions and techniques for future studies. Overall, this review provides insights into the control of AI in terms of the risk of introduction and spread of the virus.

Introduction

Avian influenza (AI) is a disease caused by influenza type A viruses. The natural reservoir for avian influenza virus (AIV) is aquatic wild birds (Poetri, 2014); however, AIV can also infect domestic poultry in addition to other avian and mammalian species. AIV also sporadically infects human beings and is, therefore, regarded as a zoonotic virus (CDC, 2010).

AIV outbreaks in the commercial poultry industry pose a continuing threat. To contain outbreaks of AIV, various control measures such as culling of the birds, quarantine, isolation, and vaccination, have been applied. Such policies, however, may lead to substantial financial losses, regardless of their effectiveness. A large number of studies have employed mathematical models to gain a better understanding of how AIV outbreaks occur, and also to facilitate determining which factors contribute to AI progression. Modeling methods are used to select cost-effective strategies for control of AIV outbreaks.

A mathematical model is a simple and quantitative representation of a real-world function. Mathematical models can provide a theoretical framework to test real-world scenarios (Siettos and Russo, 2013) or predict the output of complex systems, in which performing a real experiment is costly or impossible. Furthermore, computer simulations in conjunction with mathematical models can bring realism to the models and approximate the behavior of real systems. Mathematical models have been used in AI research (Wiratsudakul, 2014; Maseleno *et al.*, 2015) to explore patterns and dynamics of disease spread, to assess the effectiveness of interventions, and to manage containment plans during outbreaks (Dorjee *et al.*, 2013). Mathematical modeling constitutes a step in the process of knowledge discovery in databases (KDD). In general, KDD refers to a broader process of finding knowledge in data sources. Therefore, this review focuses not only on mathematical models, but also on the entire process of KDD. The present report provides a review of AI modeling methods and explains the advantages of novel methods that can be used in KDD processes in this field. In general, the existing modeling methods can be divided into two groups based on the amount of data that are required to construct them. For each group, the goal of this review is to summarize previous work, and to identify the existing gaps concerning the knowledge discovery process in addition to describing ways to address these gaps.

Knowledge discovery in databases (KDD)

KDD refers to the overall process of extracting novel and useful patterns from data sources (Fayyad *et al.*, 1996; Williams and Huang, 1996). The primary goal of KDD is to transform data from large databases into new knowledge (Qi, 2008). Currently, several data sources

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

relevant to AI outbreak detection and containment, such as sensor networks, social media, and satellite technology, are being collected and accumulated at a dramatic pace. For example, sensor networks can be used to monitor AI in poultry farms and satellite images can be used to monitor environmental changes. Such data sources can provide an opportunity to gain precise and timely knowledge required for AI containment planning. Data, however, are usually streaming, large, and in varying formats. These types of data require continuous and automatic storage, pre-processing, analysis, interpretation, and evaluation. Therefore, not only data collection and analysis, but the whole chain of KDD is required to guarantee high quality knowledge discovery for AI-related decision-making.

A general overview of the KDD process is presented in Fig. 1. According to Nakamori (2011), KDD is comprised of five phases:

(1) Problem definition: Understanding the problem domain is a necessary prerequisite for a relevant knowledge extraction task. In the case of AI, experts from different disciplines, such as epidemiology, environmental science, statistics, and computer science should collaborate on the underlying problem that is being addressed in a data analysis task. They should also determine prior knowledge, potential data sources, requirements, and project objectives. In general, without problem definition, even the most advanced techniques will be incapable of providing the desired results.

(2) Data collection and pre-processing: After obtaining a clear understanding of the problem, domain experts explore the data, create a target dataset from the available data sources, and prepare data for deriving knowledge. Traditionally, the process of data collection was paper-based (Blumenberg and Barros, 2016), which is indeed time-consuming. Recently, due to the high volume of data generated by the internet, digital devices, and computational simulations, epidemiological studies have been turning into a data-intensive discipline. To this end, recently, concepts such as ontology (Pesquita *et al.*, 2014) have been introduced to facilitate the integration, validation, searching, and sharing of epidemiological resources. Ontology is a standard description of domain concepts and their relationships (Noy and McGuinness, 2001). A potential application of ontology in the field of AI can be to define the relationship between conceptual entities, such as highly pathogenic AI (HPAI), low pathogenic AI (LPAI), AIV subtypes, and outbreak locations. Such an ontology can be used to search AI databases or aggregate data sources.

Data pre-processing is a labor-intensive step in the KDD process. This step serves several purposes including cleaning, quality improvement, and dimensionality reduction of data, in addition to managing large volumes of data that are not capable of being processed in the computer memory. Data cleaning involves several tasks, such as the removal of noise and inconsistent data, managing missing data fields, and discretization of data (RamrezGallego *et al.*, 2016), which are necessary to correct inaccurate data. Record linkage (Dusetzina *et al.*, 2014) is another pre-processing task to improve data quality and integrity. Furthermore, data transformation can be implemented by reducing the number of data elements without destroying the validity of data. During the cleaning stage, depending upon the goal of data mining, representations of data such as normalization, type converting, aggregation, or smoothing may be required.

The term ‘data instance’ or ‘data example’ describes a single object of a dataset. Instances are described by ‘feature’ vectors. A feature is a specification that defines a property of a data entity. The term feature is sometimes used synonymously with ‘attribute’ or ‘dimension’. Recent trends in data collection have resulted in datasets with enormous dimensions. This problem is called ‘curse of dimensionality’ (Bellman, 2013), and appears in datasets with hundreds or thousands of features. Curse of dimensionality increases computation cost, storage requirements, and time required for analyses.

As a solution, several feature selection (FS) methods have been proposed with the objective of finding a subset of features that are most representative, and then discarding the rest (Alpaydin, 2014; Chi, 2009). Therefore, the selected subset is a reduced representation of the initial data, meaning that it is much smaller than the initial dataset in size, but produces the same results. FS methods can be divided into three categories: filter, wrapper, and embedded (Neumann *et al.*, 2016). In filter methods, the FS step is a separate pre-processing step from the machine learning (ML) model. This group of methods assesses the properties of data using scoring metrics such as the chi-square test, mutual information, correlation coefficients, Fisher’s discriminant scores, and variance threshold. Filter methods have been used in AI studies to find the most critical features and investigate which ones are statistically significant (Herrick, 2013; Si *et al.*, 2013; Gilbert *et al.*, 2014; Wang *et al.*, 2017). Although filter methods are computationally efficient and fast, they do not involve any learning, which may affect the classification accuracy (Hira and Gillies, 2015).

Wrapper methods, however, use ML models to measure the quality of candidate subsets of features by searching in the feature space. Genetic algorithm (GA) is an example of wrapper FS methods. GAs are search techniques used for the selection of populations of solutions to a problem. GAs are inspired from the natural evolution and genetic mechanisms of living things. Although wrapper methods outperform filter methods in terms of accuracy (Neumann *et al.*, 2016), they are computationally expensive and can suffer from over-fitting. Random forest (RF), an ensemble of decision trees that has been used to identify the most significant risk factors for the prediction of AI, is an example of wrapper FS (Herrick *et al.*, 2013). Finally, embedded methods combine filter and wrapper FS methods and offer low-cost and high accuracy. Recursive feature elimination is an embedded FS method. Despite the benefits that this method offers, embedded FS methods have not been popular in AI literature.

Feature extraction is another approach to create a lower dimension of data. Feature extraction constructs a new set of features by combining original features (Alpaydin, 2014). For example, principal component analysis is a well-known feature extraction method. To the best of our knowledge, feature extraction methods have not been used in AI modeling. In studies aimed at risk factor analysis (Nishiguchi *et al.*, 2007; Busani *et al.*, 2009; Gonzales, 2012; Nguyen, 2013), feature extraction can be used as an approach for creating new covariates.

(3) Data mining (DM): This stage requires selecting a dataset and the appropriate DM algorithms for a specific mining objective. The DM algorithms then discover patterns that exist in data. ML and statistical methods are examples of many different approaches used in DM. ML approaches can be grouped into supervised learning, unsupervised learning, and semi-supervised learning methods. Unsupervised methods find useful patterns from unlabeled data. Clustering is an example

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

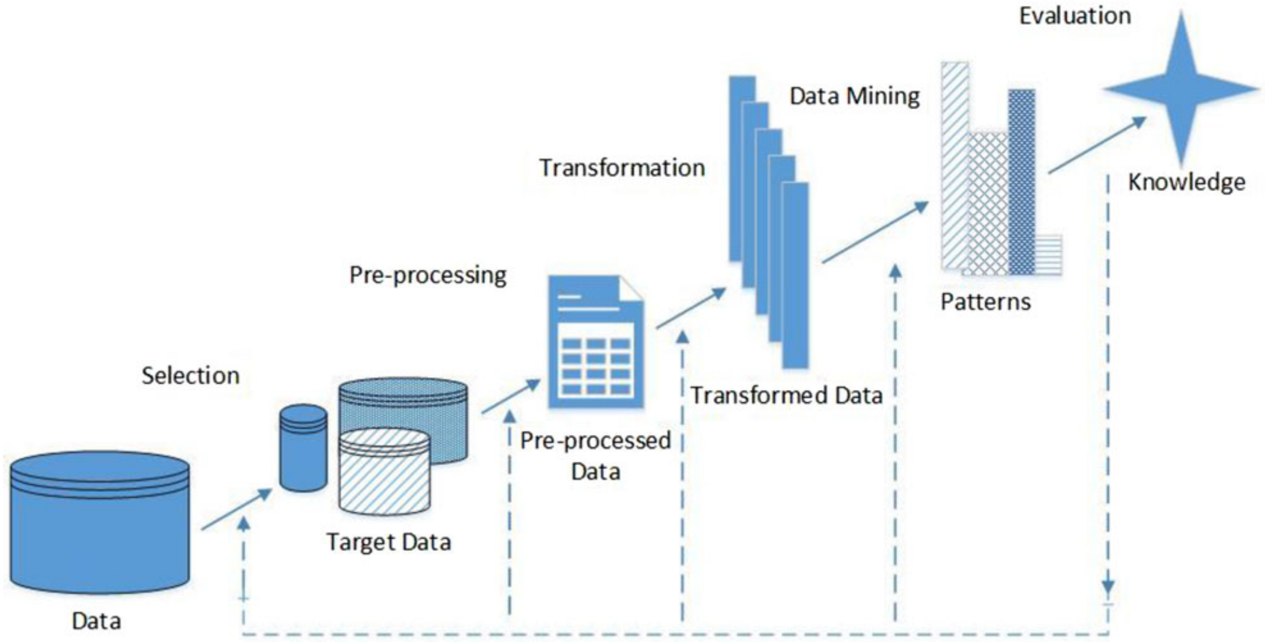

Fig. 1. KDD process.

of unsupervised ML algorithms that map data into clusters based on similarity metrics or probability density models. By contrast, supervised methods learn to map labeled data. Classification is a supervised ML task that learns from labeled data (i.e. training data) and categorizes data into one of several predefined classes. Regression falls into the supervised learning group and is a measure to determine the strength of the relationship between covariates and an independent variable. Since in many real-world applications, labeled data may be expensive and unavailable, a third group of ML has been introduced (Chapelle *et al.*, 2009). This group, known as semi-supervised learning, is trained on a combination of labeled and unlabeled data.

There have been recent attempts in epidemiological research to use semi-supervised (Zhao *et al.*, 2015), supervised (Erraguntla *et al.*, 2010; Santillana *et al.*, 2015; Valdes-Donoso *et al.*, 2017), and unsupervised (Chen *et al.*, 2016; Ghosh *et al.*, 2017; Lim *et al.*, 2017) learning approaches. In AI research, unsupervised ML algorithms such as *K*-mean are recommended for spatio-temporal profiling, outbreak detection, and surveillance studies.

(4) Post processing: After building one or more models, the next step is to interpret the obtained knowledge from the DM algorithms. The aim is to see whether suitable patterns have been discovered with respect to the goals defined in the first step. In this phase, various visualizations such as box-plots, histograms, time series plots, or two-dimensional scatter plots are used as a part of the evaluation stage.

(5) Practical use: The final goal of KDD is to use the newly obtained knowledge in real-world applications. In other words, the knowledge captured in the process needs to be organized and depicted in a way that a user or a machine can use it. Depending on the goal of a knowledge discovery process, a variety of applications may be built and provided to the user. In a potential AI decision support system, the goal of the KDD process could be: presenting reports, outbreak warnings, outbreak spread monitoring, outbreak prediction, and assessing intervention policies.

Among the five steps of KDD, the DM step is highlighted in epidemiological research. However, other steps of KDD are also essential and disregarding them may lead to inappropriate outcomes. Furthermore, if unsatisfactory results occur in any phase of the KDD process, it is possible to return to earlier stages and repeat them (Zhang *et al.*, 2010). Therefore, applying a comprehensive and iterative KDD is a factor in the success of epidemiological research as it assists in making sound decisions and finding the best possible outcome in a situation.

Data-intensive modeling

AI modeling methods may be classified into two categories: data-intensive modeling and small-data modeling. The central goal of data-intensive modeling is mining new insights from vast and diverse datasets such as click-stream, geo-location data, sensor network data, and digital health records (Marathe and Ramakrishnan, 2013).

Time-series analysis

Time-series data are a sequence of numerical data points in successive order showing how a given variable changes over time. The associated patterns obtained from time-series models are beneficial to predict future events. A commonly used time-series model in multiple previous studies (Soebiyanto *et al.*, 2010; Kane *et al.*, 2014; Permanasari *et al.*, 2015; Chadsuthi *et al.*, 2015; Ngattia *et al.*, 2016) is auto-regressive integrated moving average (ARIMA) or Box–Jenkins model (Box *et al.*, 2015). ARIMA is the combination of the auto-regressive model, the moving average model, and the auto-regressive moving average model.

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

The time-series analyses in AI research have been applied to model the temporal changes of AI incidence and to forecast possible outbreaks. For example, a non-seasonal ARIMA model was built in a study by Permanasari *et al.* (2015) to forecast future occurrences of AI. The prediction was made based on a 10-year monthly time-series of AI incidence in two regions of Indonesia. The required parameters of the ARIMA model were selected using three tests, including parameter significance, white noise, and residual normality. Similarly, ARIMA and RF time-series models have been used by Kane *et al.* (2014) to predict the future occurrence of AIV outbreaks.

In terms of data sources used in time-series analysis of AI, studies are usually limited to the temporal changes of AI incidence. While the history of disease incidence is an important factor to consider in the prediction of future outbreaks, the role of other risk factors cannot be ignored. Accordingly, in other infectious disease studies, the association between time-series of disease incidence and climate factors has been examined (Chadsuthi *et al.*, 2015; Ngattia *et al.*, 2016). Influenza outbreaks have been predicted by incorporating climate factors such as rainfall and temperature as inputs for the ARIMAX model, which is an ARIMA with additional explanatory variables. For example, Soebiyanto *et al.* (2010) showed that including climatic variables in ARIMA models leads to better performance compared to including only past case values. Additionally, Chadsuthi *et al.* (2015) showed the best performance for central regions of Thailand was obtained using the ARIMAX model that included the average temperature and the minimum relative humidity, whereas, for southern regions, minimum relative humidity as input series resulted in the best model. Similarly, Ngattia *et al.* (2016) concluded that adding rainfall factor increases the performance of the ARIMAX model.

Classical statistical models, such as ARIMA and support vector machine, are present in the literature concerning infectious disease (Zhang *et al.*, 2014; Chadsuthi *et al.*, 2015; Imai *et al.*, 2015; Song *et al.*, 2016) and AI (Kane *et al.*, 2014; Permanasari *et al.*, 2015). However, computational intelligence models such as those introduced by Ma *et al.* (2015) are not widely used, despite the fact that they have the potential to outperform classical techniques. Classical models usually require pre-defined assumptions, such as normally distributed residuals. Also, the performance of classical models may potentially be jeopardized by noisy or missing data. Models such as long short-term memory and recurrent neural network can discover non-linear and high-dimensional relationships in data (Ma *et al.*, 2015). Furthermore, the application of ensemble methods such as RF could be considered in AI time-series analysis. Ensemble methods combine multiple models to obtain a single output in order to achieve a better performance than any individual model. Recently, the potential of ensemble methods has been investigated for decision-making in infectious disease surveillance (Ray and Reich, 2018). Also, some studies related to infectious disease have discovered that most often RF results in a better prediction performance than ARIMA (Kane *et al.*, 2014; Wu *et al.*, 2017). These methods can be used in future AI research to improve the accuracy and reliability of predictions in comparison with a single model (Araque *et al.*, 2017).

Social media surveillance

Traditionally, reports from hospitals or public health centers have been used for disease surveillance (Robertson and Yee, 2016). However, these passive case reports are usually manually created, and are reported 1–2 weeks after the cases are diagnosed. This can delay subsequent actions in the case of disease outbreaks. After the invention of social media, blogging websites, and web searches, online media have been employed as surrogate data sources. To obtain data from online media, crawlers and application programing interfaces (APIs) have been utilized. Many websites offer APIs for their services, which allow third parties to query and fetch data in a convenient format. A crawler is an Internet bot used for browsing websites and social media automatically and regularly. Disease trends in social media have been employed for epidemiological purposes, as they enable authorities to track, predict, and be informed of disease emergencies.

Several studies have been carried out to examine the value of social media for human disease surveillance and to ensure its potential for being a surrogate source for the common reports of disease. For instance, the strength of relationships between disease-related posts and reports from health institutions (e.g. the Center for Disease Control and Prevention (CDC), the World Health Organization (WHO), and the World Organization for Animal Health (OIE)) have been measured using correlation.

To study social media for animal disease surveillance, there are several barriers and challenges to overcome. For instance, when analyzing social media posts regarding influenza, it is critical to differentiate between human and animal influenza. This is because users of social media are people who usually use it to communicate their daily events. Therefore, it is more likely that social media posts represent cases of human influenza. Consequently, several studies that exploit social media for the purpose of surveillance have focused on disease among human populations (Achrekar *et al.*, 2011; Szomszor *et al.*, 2011; Chen *et al.*, 2016; Sharpe *et al.*, 2017). Nevertheless, findings by Szomszor *et al.* (2011) indicate that social media users share articles from official resources. Therefore, articles regarding animal disease can be shared on social media. This provides researchers with an opportunity to employ social media for monitoring animal disease, such as AI. AI surveillance using social media has been previously attempted. Robertson and Yee (2016) introduced an online AI surveillance system to detect the AIV outbreaks. First, AI-related Twitter posts were collected, and outbreaks were identified based on anomalies in the time-series data. After comparing the detected outbreaks in Twitter with AIV outbreak reports from the OIE in the same period, a strong correlation was discovered. Also, anomalies were detected using a general linear time-series algorithm based on static and dynamic thresholds. Moreover, a latent Dirichlet allocation model was applied to the outbreak data to extract topics, concluding that the dynamic threshold leads to more meaningful topics. Further research in social media mining is required to determine whether social media can be employed as a reliable online surveillance mechanism for AI.

Another challenge in social media analysis is the volume of data. The growth of data has led to the development of new database technologies. Large amounts of data extracted from social media make analysis and daily maintenance of data difficult. Both relational (Byrd *et al.*, 2016; Jayawardhana, 2016) and NoSQL (Padmanabhan *et al.*, 2013; Wang F *et al.*, 2016) databases have been used in infectious disease surveillance using social media data. Traditional relational databases are designed to store small amounts of relevant data, while NoSQL (not only structured query language) databases are suitable for non-structured data (e.g. articles, photos, social media data, or videos). In comparison with relational databases, NoSQL databases

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

provide a number of advantages. NoSQL databases offer lower cost, easier scalability, and open source features, which make them a candidate option for AI surveillance applications that employ large social media data.

Social media data preprocessing can be challenging and time consuming. Social media contains spam messages that need to be discarded and unstructured text that needs to be transformed to an interpretable form for DM algorithms. In general, spam removal has been performed in a limited number of studies that developed surveillance systems to monitor disease from social media (Szomszor *et al.*, 2011; Kostkova *et al.*, 2014; Signorini, 2014). In the AI surveillance study (Robertson and Yee, 2016) that employed Twitter, several data cleaning operations such as stop word removal were performed, but spam removal was skipped. Removing spam, however, can enhance the accuracy of disease surveillance systems.

Furthermore, there are several gaps in the currently used methods for pre-processing of social media data for disease surveillance. For instance, manual spam removal methods, such as the link ratio calculation method used by Szomszor *et al.* (2011), are only applicable to a specific group of tweets (i.e. with a link). In addition, hand-crafted features (e.g. bag of words method) have usually been employed as the input of the spam detection classification algorithms. The manual process of feature extraction, however, involves human labor and relies on expert knowledge. Therefore, state-of-the-art methods, such as deep learning algorithms, have potential to be used for spam detection from texts. Deep learning algorithms are capable of generating word or sentence representation automatically as part of their learning process.

In general, among KDD steps related to social media surveillance research, the preprocessing step has received much attention. This is because social media text is unstructured, requiring its transformation to an interpretable form for DM algorithms. Moreover, among DM methods, correlation and classification are widely used in social media analyses.

Spatiotemporal risk prediction models

Spatiotemporal variabilities are key to reliable predictions of infectious disease (Arab, 2015). However, spatiotemporal predictions depend on the availability of relevant time- and space-related health data. Recent computational advances combined with accessibility to data containing time and space dimensions have made spatiotemporal models more popular methods (Arab, 2015). These models are used to analyze the spatiotemporal evolution of infectious diseases and assess the effect of control policies. In spatiotemporal models, clusters of disease are usually depicted on geographical maps to show the risk of disease occurrence (Gilbert and Pfeiffer, 2012).

In the AI modeling literature, considerable efforts have been put forth to find a connection between AI and environmental factors (Erraguntla *et al.*, 2010; Herrick *et al.*, 2013; Mu *et al.*, 2014). Furthermore, some studies have connected AIV transmission with migratory birds and poultry trade. For instance, Kilpatrick *et al.* (2006) determined H5N1 HPAI pathways that led to introduction of the virus into 52 countries and predicted the most likely mechanisms, including migratory bird movements and poultry trade, that facilitated the spread of AIV. In addition, the impact of agriculture and ecology, such as the presence of ducks and rice harvests, on the risk of HPAI has been explored (Gilbert *et al.*, 2007; Martin *et al.*, 2011*b*).

In addition to the risk prediction studies, attempts have been made to simulate the spread of AIV by considering a more comprehensive list of risk factors. To this end, Patyk *et al.* (2013) conducted a transmission simulation of HPAI H5N1 infection in commercial and backyard domestic poultry in South Carolina. They divided risk factor parameters into direct, indirect, and airborne. Subsequently, the North American Animal Disease Spread Model (NAADSM) was used to simulate H5N1 transmission. NAADSM is a well-established stochastic spread simulation framework designed for populations of livestock herds. Ultimately, Patyk *et al.* (2013) concluded that parameters related to indirect contact, such as people movement, vehicles, and fomites, had the highest impact on the number of infected flocks, and the duration of outbreaks.

Risk-based studies usually generate hypotheses using previous observations or from examples in the literature. Hypotheses are then defined and tested to investigate the impact of risk factors on AI outbreaks (Belkhiria *et al.*, 2018). This method may fail to identify several hypotheses. As a solution, ML methods can be employed to extract rules from observational data. The process of extracting rules can be less time-consuming than generating and testing hypotheses. Approaches such as online analytical processing, association rule mining, and sequential pattern mining have been used to find hidden rules and assess the temporal and spatial transmission of HPAI (Xu *et al.*, 2017). The outcome information from these analyses assists decision makers in understanding the spatial and temporal routes that AI will likely follow in the future.

Small-data modeling

The following section reviews research on statistical and mathematical modeling methods that rely on approaches such as questionnaire, interview, sampling, contact tracing, and direct observations. These studies can advance the understanding of AIV behaviors and dynamics.

Empirical studies

Empirical studies may be divided into two main groups: (1) studies that estimate AI transmission parameters from experimental and observational data; and (2) studies that exploit observed contact networks to make network models of AI transmission. Mathematical transmission models act as a framework to facilitate the understanding of the complex processes of disease contagion (Wiratsudakul, 2014). Once epidemiological, traffic, and biological data are imported into mathematical models, transmission patterns and parameters can be quantified (Wilasang *et al.*, 2016). In AI literature, transmission models are made based upon a series of assumptions, including equal infection susceptibility of birds, lack of pre-existing immunity in a flock, and that infected birds demonstrate similar levels of infection.

Estimation of transmission dynamics

A great number of within-flock studies have attempted to estimate the transmission parameters on a flock level (Van der Goot *et al.*, 2003; Tiensin *et al.*, 2007; Bos *et al.*, 2009; Rohani *et al.*, 2009; Bos *et al.*, 2010; Comin *et al.*, 2011; Gonzales *et al.*, 2011; Saenz *et al.*, 2012; Wang *et al.*, 2012; Nickbakhsh *et al.*, 2016). The main goal of these studies has been to extract parameter values for future containment programs of control and surveillance (Stegeman

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

et al., 2004; Savill *et al.*, 2006; Tiensin *et al.*, 2007; Bouma *et al.*, 2009).

Some of the above studies estimated AIV transmission dynamics by using generalized linear model (GLM) and 'Final Size' statistical methods (Van der Goot *et al.*, 2003; Comin *et al.*, 2011). It is thought that GLM estimation is more precise and more widely used than 'Final Size' method (Gonzales *et al.*, 2011). In a study by Gonzales *et al.* (2011), the transmission parameters of five chickens infected with a low pathogenic H7N1 virus was explored using five contact chickens. By assuming the latent period of infected birds to be a maximum of 1 day, the mean infectious period, the transmission rate (β), and the basic reproduction ratio (R_0) were estimated. Estimates like these are beneficial for building surveillance and control programs in poultry.

Small empirical studies usually work with a small number of birds (Van der Goot *et al.*, 2003; Gonzales *et al.*, 2011). The results, therefore, cannot be directly extrapolated to real-world situations. In other words, the estimated variables in empirical studies have low resolution and may not be precise enough to construct models (Pepin *et al.*, 2014). To address this issue, Saenz *et al.* (2012) estimated the dynamics of LPAI and HPAI spread using a greater number of contact turkeys compared to studies conducted by Gonzales *et al.* (2011) and Van der Goot *et al.* (2003). In the aforementioned studies, the daily number of dead birds was fitted to a stochastic Susceptible-Infectious-Recovered (SIR) model.

Back-calculation has been used by Tiensin *et al.* (2007) and Bos *et al.* (2010) to estimate required AIV transmission parameters. In the back-calculation method, mortality is measured regularly, then, the previous time-series of other classes in the SIR model are calculated according to mortality time-series. These calculations are based on several assumptions, such as a predetermined infectious period and days-to-die after infection. This method is not applicable for LPAI, as the rate of mortality for LPAI is very low. However, in order to measure HPAI H5N1 transmission dynamics within a flock, Tiensin *et al.* (2007) applied the statistical back-calculation method on 139 flocks of poultry in Thailand. Having access to time-series of recovered (R) (i.e. mortality) and infectious period in a SIR model, the time-series of susceptible (S) and infectious (I) were calculated. The obtained infection time-series was then fitted with GLM using negative binomial likelihood distribution to find the transmission parameters. Depending on the length of infectious period, R_0 was estimated between 2.26 and 2.64. These results can help to evaluate policies with simulation studies.

There are significant differences in the estimated values found between studies performed by Gonzales *et al.* (2011) and other similar studies (Van der Goot *et al.*, 2003). This difference can be attributed to the origin of the isolated viruses used in these experiments. For instance, the virus used in the study by Gonzales *et al.* (2011) originated in turkeys, which are more susceptible to LPAI. Additionally, the inconsistency of output values might have been due to the use of different AIV strains that have various transmission characteristics.

It is worth mentioning that variety of AIV strains, farm characteristics, and birds' age can lead to inconsistencies in estimated transmission parameters. In order to assess the result of variance in outcome parameters, Comin *et al.* (2011) took into account a range of values for transmission dynamics. It was concluded that the variation of R_0 plays an essential role in the outputs of an epidemic. Furthermore, the inconsistency of dynamics found in past experiments has been considered in simulation studies of LPAI in chickens (Gonzales *et al.*, 2014). In the latter study, a categorization of LPAI dynamics into low and high characteristics was introduced based on the variability in R_0.

Network models

The contact patterns among farms form a social network. Social network analysis (SNA) has been studied using both animal movement or trade networks of poultry (Van *et al.*, 2009; Martin *et al.*, 2011*a*; Hosseini *et al.*, 2013; Lee *et al.*, 2014; Moyen *et al.*, 2018) and other animals (Nöremark *et al.*, 2011; Lebl *et al.*, 2016). Networks can be presented by graphs, adjacency matrices, or a set of pairs. SNA utilizes the concepts of graph theory, which allows users to identify the essential components of a graph and find its key patterns. In fact, searching for dominant spreaders in networks is crucial in controlling epidemics. In other words, once the movement or spatial structure in an area is explained, it may disclose the implications of an infection spread throughout that area. Such insights can then assist in planning the containment policies.

Several metrics, such as centrality measures (Lee *et al.*, 2014; Moyen *et al.*, 2018) are calculated to highlight AI introduction or spread risks in a defined network among chickens or flocks. In SNA model used in AI, a node usually represents a flock, market, or trader while an edge demonstrates a connection, usually movement, between those nodes. Furthermore, when all the nodes in a graph are directly or indirectly accessible from each node in that graph, the graph is called a strong component. If a node that is a part of a strong component becomes infected, that node is likely to infect all other nodes. Furthermore, if removing an edge or node in a graph divides the graph into two separated part, the spread of disease can be curtailed. Such nodes are known as a bridge or cut-point (Martínez-López *et al.*, 2009).

The effect of network properties on the persistence of H5N1 virus has been evaluated within a poultry population (Hosseini *et al.*, 2013). A stochastic simulation was constructed using the Gillespie algorithm considering a network of flocks, traders, and markets. The findings showed that the size of flocks and frequency of interactions among flocks play a role in the persistence of H5N1 infection and the pace at which an epidemic occurs.

There are several gaps in current network models of poultry movements. Although the effect of network measures, flock size, and movement frequency has been assessed in poultry, topologies of contact networks formed by the movement of traders have been overlooked. There are four known types of theoretical contact network: random, small-world, lattice, and scale-free networks. Moreover, social network analyses that have been performed in the literature of AI have not taken into account the temporal ordering of trade links. Recently, temporal network analysis has been used in pig trade networks (Lebl *et al.*, 2016), where each connection has a time stamp denoting its occurrence time. Therefore, temporal network analysis can be used to better assessment of the impact of control measures in poultry.

Simulation studies

Under experimental conditions, an exploration of variability in transmission processes that takes place in real situations is impossible. Therefore, simulation models help to extrapolate results to field situations. For example, Reeves (2012) developed a stochastic simulation model to incorporate within-flock transmission dynamics by estimating latent, subclinical infectious, and clinical infectious stages. In this study, a within-flock simulation of HPAI

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

was performed for broiler chickens in three scenarios considering hourly timestamps, where the presence of the virus was detected based on a rise in bird mortality. It was concluded that not only could HPAI virus still spread in vaccinated flocks, but its detection time could also be delayed (silent spread). Finally, it was suggested that vaccination could be useful to reduce the degree of spread of HPAI between flocks.

Simulation studies have been performed where bird vaccination has also been included. Simulation is sometimes used to evaluate several vaccination strategies based on the obtained parameters from previous studies. For instance, Galvin *et al.* (2014) performed a simulation study of vaccination strategies and compared it with non-vaccination practices, with the main objective of finding a cost-effective strategy. In order to simulate the impact of vaccination, a Susceptible-Exposed-Infectious-Recovered-Dead (SEIRD) compartmental model, in which 'D' represented an extra health state representing 'dead', was applied. In this model, chickens vaccinated with an inactivated virus vaccine transitioned directly from 'S' to the 'R' state. By taking into the costs associated with vaccination and losses due to mortality, it was concluded that immunization of 50% of the birds within a flock with the inactivated virus vaccine is the most cost-effective strategy. In another study, a simulation was conducted to estimate the transmission parameters of AIV in an unvaccinated group, a vaccinated group, and from an unvaccinated group to a vaccinated group (Poetri *et al.*, 2009). Birds were regularly observed after vaccination before the observed data were fitted to Susceptible-Exposed-Infectious-Recovered (SEIR) simulation data by maximizing the likelihood of parameters. Finally, it was concluded that an H5N2 inactivated virus vaccine could reduce the susceptibility of chickens to HPAI H5N1 by 88%.

Behavioral-based models

This section explores compartmental and agent-based models (ABM). Compartmental models usually focus on the average behavior of a group while ABMs build detailed individual behaviors. In addition, compartmental models follow a top-down approach whereas ABMs follow a bottom-up approach. Top-down models utilize estimated parameters to simulate a process. Conversely, bottom-up models use simulated data to estimate parameters. Behavioral models can be deterministic or stochastic. Stochastic models consider random elements and run thousands of scenarios using simulation algorithms such as Gillespie. While the output of deterministic models is the same each time (Maidstone, 2012), the output from different runs of a stochastic model varies and can be summarized in various ways.

Compartmental models: Compartmental models are simple population-based models, which are extensively used in AI research. These models are known as SIR or system dynamics (Thakur *et al.*, 2015). Several SIR extensions such as Susceptible-Exposed-Infectious-Recovered, Susceptible-Infectious, Susceptible-Infectious-Susceptible, and Susceptible-Exposed-Infectious-Susceptible have also been introduced. In compartmental models, at each discrete time unit, a group of individuals may belong to one of the defined discrete classes based on the average health status of the group (Höhle and Jørgensen, 2002; Dorjee *et al.*, 2013). Simulation of disease spread using compartmental models is typically performed by differential equations. In a compartmental model, the risk of spread of an infection can be described by its basic reproduction ratio (R_0). R_0 denotes the number of cases that an infectious case can generate during its infection. For R_0 values greater or equal to one, an outbreak can take place and reach a peak while for R_0 values of less than one, there is no chance of major outbreak (Coburn *et al.*, 2009).

Agent-based models: A more recent and sophisticated group of models are known as stochastic individual-based, individual-centric, or ABMs. In these models, the behavior, histories, and properties (e.g. mobility) of every individual is taken into account. In addition, the population is heterogeneous, and the spatial structure of the population could be incorporated into the model. In stochastic models, the uncertainty and randomness of the real-world are denoted with probabilities. Therefore, stochastic ABM simulations produce a range of possible outcomes and contribute to the development of decision support tools (Taylor, 2003). ABMs have the potential to generate large amounts of data and the processing of such data may be challenging. Therefore, ABMs are expected to run slower (e.g. weeks) than compartmental models on a computer (Maidstone, 2012). Similar to agent-based modeling studies of infectious disease in pigs, ABMs in the spread of AI (Patyk *et al.*, 2013; Lewis *et al.*, 2017) have been performed using the NAADSM conceptual framework.

Component-based simulation

This section divides the simulation models into within-flock and between-flock models based on the resolution that can be accounted in the models. Between-flock transmission models, which consider a flock as the unit of interest have a lower resolution than within-flock transmission models.

Within-flock transmission models: Within-flock transmission of AIV refers to transmission of the virus among birds within a single flock. Within-flock transmission simulations have been performed in poultry flocks (Reeves, 2012; Weaver *et al.*, 2012) using stochastic state transition conceptual models. In fact, the transmission equation calculates the number of birds transitioning between states of disease in a time period. The transmission model is then used in conjunction with a simulation model to allow for a scenario-based understanding of disease spread in a flock. For example, within-flock simulation models are used to assess the impact of vaccination or virus strain on transmission.

There are several gaps that can be addressed in within-flock transmission simulation studies. The output of simulation models may not be representative of real-field data because the experimental settings might be different from one flock to another due to differences in flock characteristics such as housing systems and flock management, in addition to differences in virus strains. To address this, field data can be collected by building wireless sensor networks to track virus transmission behaviors for each flock. Furthermore, a number of transmission dynamics such as temperature, wind direction, ventilation system, and humidity have been overlooked in within-flock transmission studies. Sound results generated by within-flock transmission models can then be utilized in the development of parameters of between-flock transmission models.

Between-flock transmission models: Between-flock transmission of AIV refers to a direct (e.g. bird movement and bird trade) or an indirect transmission (e.g. human contact, shared trucks, and dust) of AIV among poultry flocks. According to Pepin *et al.* (2014), performing between-flock experimental studies is impossible, as it is expensive and life-threatening.

Therefore, transmission models have been used in AI modeling to generate hypotheses on the impact of control measures and find optimal prevention solutions (Mannelli *et al.*, 2007; Mulatti *et al.*, 2010; Lee *et al.*, 2014; Backer *et al.*, 2015).

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

Between-flock transmission models in AI have used probability-based (Dorigatti *et al.*, 2010; Ssematimba *et al.*, 2012; Backer *et al.*, 2015), agent-based (Patyk *et al.*, 2013; Lewis *et al.*, 2017), and network-based (Van *et al.*, 2009; Lee *et al.*, 2014) approaches. Probability-based methods follow a top-down approach to estimate a kernel function that usually combines disease dynamics and distance between farms. This is due to the lack of detailed information on the level of contribution of each factor to an outbreak. On the other hand, agent-based and network-based approaches usually follow a bottom-up procedure to assess the effectiveness of control strategies. Network-based model are suitable when network characteristics of a set of flocks and their biosecurity indicators need to be considered as risk factors for AI outbreak prediction (Martin *et al.*, 2011*b*). Notably, the above approaches are not necessarily mutually exclusive, meaning a model can be built based on more than one approach.

In a study by Mulatti *et al.* (2010), a top-down approach was followed to find the best intervention policies for reduction of virus transmission between flocks. In the study conducted by Mulatti *et al.* (2010), data from four previous LPAI epidemics with different interventions were fitted to a Susceptible-Infectious-Depopulated model. Subsequently, using univariate and multivariate analysis, the risk ratio and risk reproduction number (R_0) were estimated to identify the most effective policies.

In the Netherlands, Ssematimba *et al.* (2012) studied the role of downwind dust in the spread of HPAI H7N7 between poultry flocks. Particle deposition and virus decay were included when assembling the dispersion model. It was concluded that wind-borne pathogen transmission alone is not enough to explain the incidence of AI. However, for nearby surroundings, the wind-borne route plays a substantial role.

In this review, simulation studies are placed in the category of small-data modeling. However, it is worth noting that these studies could be considered as data-intensive modeling methods when the parameter space is large. In this case, as a wide range of values is assigned to parameters, the timeliness of processes needs to be taken into account as well. Optimization algorithms, such as GA, can provide an effective search in the parameters space. In addition to a large parameter space, simulation approaches generally result in a large volume of output data. Extracting meaningful patterns from such data can be a computationally expensive task. Therefore, ML algorithms and big data stream processing techniques need to be considered in future transmission simulation models pertaining to AI.

Additional recommendations

There are several limitations regarding data sources that have been exploited in recent studies focused on AI modeling. To begin, there are specific locations that have received more attention than others due to the availability of data, or a high number of confirmed cases in a specific area. For instance, a field survey in Phitsanulok province in Thailand has been used by several authors for AI modeling (Wiratsudakul *et al.*, 2014; Wilasang *et al.*, 2016). Another example is a dataset from an outbreak of H7N7 in the Netherlands in 2003, which has been used several times in the literature (Stegeman *et al.*, 2004; Boender *et al.*, 2007; Bavinck *et al.*, 2009; Bos *et al.*, 2009). Such data lead to findings that may not be generalizable to other locations and different virus strains. The above-mentioned retrospective studies infer insights about previous outbreaks in specific regions. However, poultry health authorities need to gain global knowledge about the underlying mechanisms of AI outbreaks. As a result, generalizing the insights gained from studies that focus on one specific time and region to other times and regions is still challenging. In addition, it is of interest to know that Twitter has been the center of interest in digital surveillance studies. However, to the best of our knowledge, the potential of blogs, search engines, and news feeds have been overlooked in AI surveillance studies. Furthermore, AI risk-based studies define and test hypotheses to investigate the impact of risk factors on AI outbreaks. The hypotheses are usually generated based on past observations or the literature. FS, a pre-processing technique in KDD, can generate other hypotheses that may represent a more precise behavior of AI.

The data cleaning step of KDD seems to be more commonly practiced in AI modeling studies compared to other pre-processing techniques including data integration, data transformation, and data reduction. Syndromic surveillance studies in social media, for instance, use natural language processing methods such as tokenization, stemming, lemmatization, and stop word removal to clean text data (Lee *et al.*, 2013; Chen *et al.*, 2016; Ghosh *et al.*, 2017).

Pre-processing of data is considered a more time-consuming phase of KDD compared to other phases (Tsumoto, 2000; García *et al.*, 2016). It is estimated that pre-processing takes about 80% of the entire time allocated to a project (Duhamel *et al.*, 2003; Pérez *et al.*, 2015). As a result, to save time, performing this step simultaneously with data collection is recommended.

An important consideration is that decisions during AI emergencies need to be timely and rapid. Simultaneously, there is a rise in new and large digital data sources in epidemiology (Salathe *et al.*, 2012). Therefore, parallel and distributed KDD methods may be used to enhance the performance of knowledge extraction from large datasets. In the current era, with advancements in computing power, traditional algorithms of DM need to be adjusted in order to fit cutting-edge computing approaches, such as those being used in Hadoop (White, 2012).

Conclusions

The work presented here provides an overview of the modeling methods that have been proposed for control of AI. Furthermore, the present survey has highlighted AI research limitations with regard to the KDD process. As new technologies improve, AI modeling is turning into a data-intensive and multidisciplinary field, with high volume, variety, and velocity of data. Therefore, small data methods introduced here, in particular, need to be adapted to state-of-the-art analytical approaches to reveal new patterns that have previously been overlooked. This might consequently minimize the financial, animal health, and public health impacts of AI.

Acknowledgments. This work was funded by Egg Farmers of Canada, Chicken Farmers of Saskatchewan, and the Canadian Poultry Research Council. This research is supported in part by the University of Guelph's Food from Thought initiative, and the authors acknowledge funding from the Canada First Research Excellence Fund.

References

Achrekar H, Gandhe A, Lazarus R, Yu SH and Liu B (2011) Predicting flu trends using Twitter data. In: *Proceedings of 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, P.R. China*, pp. 702–707.

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

Alpaydin E (2014) Introduction to machine learning. Cambridge, MA, USA: MIT press.

Arab A (2015) Spatial and spatio-temporal models for modelling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health* **12**, 10536–10548.

Araque O, Corcuera-Platas I, Sanchez-Rada JF and Iglesias CA (2017) Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications* **77**, 236–246.

Backer JA, van Roermund HJW, Fischer EAJ, van Asseldonk MAPM and Bergevoet RHM (2015) Controlling highly pathogenic avian influenza outbreaks: an epidemiological and economic model analysis. *Preventive Veterinary Medicine* **121**, 142–150.

Bavinck V, Bouma A, Van Boven M, Bos MEH, Stassen E and Stegeman JA (2009) The role of backyard poultry flocks in the epidemic of highly pathogenic avian influenza virus (H7N7) in the Netherlands in 2003. *Preventive Veterinary Medicine* **88**, 247–254.

Belkhiria J, Hijmans RJ, Boyce W, Crossley BM and Martínez-López B (2018) Identification of high risk areas for avian influenza outbreaks in California using disease distribution models. *PLoS ONE* **13**, e0190824.

Bellman R (2013) *Dynamic Programming*. Princeton, NJ: Courier Corporation, Princeton University Press.

Blumenberg C and Barros AJD (2016) Electronic data collection in epidemiological research. *Applied Clinical Informatics* **7**, 672–681.

Boender GJ, Elbers ARW and de Jong MCM (2007) Spread of avian influenza in the Netherlands: identifying areas of high-risk. *Veterinaria Italiana* **43**, 605–609.

Bos MEH, Nielen M, Koch G, Bouma A, De Jong MCM and Stegeman A (2009) Back-calculation method shows that within-flock transmission of highly pathogenic avian influenza (H7N7) virus in the Netherlands is not influenced by housing risk factors. *Preventive Veterinary Medicine* **88**, 278–285.

Bos MEH, Nielen M, Toson M, Comin A, Marangon S and Busani L (2010) Within-flock transmission of H7N1 highly pathogenic avian influenza virus in turkeys during the Italian epidemic in 1999–2000. *Preventive Veterinary Medicine* **95**, 297–300.

Bouma AM, Claassen I, Natih K, Klinkenberg D, Donnelly CA, Koch G and Van Boven M (2009) Estimation of transmission parameters of H5N1 avian influenza virus in chickens. *PLoS Pathogens* **5**, e1000281.

Box GEP, Jenkins GM, Reinsel GC and Ljung GM (2015) *Time Series Analysis: Forecasting and Control*, 5th Edn. Hoboken, New Jersey, United States: John Wiley & Sons.

Busani L, Valsecchi MG, Rossi E, Toson M, Ferre N, Dalla Pozza M and Marangon S (2009) Risk factors for highly pathogenic H7N1 avian influenza virus infection in poultry during the 1999–2000 epidemic in Italy. *The Veterinary Journal* **181**, 171–177.

Byrd K, Mansurov A and Baysal O (2016) Mining Twitter data for influenza detection and surveillance. In: *Proceedings of IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS), Austin, Texas*, pp. 43–49.

CDC (2010) Centers for Disease Control and Prevention. Available at https://www.cdc.gov/flu/avianflu.

Chadsuthi S, Iamsirithaworn S, Triampo W and Modchang C (2015) Modelling seasonal influenza transmission and its association with climate factors in Thailand using time-series and ARIMAX analyses. *Computational and Mathematical Methods in Medicine* **2015**, 436495.

Chapelle O, Scholkopf B and Zien A (2009) Semi-supervised learning. *IEEE Transactions on Neural Networks* **20**, 542–542.

Chen L, Hossain KSMT, Butler P, Ramakrishnan N and Prakash BA (2016) Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery* **30**, 681–710.

Chi CL (2009) Medical Decision Support Systems Based on Machine Learning Methods (Ph.D. thesis). The University of Iowa.

Coburn BJ, Wagner BG and Blower S (2009) Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Medicine* **7**, 30.

Comin A, Klinkenberg D, Marangon S, Toffan A and Stegeman A (2011) Transmission dynamics of low pathogenicity avian influenza infections in Turkey flocks. *PLoS ONE* **6**, e26935.

Dorigatti I, Mulatti P, Rosà R, Pugliese A and Busani L (2010) Modelling the spatial spread of H7N1 avian influenza virus among poultry farms in Italy. *Epidemics* **2**, 29–35.

Dorjee S, Poljak Z, Revie CW, Bridgland J, McNab B, Leger E and Sanchez J (2013) A review of simulation modelling approaches used for the spread of zoonotic influenza viruses in animal and human populations. *Zoonoses and Public Health* **60**, 383–411.

Duhamel A, Nuttens MC, Devos P, Picavet M and Beuscart R (2003) A preprocessing method for improving data mining techniques. Application to a large medical diabetes database. *Studies in Health Technology and Informatics* **95**, 269–274.

Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L and Carpenter WR (2014) Linking data for health services research: a framework and instructional guide [Internet]. Rockville, USA: Agency for Healthcare Research and Quality. Available at http://www.ncbi.nlm.nih.gov/books/NBK253312/.

Erraguntla M, Ramachandran S, Wu CN and Mayer RJ (2010) Avian influenza data mining using environment, epidemiology, and etiology surveillance and analysis toolkit (E3SAT). In: *Proceedings of 43rd Hawaii International Conference on System Sciences (HICSS), Honolulu, Hawaii*, pp. 1–7.

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996) From data mining to knowledge discovery in databases. *AI Magazine* **17**, 37.

Galvin CJ, Rumbos A, Vincent JI and Salvato M (2014) Modeling the effects of avian flu (H5N1) vaccination strategies on poultry. *CODEE Journal* **10**, 1.

García S, Ramírez-Gallego S, Luengo J, Benítez JM and Herrera F (2016) Big data preprocessing: methods and prospects. *Big Data Analytics* **1**, 9.

Ghosh S, Chakraborty P, Nsoesie EO, Cohn E, Mekaru SR, Brownstein JS and Ramakrishnan N (2017) Temporal topic modelling to assess associations between news trends and infectious disease outbreaks. *Scientific Reports* **7**, 40841.

Gilbert M and Pfeiffer DU (2012) Risk factor modelling of the spatio-temporal patterns of highly pathogenic avian influenza (HPAIV) H5N1: a review. *Spatial and Spatio-Temporal Epidemiology* **3**, 173–183.

Gilbert M, Xiao X, Chaitaweesub P, Kalpravidh W, Premashthira S, Boles S and Slingenbergh J (2007) Avian influenza, domestic ducks and rice agriculture in Thailand. *Agriculture, Ecosystems & Environment* **119**, 409–415.

Gilbert M, Golding N, Zhou H, Wint GRW, Robinson TP, Tatem AJ, Lai S, Zhou S, Jiang H and Guo D (2014) Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nature Communications* **5**, 4116.

Gonzales JL, Boender GJ, Elbers ARW, Stegeman JA and de Koeijer AA (2014) Risk-based surveillance for early detection of low pathogenic avian influenza outbreaks in layer chickens. *Preventive Veterinary Medicine* **117**, 251–259.

Gonzales JL, Van Der Goot JA, Stegeman JA, Elbers ARW and Koch G (2011) Transmission between chickens of an H7N1 low pathogenic avian influenza virus isolated during the epidemic of 1999 in Italy. *Veterinary Microbiology* **152**, 187–190.

Gonzales Rojas JL (2012) Surveillance of Low Pathogenic Avian Influenza in Layer Chickens: Risk Factors, Transmission and Early Detection (Ph.D. thesis). Utrecht University.

Herrick KA (2013) Predictive Modelling of Avian Influenza in Wild Birds (Ph.D. thesis). University of Alaska Fairbanks (UAF).

Herrick KA, Huettmann F and Lindgren MA (2013) A global model of avian influenza prediction in wild birds: the importance of northern regions. *Veterinary Research* **44**, 42.

Hira ZM and Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics* **2015**, 198363.

Höhle M and Jørgensen E (2002) *Estimating Parameters for Stochastic Epidemics*. Dina Research Report 102, Danish Institute of Agricultural Sciences, Tjele, Denmark.

Hosseini PR, Fuller T, Harrigan R, Zhao D, Arriola CS, Gonzalez A, Miller MJ, Xiao X, Smith TB and Jones JH (2013) Metapopulation dynamics enable persistence of influenza A, including A/H5N1, in poultry. *PLoS ONE* **8**, e80091.

Imai C, Armstrong B, Chalabi Z, Mangtani P and Hashizume M (2015) Time series regression model for infectious disease and weather. *Environmental Research* **142**, 319–327.

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

Jayawardhana UK (2016) An Ontology-Based Framework for Formulating Spatio-Temporal Influenza (flu) Outbreaks from Twitter (Ph.D. thesis). Bowling Green State University.

Kane MJ, Price N, Scotch M and Rabinowitz P (2014) Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics* **15**, 276.

Kilpatrick AM, Chmura AA, Gibbons DW, Fleischer RC, Marra PP and Daszak P (2006) Predicting the global spread of H5N1 avian influenza. *Proceedings of the National Academy of Sciences* **103**, 19368–19373.

Kostkova P, Szomszor M and St Louis C (2014) The use of Twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. *ACM Transactions on Management Information Systems* **5**, 8.

Lebl K, Lentz HHK, Pinior B and Selhorst T (2016) Impact of network activity on the spread of infectious diseases through the German pig trade network. *Frontiers in Veterinary Science* **3**, 48.

Lee K, Agrawal A and Choudhary A (2013) Real-time digital flu surveillance using twitter data. In: *Proceedings of the Second Workshop on Data Mining for Medicine and Healthcare, Austin, Texas*, pp. 19–27.

Lee HJ, Suh K, Jung NS, Lee IB, Seo IH, Moon OK and Lee JJ (2014) Prediction of the spread of highly pathogenic avian influenza using a multifactor network: part 2 – comprehensive network analysis with direct/indirect infection route. *Biosystems Engineering* **118**, 115–127.

Lewis N, Dorjee S, Dubé C, VanLeeuwen J and Sanchez J (2017) Assessment of effectiveness of control strategies against simulated outbreaks of highly pathogenic avian influenza in Ontario, Canada. *Transboundary and Emerging Diseases* **64**, 938–950.

Lim S, Tucker CS and Kumara S (2017) An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics* **66**, 82–94.

Ma X, Tao Z, Wang Y, Yu H and Wang Y (2015) Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research, Part C: Emerging Technologies* **54**, 187–197.

Maidstone R (2012) Discrete event simulation, system dynamics and agent based simulation: discussion and comparison. *System* **2012**, 1–6.

Mannelli A, Busani L, Toson M, Bertolini S and Marangon S (2007) Transmission parameters of highly pathogenic avian influenza (H7N1) among industrial poultry farms in northern Italy in 1999–2000. *Preventive Veterinary Medicine* **81**, 318–322.

Marathe MV and Ramakrishnan N (2013) Recent advances in computational epidemiology. *IEEE Intelligent Systems* **28**, 96–101.

Martin V, Zhou X, Marshall E, Jia B, Fusheng G, France Dixon MA, DeHaan N, Pfeiffer DU, Magalhães RJS and Gilbert M (2011*a*) Risk-based surveillance for avian influenza control along poultry market chains in South China: the value of social network analysis. *Preventive Veterinary Medicine* **102**, 196–205.

Martin V, Pfeiffer DU, Zhou X, Xiao X, Prosser DJ, Guo F and Gilbert M (2011*b*) Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. *PLoS Pathogens* **7**, e1001308.

Martínez-López B, Perez AM and Sánchez-Vizcaíno JM (2009) Social network analysis. Review of general concepts and use in preventive veterinary medicine. *Transboundary and Emerging Diseases* **56**, 109–120.

Maseleno A, Hasan MM, Tuah N and Tabbu CR (2015) Fuzzy logic and mathematical theory of evidence to detect the risk of disease spreading of highly pathogenic avian influenza H5N1. *Procedia Computer Science* **57**, 348–357.

Moyen N, Ahmed G, Gupta S, Tenzin T, Khan R, Khan T, Debnath N, Yamage M, Pfeiffer DU and Fournie G (2018) A large-scale study of a poultry trading network in Bangladesh: implications for control and surveillance of avian influenza viruses. *BMC Veterinary Research* **14**, 12.

Mu JE, McCarl BA, Wu X and Ward MP (2014) Climate change and the risk of highly pathogenic avian influenza outbreaks in birds. *British Journal of Environment and Climate Change* **4**, 166–185.

Mulatti P, Bos MEH, Busani L, Nielen M and Marangon S (2010) Evaluation of interventions and vaccination strategies for low pathogenicity avian influenza: spatial and space-time analyses and quantification of the spread of infection. *Epidemiology & Infection* **138**, 813–824.

Nakamori Y (2011) Knowledge science: modelling the knowledge creation process. In: *Proceedings of the 55th Annual Meeting of the ISSS 2011, Hull, UK*, pp. 17–22.

Neumann U, Riemenschneider M, Sowa JP, Baars T, Kälsch J, Canbay A and Heider D (2016) Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Mining* **9**, 36.

Ngattia AK, Coulibaly D, Nzussouo NT, Kadjo HA, Chérif D, Traoré Y, Kouakou BK, Kouassi PD, Ekra KD and Dagnan NS (2016) Effects of climatological parameters in modeling and forecasting seasonal influenza transmission in Abidjan, Cote d'Ivoire. *BMC Public Health* **16**, 972.

Nguyen VL (2013) The Epidemiology of Avian Influenza in the Mekong River Delta of Viet Nam: A Dissertation Presented (Ph.D. thesis). New Zealand: Massey University.

Nickbakhsh S, Hall MD, Dorigatti I, Lycett SJ, Mulatti P, Monne I, Fusaro A, Woolhouse MEJ, Rambaut A and Kao RR (2016) Modelling the impact of co-circulating low pathogenic avian influenza viruses on epidemics of highly pathogenic avian influenza in poultry. *Epidemics* **17**, 27–34.

Nishiguchi A, Kobayashi S, Yamamoto T, Ouchi Y, Sugizaki T and Tsutsui T (2007) Risk factors for the introduction of avian influenza virus into commercial layer chicken farms during the outbreaks caused by a low-pathogenic H5N2 virus in Japan in 2005. *Zoonoses and Public Health* **54**, 337–343.

Nöremark M, Håkansson N, Lewerin SS, Lindberg A and Jonsson A (2011) Network analysis of cattle and pig movements in Sweden: measures relevant for disease control and risk-based surveillance. *Preventive Veterinary Medicine* **99**, 78–90.

Noy NF and McGuinness DL (2001) Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05, Stanford University, Palo Alto, Stanford, CA.

Padmanabhan A, Wang S, Cao G, Hwang M, Zhao Y, Zhang Z and Gao Y (2013) FluMapper: an interactive CyberGIS environment for massive location-based social media data analysis. In: *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, pp. 33.

Patyk KA, Helm J, Martin MK, Forde-Folle KN, Olea-Popelka FJ, Hokanson JE, Fingerlin T and Reeves A (2013) An epidemiologic simulation model of the spread and control of highly pathogenic avian influenza (H5N1) among commercial and backyard poultry flocks in South Carolina, United States. *Preventive Veterinary Medicine* **110**, 510–524.

Pepin KM, Spackman E, Brown JD, Pabilonia KL, Garber LP, Weaver JT, Kennedy DA, Patyk KA, Huyvaert KP and Miller RS (2014) Using quantitative disease dynamics as a tool for guiding response to avian influenza in poultry in the United States of America. *Preventive Veterinary Medicine* **113**, 376–397.

Pérez J, Iturbide E, Olivares V, Hidalgo M, Martínez A and Almanza N (2015) A data preparation methodology in data mining applied to mortality population databases. *Journal of Medical Systems* **39**, 152.

Permanasari AE, Utami IK, Hidayah I and Kusumawardani SS (2015) Forecasting avian influenza incidence in Java and Madura area. In: *Proceedings of 2015 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Indonesia*, pp. 212–216.

Pesquita C, Ferreira JD, Couto FM and Silva MJ (2014) The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. *Journal of Biomedical Semantics* **5**, 4.

Poetri ON (2014) Towards an Improved Vaccination Programme Against Highly Pathogenic Avian Influenza in Indonesia (Ph.D. thesis). Utrecht University.

Poetri ON, Bouma A, Murtini S, Claassen I, Koch G, Soejoedono RD, Stegeman JA and Van Boven M (2009) An inactivated H5N2 vaccine reduces transmission of highly pathogenic H5N1 avian influenza virus among native chickens. *Vaccine* **27**, 2864–2869.

Qi L (2008) Advancing knowledge discovery and data mining. In: *Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, Adelaide, SA, Australia*, pp. 3–5.

RamrezGallego S, Garca S, MourioTaln H, MartnezRego D, BolnCanedo V, AlonsoBetanzos A, Bentez JM and Herrera F (2016) Data discretization: taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **6**, 5–21.

Ray EL and Reich NG (2018) Prediction of infectious disease epidemics via weighted density ensembles. *PLoS Computational Biology* **14**, e1005910.

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press

Reeves A (2012) Construction and Evaluation of Epidemiologic Simulation Models for the Within-and Among-Unit Spread and Control of Infectious Diseases of Livestock and Poultry (Ph.D. thesis). Colorado State University.

Robertson C and Yee L (2016) Avian influenza risk surveillance in North America with online media. *PLoS ONE* **11**, e0165688.

Rohani P, Breban R, Stallknecht DE and Drake JM (2009) Environmental transmission of low pathogenicity avian influenza viruses and its implications for pathogen invasion. *Proceedings of the National Academy of Sciences* **106**, 10365–10369.

Saenz RA, Essen SC, Brookes SM, Iqbal M, Wood JLN, Grenfell BT, McCauley JW, Brown IH and Gog JR (2012) Quantifying transmission of highly pathogenic and low pathogenicity H7N1 avian influenza in turkeys. *PLoS ONE* **7**, e45059.

Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S and Mabry PL (2012) Digital epidemiology. *PLoS Computational Biology* **8**, e1002616.

Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO and Brownstein JS (2015) Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology* **11**, e1004513.

Savill NJ, St Rose SG, Keeling MJ and Woolhouse MEJ (2006) Silent spread of H5N1 in vaccinated poultry. *Nature* **442**, 757.

Sharpe D, Hopkins R, Cook RL and Striley CW (2017) Using a Bayesian method to assess Google, Twitter, and Wikipedia for ILI surveillance. *Online Journal of Public Health Informatics* **9**, e26.

Si Y, de Boer WF and Gong P (2013) Different environmental drivers of highly pathogenic avian influenza H5N1 outbreaks in poultry and wild birds. *PLoS ONE* **8**, e53362.

Siettos CI and Russo L (2013) Mathematical modelling of infectious disease dynamics. *Virulence* **4**, 295–306.

Signorini A (2014) Use of Social Media to Monitor and Predict Outbreaks and Public Opinion on Health Topics (Ph.D. thesis). University of Iowa.

Soebiyanto RP, Adimi F and Kiang RK (2010) Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PLoS ONE* **5**, e9450.

Song X, Xiao J, Deng J, Kang Q, Zhang Y and Xu J (2016) Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011. *Medicine* **95**, e3929.

Ssematimba A, Hagenaars TJ and De Jong MCM (2012) Modelling the wind-borne spread of highly pathogenic avian influenza virus between farms. *PLoS ONE* **7**, e31114.

Stegeman A, Bouma A, Elbers ARW, de Jong MCM, Nodelijk G, de Klerk F, Koch G and van Boven M (2004) Avian influenza A virus (H7N7) epidemic in the Netherlands in 2003: course of the epidemic and effectiveness of control measures. *The Journal of Infectious Diseases* **190**, 2088–2095.

Szomszor M, Kostkova P and Louis CS (2011) Twitter informatics: tracking and understanding public reaction during the 2009 swine flu pandemic. In: 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Lyon, France, pp. 320–323.

Taylor N (2003) Review of the use of models in informing disease control policy development and adjustment. A Report for the Department for Environmental, Food, and Rural Affairs (DEFRA), UK.

Thakur KK (2015) Simulation models for between farm transmission of PRRS virus in Canadian swine herds. Ph.D. thesis, University of Prince Edward Island.

Tiensin T, Nielen M, Vernooij H, Songserm T, Kalpravidh W, Chotiprasatintara S, Chaisingh A, Wongkasemjit S, Chanachai K and Thanapongtham W (2007) Transmission of the highly pathogenic avian influenza virus H5N1 within flocks during the 2004 epidemic in Thailand. *The Journal of Infectious Diseases* **196**, 1679–1684.

Tsumoto S (2000) Clinical knowledge discovery in hospital information systems: two case studies. In: *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France*, pp. 652–656.

Valdes-Donoso P, VanderWaal K, Jarvis LS, Wayne SR and Perez AM (2017) Using machine learning to predict swine movements within a regional program to improve control of infectious diseases in the US. *Frontiers in Veterinary Science* **4**, 2.

Van der Goot JA, De Jong MCM, Koch G and Van Boven M (2003) Comparison of the transmission characteristics of low and high pathogenicity avian influenza A virus (H5N2). *Epidemiology & Infection* **131**, 1003–1013.

Van Kerkhove MD, Vong S, Guitian J, Holl D, Mangtani P, San S and Ghani AC (2009) Poultry movement networks in Cambodia: implications for surveillance and control of highly pathogenic avian influenza (HPAI/H5N1). *Vaccine* **27**, 6345–6352.

Wang RH, Jin Z, Liu QX, van de Koppel J and Alonso D (2012) A simple stochastic model with environmental transmission explains multi-year periodicity in outbreaks of avian flu. *PLoS ONE* **7**, e28873.

Wang F, Wang H, Xu K, Raymond R, Chon J, Fuller S and Debruyn A (2016) Regional level influenza study with geo-tagged Twitter data. *Journal of Medical Systems* **40**, 189.

Wang X, Wang Q, Cheng W, Yu Z, Ling F, Mao H and Chen E (2017) Risk factors for avian influenza virus contamination of live poultry markets in Zhejiang, China during the 2015–2016 human influenza season. *Scientific Reports* **7**, 42722.

Weaver JT, Malladi S, Goldsmith TJ, Hueston W, Hennessey M, Lee B, Voss S, Funk J, Der C, Bjork KE, Clouse TL and Halvorson DA (2012) Impact of virus strain characteristics on early detection of highly pathogenic avian influenza infection in commercial table-egg layer flocks and implications for outbreak control. *Avian Diseases* **56**: 905–912.

White T (2012). *Hadoop: The Definitive Guide*, 3rd Edn. Sebastopol, CA, USA: O'Reilly Media, Inc.

Wilasang C, Wiratsudakul A and Chadsuthi S (2016) The dynamics of avian influenza: individual-based model with intervention strategies in traditional trade networks in Phitsanulok province, Thailand. *Computational and Mathematical Methods in Medicine* **2016**, 198363. doi: 10.1155/2015/198363

Williams GJ and Huang Z (1996) A case study in knowledge acquisition for insurance risk assessment using a KDD methodology. In: *Proceedings of the Pacific Rim Knowledge Acquisition Workshop, Dept. of AI, Univ. of NSW, Sydney, Australia*, pp. 117–129.

Wiratsudakul A (2014) Mathematical Modelling of the Infectious Spread of Avian Influenza on a Backyard Chicken Production Chain in Thailand (Ph.D. thesis). Université Blaise Pascal Clermont-Ferrand II.

Wiratsudakul A, Paul MC, Bicout DJ, Tiensin T, Triampo W and Chalvet-Monfray K (2014) Modelling the dynamics of backyard chicken flows in traditional trade networks in Thailand: implications for surveillance and control of avian influenza. *Tropical Animal Health and Production* **46**, 845–853.

Wu H, Cai Y, Wu Y, Zhong R, Li Q, Zheng J, Lin D and Li Y (2017) Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression. *Bioscience Trends* **11**, 292–296.

Xu Z, Lee J, Park D and Chung Y (2017) Multidimensional analysis model for highly pathogenic avian influenza using data cube and data mining techniques. *Biosystems Engineering* **157**, 109–121.

Zhang Q, Segall RS and Cao M (2010) *Visual Analytics and Interactive Technologies: Data, Text and web Mining Applications*. Hershey, PA: IGI Global, p. 113.

Zhang X, Zhang T, Young AA and Li X (2014) Applications and comparisons of four time-series models in epidemiological surveillance data. *PLoS One* **9**, e88075.

Zhao L, Chen J, Chen F, Wang W, Lu CT and Ramakrishnan N (2015) Simnest: social media nested epidemic simulation via online semi-supervised deep learning. In: *Proceedings of 2015 IEEE International Conference on Data Mining (ICDM), Atlantic City, New Jersey, USA*, pp. 639–648.

https://doi.org/10.1017/S1466252319000033 Published online by Cambridge University Press