Chapter 11
Rescuing Public Health Data

Marc Choisy, Philavanh Sitboulang, Malyvanh Vongpanhya, Chantalay Saiyavong, Bouaphanh Khamphaphongphanh, Bounlay Phommasack, Fabrice Quet, Yves Buisson, Jean-Daniel Zucker, and Wilbert van Pahuis

Abstract Modeling approaches in science can be dichotomized between the statistical versus the mathematical models. The former are strongly data oriented (experimental or field data) and can be used for quantitative predictions. The latter are more qualitative and conceptual and focus more on explaining the mechanisms ruling the phenomena under study. A powerful approach that has been developed recently aims at combining the advantages of both methods by fitting mathematical models to real data. Modern computers allow to simulate models that are more and more complex. Furthermore, recent statistical developments and algorithms allow to fit models to data that are importantly noisy and generated from natural systems that can be strongly nonlinear. A requirement is to have numerous enough data containing enough information. These technical advances bring new opportunities

M. Choisy (✉)
MIVEGEC (UMR IRD, CNRS, Universités Montpellier 1 & 2), Montpellier, France
e-mail: mchoisy@gmail.com

P. Sitboulang • M. Vongpanhya • F. Quet • Y. Buisson
Institut de la Francophonie pour la Médecine Tropicale (AUF-IFMT), Vientiane, Laos

C. Saiyavong
Epidemiology Unit, Vientiane Capital Department of Health, Vientiane, Laos

B. Khamphaphongphanh
Epidemiology Division, National Center for Laboratory and Epidemiology (NCLE), Ministry of Health, Vientiane, Laos

B. Phommasack
National Emergent Infectious Disease Coordination Office (NEIDCO), Ministry of Health, Vientiane, Laos

J.-D. Zucker
IRD, UMI 209, UMMISCO, IRD France Nord, F-93143 Bondy, France

Sorbonne Universités, Univ Paris 06, UMI 209, UMMISCO, F-75005 Paris, France

ICAN, AP-HP, Pitié-Salpêtrière hospital, F-75013 Paris, France

USTH, MSI-ICT Lab, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

W. van Pahuis
Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

S. Morand et al. (eds.), *Socio-Ecological Dimensions of Infectious Diseases in Southeast Asia*, DOI 10.1007/978-981-287-527-3_11

to the scientific method. From a practical point of view, this method extends the possibilities of prediction by extrapolation. We here present how such methodology can be applied in epidemiology. Human infectious diseases have been routinely monitored by health authorities for a long time in a number of countries around the world. Yet, until recently, such data have rarely been exploited neither for scientific nor for public health purposes, the main reason being a quality of data often judged too poor (bias, missing values, etc.). On the other hand, these data are impressively abundant and offer unique opportunity to study infectious diseases over broad spatial and temporal ranges. We show how the abundance of data can partially compensate their quality issues and how biases can be dealt with in an efficient manner. Such public health data have been extensively collected in many parts of the world but have rarely been exploited so far. Furthermore, most of these data are currently in paper forms without any copy and thus prone to destruction. One aim of the Bill & Melinda Gates Foundation-funded Vaccine Modeling Initiative is to collect these data and convert them to electronic databases that are both safe and open for the whole scientific community for thorough exploitation. We present the achievement of this initiative on dengue fever in Southeast Asia.

11.1 Introduction

Tycho Brahe (1546–1601) was born 3 years after Nicolaus Copernicus (1473–1543)'s death and 18 years before Galileo Galilei (1564–1642)'s birth. One of his major occupations was astronomy, and he spent a substantial part of his life laboriously recording the positions of celestial objects. It is by using this enormous amount of data – of exceptional quality for that time – that Johannes Kepler (1571–1630), one of his very last assistants (he had more than 100 during his career), derived his laws of planetary motion between 1609 and 1619. These laws largely influenced Isaac Newton (1642–1727) in the elaboration of his theory of universal gravitation a century after Tycho Brahe started collecting data (1687).

This famous example, from the time when modern science really emerged (notably with Galileo Galilei and his controversial defense of heliocentrism), illustrates perfectly the constant dialog between data analyses and theoretical developments that has been at the basis of the scientific method since then. Astronomy has this in common with epidemiology that experimentation is impossible. Data come uniquely from observations of the natural system. This has both advantages and disadvantages. The disadvantages are obvious and pertain to the issues of data quality and control of potential confounding effects. But observation data are also strong assets that are rarely recognized as they should: they are the real data of the natural system and not data of an artificial experimental system.

Epidemiology is the science studying the distribution of diseases in space and time. Modern epidemiology is considered to have emerged with the pioneering investigation of John Snow (1813–1858) on the causes of cholera in London during the third pandemic (1846–1861). By carefully mapping the addresses of all recorded cholera cases, John Snow could identify clusters of cases and thus hypothesize

on the role of one water distribution pump in the infection of the inhabitants of the neighbor. By removing the handle of the pump, he famously ended the epidemic. This event remains today as a classical example of successful public health intervention. More than that, these observations and his intervention proved the role of water in the transmission of the disease, which was eagerly debated at the time, 30 years before the identification of the etiologic agent by Robert Koch (1843–1910).

This chapter is about the key role that public health surveillance data play in the science of epidemiology. For long, the purpose of surveillance data collection was limited to the purpose of local or national public health monitoring only, and epidemiological investigations could be envisaged only in the context of well-defined cohort studies. The recent recognition that unique and extremely valuable information could also be drawn from surveillance data, despite their inherent quality issues, opens totally new avenues of investigation. But at the same time that the value of such data is recognized, their persistence is more than ever threatened, and it is likely that if no action is taken, a large proportion of it will be lost before having the opportunity to be analyzed by scientists. That would constitute an important loss for science but also for global and public health. In this chapter, we review the pros and cons of surveillance data and show how these can be efficiently analyzed to get an understanding of epidemiological systems that could have not be reached by any other classical epidemiological study. After a review of the major surveillance data-based results obtained over the last decade, we will present the challenges we are faced with concerning preservation and sharing of public health data and the initiatives that are currently undertaken to preserve this global source of historical information and improve its quality for the future. The end of the chapter will be illustrated by the initiative carried out for dengue syndromic data in Southeast Asia.

11.2 The Scientific Method

The scientific reasoning grounds on the hypothetico-deductive method in which theoretical hypotheses are formulated and empirically tested for possible refutation. The experimental approach has long been and still holds as the gold standard for hypothesis testing. A well-designed experimental setup allows, first, to control for effects that are of no interest to the study but still can affect the results (confounding variables) and, second, to produce data amenable to proper and efficient statistical treatments. The control of confounding variables addresses the "everything else being equal" prerogative by ensuring that individuals of the sample are as homogeneous as possible for the factors that are not the focus of the analysis. The classical statistical theory and scientific controlled experiments developed hand in hand in the field of agronomy during the first half of the twentieth century, during which most of the classical tests and models (t-test, F-test, χ^2-test, analysis of variance, linear regression, etc.) were developed by researchers such as Karl Pearson (1857–1936), Ronald Fisher (1890–1962), and Jerzy Neyman (1894–1981), to cite only the most famous of them. This statistical framework is extremely powerful but has the

inconvenience of being also extremely restrictive. The famous three assumptions of the linear regression (normality, independence, and homoscedasticity of data) illustrate such strong constraints of the statistical theory, and researchers strive hard to comply with them. Experimental design is the most appropriate way to do so. Out of despair, others sometimes resort to ad hoc data transformation.

If experimental designs allow to efficiently control for confounding effects to statistically test for effects under study, they also have major drawbacks that tend, too often, to be eluded. Indeed, one should not lose sight of the fact that his/her experimentally controlled results, however statistically strong they may be, are true for the specific system under study, in that case the experimental setup, which is generally far from a real natural system. What is true in the laboratory may not be true in nature, and vice versa. This is all too much known by pharmacologists who design new medicines. A new engineered molecule has to go through a series of successive tests and filters before being granted for release in the general population. One of the final of such steps – clinical trial – aims at verifying that the effects of the molecule in the general population is not different from those observed in the laboratory. A number of molecules have failed this last stage, with excellent results in the laboratory and highly detrimental effect in real field situations. Some antimalaria molecules are thus extremely efficient in controlled laboratory situations and yet totally inefficient in natural for reasons not always understood (Nacher 2001, 2006). The experimental method, despite its rigor and powerfulness, has thus a major drawback, which is that the system under study is not exactly the real natural system.

As reminded in the introduction, experimentation is impossible in epidemiology for obvious ethical reasons. Thus, the abovementioned problem of representativity of nature does not hold in epidemiology. However, other problems naturally arise, related to the ability of analyzing data that are highly variable, biased, incomplete, and complex. Mathematical modeling is of great use to this.

11.3 Mathematical and Computational Resources

Modeling is more and more used in biology in general and in health sciences in particular, as attested by the growing number of scientific publications including modeling work (Levin et al. 1999; Cohen 2004). However, behind the word of modeling are a large number of different practices that differ in their approaches and their aims and that are quite often mixed up (Hilborn and Mangel 1997). For example, a major distinction stands between mathematical modeling and statistical modeling. Statistical modeling is basically what has been treated in the previous paragraph. Such modeling is by essence data oriented and is interested in the relationships between variables. Its purposes include hypotheses testing and predictions. As explained above, this modeling approach is efficient as long as the data in question comply with the underlying assumptions of the statistical theory at use. At the opposite of this approach is mathematical modeling, which is typically

disconnected from data. Its fundamental aim is to understand the mechanics of the system that generate the observed relationships between variables. The major limitation of mathematical models, besides being disconnected from any form of real data, is that mathematical tractability often imposes oversimplifications of the system. Statistical and mathematical models oppose in many other respects: whereas statistical models are intrinsically quantitative, mathematical models are often qualitative; whereas statistical models are necessarily phenomenological, mathematical models can be mechanistic; whereas statistical models are by nature very precise and specific (to the sample or the population from which this sample is drawn), mathematical models are more general but also more vague.

More and more, statistical and mathematical models are considered as extremes of a same continuum, and the emerging trend in science is now to adopt a modeling approach that combines the advantages of both statistical and mathematical modeling (Fig. 11.1).

This is made possible by both the availability of ever more data and the raising computational ability to process them. With the development of the statistical theory of likelihood and, above all, its practical application to complex systems made possible by the availability of huge computer power, the idea now is to develop mathematical models both mechanistic and specific to a real system and to estimate its parameters by comparing its predicted variable values with real data. A model, be it mathematical or statistical, is made of two kinds of entities: the variables and the parameters (Fig. 11.2).

The variable, as its name indicates, is a quantity that varies and that can be measured directly. A number of diseased individuals is an typical example of a variable. A parameter is a quantity that is fixed by the modeler (or a program), which determines the fate of the variables' values and which cannot be measured directly. The major difference between phenomenological statistical and mechanistic mathematical models is that parameters do not necessarily – and most of the time they do not – have a biological meaning for statistical models, whereas they always have for mathematical models. The slope of a linear regression is a typical parameter for a statistical model: it does not mean much biologically besides giving an idea of

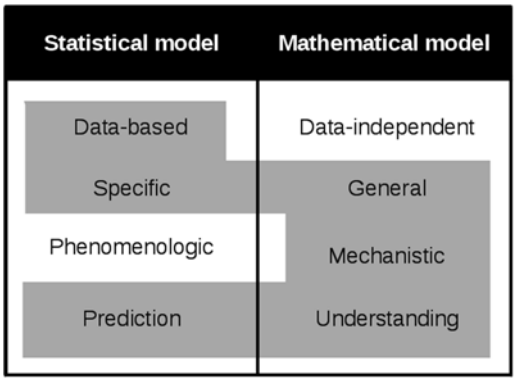

Fig. 11.1 Comparison of statistical and mathematical models as classically used. The trends today, thanks to the availability of enormous amount of data, but also to computer power, are to develop modeling approaches that combine the advantages of both statistical and mathematical models

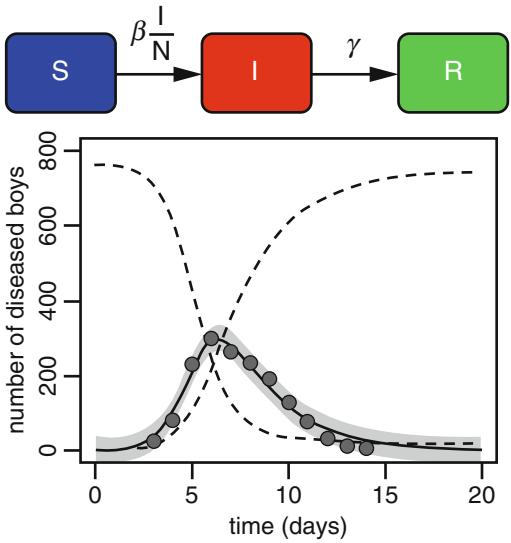

Fig. 11.2 The classical SIR compartment model and its fit to a local influenza epidemics. The total host population is partitioned into three compartments according to their clinical status, susceptibles (S), infected (I), and recovered I, the three state variables of the model. The two parameters are the contact rate (β) and the recovery rate (γ). The *dots* are the data, i.e., number of infected children in an English school board from day 3 to day 14. We use the information in this data set to fit the model-predicted prevalence (*full line curve*) as good as possible. This allows to make inference on the contact rate (1.67/day), the recovery rate (0.43/day), and the dynamics of the number of susceptibles and recovered (*dashed curves*, respectively), two parameters, and two variables that could not be measured directly (Source of data: 4th March edition of the British Medical Journal, 1978)

the association between two variables. A recovery rate is a typical example of a parameter for a mathematical model: the parameter necessarily has a biological meaning, by construction. These parameters always have a clear biological meaning, and hence they may be difficult or even impossible to measure in practice. It is here that the method of maximum likelihood plays its powerful role in allowing to estimate the values of the model's parameters.

A likelihood function depends on both data and parameters and expresses the probability that the data could have been generated by the model with specified parameters' values. For that, it compares the values of the model's variables for which data are available with their values that are actually measured. The comparison is made probabilistically. With such a likelihood function, the game is easy in principle: looking for the combination of parameters' values that yields the maximum likelihood. This combination gives the maximum-likelihood estimates of the parameters (Fig. 11.3).

All the powerfulness of the method lies here: by using information on a variable we can measure directly (such as the number of diseased individuals), we can

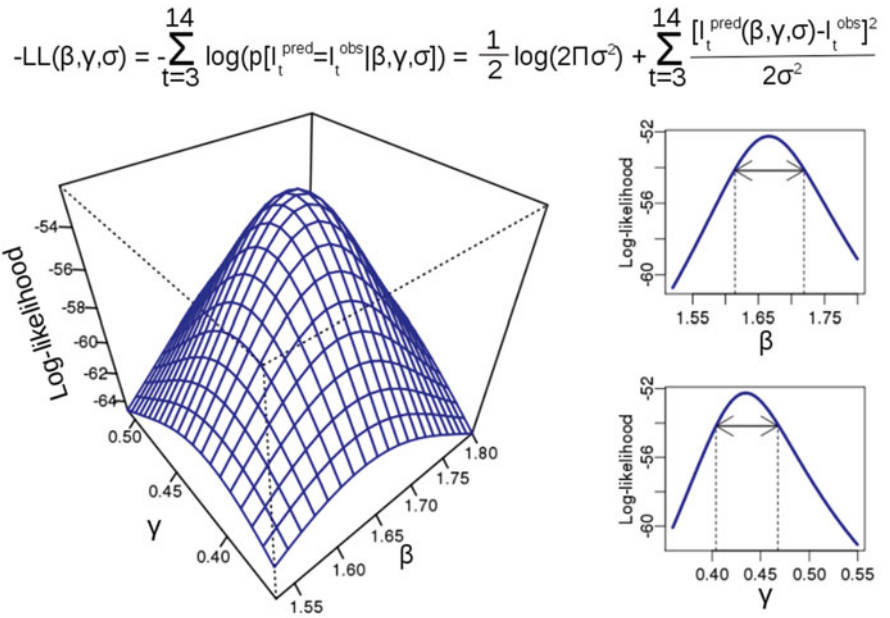

Fig. 11.3 The likelihood function applied to the model of Fig. 11.2. The *left* 3D plot shows the log-likelihood surface as a function of the value of the two parameters β (contact rate) and γ (recovery rate). The summit of this surface defines the maximum-likelihood estimates of the parameters (1.67 and 0.43, respectively). The two panels on the right show log-likelihood profiles and the use of the likelihood ratio test to estimate confidence intervals of the estimations

estimate the value of parameters of biological importance that can be out of reach by direct measure, such as a recovery rate. If simple in principle, the method can quickly become complicated in practice, especially for models with a large number of variables and parameters. More and more efficient searching algorithms and increasing computer power make such task more and more accessible. Thanks to their high flexibility, mathematical models allow to analyze complex data that would not be possible to analyze with classical statistical tools. They also allow one to analyze it in a powerful manner by using as much information from the data as possible. Given that data collection can be extremely expensive, these two points are of prime value.

The use of mathematical models and the fit of their parameters' values to data are a major shift in the scientific methodology. Instead of forcing the nature to comply with a predefined restrictive – though powerful – statistical theoretical framework, the effort is now made on the theoretical side in developing mathematical models that allow to analyze any form of data, however complicated and unorthodox they may be. By being able to draw valuable information from any form of data, we thus broaden the possibilities of scientific investigation by several orders of magnitude.

11.4 Surveillance Data and Their Quality

Any patient who enters a hospital is, at some stage, recorded with basic information regarding his gender, age, address, symptoms, and sometimes diagnosis. The aim of such records is to monitor the hospital activity. Sometimes, hospitals communicate with each other and can possibly share or exchange information from their records. It is through such practices that coincidental apparitions of rare diseases in the early 1980s on both coasts of the USA first alerted public health service about a new epidemic that soon after would be known all around the world as HIV/AIDS (Montagnier 2002).

Even long before the start of the first large-scale vaccination campaigns, infectious diseases, especially the most prevalent ones, have been the object of surveillance, through networks of health professionals and institutions, either on a voluntary or a compulsory basis (Rohani and King 2010). The primary aim of such surveillance was public health monitoring, and it became readily used after the start of the first vaccine policies in the 1940s and the 1950s in Europe and North America as a way of assessing the efficacy of disease prevention. However, the use of such data for scientific purpose remained very limited. Indeed, researchers have long refrained from using surveillance data arguing on their poor quality.

It is a fact that surveillance data such as the ones described above suffer from a number of quality issues. The first one is bias. Contrary to sentinel surveillance where a sample from the general population is chosen at random and actively tested for the disease under study (by serology or PCR), the surveillance data we are dealing with here are passively recorded from the people who do check for medical aid, and this is the source of a number of potential biases. First, people looking for medical assistance are clearly not representative of the general population, and this is strongly affected by socioeconomic factors. The high sex ratio bias in favor of males among tuberculosis-diagnosed patients is interpreted in some countries as the result of social pressures or habits where females seek for medical care less than males, especially in the case of stigmatizing diseases such as tuberculosis (Neyrolles and Quintana-Murci 2009). Second, asymptomatic carriers obviously do not seek for medical care, and yet they can play an important role in the epidemiological dynamics of the disease. Third is the problem of diagnostic: contrary to active sentinel studies where sample sizes are small enough to allow the use of sensitive and specific molecular diagnosis methods, the diagnosis carried out in case of passive surveillance systems mostly relies on symptoms, and the criteria are rarely consistent neither in space (despite WHO's efforts to homogenize it) nor in time, not speaking of the subjectivity of the examiner (HMN 2008). In addition to these biases, passive surveillance data also often suffer from frequent errors or missing values. There can be also underreporting for very small incidences (where the medical staff is not alert enough about the risk of a particular disease) or for very large incidences (where the medical staff get overwhelmed with too many patients).

Surveillance data thus suffer from a number of serious quality issues. However, they still remain the unique source of information over large spatial and temporal

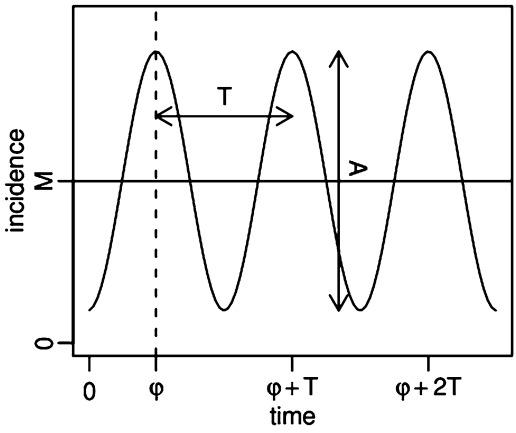

Fig. 11.4 The anatomy of a periodic time series can be summarized by four statistics. Two of them are quantitative (the average M and the amplitude A) and are potentially affected by biases in the reporting rate. Two are qualitative (the period T and the phase φ) and robust to biases in the reporting rate

ranges. We thus have to find clever ways to use this valuable source of information. There are basically two ways to deal with biases. The first one is to use methods that are robust to biases; the second one is the correct for biases. For someone interested in the epidemiological dynamics of infectious diseases, in their seasonality of recurrence over multiyear periods, it is a fact that bias would largely affect the quantitative characteristics of such dynamics such as average incidences or amplitudes. However, it is as much a fact that biases will have only a very limited impact on qualitative features of the dynamics such as periodicity of epidemic peak recurrence (period) or the timing of these peaks (phase) (see Fig. 11.4).

There are a number of powerful statistical methods that allow to extract the qualitative statistics of the time series and to perform a number of scientific analyses which are robust to biases. The next section reviews some of them. The second method to deal with bias is to correct it or, more correctly, to account for it. This implies having some source of information on the potential biases. It can come from the surveillance recommendations, such as WHO's criteria for diagnosis, or from other complementary studies performed at much smaller scales such as sentinel surveillance. By taking advantage of the flexibility of mathematical models, as exposed in the previous section, one can incorporate these information into his/her model, thus explicitly accounting for possible major historical shift in diagnosis methods or any other source of bias. Even more powerful than that can be the situation where we suspect the specific bias to exist but we are not able to assess it by any means. In the previous section, we emphasized the powerfulness of mathematical modeling interfaced with real data within the maximum-likelihood framework. We indeed explained how parameters with clear biological meaning could be estimated by using the information that can be measured on variables. We can thus adopt this very approach here, and the bias, which would be one parameter of our model, could be estimated by maximum likelihood exactly the same way as any other parameter. This has been successfully applied recently on cholera in the state of Matlab in India. Since the pioneering work of John Snow (see introduction), it has been known that cholera can be transmitted either directly from person to person

or indirectly through contaminated water. However, the respective weights of both routes of transmission remain lively debated, and a central topic of the debate has to do with the possibility of asymptomatic carriers who would not be counted in the incidence but yet who would play an important role in the incidence's dynamics. Using a mathematical model allowing the possibility of such asymptomatic carriers to analyze long-term cholera incidence time series within a maximum likelihood, Aaron King and his colleagues managed to demonstrate the existence of such asymptomatic carriers and to estimate their prevalence in the population (King et al. 2008). In the next section, we briefly review a number of studies that have been particularly influential to the scientific use of surveillance data, despite all the acknowledged quality issues.

11.5 Major Studies Based on Infectious Disease Surveillance Data

Surveillance records of major infectious diseases started at the end of the nineteenth century, but it is not before the early 2000s that their analysis with the approach exposed above really started. The work of Bryan Grenfell on measles and other childhood diseases in England and Wales was particularly influential. The extremely simple life cycle (high force of infection and permanent immunity after recovery) of childhood diseases makes them particularly amenable to analysis with mathematical models and parameter estimation with maximum likelihood. These diseases also often display specific symptoms (e.g., measles, pertussis, chicken pox), rendering their symptom-based diagnosis reliable, and they often display high incidences – at least in the pre-vaccine eras – and regular epidemics, easing the study of their dynamics. In addition to these surveillance data sets of exceptional quality and spatiotemporal coverage available in England and Wales, demographic information of similar quality and resolution are also available. Vaccine coverage is also available most of the time in the vaccine era. The study of more than six decades of such data allowed an unprecedented opportunity to investigate the role of demographic transitions and vaccination on the epidemiological dynamics of infectious diseases and to understand the laws that govern their diffusion in space. It has thus been shown that the recurrence of childhood diseases (annual, biannual, triannual, or any other multi-annual regimen), however complicated it may be, can be efficiently predicted simply from birth rate and vaccine coverage (Earn et al. 2000; Grenfell et al. 2002; Bjørnstad et al. 2002). The study of the timing of epidemics of measles in different localities of England and Wales, both before and after vaccine policies, revealed the mechanism of spatial diffusion where large cities lead the nationwide epidemiology according to a gravity-like process in which distances and population size are the main two predictors of spatial dynamics (Grenfell et al. 2001; Xia et al. 2004). This mechanism has been later on successfully verified on a variety of infectious diseases, childhood (pertussis), and others (influenza, Viboud et al. 2006).

Besides geography, demography, and vaccination, there have been investigations on the role of other mechanisms in driving infectious epidemiological dynamics. Among the most notable are climatic, socioeconomic, and immunological factors. Investigations on the putative impact of climatic conditions on the transmission of infectious diseases have been triggered by the general growing concern about global climate changes (IPCC, International Panel on Climate Change).

Climatic conditions are expected to affect disease transmission for different reasons. For diseases transmitted directly through aerial droplets, it is plausible that the survival of viral particles in these droplets and thus their infectiousness depend on climatic conditions such as humidity or temperature, as demonstrated for influenza virus both experimentally (Shaman and Kohn 2009) and empirically (Alonso et al. 2007; Shaman et al. 2010). The effect of climatic conditions is however expected to be even stronger on diseases that are either environmentally transmitted such as cholera or transmitted by vectors such as malaria or dengue fever. Cholera is a disease caused by the bacteria *Vibrio cholerae*. This bacterium naturally thrives on the surface of estuarine copepods with which it maintains a symbiotic interaction. Modifications in sea water temperature can stimulate the development of resources on which the cholera-carrying copepods feed, thus stimulating its development, which rises to several order of magnitude the probability of the copepods and the *V. cholerae* they carry to get into contact with human beings. Once such a contact has happened, an epidemic can start and spread in a human population at an incredible pace. In these conditions, we expect cholera epidemics from surveillance data to be synchronized with sea surface temperature that can be easily estimated from satellite images. This has been verified not only on the seasonal scale but also on the longer periods (3 to 7 years) of the El Niño Southern Oscillation (ENSO) (Pascual et al. 2000, 2002). Vector-borne diseases are the other group of diseases for which climatic factors are expected to influence the epidemiological dynamics. Indeed, most vector animal species, particularly the arthropod ones, have a population dynamics extremely dependent on climatic conditions for metabolic reasons (temperature) or for physical reason (rainfalls creating breeding niches). The role of temperature and rainfall on the development rate of dengue and malaria vectors has been demonstrated in the laboratory, and presence/absence field data tend to confirm this (Craig et al. 1999). However, this has never been precisely quantified in the field so far, for what concerns the timing of epidemics in one season and its severity.

The long underestimated effects of behavioral and socioeconomic factor effects on the epidemiological dynamics of major infectious diseases become more and more documented. Soper in 1929 (Soper 1929) was one of the very first to recognize the strong forcing that the alternation of vacations and school terms could have on the seasonality of measles among English children. This has been largely confirmed in recent studies where the precise school calendar can be included in the model (Keeling and Grenfell 1997). The structure of social contact is also a strong determinant of disease dynamics (Keeling and Eames 2005). Age structure is the most obvious one (Mossong et al. 2008) and has been recently put forward to explain the mysterious reemergence of pertussis in high-coverage vaccinated countries such as

Denmark (Rohani et al. 2010). Sexually transmitted diseases are the class of diseases where such social structure is intuitively expected to play the most important role. It is a well-known fact now that diseases such as HIV/AIDS have a totally different epidemiology in the homosexual and the heterosexual populations, among other factors (Keeling and Rohani 2008). The precise nature of the contact network here plays a key role (Eames and Keeling 2002). On the more economic side, it is only recently that links between economic welfare and infectious disease epidemiology have started to be investigated, with very promising possibilities (Bonds and Rohani 2009; Bonds et al. 2009).

The picture would not be complete without mentioning the immunological factors. This concerns diseases caused by pathogens that alternate (e.g., dengue serotypes) or succeed to each other (e.g., influenza strains) in time. A number of organisms causing infectious diseases have this particularity that the time scales of their molecular evolution and of their epidemiological dynamics are of the same order (Earn et al. 2002), allowing to study the interactions between these two sorts of mechanisms (Grenfell et al. 2004; Bedford et al. 2010).

The studies briefly reviewed above revealed the richness of mechanisms that can explain the epidemiological dynamics of infectious diseases. These can be immunological, climatic, environmental, behavioral, socio-economical, demographic, etc. Interestingly, it has also been proposed that infectious diseases can affect each others' dynamics, and this has been shown on childhood measles and whooping cough which infect the same cohort of children. Any child having one of these diseases is usually kept at home, which makes him/her unavailable for infection by the other disease, thus leading to some interference between the two diseases' dynamics (Rohani et al. 1998, 2003). All these studies that we presented in this section have thus contributed to some of the major changes in the history of infectious disease epidemiology. None of them would have been possible without the availability of long-term time series of surveillance notifications. The qualities of these data are unequal, but their strength resides in their huge quantity, allowing comparative and long-term studies that unravel singular mechanism that could not be detected otherwise. The availability of such data opens totally unexplored fields in epidemiology, and this is a paradigm shift that can be compared to the one that happened in astronomy with the use of the very first telescopes, which were of terrible quality too! It is likely that the transmission of diseases is multifactorial. One of the major challenges in the future would be to quantify the respective weights of the different factors that can affect infectious disease transmission. Indeed, a number of debates currently revolve around the major drivers of infectious disease dynamics. This is the case, for example, for dengue in Thailand where demographic factors have been put forward by some researchers (Cummings et al. 2004), whereas climatic factors have been claimed to play the most important role (Cazelles et al. 2005) and immunological interactions between the four serotypes by other groups (Adams et al. 2006; Wearing and Rohani 2006). Similarly, the respective importance of demographic (Viboud et al. 2006), climatic (Shaman et al. 2010), and immunological factors (Bedford et al. 2010) on the epidemiological dynamics of seasonal human influenza is still largely unresolved. Even measles, the paradigmatic model on the influence of

demographic factors on epidemiological dynamics, seems also to be affected, to some extent, by climatic factors (Ferrari et al. 2008). Model confrontations in the maximum-likelihood framework and comparative analyses will be the ways to resolve these debates. Hence, there is a need for more and more historical surveillance data.

11.6 The Value of and Threats on Public Health Data

It is a fact that infectious disease data suffer from a number of quality issues. However, they also contain a unique source of valuable information that could not be generated by any other means. We hope that the above paragraph has convinced the reader that despite their limitations, surveillance data, when analyzed with appropriate mathematical and statistical tools, can lead to results of major significance, both from the basic scientific and the applied public health perspectives.

Scientific data are expensive to generate. Look at any research project and you will see that the largest chunk of its budget is devoted to data collection. And yet, most of the time, these data are analyzed only once, in the very study that did collect them. These data are stored for a while, and soon, the location of their repository is forgotten, and eventually the data are just lost. A foundation of modern science is repeatability, including repeatability of data analysis. This latter clearly cannot be ensured any more if the data disappear. Furthermore, there is often much more information in a data set than the information used for the purpose of their collection. But this information cannot be used if the data are not made available to other researchers with other research inquiries (Bolker 2005).

The recognition of these issues has recently encouraged researchers to make their data more available. It has been now 30 years that publication of study based on the analysis of molecular sequences requires the publication of original data on the open-access electronic database GenBank (www.ncbi.nlm.nih.gov/genbank). Scientific research founders now more and more require the produced original data to be made freely available to the rest of the scientific community. It also becomes common practice for international scientific journals to require that the data should be made available from electronic repositories such as Dryad (datadryad.org).

These changes in the way to communicate and share scientific data and results are timely and will certainly improve the situation in the future. Historical data that have been collected over long time periods, such as public health surveillance data, all share the same characteristics: (1) they are unique and cannot be collected again (contrary to experimental data for which the experiments can always be rerun); (2) they represent an enormous amount of valuable information that have been rarely really analyzed; (3) they most of the time exist only in paper format, and their number of copies is often low. The studies that have been presented in the previous paragraph were carried out on a very small number of data sets: measles and whooping cough in England and Wales, dengue in Thailand, measles in Niger, and influenza in the USA. This is only the tip of the iceberg of such existing data. This is also the very tiny proportion of such data are currently available in electronic format. Much

Fig. 11.5 Data collection on the field in Laos. (**a**) Particularly unordered public health data threatened by destruction. (**b**) A typical communal health center in rural Laos (Savannakhet province). (**c**) Public health record scanning. (**d**) An exceptionally well-ordered and well-preserved ensemble of public health data (Photos: Marc Choisy)

more of such data exist all around the world in a number of different health centers, hospitals, or ministries. The vast majority of them unfortunately exist only on paper format, in one unique copy, stored in some highly vulnerable place (Fig. 11.5).

These data are historical data spanning several decades, which means that if these data were to disappear, there would be no mean to regenerate them as we could do for classic experimental data. That would thus represent a huge loss for the scientific community, as well as for the public and international health. There is thus an urgent need to secure such data from destruction that can happen at any time by office relocation, flooding, fire, etc.

11.7 Data Rescuing Programs in Southeast Asia and Challenges

The Vaccine Modeling Initiative (VMI, www.vaccinemodeling.org) is a Bill & Melinda Gates Foundation-funded international project, the aim of which is to strengthen the links between mathematical and computational models and public

health data. To this purpose, one of its activities consists in creating open-access electronic infectious disease surveillance databases, in the very same spirit as what has been done for molecular sequences with GenBank since the early 1980s. The first such database that has been created is the Tycho database (www.tycho.pitt.edu), named after Tycho Brahe (see introduction). This open-access electronic database is the result of digitization and manual double-blinded entering of the US weekly reports on infectious diseases. It represents a total of 6300 weekly reports for 55 infectious diseases from 1888 to present with data spatially aggregated by states (50) and cities and towns (1,500). In total, this database contains 100 million cases and 4 million deaths, and it took 90-man-year full-time employment to manually enter this database (Van Panhuis et al. 2013).

A similar program of the VMI is currently undertaken in Southeast Asia, focusing primarily on gathering dengue syndromic monthly surveillance notifications. As of today, it contains monthly data (since 1997) aggregated by provinces (189) for Malaysia, Thailand, Vietnam, Cambodia, and Laos, covering a population of 209 million people on 1.6 million km^2 (see Fig. 11.6).

The buildup of such an international database was eased by the fact that these province-aggregated data were already centralized at the level of each country

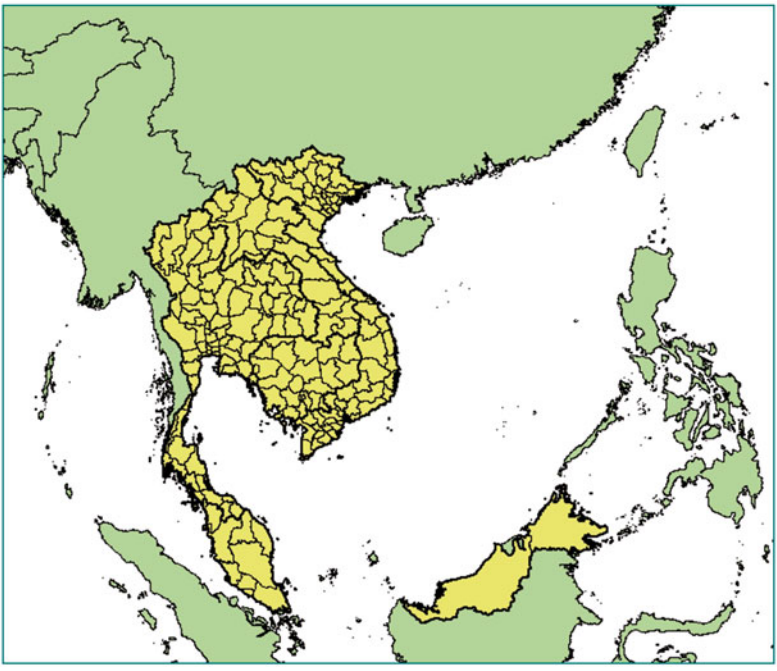

Fig. 11.6 Dengue database in Southeast Asia. Syndromic surveillance notifications have been collected and aggregated by month and province (189) for Malaysia, Thailand, Vietnam, Cambodia, and Laos for the last 15 years. This represents a total population size of 208 million individuals on an area of 1.6 million km^2

(usually the ministries of health). Disaggregated raw data are unfortunately not centralized, and their collection thus demands an important amount of field work.

In most of the countries of the region, public health surveillance networks are hierarchically organized from communal health centers to upper levels, districts, and provinces, up to the ministry of health (Fig. 11.7).

Being able to collect data from each of these levels will allow to assess the quality of data transfer along the national surveillance network, identify major weaknesses, and propose solutions of improvement. A first limitation is related to the lack of homogeneity in notification criteria and human resources, not only between countries but also within countries. The World Health Organization have developed electronic-based surveillance systems that are consistent between countries (HMN 2008), but the lack of equipment, internet access, and computer training for the medical staff unfortunately makes the use of such systems anecdotal in practice. The second major limitation pertains to the quality of information flow along the surveillance network. At each level of the network, information is received from the level below, processed and aggregated by time and space before being transferred to the next level. All this necessarily implies loss of information.

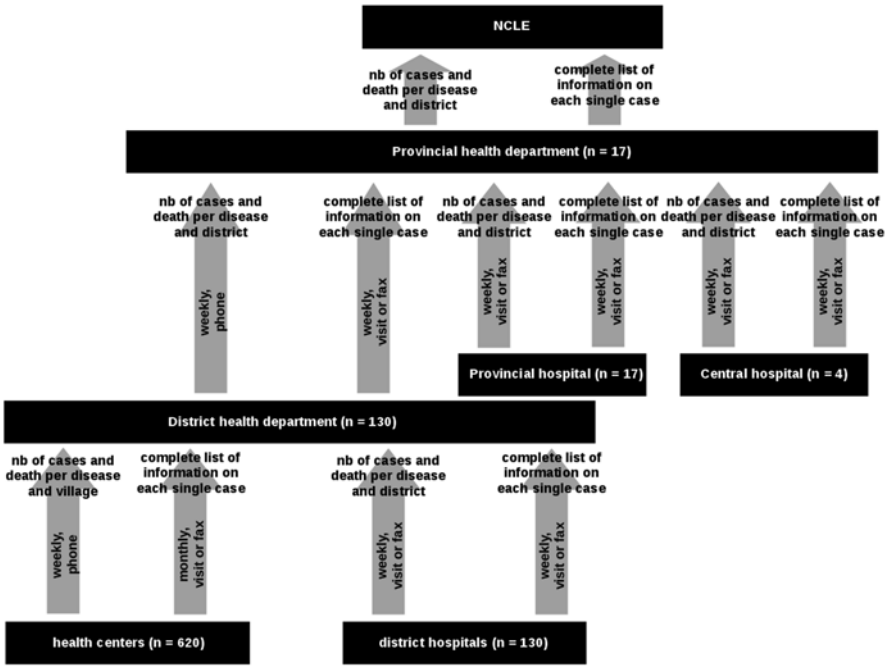

Fig. 11.7 A typical infectious disease surveillance network as implemented in Laos. Numbers in brackets refer to the numbers of health offices at each level. *Arrows* show the flow of information from the lower levels to the higher ones, up to the ministry of health (NCLE, National Center for Laboratory and Epidemiology) (note that all intermediate levels both gather data directly and aggregate data from lower levels before transmitting the total to the upper level)

Moreover, this data processing and aggregation is mostly performed by hand with all the error sources that this involves. In the absence of quality check, errors accumulate and propagate along the network. In the most remote areas, it may also happen that communication between levels is interrupted for various hazard sources.

Collecting disaggregated data at all the levels of the surveillance network will thus allow to identify the major bottleneck in information transfers and the major sources of data quality corruption. Only data aggregated by the province is maintained over a long term at the highest national level (ministry of health). At all other levels, after having processed the data from lower levels and transferred it to the higher level, the recommendation is to keep it for a minimum of 5 years. There are naturally no incentives to destroy the data after 5 years, but, in practice, because of storage space shortage, it is very rare to find disaggregated data older than 5 years. This represents an enormous loss of scientific and public health information. Only an electronic surveillance system would allow to cope with most of the issues raised in this section. By automating aggregation calculation and data transfer, it would reduce errors due to these two processes to its minimum. Furthermore, backup drives at each level would ensure the long-term preservation of raw data without requiring too much physical space. Backup in each health center of the network would also ensure that data are constantly saved in several different places. If one of the centers were to disappear, the data would be preserved in any upper level or could be reconstructed from any lower level.

11.8 Conclusion

Dengue is the first human arbovirus in the world in terms of affected population and population at risk (3.5 billion people, 55 % of the world population is estimated to be at risk by WHO, Beatty et al. 2007). It is primarily affecting the intertropical regions of the world, with a special high and ancient burden in Southeast Asia, and it has become a major international public health issue due to an increase in its worldwide distribution (Gibbons and Vaughn 2002; Guzman and Kouri 2002). In the absence of vaccine, the sole mean of dengue prevention is through vector control. A live-attenuated, tetravalent, chimeric yellow fever dengue vaccine has been in development for many years, and its commercial availability is announced for 2016 (phase III trials started in December 2010) (Guy et al. 2011). Yet, the public health services still have no clue of how to implement the best vaccine policy. It thus becomes timely to start thinking about efficient strategies that will involve both vaccine use and vector control (WHO-VMI Dengue Vaccine Modeling Group 2012). As reminded earlier, experimentation is impossible in epidemiology for obvious ethical reasons. Mathematical and computational modeling thus constitutes the only means we have to explore the efficiency of various vaccine policy scenarios (Ferguson et al. 2005). However, such prospective explorations are possible only with realistic enough models, and this depends strongly on the data available for parameter fitting. In the context of applying a vaccine policy to the Southeast Asia

region, a number of questions arise respective to the spatial dynamics of the disease at between-country level. Since the spatial dynamics of a disease at local scales largely determine its persistence at larger scales (Grenfell and Harwood 1997; Earn et al. 1998), it appears most important to understand its major drivers among demographic (population size and birth rate), climatic (affecting vector population dynamics), administrative (contact networks), and immunological (serotype interaction) factors (Racloz et al. 2012). Only historical surveillance data would allow to develop and fit realistic enough models of practical use for the design of an optimal vaccine policy.

References

Adams B, Holmes EC, Zhang C, Mammen MPM Jr, Nimmannitya S, Kalayanarooj S et al (2006) Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. Proc Natl Acad Sci U S A 103:14234–14239

Alonso WJ, Viboud C, Simonsen L, Hirano EW, Daufenbach LZ, Miller MA (2007) Seasonality of influenza in Brazil: a traveling wave from the Amazon to the subtropics. Am J Epidemiol 165:1434–1442

Beatty M, Letson W, Edgil D, Margolis H (2007) Estimating the total world population at risk for locally acquired dengue infection. Am J Trop Med Hyg 77(Suppl 5):170–257

Bedford T, Cobey S, Beerli P, Pascual M (2010) Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). PLoS Pathog 6:e1000918

Bjørnstad ON, Finkenstädt BF, Grenfell BT (2002) Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. Ecol Monogr 72:169–184

Bolker B (2005) Other people's data. BioScience 55:550–551

Bonds MH, Rohani P (2009) Herd immunity acquired indirectly from interactions between the ecology of infectious diseases, demography and economics. J R Soc Interface 7:541–547

Bonds MH, Keenan DC, Rohani P, Sachs JD (2009) Poverty trap formed by the ecology of infectious diseases. Proc R Soc Lond B 277:1185–1192

Cazelles B, Chavez M, McMichael AJ, Hales S (2005) Nonstationary influence of El Niño on the synchronous dengue epidemics in Thailand. PLoS Med 2:e106

Cohen JE (2004) Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. PLoS Biol 2:e439

Craig MH, Snow RW, le Sueur DA (1999) Climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitol Today 15:105–111

Cummings DAT, Irizarry RA, Huang NE, Endy TP, Nisalak A, Ungchusak K, Burke DS (2004) Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. Nature 427:344–347

Eames KTD, Keeling MJ (2002) Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. Proc Natl Acad Sci U S A 99:13330–13335

Earn DJ, Rohani P, Grenfell BT (1998) Persistence, chaos and synchrony in ecology and epidemiology. Proc R Soc Lond B 265:7–10

Earn DJ, Rohani P, Bolker BM, Grenfell BT (2000) A simple model for complex dynamical transitions in epidemics. Science 287:667–670

Earn DJD, Dushoff J, Levin SA (2002) Ecology and evolution of the flu. Trends Ecol Evol 17:334–340

Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A et al (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437:209–214

Ferrari MJ, Grais RF, Bharti N, Conlan AJK, Bjørnstad ON, Wolfson LJ et al (2008) The dynamics of measles in sub-Saharan Africa. Nature 451:679–684
Gibbons RV, Vaughn DW (2002) Dengue: an escalating problem. Br Med J 324:1563–1566
Grenfell BT, Harwood J (1997) (Meta)population dynamics of infectious diseases. Trends Ecol Evol 12:395–399
Grenfell BT, Bjørnstad ON, Kappey J (2001) Travelling waves and spatial hierarchies in measles epidemics. Nature 414:716–723
Grenfell BT, Bjørnstad ON, Finkenstädt BF (2002) Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. Ecol Monogr 72:185–202
Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA et al (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–332
Guy B, Almond J, Lang J (2011) Dengue vaccine prospects: a step forward. Lancet 377:381–382
Guzman MG, Kouri G (2002) Dengue: an update. Lancet Infect Dis 2:33–42
Hilborn R, Mangel M (1997) The ecological detective. Confronting models with data. Princeton University Press, Princeton
HMN (2008) Framework and standards for country health information systems. WHO, Geneva
Keeling MJ, Eames KTD (2005) Networks and epidemic models. J R Soc Interface 2:295–307
Keeling MJ, Grenfell BT (1997) Disease extinction and community size: modeling the persistence of measles. Science 275:65–67
Keeling MJ, Rohani P (2008) Modelling infectious diseases in humans and animals. Princeton University Press, Princeton
King AA, Ionides EL, Pascual M, Bouma MJ (2008) Inapparent infections and cholera dynamics. Nature 454:877–880
Levin BR, Lipsitch M, Bonhoeffer S (1999) Population biology, evolution, and infectious disease: convergence and synthesis. Science 283:806–809
Montagnier L (2002) Historical essay. A history of HIV discovery. Science 298:1727–1728
Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R et al (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5:e74
Nacher M (2001) Malaria vaccine trials in a wormy world. Trends Parasitol 17:563–565
Nacher M (2006) Worms and malaria: resisting the temptation to generalize. Trends Parasitol 22:350–351
Neyrolles O, Quintana-Murci L (2009) Sexual inequality in tuberculosis. PLoS Med 6:e1000199
Pascual M, Rodó X, Ellner SP, Colwell R, Bouma MJ (2000) Cholera dynamics and El Niño-Southern Oscillation. Science 289:1766–1769
Pascual M, Bouma MJ, Dobson AP (2002) Cholera and climate: revisiting the quantitative evidence. Microbes Infect 4:237–245
Racloz V, Ramsey R, Tong S, Hu W (2012) Surveillance of dengue fever virus: a review of epidemiological models and early warning systems. PLoS Negl Trop Dis 6:e1648
Rohani P, King AA (2010) Never mind the length, feel the quality: the impact of long-term epidemiological data sets on theory, application and policy. Trends Ecol Evol 25:611–618
Rohani P, Earn DJ, Finkenstädt B, Grenfell BT (1998) Population dynamic interference among childhood diseases. Proc R Soc Lond B 265:2033–2041
Rohani P, Green CJ, Mantilla-Beniers NB, Grenfell BT (2003) Ecological interference between fatal diseases. Nature 422:885–888
Rohani P, Zhong X, King AA (2010) Contact network structure explains the changing epidemiology of pertussis. Science 330:982–985
Shaman J, Kohn M (2009) Absolute humidity modulates influenza survival, transmission, and seasonality. Proc Natl Acad Sci U S A 2009(106):3243–3248
Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M (2010) Absolute humidity and the seasonal onset of influenza in the continental United States. PLoS Biol 8:e1000316
Soper HE (1929) The interpretation of periodicity in disease prevalence. J R Stat Soc A 92:34–61
Van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H et al (2013) Contagious diseases in the United States from 1888 to the present. N Engl J Med 369:2152–2158

Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312:447–451

Wearing HJ, Rohani P (2006) Ecological and immunological determinants of dengue epidemics. Proc Natl Acad Sci U S A 103:11802–11807

WHO-VMI Dengue Vaccine Modeling Group (2012) Assessing the potential of a candidate dengue vaccine with mathematical modeling. PLoS Negl Trop Dis 6(3):e1450

Xia Y, Bjørnstad ON, Grenfell BT (2004) Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. Am Nat 164:267–281