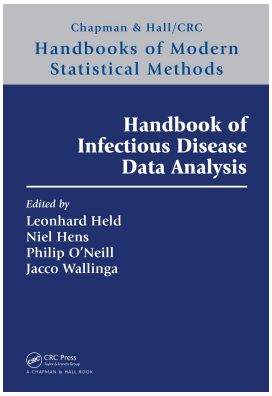

Handbook of Infectious Disease Data Analysis

Leonhard Held, Niel Hens, Philip O'Neill, Jacco Wallinga

Infectious Disease Data from Surveillance, Outbreak Investigation, and Epidemiological Studies

Publication details
https://www.routledgehandbooks.com/doi/10.1201/9781315222912-3
Susan Hahné, Richard Pebody
Published online on: 04 Nov 2019

3

Infectious Disease Data from Surveillance, Outbreak Investigation, and Epidemiological Studies

Susan Hahné and Richard Pebody

CONTENTS

3.1 Introduction

Before deciding the most appropriate analytical method and then applying it to infectious disease data, a thorough understanding of the different types of data which are available, the way in that they are collected, and the limitations inherent to them is essential. This knowledge will influence the approach you might take and then what conclusions you may or may not be able to draw from your analyses.

In this chapter, we first provide a general introduction to the available types of data on infectious diseases, which we divide into two kinds: epidemiological and microbiological data. 'Big data' does not necessarily fall into these two categories and is discussed in the context of surveillance (Section 3.3). Subsequently, in Sections 3.3 and 3.4, we address the main methods of how data on infectious diseases can be collected: by surveillance, outbreak investigation and by dedicated studies. In these sections, more details on specific types of data derived from these methods will be provided. Examples of the kinds of data and study designs are given throughout the text.

The overall aim of the current chapter is to provide the reader with a sense of the types of data available on infectious diseases, how they are generated, their strengths and limitations, and to guide readers towards where data may be obtained.

3.2 Infectious Disease Data: General Aspects

3.2.1 Epidemiological data

Basic epidemiological data on infectious diseases in individuals or populations includes information on (possible) exposure to infectious agents, other determinants of infection (risk factors), the occurrence of symptoms and microbiological evidence of an infection, demographic characteristics of the host, and, in cohort studies, on the time of follow up of individuals in the study and the reasons for loss to follow up. More specific epidemiological data such as contact pattern data are discussed elsewhere (in Chapter 6).

In infectious disease epidemiology, it is important to realize that the definition of what constitutes an exposure to an infectious agent varies depending on the pathogen, and what is known about its routes of transmission and infectiousness. An individual who has been present in a room with an infectious case of measles can be classified as exposed to measles, whereas merely being present in a room with a person who is infected with hepatitis B does not indicate exposure to hepatitis B. The definition of what constitutes an effective exposure may change over time when more knowledge of the pathogen becomes available. The concept of exposure is central to the definition of what constitutes an epidemiological link: An individual can be classified as having an epidemiological link if exposed to a person in whom the infection is diagnosed. Establishing whether a person has an epidemiological link needs to take into account the characteristics of the pathogen, including its infectiousness and routes of transmission and having detailed information on the exposure of the individual.

Data on the presence of determinants for infection other than exposure to a pathogen is usually also collected in epidemiological studies, e.g., to study the association between these determinants and an outcome or to assess the representativeness of the study population. Determinants can include behavioral, environmental, and demographic characteristics, and information on these determinants often is collected through questionnaires that are completed through an interview or by participants to a study. The quality of the questionnaire and the way it is completed is a main determinant of the validity and usefulness of the resulting data. On-line questionnaires completed by the study participant avoid the need for data entry by the research team but can of course also contain data entry errors. When data is entered from paper questionnaires, double data entry is good practice to limit this.

A central step in the generation of epidemiological data on infectious diseases is to decide which individuals are classified as infected. This classification is most systematically done by a priori agreeing on a case definition. Case definitions in infectious disease epidemiology usually consist of a set of clinical, epidemiological, and/or microbiological criteria. The

clinical criteria in a case definition usually consist of the typical symptoms of the disease. The epidemiological criteria are defined by what constitutes an epidemiological link (see above) while microbiological criteria are defined by laboratory testing (see Section 3.2.2). International organizations such as the World Health Organization (WHO) the European Center for Disease Prevention and Control (ECDC), and the Brighton Collaboration have established case definitions to aid standardized data collection. The importance of case definitions for surveillance and outbreak investigations is outlined in Section 3.3 and 3.4, respectively.

3.2.2 Microbiological data

Exposure to micro-organisms (bacteria, viruses, fungi, or parasites) of an individual can result in a symptomatic or asymptomatic infection, which, for certain pathogens (such as human immunodeficiency virus (HIV) or hepatitis B), can result in a carrier status. Exposure to certain pathogens may also merely lead to colonization, in which the pathogen is present in certain non-sterile body compartments such as the skin or mucosa. Infection, carriage and colonization all may result in transmission to others.

For most infections, a definitive conclusion on whether it has occurred in an individual depends on confirmatory testing in a microbiology laboratory. There are only a few pathogens in which certain clinical symptoms are thought to be pathognomonic of infection. An example of this is measles, characterized by fever and rash, whereby the appearance of so-called Koplik spots on the patient's oral mucosa is highly specific for the diagnosis. Since Koplik spots only occur in about 70 percent of measles cases and are often not recognized, the sensitivity of this symptom is only moderate: Many measles cases will be missed if only cases with Koplik spots are considered true cases.

Data on microbiology testing aiming to diagnose an infection typically consists of the type of test that was used (e.g., polymerase chain reaction (PCR)), the type of sample which was tested (e.g., a nose swab), and the test result. The latter can be qualitative (e.g., positive, negative, equivocal), semi-quantitative (e.g., intensity of a result), or quantitative (a numerical value). Some laboratory tests can only provide a qualitative test result. An example of this is culture of pathogens, which only indicates whether or not a certain pathogen is present in a sample. Quantitative test results can be transformed into qualitative data by using thresholds for what constitutes a negative, equivocal, or positive result. These thresholds often are defined by the manufacturer of the test, reflecting a specific aim of testing, e.g., diagnosing infection with a high degree of specificity. This aim may differ from the needs of a specific study, and hence pre-defined kit thresholds may not be optimal. Therefore, it is preferable to use quantitative data when available.

Laboratory testing to diagnose an infection can further be distinguished into two types: tests aimed at detecting the pathogen (e.g., PCR) and tests to identify the immunological reactions generated by the body upon encountering a pathogen (e.g., serological tests). The sensitivity of these methods highly depends on the timing of sampling relative to the onset of symptoms. To detect pathogens in acute infections, samples need to be taken relatively soon after onset of symptoms, while full characterization of immunological reactions requires later samples.

After diagnosing a pathogen at the genus level (e.g., Salmonella), further characterization at the species and subspecies level can be done by applying specific methods. Determination of, e.g., strains, clones, and sequence types currently usually requires characterizing part of the pathogen's genome by sequencing it. Recent advances in sequencing methods have reduced the time and costs needed for this greatly, and hence genomic data, including whole genome sequencing (WGS) data is becoming increasingly available. A relatively new area of molecular microbiology, metagenomics, involves characterization of the entire genomic

content of a biological sample. High-throughput sequencing methods, labeled as next generation sequencing (NGS) are needed for this purpose. Metagenomics is applied to studying microbiomes, which can be defined as the population of microorganisms that inhabit a certain location (e.g., the nasopharynx or the gut). Characterizing such populations requires methods which do not depend on culturing micro-organisms, since many pathogens cannot be grown in culture yet. In addition to assessing their genomic content, microbiomes also can be characterized by studying small molecules or proteins present in the sample [1].

A separate area of microbiological testing is aimed at documenting indirect evidence of infection by assessing the body's immunological response to it. These methods also are used to assess vaccine induced immunity. The immune response to infection (and vaccination) usually consists of two different mechanisms: a cellular and a humoral response. In the cellular response, immune cells directly attack the pathogen, while the humoral response acts through antibodies. Microbiological tests aimed at diagnosing infections by assessing the immune response are focused mostly on testing the presence of humoral rather than cellular immunity, since standardized assays for the latter are lacking. Data resulting from microbiological tests of humoral immunity can give qualitative and quantitative results about the presence of antibodies specific for a certain pathogen. It also can give an indication of when the infection was acquired by performing avidity testing, which is used in, e.g., HIV diagnosis and surveillance [2].

In addition to diagnosing infections and characterizing pathogens, microbiological testing also can provide other data relevant from a public health perspective. This includes data to assess the infectiousness of an infected individual and whether the pathogen causing the infection is resistant to antimicrobial drugs. Assessing infectiousness can be done by determining the pathogen load (e.g., the viral load, which represents the number of copies of the virus present in the sample). Testing for antimicrobial resistance can be done by microbiological tests in which cultured bacteria or fungi are exposed to a panel of antimicrobial agents to assess their effect on growth of the bacteria. Applying certain clinical breakpoint criteria when assessing pathogen growth results in a qualitative test result (e.g., 'resistant'). The internal and external validity of results depends on the quality of the laboratory procedures and on the use of standard clinical breakpoint criteria. The presence of antimicrobial resistance in micro-organisms also can be detected by genotypic methods sequencing genes coding for resistance traits. However, the presence of such genes does not necessarily indicate the presence of resistance since this also depends on the level of gene expression.

3.2.3 Data errors and bias

When analyzing data, it is important to be aware of potential errors which may be inherent, since ignoring these may lead to incorrect conclusions. Errors in epidemiological or laboratory data can be classified into two main categories: those arising from random errors and those arising from systematic errors. Random errors are due to chance driven variation in measurements or sampling. The size of random error can be reduced by increasing the sample size of a study, which will increase the precision of the measurement. Systematic errors in data can result from non-representative sampling procedures (e.g., working with self-selected participants who are likely to differ from randomly selected participants) and from systematic measurement errors. Increases in sample size will not reduce the size of systematic errors.

Systematic data errors only lead to bias when they are differential, i.e., when the extent of it depends on study participants' characteristics and their outcomes. Bias is defined as a systematic deviation of results or inferences from truth [3]. In general, we can try to avoid bias by having an optimal design, collection, analysis, interpretation, and reporting of a

study. Bias in epidemiological studies can arise in many ways, and dozens of types of biases have been described. The two major types of bias we will describe here are selection bias and information bias. We will not discuss confounding bias, since it is mainly a problem with epidemiological interpretation of results of studies into effects of determinants, while selection bias and information bias are also applicable to broader use of data.

Selection bias occurs when the chances of individuals to be included (or stay) in the study population are related to their level of exposure and the occurrence of the outcome of interest. When selection bias is present, study participants differ from non-participants (who are eligible for the study) in terms of the relation between the exposure and the outcome. An example of selection bias distorting a vaccine effectiveness study is when vaccinated cases who have become ill are more likely to participate than vaccinated individuals who have not become ill. This bias would result in an underestimate of vaccine effectiveness.

Information bias refers to flaws in measuring exposure, covariate, or outcome variables that result in different accuracy of information between comparison groups [3]. Invalid information can lead to misclassification of the exposure status of individuals. When the study aim is to assess the effect of an exposure on an outcome, it is of key importance to assess whether this misclassification is differential (i.e., differing between cases and non-cases) or non-differential. Non-differential misclassification results in underestimating an effect, while differential misclassification can lead to over- or underestimating effects. In a vaccine effectiveness study, information bias can arise when cases of the infection of interest are more likely to remember that they were vaccinated than people who have not become ill. This would result in an underestimation of the protective effect of vaccination.

Systematic errors in microbiological data can result from biased (non-representative) sampling procedures, systematic mistakes in handling of samples, and laboratory procedures (e.g., poor quality growth medium). It is key to understand the sensitivity and specificity of testing, and whether cross-reactions occur, to interpret test results well. The application of different testing algorithms or expert rules also can result in biased results. If, e.g., antimicrobial susceptibility testing for antimicrobial agent B is only performed when the pathogen is resistant against agent A, test results for agent B are not representative of the entire population.

3.3 Data from Surveillance of Infectious Diseases

3.3.1 Introduction

Surveillance was first used in public health in the 14th century in southern Europe, when it was carried out to detect cases of plague among people who had been placed in quarantine, usually after travel abroad. Early detection of cases and their subsequent isolation was a direct tool to control the spread of infectious diseases. It is used still in a similar fashion currently as the main method to control serious infectious diseases such as ebola, for which until recently no effective vaccine existed. Surveillance of diseases rather than of individuals was first systematically performed by John Graunt (1620–1674), who analyzed the London Bills of Mortality to understand patterns of death and in particular to understand the extent of the problem with plague. This work was further developed by William Farr (1807–1883), a statistician who introduced a certificate for cause of death and used mortality data to understand the course of infectious disease epidemics in order to identify opportunities for control [4]. In the 1950s, Alexander Langmuir, generally considered the father of modern surveillance, defined disease surveillance as "the continued watchfulness

over the distribution and trends of incidence through the systematic collection, consolidation and evaluation of morbidity and mortality reports and other relevant data together with the timely and regular dissemination to those who need to know" [5]. In 1968, the World Health Assembly summarized this definition as "surveillance is information for action" [6]. Surveillance is, together with outbreak investigation and epidemiologic studies, a main tool of applied infectious diseases epidemiology. It is aimed at generating evidence to support infectious disease prevention and control by identifying trends and outbreaks, aiding public health prioritization of disease control and evaluating interventions. It also can be used to generate hypotheses for further research.

The aim of this section is to provide a general overview of surveillance methods, examples of surveillance data, and information on where it may be accessed. The sections on surveillance methods do not aim to instruct the reader on how to set up, perform, or evaluate surveillance systems, but rather to aid access to and appropriate interpretation of the data that is generated by such systems. Liaising closely with surveillance and disease control experts at an early stage when considering using surveillance data is important to ensure optimal data is obtained and to avoid misinterpretation.

3.3.2 The population under surveillance

To understand the data resulting from an infectious disease surveillance system, it is important to know what was defined as the population covered by the system. In national, population-based surveillance, the population under surveillance is usually the entire population of a country. However, in many instances this type of comprehensive surveillance may not be feasible or necessary. An alternative to comprehensive surveillance is sentinel surveillance, whereby data is coming from a selection of hospitals or general practitioner (GP) practices only (see Section 3.3.4.1). In this situation the size and characteristics of the population under surveillance are important to understand to ensure, e.g., that accurate age-specific incidence rates can be calculated. In the surveillance of zoonoses, animals may be the population under surveillance. To be able to interpret surveillance data, its representativeness for the population under surveillance needs to be carefully assessed: are population subgroups such as undocumented migrants included? Is there a bias towards reporting cases in certain age groups or with certain (more severe) disease manifestations? To assess this is usually problematic, since a gold standard dataset often is not available. It requires close liaison and discussion with surveillance and disease experts.

3.3.3 The use of case definitions in surveillance

Infectious disease surveillance systems usually aim to identify people or animals infected with a certain pathogen in order to assess the incidence rate of the infection of interest in a population. Since asymptomatic infections are difficult to detect unless serological assessment is undertaken of a probabilistic sample of the entire population, this means in practice identifying people or animals with disease symptoms that may be caused by the infection. To allow comparison of the incidence rate in one population with the rate in other populations, consistent use of the same criteria to decide who is a case is a prerequisite. The set of criteria used to define a case is called a case definition. Case definitions are not only essential in surveillance, but in many epidemiologic studies including outbreak investigations.

Case definitions in infectious disease epidemiology usually consist of a set of clinical, epidemiological, and/or microbiological criteria, a description of the population that is being studied, and, particularly in outbreak investigations, a point in time from which onwards cases are counted. International organizations such as ECDC and the WHO and many

countries have established a set of standard case definitions which aid standardized data collection. An example of this is the European Union (EU) case definition for measles (see Box 3.1).

BOX 3.1 EU CASE DEFINITION FOR MEASLES

Clinical Criteria: Any person with fever AND Maculo-papular rash AND at least one of the following three: Cough, Coryza, Conjunctivitis

Laboratory Criteria: At least one of the following four: Isolation of measles virus from a clinical specimen, Detection of measles virus nucleic acid in a clinical specimen, Measles virus specific antibody response characteristic for acute infection in serum or saliva, Detection of measles virus antigen by DFA in a clinical specimen using measles specific monoclonal antibodies. (DFA = direct fluorescence antibody)

Epidemiological criteria: An epidemiological link by human to human transmission

Case Classification

- Possible case: Any person meeting the clinical criteria
- Probable case: Any person meeting the clinical criteria and with an epidemiological link
- Confirmed case: Any person not recently vaccinated and meeting the clinical and the laboratory criteria

Source: `https://ecdc.europa.eu/en/infectious-diseases-public-health/surveillance-and-disease-data/eu-case-definitions`

The design of a case definition for surveillance purposes needs to consider the level of sensitivity, specificity, and positive predictive value required to achieve the aim of the surveillance. It also needs to reflect the diagnostic practices and health care seeking behaviour in a certain setting. If microbiological testing is rarely performed for a certain condition, a surveillance case definition including only microbiological criteria is inappropriate. An example of this is influenza, the incidence of which is usually assessed by counting the number of people presenting with influenza like illness (ILI) in primary care, irrespective of any microbiological diagnosis.

To allow some flexibility in the sensitivity, specificity, and positive predictive value of a case definition, often layered, hierarchical definitions are proposed for suspected and confirmed cases (see Box 3.1). The criteria for a suspected case usually include only a set of clinical symptoms, while confirmed cases have a higher degree of certainty to have the disease of interest. This certainty can be based on the results of microbiologic testing and/or on epidemiologic information increasing the likelihood that the symptoms of disease are due to a certain infection (see information on epidemiologic link in Section 3.2).

3.3.4 Data for surveillance of infectious diseases

The data sources used in surveillance of a particular disease are determined by the aim of the surveillance, which data sources are (routinely) available, the required quality of the data, the resources required to implement the system, and the feasibility of running the surveillance system. For the system to be sustainable, the requirements for those contributing the data and maintaining the system need to be as simple as possible. This usually compromises the quality of the data, leading to suboptimal timeliness, completeness, or accuracy. To compensate for this, it is useful to include several data sources in a surveillance system,

so that by triangulation the number and validity of conclusions that can be drawn from the surveillance can be enhanced (see Box 3.2 and Figure 3.1). Below, main data sources used in infectious disease surveillance are discussed.

BOX 3.2 USE OF MULTIPLE DATA SOURCES FOR THE SURVEILLANCE OF ROTAVIRUS IN THE NETHERLANDS

Rotavirus (RV) causes acute gastro-enteritis (AGE), which is in most cases short-lived and does not lead to health-care visits. RV infection is not notifiable in the Netherlands. The main source of data for RV surveillance in the Netherlands is the weekly reporting of the number of RV diagnoses by a number of virological laboratories. This system does not yet fully capture the number of laboratory tests performed. A decrease in the number of RV diagnosis can therefore reflect a genuine decrease in transmission of the virus, but also a decrease in how frequent cases with AGE are tested for RV. Considering an additional, independent data source such as the frequency of general practitioner (GP) visits for AGE facilitates drawing conclusions on the incidence of RV infection. The absence of a clear RV peak in winter 2014 observed in data from the virologic laboratories was confirmed by data from GP surveillance [6]. See Figure 3.1.

3.3.4.1 Data generated in the health service

Health service providers are an important source of data for infectious disease surveillance because they routinely record health events which may be of interest to public health.

Notifiable diseases. Almost every country has a list of infectious diseases which clinicians and/or laboratories are obliged by law to report to their public health authorities.

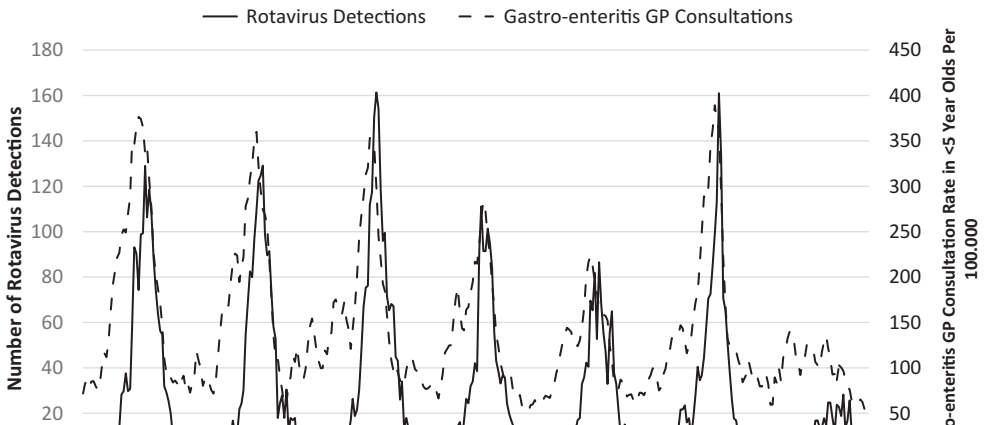

FIGURE 3.1
Number of rotavirus detections by a number of virological laboratories and frequency of GP visits for acute gastro-enteritis in <5 year olds by week, The Netherlands, August 2007 – August 2014. (Adapted from Hahné, S. et al., *Eurosurveillance*, 19, 20945, 2014.)

These notifiable diseases are usually severe with a potentially large impact on public health. Together with the report of the diagnosis of the disease, minimal information that is usually collected includes the date of onset, sex, age, place of residence of the patient, and some information on risk factors for infection such as recent travel history. The usefulness of data on notifiable diseases first depends on whether a clear and practicable case definition was applied when selecting which patients are notified. When the case definition is only based on clinical symptoms, its likelihood to indicate a true case of a certain infectious disease (i.e., the positive predictive value) depends on the prevalence of the disease in the population. Second, the usefulness of notification data depends on the completeness of reporting. Usually, notified cases represent only the tip of the iceberg of all patients with the notifiable condition, since not all patients consult health care, and reporting by health care providers is usually also incomplete. This may not necessarily be a problem for public health decision making, depending on how constant and representative the reporting is. Changes in completeness of reporting over time can be caused by changes in the health care seeking behavior of patients (e.g., due to media coverage of an outbreak) or by changes in diagnostic practices or the availability of a new test. Completeness of reporting of notifiable diseases can be improved by obliging microbiological laboratories to report patients with a notifiable disease, since for most notifiable diseases a confirmation by microbiological testing is part of the case definition. This reporting can be potentially supported by automated extractions from laboratory information management systems. Timeliness is another attribute of surveillance data which influences its usefulness for public health decision making. In rapidly emerging outbreaks, delays in the time between onset of the infection and notification can make it difficult to assess whether the outbreak is ongoing or subsiding. Using techniques to adjust for such reporting delays and to create thresholds based on historical incidence, it is possible to create automated alerts when significantly more notifications are being seen than expected. This is the basis for further investigations to determine whether the increase is real and what might be explaining it. In addition to compulsory notification, voluntary reporting of health events by health care providers and/or microbiological laboratories can be organized through reporting schemes.

Sentinel surveillance. For infectious diseases which are common and which do not require a public health response for every case (such as influenza), reporting of all cases in a country usually is not necessary for public health decision making. In these situations it may be more efficient to establish a sentinel surveillance system where only a selection of health service providers or laboratories report cases. Having only a subset of health service providers or laboratories involved in reporting can make it feasible to improve the quality of data through training of data providers and to enhance the data by collecting additional information or laboratory testing of cases, as outlined earlier. An example of sentinel surveillance established by many countries is the GP sentinel influenza surveillance, where GPs report patients consulting with acute ILI and take a respiratory swab from a sample of patients to test in a laboratory for influenza. This type of surveillance has proved invaluable to monitor the intensity of influenza transmission, the dominant circulating strains, and to act as a platform for vaccine effectiveness studies to measure the performance of seasonal influenza vaccine and inform optimal selection of vaccine strains by the annual WHO Vaccine composition meeting.

Routinely recorded health care and mortality data. Routinely recorded health service data on medical encounters (i.e., electronic health records) also can be a useful source of data for infectious diseases surveillance. This data includes the routine registration of health events in primary care, specialist care (e.g., sexually transmitted infection (STI) clinics), hospitals, by health insurance companies, and death registration (see Box 3.3). It also includes pharmacotherapeutical data, e.g., on the number of prescriptions of antibiotics in a certain population.

Routine health service data is most useful when information on health events is coded, e.g., using the International Classification of Diseases (ICD) developed by the WHO. An overview of coding systems is available from https://www.nlm.nih.gov/research/umls/sourcereleasedocs/#.

Most coding systems do not require or include information on laboratory confirmation, which limits the specificity and therefore usefulness of the data for infectious disease surveillance and research. Furthermore, there may be regional and temporal differences and changes in the way that coding is undertaken. Another limitation of routinely collected health data is that there often are significant delays between the occurrence of the health event and the availability of the data, although methods are available to adjust for this.

BOX 3.3 ROUTINELY RECORDED MORTALITY DATA FOR PUBLIC HEALTH SURVEILLANCE

In Europe, a project for monitoring of excess mortality for public health action (EuroMOMO) assesses weekly 'real-time' all-cause age-specific excess mortality in countries in Europe. By using a standardised approach, results can be pooled. Through this monitoring a significant peak in mortality in adults >65 years of age in 2014/2015 was detected [7]. See Figure 3.2.

Surveillance using data on only clinical symptoms is an example of syndromic surveillance. The main advantage of syndromic surveillance is its timeliness, since it uses data which may be available before microbiologic diagnoses are established. It can include a wide range of health data such as GP or emergency room consultation data, telephone consultation data, and data on medical purchases (e.g., other-the-counter drugs). It also can include non-health-data, as outlined in 3.3.4.2. An example of syndromic surveillance data is provided in Box 3.3 and Figure 3.2.

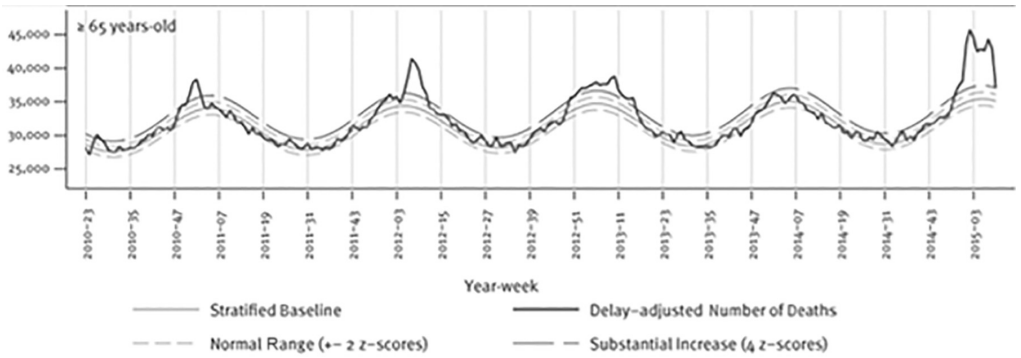

FIGURE 3.2
Number of deaths by week and modelled baseline in persons >65 years of age, obtained from pooled analysis of data from EuroMOMO countries, week 23, 2010 – week 9, 2015 ($n = 14$ countries). (Adapted from Mølbak, K. et al., *Eurosurveillance*, 20, 2015.)

Laboratory surveillance. Obtaining data on infections in humans directly from microbiological laboratories for surveillance purposes is attractive since a laboratory diagnosis is usually highly specific and data is often electronically available. Laboratory surveillance is the key method for surveillance of antimicrobial resistance. Data from laboratory surveillance may include basic demographic information such as age, gender, and date of sample and results of individual patients' microbiological tests. More information on the types of laboratory data that may be available is provided in Section 3.2.2. The interpretation of laboratory data for infectious disease surveillance purposes is greatly facilitated if, in addition to data on the number of positive tests in a certain period of time, also data on the number of tests that were performed in that period is available. This information allows disentangling whether any observed changes in the incidence of an infection are due to increased testing (such as following the introduction of near-patient testing) or likely to reflect a genuine increase in the number of infected individuals.

The validity and precision of laboratory data first depends on the quality and type of laboratory test used. Random errors in the data can result from measurement variation. Systematic errors can result from certain decisions during the collection, processing, and analyses of the laboratory data (3.2.3). Limitations of laboratory data include that usually information on the symptoms of the patient and the clinical outcome are not available. Another difficulty can be that results of repeat testing of the same patient are included, which makes it difficult to distinguish separate infection episodes. Many systems have developed approaches to de-duplicate such repeats. However to optimize comparability it is important that standard approaches are employed.

Surveillance of zoonoses. Surveillance of zoonoses (infections which can spread between vertebrate animals and humans) usually requires dedicated surveillance in not only humans, but also animal reservoirs, vectors and sometimes the environment (so called one health surveillance). Birds are an important reservoir for a large number of influenza viruses. Some of these viruses occasionally take on the ability to spread from birds to humans, and occasionally then from person to person (which can then develop into a pandemic). Early detection of viruses that may have acquired the potential for spread from bird to human is important as part of pandemic preparedness. A recent example is A(H7N9), which is currently circulating in poultry flocks in China, and has acquired the ability to spread from bird to human. Avian surveillance has been important to detect if the virus has spread elsewhere in China and to neighboring countries.

International surveillance. The importance of international surveillance is being increasingly recognized, as infectious diseases do not respect borders and can easily spread as was seen with SARS in 2003. The International Health Regulations are a surveillance tool whereby countries are compelled to report events of international public health importance at the earliest possible stage. The sharing of such information at an early stage helps to support control and prevention efforts as was well illustrated with emergence of Middle East Respiratory Syndrome-Corona Virus (MERS-CoV) in 2011. International organizations such as WHO and ECDC have international surveillance as one of their main tasks.

Surveillance pyramid. The previously listed sources of data generated in the health service can be visualized in the form of the surveillance pyramid (Figure 3.3). To understand the validity of data resulting from surveillance, it is necessary to understand to what extent individuals in a certain level of the pyramid are representative of the individuals in a lower level. Since surveillance data is derived from real-life situations rather than dedicated studies, the determinants of why certain individuals go from one level of the pyramid to the next are usually not random, and this can result in important biases. When only severely ill people are tested and the severity of the infection varies by age, the age distribution of laboratory positive cases is not representative of all individuals with the infection of interest.

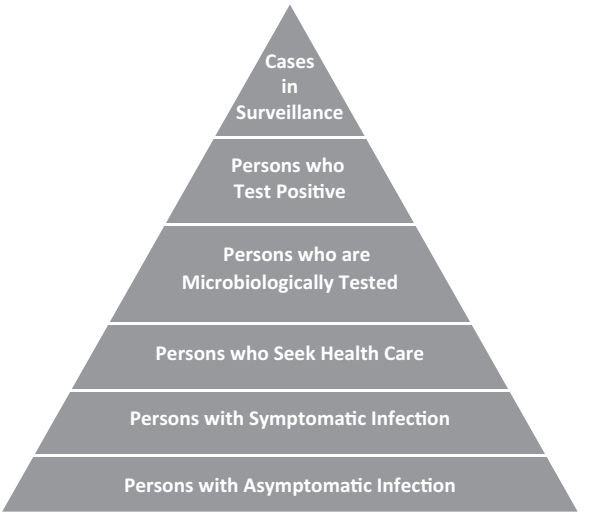

FIGURE 3.3
The surveillance pyramid.

When many cases with an infection die at home before reaching health care and diagnostic facilities, the burden of disease and case fatality may be underestimated.

3.3.4.2 Non-health data for the surveillance of infectious diseases

Denominator data. A key source of non-health data which is important for surveillance is demographic data on population denominators. In many countries this is collected and made available by a national statistics bureau, either derived from a population census or, in some countries, from a population register. Denominator data is crucial to be able to calculate incidence rates and to assess how representative surveillance data is of the general population.

Citizen science. The growing use of the internet has enabled a form of surveillance where members of the public are encouraged to report certain health issues. Examples of this are the tick bite radar in the Netherlands (www.tekenradar.nl) and the Flusurvey in the UK (https://flusurvey.org.uk/). Data often is presented on-line to provide feedback to the population. These surveys can be tailored to rapidly answer acute questions. These initiatives can be classified as citizen science, a phrase which encompasses a wide range of research involving active citizen participation. Biases in data generated by citizen science may arise since the data is generated by a self-selected part of the population, which is generally better educated and younger. Whether any biases are important for the interpretation of the study results depends mainly on the study question.

Big data. The availability of digital data has enormously increased in the past 15 years or so, mainly due to growing use of the internet, social media, and mobile phones. This data often labeled as "big data," a term which is applied to data of large volume, rapid availability (velocity), and large variety, with some other "V" words being added to this description (veracity validity, value, volatility). The use of big data for public health surveillance lagged behind its use for, e.g., marketing purposes, but is increasingly recognized. It includes analyses of data generated through the use of search engines on the internet, such as Google. The advantages include the fast availability of the data, and that it includes information from individuals who have not consulted formal health care. However, there are many potential

biases that need to be borne in mind consequent upon the way that the population interact with the internet and algorithms are fitted. For example, Google's "Flu and Dengue Trends" initiatives were ceased after results were disappointing, due to reasons including overfitting of the underlying algorithms. In general, it also is important to realize that disease-related internet search queries or messages in social media Twitter or Facebook may reflect a variety of situations which are unrelated to the real occurrence of infections.

3.3.5 Confidentiality and privacy in surveillance

In most countries, laws exist which grant special rights to public health authorities to collect data on infectious diseases in humans, animals, and institutions for public health surveillance. Unlike research, the data collection for the purpose of public health surveillance usually is not overseen by an ethical review board, and does not involve obtaining informed consent from participants. This arrangement creates ethical questions since it infringes on the privacy of individuals. To protect privacy, public health organizations, therefore, are bound by laws on how data should be handled in a confidential manner. A key measure to improve confidentiality is to remove personal identifying information wherever this is possible. However, some personal identifiers remain necessary, especially at local levels, to obtain follow-up information or to de-duplicate data. To improve confidentiality, surveillance data at national level is usually anonymized. However, by linking data sources (e.g., by probabilistic linkage methods), the usefulness of the data may be much increased but at the same time individuals may become identifiable. Linking data, therefore, is a controversial area in surveillance. Professionals involved in surveillance are usually bound by certain rules of conduct aimed at further improving confidentiality.

The release of surveillance data is in some countries also governed by dedicated laws, e.g., granting the public access to all information which was collected by using collective funds. This again may infringe on privacy, so personally identifiable surveillance data is usually exempt from this. Surveillance data should be made available to the public in such a way that individuals cannot be identified even through deductive disclosure. The rules for this, the minimum number of people in cells of tabulated data, are becoming increasingly well defined.

When data collection is done for research purposes, ethical guidelines usually require a review by an ethical review board and often involve obtaining informed consent from participants [8]. The informed consent procedure should include at least a clear explanation of the purposes of the data collection, what it will be used for, who can access the data, and how long it will be stored.

3.3.6 Where to access surveillance output and data?

Typically, public health departments responsible for infectious disease surveillance routinely produce reports presenting surveillance indicators such as incidence and prevalence of a selected group of infectious diseases or syndromes, aggregated by demographic variables, risk factors, or geographical regions. These reports contain aggregate data, usually without detailed information on age, sex, or location of the patient. These reports are regularly shared with stakeholders contributing data, public health policy makers, and often with a wider audience.

More basic outputs of surveillance, such as the weekly number of notified cases of infectious diseases, usually is available also from public health institutes' websites. This data is usually still in aggregated form, often in a pdf which cannot directly be used for data analyses.

Researchers usually require more detailed, disaggregated data for analyses, ideally available in an electronic format which is directly useable. The availability of this type of data is increasing, as a result of specific projects (e.g., Project Tycho (http://www.tycho.pitt.edu/) and efforts of public health institutes (e.g., the German 'survstat' https://survstat.rki.de/default.aspx, the English Fingertips project http://fingertips.phe.org.uk/profile/health-profiles/data#page/0, and the European Atlas of Infectious Disease https://ecdc.europa.eu/en/surveillance-atlas-infectious-diseases). Availability of disaggregated data is further increased by the requirement of an increasing number of scientific journals authors to make it available upon publication of a research article. Despite these efforts, unfortunately, many barriers to sharing health data remain.

3.4 Data from Observational Epidemiological Studies and Outbreak Investigations of Infectious Diseases

3.4.1 Introduction

Epidemiology is the study of the occurrence and distribution of health-related states or events in specified populations, including the study of determinants influencing such states, and the application of this knowledge to control the health problems [3]. Epidemiological studies can be experimental or observational. Examples of experimental studies are clinical trials used to study the effects, safety, and impact of pharmaceutical treatments or vaccination.

In this section, we focus mainly on observational epidemiological studies, in which the researcher does not interfere with the conditions or determinants that are studied. Surveillance as a specific type of observational epidemiological study is discussed in 3.3. Observational studies can be used to generate or test hypotheses. The former is usually done by descriptive epidemiology while testing hypotheses requires analytical epidemiological designs and methods. Analytic epidemiological studies can have individuals or populations as the unit of observation. For the study of individuals, three main study designs exist: cohort (prospective) studies, case-control studies, and cross-sectional studies. Observational studies in which populations rather than individuals are studied are called ecological studies. All of these study designs, except the cross-sectional design, can be applied both prospectively and retrospectively, depending on whether the exposure measurement occurs before or after the disease occurrence.

In this section, we will describe the designs and methods used in these different types of observational epidemiological studies. In the last sections, we will discuss methods of the investigation of outbreaks and of emerging diseases, both particularly relevant for the epidemiological study of infectious diseases. Throughout the text, we will provide examples of data derived from some of these epidemiological study designs. The main aim of this section is to provide background information about where and how epidemiological data may be generated and inherent limitations which may be arising from this.

3.4.2 Cohort studies

Of all epidemiological observational study designs, the cohort study is the main design to make inferences on causality, since it follows people over time from exposure(s) to outcome(s) and hence provides information on the temporal relation between these two. In a cohort

study, a group of people, who differ in the extent to which they are exposed to a potential determinant of the disease(s) of interest, is followed up in time whereby ascertaining the occurrence of health outcome(s) of interest and, ideally, the duration of time of follow-up and reasons for drop-out. People who are lost to follow-up no longer contribute to the data, and are called censored observations: their disease outcome status remains unknown. When analyzing a cohort study in real-time (at several moments when the cohort progresses in time), the outcome for many individuals in the cohort may be yet unknown and these then are classified also as censored observations. In a cohort study, the association between exposure and health outcome is quantified by calculating a measure of association which compares the occurrence of the health outcome between groups of individuals with a certain exposure status. The measure of association can be a relative risk (not taking the time of follow-up into account), a rate ratio (taking the duration of follow-up into account) or a hazard-ratio (taking both the duration of follow-up and the timing of the health event into account).

Data of cohort studies typically contains individual information on certain exposures, other determinants influencing the risk of certain health outcomes (confounders which may be adjusted for in the analysis), the occurrence and timing of certain health outcomes, the time of follow-up, and reasons for droping-out of the study.

In this described design of cohort studies, individuals are typically included in the cohort irrespective (i.e., unconditional) of their exposure to an infection of interest. In epidemiological studies of infectious diseases, study designs can be used in which cohorts are followed up which are more or less uniformly exposed to a certain infection. These studies are called conditional on exposure and are, of course, only ethical when effective post-exposure prophylaxis and/or treatment are made available to participants. The main advantage of this type of study is that efficiency is increased (the incidence of infection is higher in these cohorts). An example of a cohort study design conditional on exposure is the household follow-up study, in which household members of cases of an infectious disease are followed-up over time whereby ascertaining new cases of the infection of interest in household members. Household studies are a specific type of a contact-tracing study. In contact tracing studie, individuals who have an epidemiological link to a case (see Section 3.2.1), i.e., contacts, are included as study participants. Data from contact tracing studies typically includes individual data on age, gender, infection status (and, in case of a symptomatic infection, the disease onset date), some identification of the most likely source, date(s) of exposure, and exposure duration [9].

Bias in data from cohort studies can occur in several ways. First, when people who are lost to follow-up differ in terms of determinants or characteristics from those who remain in the study, the resulting data is not representative of the initial cohort. This outcome can lead to a biased assessment of the effects of a certain determinant, when the loss to follow-up is related to both to the determinant and the outcome of interest. A second cause of bias can result from unequal assessment of health outcome status between exposed and unexposed individuals, or, vice versa, from unequal assessment of the disease status between exposed and non-exposed individuals. Another important source of bias occurs when the exposure status of study participants is dependent on factors which also are related to the health outcome of interest (confounding by indication, e.g., frailty bias in influenza vaccine studies). Other sources of bias, e.g., those discussed in 3.2.3, can also occur in cohort studies.

3.4.3 Case-control studies

In a traditional case-control study, cases of a certain disease are compared to controls (non-diseased individuals) in terms of their exposure to a certain determinant. The major

advantage over cohort methods is the efficiency achieved by the reduction in sample size. However, choosing suitable controls, which should be representative of the population that gave rise to the cases, can be problematic and may introduce bias. The measure of association which is calculated in case-control studies is the odds ratio, comparing the odds of exposure between cases and controls. Depending on the sampling method, the odds ratio can be directly interpreted as a risk of rate ratio [10].

Ideally controls should meet two criteria – first, their exposures should be representative of the population from which the cases arise and second assessment of determinants, e.g., vaccine status, in both cases and controls should be the same. The key principal to avoid bias in case-control studies is to ensure that cases and controls are derived from the same source population. Biased estimates of the odds ratio occur when cases systematically differ from controls in characteristics, that are related to both the exposure and health outcome of interest.

Different sources of controls can be used (e.g., from the community or other persons admitted to hospital), but it is often challenging to meet these two requirements. In attempts to reduce the occurrence of bias in case-control studies, alternatives to selecting healthy individuals as controls are being used. The test-negative design (TND) is a specific form of a case-control study design in which controls are individuals who present with the same clinical syndrome (e.g., influenza-like illness) but who have a negative microbiologic test result for the infection of interest. The main advantage of the TND is that biases arising from different health care seeking behavior (propensity to consult) between cases and controls is reduced [11]. Another variant of the traditional case-control study, is the case-case study. Here, the referent group consists of patients of another (form) of the disease.

Data generated during case-control studies is similar to that in cohort studies (see previous) except for that time of follow-up and censoring are not relevant. In addition to biases mentioned previously, biases listed in 3.2.3 also may apply.

3.4.4 Cross-sectional studies

In a cross-sectional epidemiological design, a group of individuals is studied by ascertaining exposures and outcomes at the same point in time. A prevalence study is an example of a cross-sectional study which aims to estimate the prevalence of a certain health status. In analyses of cross-sectional studies, groups of individuals with a different status regarding their exposure to determinants of infections can be compared in terms of their prevalence of infection. Alternatively, the odds of exposure can be compared between cases and non-cases. An important limitation of cross-sectional studies is that the sequence in time between a potential determinant and the outcome of interest is unknown, unless historical information is ascertained, without such longitudinal data it is not possible to ascertain a causal relationship. Biases occurring in cross-sectional study data can arise through overrepresentation of cases with a long duration of illness.

An example of a cross-sectional study design is a serological survey (see Box 3.4). This type of survey can be used in the evaluation of a vaccination program. It also can be used to study risk factors for infection, provided the infection gives rise to long-lasting antibodies (see Chapter 16) [12, 13]. Modeling approaches with underlying assumptions can be used to make estimates of incidence from such prevalence surveys. Biases in data from serological studies can arise due to the sampling of individuals who may not be representative of the population of interest. Cross-sectional seroprevalence data often is studied by age. When doing so, it is important to realize that an increase in seroprevalence by age can reflect both an ongoing and a decreasing risk of infection over time, representing an age and cohort effect, respectively. In one-off cross-sectional studies, it is impossible to disentangle these effects (Figure 3.4).

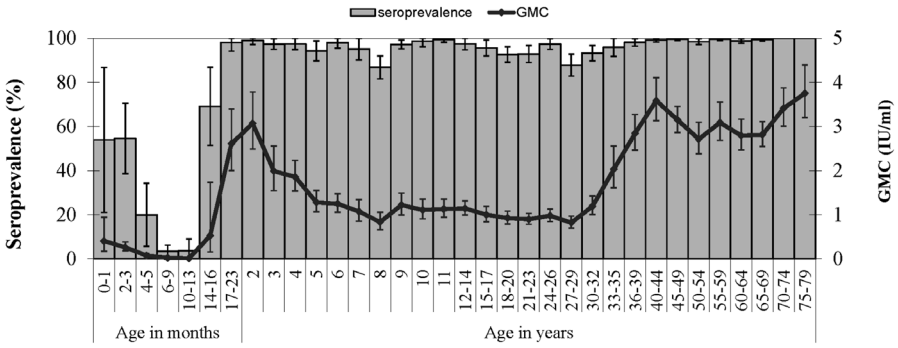

FIGURE 3.4
Weighted age-specific seroprevalence and geometric mean concentrations (GMC) of measles IgG antibody (with 95% confidence intervals) in the general population. (Adapted from Mollema, L. et al., *Epidemiol Infect.*, 42, 1100–1108, 2014.)

BOX 3.4 AN EXAMPLE OF A CROSS-SECTIONAL STUDY: MEASLES SEROPREVALENCE IN THE NETHERLANDS

Measles virus infection gives rise to an immune reaction, which includes the generation of antibodies that persist for life. This outcome also holds for measles vaccination, although a small proportion of vaccinated persons does not respond to the vaccine and antibodies wane over time. Measles virus specific antibodies can be quantified in serum, and when a threshold level is applied, the seroprevalence of measles antibodies can be assessed in a population [14].

In the Netherlands, a large survey of was undertaken in 2006/2007 which was used to assess measles immunity in the general population [14]. Over 6,000 individuals participated by giving blood and filling out a questionnaire. Serum samples were tested for measles virus specific IgG, and seroprevalence and geometric mean concentrations (GMC) calculated. Results are displayed in Figure 3.4 showing a gap in immunity in infants below the age of the first dose of measles containing vaccine (MCV) (14 months), a decline in antibody levels after MCV-1, and relatively high GMCs in older adults. The GMC and seroprevalence results by age need to be interpreted in the context of (historic) measles vaccination schedules, their uptake, and information on the incidence of measles virus infection over time.

3.4.5 Ecological studies

Studies in which the units of observation are groups of people rather than individuals are called ecological (or aggregate) studies. Data collected in ecological studies are measurements averaged over individuals, e.g., an incidence of an infectious disease in the population or the prevalence of uptake of preventive measures. Importantly, the degree of association between these population averages of disease and exposure may differ from the associations between the two assessed at an individual level. An oversight of this phenomenon is referred to as the ecological fallacy. Ecological studies also are prone to confounding, since populations studied are likely to differ in many factors potentially related to the disease being

studied. Because of these limitations, ecological studies are more useful for hypothesis generation than for hypothesis testing. An example of a hypothesis generated by an ecological study design is that the zika virus caused microcephaly: this idea came from the observation that the incidence of both microcephaly and zika virus infection increased at the same time in 2015 in Brazil. To prove the causal relation, other studies were necessary.

3.4.6 Investigation of an emerging infectious disease: "First-few-hundred" studies

The emergence of a new pathogen raises important questions regarding the key characteristics of the newly identified organism as part of severity assessment, such as its transmissibility to close contacts and the likelihood that someone who is infected will developed symptoms and suffer more severe consequences. First-few-hundred studies are longitudinal cohorts that involve the follow-up of the first cases and their close household and other contacts together with gathering clinical and biological data to ascertain whether there is serological or virological evidence of infection. These data can be used to derive parameters such as the secondary household attack rate, the reproductive number, the incubation period, and the proportion symptomatic [15].

Differential follow-up depending on key exposures or outcomes has the potential to introduce significant bias in this type of study. Furthermore, results of first-few-hundred studies are highly dependent on the characteristics of the selected patients, their region and, e.g., their economic status. Results based on studies of a relatively small number of patients from specific regions may not be representative of results on a global scale. An example of this occurred when a new pandemic influenza strain emerged in 2009 in Mexico. This influenza virus was first thought to cause relatively severe disease, based on an initially small sample of infected patients.

3.4.7 Outbreak investigation

A key characteristic of many infectious diseases is that they can cause outbreaks. Outbreaks can occur due to infections which are communicable (i.e., with a capacity to spread from person-to-person). Non-communicable pathogens, such as, e.g., Legionella, can also cause outbreaks, typically in the form of a common source outbreak where people are exposed to a common source that is contaminated. When this common source is limited to a certain geographical location and was present only for a short period of time, a point source outbreak can result. When the source of infection continues to be contaminated and people continue to be exposed, an ongoing common source outbreak can be the result. Depending on whether the pathogen is communicable, common source outbreaks can be propagated by subsequent person-to-person spread. Descriptive epidemiological data can be used to distinguish these types of outbreaks (see Chapter 23).

The words outbreak and epidemic are synonymous. The definition of an outbreak is "the occurrence in a community of region of cases of an illness, specific health-related behaviour, or other health-related events clearly in excess of normal expectancy" [3]. The implication of this definition is that in order to declare an outbreak, one needs to have information on what the normal frequency of the occurrence of the illness is.

From a public health perspective, the primary aim of investigating outbreaks is to document evidence for interventions to stop the outbreak, limit its impact, and prevent the occurrence of similar ones in the future. An additional aim can be to find evidence for early identification of cases to allow appropriate treatment (i.e., secondary prevention). Investigating outbreaks also can aid in detecting gaps in health programs aimed at preventing and controlling infectious diseases. In addition to these aims directly related to

outbreak control and secondary prevention, the investigation of outbreaks can help to increase knowledge about the pathogen, about, e.g., risk factors for transmission, host-related risk factors, the natural course of disease, and effectiveness of interventions. Outbreaks can be considered as natural experiments, providing an opportunity to document evidence by performing an outbreak investigation.

Outbreak investigation refers to the context of the epidemiological study, rather than to a distinct epidemiological design or method. Epidemiological outbreak investigations for disease control are usually done following a number of specific steps [16]. After the outbreak is confirmed and a representative sample or all cases has been identified, descriptive epidemiological analyses are a key step to generate hypotheses. Descriptive epidemiology in outbreak investigation and surveillance require identical analyses: a description of cases by time, place, person, and type. A description of cases by time, i.e. the "epidemic curve" can be particularly informative to hypothesize whether the outbreak is due to a point source. Hypotheses testing about the cause of the outbreak usually requires an analytical epidemiological study, e.g., a cohort, case-control, cross-sectional, or ecological design (see previous discussion).

References

[1] GM Weinstock. Genomic approaches to studying the human microbiota. *Nature*, 489(7415):250, 2012.

[2] G Murphy and JV Parry. Assays for the detection of recent infections with human immunodeficiency virus type 1. *Eurosurveillance*, 13(36):18966, 2008.

[3] M Porta. *A Dictionary of Epidemiology*. Oxford: Oxford University Press, 2008.

[4] S Declich and AO Carter. Public health surveillance: Historical origins, methods and evaluation. *Bulletin of the World Health Organization*, 72(2):285, 1994.

[5] AD Langmuir. The surveillance of communicable diseases of national importance. *New England Journal of Medicine*, 268(4):182–192, 1963.

[6] S Hahné, M Hooiveld, H Vennema, A Van Ginkel, H De Melker, J Wallinga, W Van Pelt, and P Bruijning-Verhagen. Exceptionally low rotavirus incidence in the Netherlands in 2013/14 in the absence of rotavirus vaccination. *Eurosurveillance*, 19(43):20945, 2014.

[7] K Mølbak, L Espenhain, J Nielsen, K Tersago, N Bossuyt, G Denissov, A Baburin, M Virtanen, A Fouillet, T Sideroglou, et al. Excess mortality among the elderly in European countries, December 2014 to February 2015. *Eurosurveillance*, 20, 2015.

[8] World Health Organization, Council for International Organizations of Medical Sciences et al. *International Ethical Guidelines for Health-Related Research Involving Humans*. Geneva, Switzerland: Council for International Organizations of Medical Sciences, 2016.

[9] L Soetens, D Klinkenberg, C Swaan, S Hahné, and J Wallinga. Real-time estimation of epidemiologic parameters from contact tracing data during an emerging infectious disease outbreak. *Epidemiology*, 29(2):230–236, 2018.

[10] MJ Knol, JP Vandenbroucke, P Scott, and M Egger. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *American Journal of Epidemiology*, 168(9):1073–1081, 2008.

[11] EW Orenstein, G De Serres, MJ Haber, DK Shay, CB Bridges, P Gargiullo, and WA Orenstein. Methodologic issues regarding the use of three observational study designs to assess influenza vaccine effectiveness. *International Journal of Epidemiology*, 36(3):623–631, 2007.

[12] H Wilking, M Thamm, K Stark, T Aebischer, and F Seeber. Prevalence, incidence estimations, and risk factors of toxoplasma gondii infection in Germany: A representative, cross-sectional, serological study. *Scientific Reports*, 6:22551, 2016.

[13] TB Hallett, J Stover, V Mishra, PD Ghys, S Gregson, and T Boerma. Estimates of HIV incidence from household-based prevalence surveys. *AIDS (London, England)*, 24(1):147, 2010.

[14] L Mollema, GP Smits, GA Berbers, FR Van Der Klis, RS Van Binnendijk, HE De Melker, and SJM Hahné. High risk of a large measles outbreak despite 30 years of measles vaccination in the Netherlands. *Epidemiology & Infection*, 142(5):1100–1108, 2014.

[15] E McLean, RG Pebody, C Campbell, M Chamberland, C Hawkins, JS Nguyen-Van-Tam, I Oliver, GE Smith, C Ihekweazu, S Bracebridge, et al. Pandemic (h1n1) 2009 influenza in the UK: Clinical and epidemiological findings from the first few hundred (ff100) cases. *Epidemiology & Infection*, 138(11):1531–1541, 2010.

[16] Outbreak investigations. https://wiki.ecdc.europa.eu/fem/w/wiki/387.outbreak-investigations. Accessed: 16 March 2018.