Patterns and scales in infectious disease surveillance data: an exploratory data analysis approach

Soetens, L.C.

1

General introduction

Motivation

This thesis was written in a data-rich era. The amount of data in the world is increasing exponentially. Due to great advances in technology, data are collected and stored everywhere; ranging from data on internet use to visits to a general practitioner. This increasing amount and complexity of data has impact on all fields of science, including the infectious disease epidemiology domain. Due to automated surveillance programs it has become easier to collect large amounts of data on a patient's whereabouts, general characteristics, disease status and risk behaviour. Advances in sequencing technology have resulted in an enormous increase in resolution of data on genetic information of the pathogen the patient is infected with. Advances in data management and data storage capacity allow connecting all this information at an individual patient level. Taken together, these developments have led to an increase in resolution of infectious disease data. Not only do we know more (more variables and more observations for each variable), but also at a more detailed level (for example, the sequence of the pathogen instead of the genotype). All these developments give rise to new opportunities for disease control, but with that also create new questions. Such as, how to detect and investigate outbreaks if not only information on the time or place of the patients is known, but also information on the genetic structure of pathogens? How to combine all these data types? Do they contribute to statistical evidence with equal weights or is there a leading data type? Also, concerning data resolution, at which scales should we look for patterns in for example surveillance data? Is this always at the highest resolution level, or are we then loosing ourselves in details while overlooking the bigger picture?

Research in infectious disease epidemiology is mainly focused on confirmatory data analysis (CDA) and on method development for this type of analysis. However, the above-mentioned questions are not easily answered with such a traditional approach, since they are not concerning hypothesis testing. They concern issues regarding structure and patterns of data, issues that are typically addressed in exploratory data analysis (EDA). However, as most research has been focused on CDA, method development naturally focused on this field as well, leaving a gap for EDA method development in the domain of infectious disease epidemiology. Currently, EDA steps in infectious disease epidemiology are often overlooked. For example, when geographical information of a patient is available it is often displayed on a map, and, little attention is paid to the choices how to display this information on a map, while these choices determine the ability to discover a pattern in the data and are therefore of uttermost importance. Presently, there are hardly any epidemiological methods available providing an evidence base for making such choices.

That is not to say that there has been no method development for EDA at all. With the advances in data development, a new research field has emerged: data science. Although

no formal definition exists of data science, generally, it involves principles, processes, and techniques for understanding phenomena via the extraction of knowledge from data [1]. Data science requires skills of various domains including computer science, statistics, mathematics, machine learning, domain expertise, data visualization, and communication and presentation skills [2]. EDA has a central place in data science, as to extract knowledge from data in data science, data has to be studied intensively, until its contents, structure, patterns and pitfalls are fully understood. The latter is typically achieved through EDA.

In this thesis, we focus on method development for EDA in infectious disease epidemiology, specifically infectious disease surveillance. We explicitly address issues concerning recent data developments in this domain related to increasing resolution in routinely collected infectious disease data for surveillance purposes. We will do so by using techniques and concepts from the data science field. This thesis is therefore targeted at infectious disease epidemiologists, who would like to broaden their view beyond traditional CDA approaches.

Organisation of this introductory chapter

In this introductory chapter we start with an extensive description of the origin and structure of routinely collected surveillance data on infectious diseases in the Netherlands. We follow by an introducing overview of EDA, its contrast with CDA, how EDA is currently applied in infectious disease epidemiology and the questions that arise from this application. Next, a brief overview of the data science process is given, highlighting central ideas that might be of use for solving previously mentioned questions. This then results in a description of the aim and objectives of this thesis.

Routinely collected data on infectious diseases in the Netherlands

Data collection

Data on infectious diseases are gathered for three main public health purposes: surveillance, outbreak investigation and scientific research. In this thesis, we consider only routinely collected data for surveillance and outbreak investigation purposes. Public health surveillance is defined by the World Health Organization (WHO) as "an ongoing, systematic collection, analysis and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice" [3]. In the Netherlands, infectious disease surveillance data is collected in various settings and through various systems [4]. To describe them all would be beyond the scope of this thesis. However, since most of the studies in this thesis use data from the national notifiable disease surveillance system, we provide a detailed description of this system. For certain infectious diseases, it is mandatory for physicians or laboratories to notify laboratory confirmed cases to the municipal health services (MHS), who then report to the national institute for public health and the environment (RIVM). A schematic overview of the information flow of the notification process is depicted in Figure 1.1.

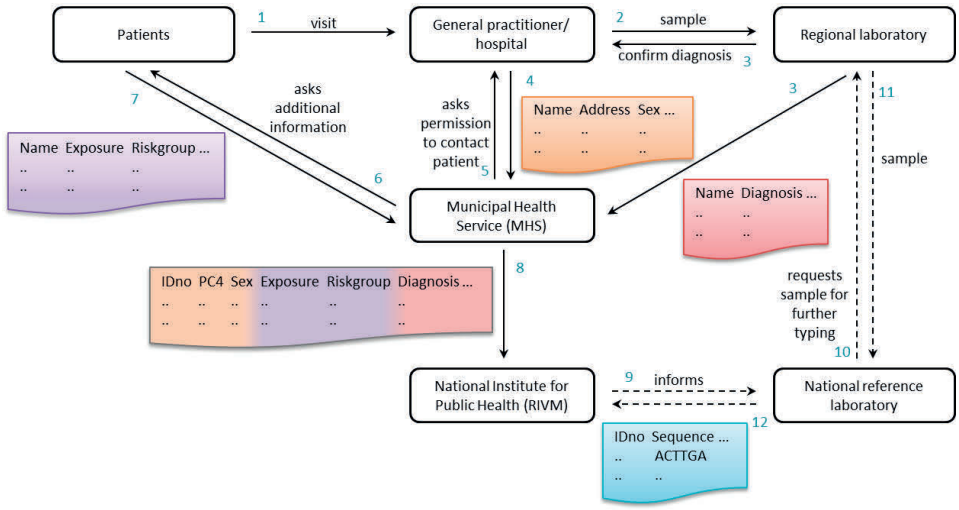

Figure 1.1. Information flow of notification process of notifiable infectious diseases in the Netherlands

A sick person visits a general practitioner or hospital, where the physician suspects a certain notifiable disease. The physician sends a sample (for example blood or urine) for confirming his or her initial suspicion to the regional laboratory. The laboratory reports

back to the physician and the MHS with or without confirmation of the diagnosis. When confirmed, general data (name, address, sex etc.) on the patient is send to the municipal health service (MHS). For notifiable diseases, this is mandatory in the Netherlands. The MHS asks permission from the physician to contact the patient for further information. The patient is then asked questions by the MHS on possible sources for exposure, risk groups, profession or any other relevant information. All this data gathered by the MHS is then anonymized and send to the national institute for public health (RIVM). The RIVM monitors the notifications to identify outbreaks and trends. For certain diseases it is also possible to get further information on the pathogen: national reference laboratories are informed by the RIVM on the notification and subsequently request the sample for further typing from the regional laboratory. More detailed information on the pathogen, such as a DNA/RNA sequence, is then reported back to the RIVM. In an outbreak situation, the information flow is roughly similar and often involves a more extensive inquiry of the patient by the MHS.

Data types

Generally, this routinely collected data on infectious diseases can be described by four data types: time, place, person and pathogen (Figure 1.2).

1. The person data type describes data characteristics of the person infected by the disease, such as its age and sex, but also risk factors of contracting this disease, like a person's profession or sexual orientation.
2. The time data type is describing when certain events happened, such as a date of onset of symptoms or a date of diagnosis. It can also contain information on when an infected person has met a certain contact during a contact tracing investigation.
3. The place data type is concerned with the geographical location of events related to the infected person. One can think of a persons living address, the geographical location of exposure to a source, etcetera.
4. The pathogen data type describes (genetic) characteristics of the pathogen a person is infected with. For instance, a sequenced sample of the DNA of the virus or bacterium the person is infected with.

So far, we have seen four data types for every patient. However, if groups of cases are considered, for example in an outbreak setting, there is also the data type of connection between the cases, describing how cases are linked. The characteristics of a link itself can be described by time, place, and person characteristics. For example, an epidemiological link between two cases often consists of a certain time, place and risk setting where they have met and possibly have transmitted the disease.

Data resolution

As shown in Figure 1.2 these types are all connected, describing a dynamic process. Over the past decades, each data type has increased in size and complexity. Data on all types has become available on more and higher resolution scales. Not only a patient's zip code of

his or her living address is known, but as mobile phone with GPS functions are widely used nowadays, data on various geographic locations and trajectories of a patient may be more easily and precisely gathered. In addition, advances in technology have made it possible to extract and sequence DNA or RNA of viruses and bacteria on a higher resolution scale, resulting in a tremendous increase in the available data on the pathogen-type data type. As for the person data type, a higher resolution can be gained by gathering more detailed information on the person's characteristics, and for time, by moving from year of symptom onset to day or hour of symptom onset.

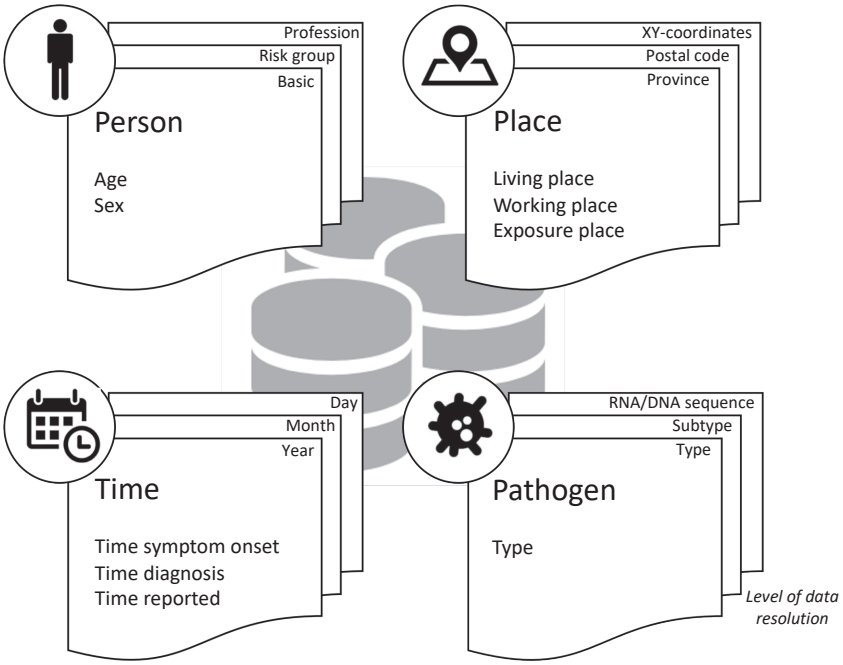

Fig. 1.2. Data types in infectious disease surveillance

Combining data

Another matter of complexity is that data on the different types of a certain infectious disease are usually not stored in the same database or system. For example, detailed sequence information as provided by the national reference laboratories is stored separately from the notifiable disease database. To study an infectious disease comprehensively requires considerable data management and processing efforts.

Exploratory data analysis

History

Exploratory data analysis (EDA) was first described by John Tukey in 1977 [5]. He describes EDA as numerical and graphical detective work, for which tools and understanding are needed. He compared EDA to the search of the evidence by a detective and confirmatory data analysis to the evaluation of the evidence's strength by a judge or jury: "Unless the detective finds the clues, judge or jury has nothing to consider. Likewise, unless EDA uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider." As such, he stated that "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step". In his book he describes tools and techniques for the detective to explore numbers, among which are the still popular stem-and-leaf plot and the box-and-whisker plot. Although he was the first to give this type of analysis a name, he was not the first to actually perform EDA. If we look back in the history of epidemiology, we can see several earlier examples of EDA. In 1854 the British nurse Florence Nightingale was send to Russia to aid in the nursing of the soldiers during the Crimean war. Thereupon arrival, she discovered that it was not due to war injuries that most soldiers died, but due to lack of hygiene. The concept of hygiene was not yet fully understood in that time, so in order to convince the British government, she invented and produced a polar diagram to present the relevant statistics (Figure 1.3) [6]. In this figure it can clearly be seen that only a small fraction of all deaths (red area) was related to war injuries, and that the great majority of deaths (blue area) was related to preventable infectious diseases.

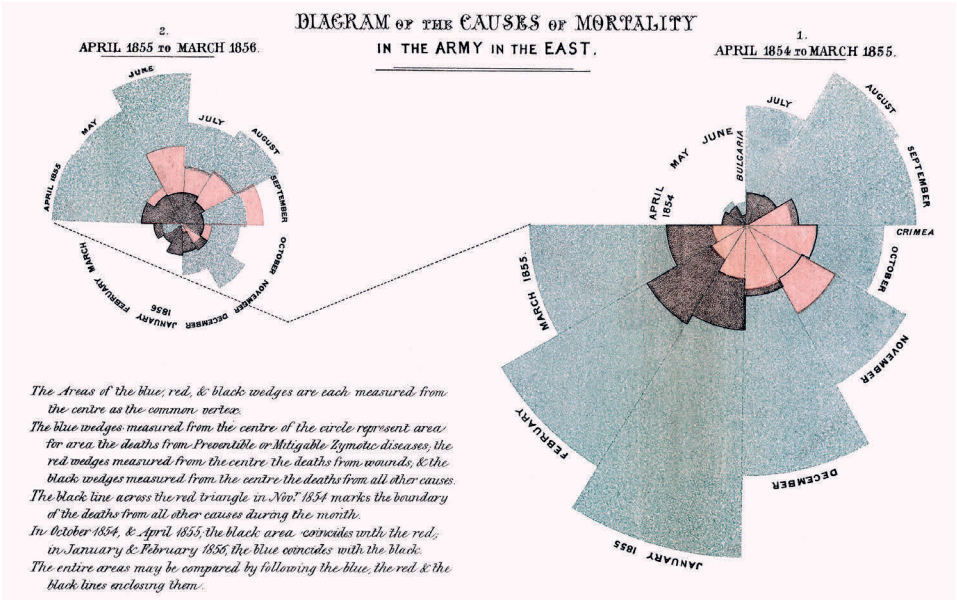

Figure 1.3. Polar diagram of Florence Nightingale (adapted from [6])

Another example of early EDA in epidemiology is the well known story of John Snow and the cholera epidemic in London in 1854 [7]. By plotting the cholera cases on a map of London (Figure 1.4), he discovered that most cases (black dots) were concentrated around a public water pump at Broad street, which contributed to the discovery that cholera was transmitted by infected water. Like the polar diagram of Florence Nightingale, the map of John Snow is also a good example of EDA: both have acted as evidence in unsolved mysteries at that time.

Since the emergence of Tukey's book, EDA has slowly developed to a field on its own, and profited from collaborations with other fields such as computer science. One of the more recent major contributions is Wilkinsons' 'Grammar of Graphics' [8], which provides theory for statistical graphics, showing the conversion from data in their original state to their graphical representation. Many modern software programs for visualizing data, such as the ggplot package by the R statistical software [9], are based on this framework.

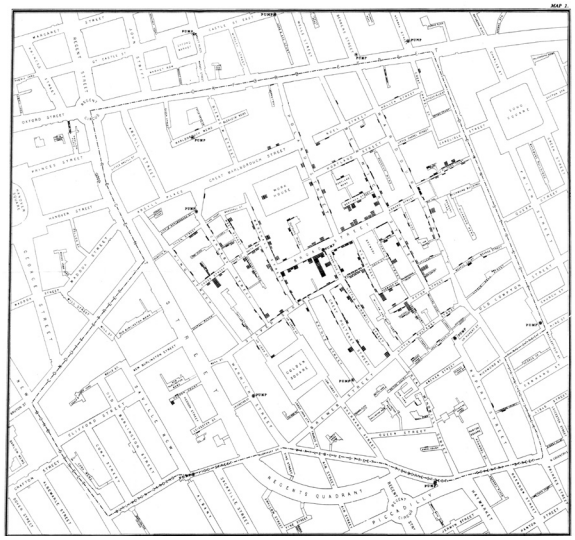

Figure 1.4. John Snow's map of London with cholera cases during the epidemic in 1854 (adapted from [7])

What is EDA?

Although John Tukey has written an entire book about the concepts of EDA, he provides no formal definition. This was also noted by Jebb et al [10], who in their research paper put together several perspectives and came to the following: "EDA is best described as an overarching analytic attitude characterized as detective work designed to reveal the structure or patterns in the data" (Haig 2005, Tukey 1980). The goal of EDA is to understand the structure of the data and discover their patterns [11]. There are several general analytical goals related to this overall goal:

- checking whether statistical assumptions are met;

- identifying outliers in the data;
- formulating new research hypotheses;
- uncover empirical relationships;
- formulating models for the data.

The techniques for applying EDA can look highly different depending on the research field it is applied to. However, one important technique in applying EDA is similar across fields and that is the visualization of data. The human vision system is extremely good in detecting patterns [12], which makes data visualization a highly suitable and therefore central technique in EDA.

EDA and CDA

To get a better grip on the characteristics of EDA it might best be contrasted with the more familiar confirmatory data analysis (CDA). In Table 1.1 the most important differences are outlined [13]. An important difference with respect to this thesis is the type of data they use. EDA is applied to observational data, which is collected without a well-defined hypothesis. The surveillance data in the Netherlands we mentioned in a previous paragraph are such a data collection. For CDA data has to be collected through formally designed experiments or carefully designed observational studies, and is therefore more applicable to research settings. EDA and CDA are complementary and follow each other. Often EDA is applied to an observational data set and a hypothesis is generated. With this hypothesis in mind, an experiment is designed and new data is collected to test this hypothesis through CDA.

Table 1.1. Differences between EDA and CDA (adapted from [13])

	Exploratory data analysis (EDA)	Confirmatory data analysis (CDA)
Reasoning type	Inductive	Deductive
Goal	Pattern recognition and hypothesis generation	Estimation, modeling, hypothesis testing
Applied data	Observational data (data collected without well-defined hypothesis)	Experimental data (data collected through formally designed experiments)
Techniques	Descriptive statistics, data visualization, cluster analysis	Traditional statistical techniques of inference, significance and confidence
Advantages	• No assumptions required • Promotes deeper understanding of the data	• Precise • Well-established theory and methods
Disadvantages	• No conclusive answers • Difficult to avoid bias produced by overfitting	• Required unrealistic assumptions • Difficult to notice unexpected results

EDA and infectious disease epidemiology

The techniques and methods used for EDA are highly dependent on the field it is applied to. In this section an overview is given of how EDA is currently applied in infectious disease epidemiology. We already saw that EDA is about revealing the structure of patterns in the data. In addition, we also already know infectious disease data can be structured along four data types, time, place, person and pathogen, and these data types therefore also determine how EDA is applied in this field.

Time

In infectious disease epidemiology, there are numerous ways of revealing patterns in time data. In an outbreak, one of the most fundamental graphs in infectious disease epidemiology is the epidemic curve (or epicurve) [14]. It is a graph showing the distribution of symptom onset dates of cases in an outbreak. On the x-axis time is typically displayed as date of symptom onset of the cases and on the y-axis the number of cases is usually depicted. It is basically a histogram of cases with a time dimension on the x-axis. In EDA we are looking for patterns in the data, and the pattern or shape of the epidemic curve will tell us much about the nature of the outbreak and its mode of transmission. Generally two modes of transmission can be discerned: transmission from a source in the environment, which can be either one-time only, intermittent or continuous, and person-to-person transmission. When the outbreak is concerning a one-time only point source, the epidemic curve is characterized by a very sharp peak in cases, which gradually declines (Figure 1.5a). When an intermittent environmental source is spreading the disease, the epidemic curve has a more flat pattern with a more continuous number of cases (Figure 1.5b). If it reflects person-to-person transmission the epidemic curve it characterized by waves of cases which are typically a generation period apart (the time period between symptom onset of a secondary case and symptom onset of its primary case) (Figure 1.5c). One important aspect considering epidemic curves is the binwidth, or resolution scale, that is chosen on the x-axis: this can be hours, days, weeks or even years. It is important to choose the binwidth as such that it will show you the patterns in the data. Although there is no official rule for determining the binwidth, it has been found that the ideal reporting interval is the mean generation time [15], the time between symptom onset of two successive cases. The ratio of cases in successive intervals then yields the effective reproduction number, a key variable that characterizes transmissibility with time [16]. It is defined as the average number of secondary cases per primary case. If it is above 1, the outbreak is increasing in size, and below 1 the outbreak is decreasing in size.

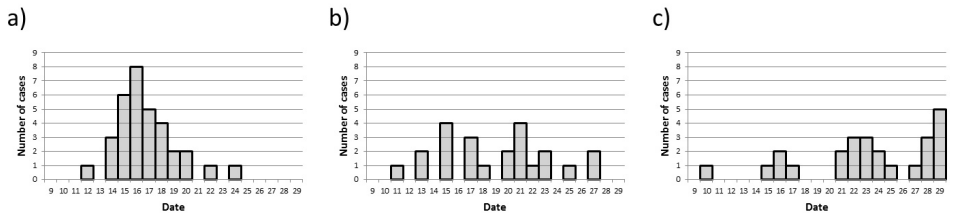

Figure 1.5. Typical epidemic curves of a point-source outbreak (a), continuous-source outbreak (b) and a person-to-person outbreak (c). (Adapted from Giesecke, 2016 [14])

Infectious disease cases are not only depicted over time during an outbreak, but also in a surveillance setting. The goal is then to monitor any aberrations (or unusual patterns) in the data. The current number of cases is then often depicted against a historical average of the

past X years. The amount of algorithms to detect aberrations in time is numerous [17-21]. Other analysis of time patterns (or time series analysis) in infectious disease epidemiology can consist of decomposing time series into trend, seasonal and remainder components, to look for unusual patterns in these subseries. Again, there are many time series analysis techniques, to discuss them all would go beyond the scope of this thesis.

Place

One could consider a table of postal codes, but as demonstrated by John Snow's map with cholera cases, a map is a much more powerful tool to study the distribution of phenomena in geographical space. The study of displaying phenomena on a map is called cartography and is a field of science on its own. One of the major contributors to cartography is Jacques Bertin (1918 – 2010), who with his book 'Semiologie Graphique' [22] provided the theoretical foundation for what is later called information visualization. It would go beyond the scope of this thesis to discuss the foundations of cartography in detail. Instead, we will discuss the most relevant and used techniques in infectious disease epidemiology. There are three types of maps that are often used, all with different purposes. A useful overview of functions of maps and other spatial methods in different stages of an outbreak investigation is given by Smith et al [23].

Figure 1.6. Examples of common map types: a) dot map, b) proportional symbol map, c) choropleth map (adapted from [24])

The first map type to be discussed is the dot map (or point map, location map) (Figure 1.6a). John Snow's map (Figure 1.4) is a typical example of such a map. In this map type each phenomenon, in the field of infectious disease epidemiology often a case with a certain infectious disease, is displayed as a single dot at an exact location on the map, often a home address. To establish the existence of an outbreak the case distribution can be visualized using a dot map. A dot map at various stages during the outbreak can also be used to describe the progression of an outbreak. The place and time data dimensions are then combined. Dot maps can also have a function in surveillance activities, mainly to signal potential outbreaks. When the disease under surveillance is quite rare, dot maps can be used to monitor any unusual occurrence or geographic clustering of disease cases.

The other two map types show quantities for a certain area (or polygon) on a map. Examples of such areas are municipalities, postal codes, provinces or municipal health regions. These quantities can either be absolute or relative. When absolute quantities are displayed a proportional symbol map is used (Figure 1.6b). On such a map, a symbol (often a circle) proportional to the quantity it is referring to is displayed in the center of the related area. Such a map could for example display the number of infectious disease cases per municipal health region. When referring to relative quantities, we use a choropleth map (or rate map, thematic map) (Figure 1.6c). These are probably the most used maps in infectious disease epidemiology. It is showing a relative quantity (for example the number of infectious disease cases per 100,000 population) of a certain area. This is typically done by shading the areas on the map along a color scale conform its relative quantity. Choropleth maps can be used to visualize the distribution of cases in relation to known or potential risk factors to aid in source finding. Or for surveillance purposes, when the disease under surveillance is very common, disease occurrence is affected by population density and choropleth maps indicating the number of cases by population density of a certain area are then most informative in signaling any unusual events.

It is not straightforward to objectively display phenomena on maps, especially when it concerns choropleth maps. The (arbitrary) choices of the boundaries of the areas, the resolution (the size of the areas), the classes for the color scale and the colors themselves can heavily influence the appearance of the map and therefore the message the map is giving. It is therefore quite easy to manipulate the reader with this map type.

In addition to visualizing cases on a map, there are quite a number of possibilities for spatial analysis of data, which can be used in infectious disease epidemiology. Geographic information systems (GIS) have increased the availability and range of tools that can be used to analyse outbreaks and surveillance data. A GIS is a database designed to handle geographically-referenced information complemented by software tools for the input, management, analysis and display of data [25]. Smith et al [23] give an overview of spatial analysis possibilities in their review on spatial analysis methods in outbreak settings. One such analysis is the analysis of the distance of cases to possible sources of infection to locate the common source of infection. An example of such an application in the Netherlands is the Q fever outbreak in 2009 [26], where analysing the distance of cases to possible sources identified goat farms as the common source. Another type of spatial analysis important for finding patterns in infectious disease data is cluster identification. There are numerous clustering algorithms for detecting clusters in space or in space-time [27-31], of which the Kulldorff's spatial/ spatiotemporal scan statistic [29] is most used in outbreak studies [23]. This statistic scans for deviations from a random (Poisson) spatial distribution of cases.

Person

EDA in the persons dimension is mainly used to gain insight into the characteristics of the infected persons. Its main purpose is to get an understanding of 'who is infected', whether or not in contrast to 'who is not infected' or relative to the average characteristics in the population. Here again visual graphs can be of great help, such as a histogram showing the age or sex distribution of cases. Likewise, the distributions of certain risk factors, such as the professions or sexual orientations of the infected persons can be visualized by various graphs, such as histograms, bar plots and others. This data type is also quite easily combined with the former two, to answer questions as 'Has the age distribution of infected persons changed over time?' or 'Is there a different risk profile of infected persons in different areas of the Netherlands?' To examine such questions, graphs can be combined with time series or maps; it will depend on the question what is the best way to do this.

Pathogen

This data type describes the (genetic) characteristics and molecular sequence of the particular pathogens the persons are infected with. At a population level, studying the structure of the genetic characteristics of these pathogens will provide us with information on how the disease has spread through the population, and might give us information on who has infected whom. Of all data dimensions, this dimension has known the greatest advances over the past decades, due to the advances in sequencing technology, immensely increasing resolution on this data type. This information contains at a lower resolution level merely the name or subtype of the bacterium or virus the patient is infected, but can range to complete patterns (sequence) of the genome, the carrier of genetic information, of that bacterium or virus at the highest resolution level. In most organisms, the genome consists of deoxyribonucleic acid (DNA), while for some viruses this is ribonucleic acid (RNA).

Starting at the lowest resolution level when only the name of the (sub)type of the pathogen is known, EDA in this data domain is much like EDA in the person domain, and can consist of describing the distribution of the pathogen (sub)types using a bar plot or some other graph. This information will tell us which (sub)types are circulating in the population. When combining this information with information from the time or place dimensions, it will provide us with information on how the circulating (sub)types change over time or if they are geographically clustered [32-34].

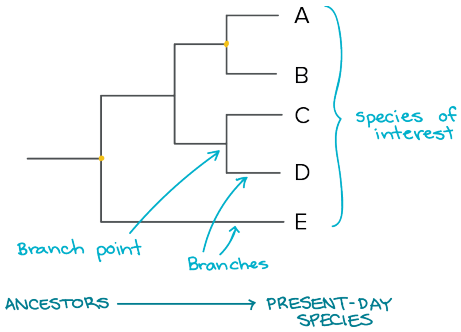

Figure 1.7. Example of phylogenetic tree

At a higher resolution level, a part of the genome of the pathogen is known by using sequencing technology. At the highest resolution level, the complete genome of the pathogen is obtained through sequencing technology, also called whole genome sequencing (WGS). In both situations, the data we are working with is DNA or RNA sequences, the genetic code of the pathogen. This consists of a chain of smaller molecules called nucleotides. DNA consists of adenine (A), guanine (G), cytosine (C) and thymine (T) nucleotides, RNA consists of A, G, C and uracil (U) nucleotides. A nucleotide sequence is thus represented by a contiguous stretch of the four letters A, G, C and T/U. DNA or RNA sequences of pathogens cannot be compared anymore by simply listing them, as they are sometimes thousands of nucleotides long. It requires visualization techniques to discover patterns or anomalies in such large datasets. This is typically done through phylogenetic analysis [35]. Phylogenetic analysis establishes the relationships between genes or gene fragments, by inferring the common history of the genes or gene fragments. This relationship is schematically represented by a tree, illustrating the evolutionary history (phylogenies) of a gene or pathogen (Figure 1.7). The relationship is often quantified in terms of similarity between sequences of the pathogens. In general, the more similar the sequences, the more closely related the pathogens are, and the closer they are in the phylogenetic tree. There are many methods of inferring a phylogenetic tree from sequence data. We can discriminate methods that use the discrete nucleotide characters in a sequence (typically tree evaluation methods such as Maximum Parsimony, Maximum Likelihood or Bayesian methods), and methods that use a distance matrix of the pairwise dissimilarities between sequences (typically clustering methods such as UPGMA, Neighbour-joining). More details of these methods can be found in Lemey et al [35]. To consider these methods in detail will be beyond the scope of this introduction.

When we study the phylogenetic tree of a group of patients, we can learn how the disease has spread through the population, as patients who are closer together in the phylogenetic tree are also likely to be closer together in the transmission chain (clusters). In a closed setting, when we are confident we have all infected patients, we might even infer who has infected whom. In addition, we might learn from studying the phylogenetic tree if there is a

situation with ongoing endemic transmission of the pathogen (when most sequences are roughly similar) or whether there have been introductions of the pathogen from for example abroad (when there is much variation in circulating sequence types). An example of the latter is the introduction of meningococcal W disease in the Netherlands, most probably originating from the United Kingdom [36].

Phylogeny can be combined with the other data dimensions. When we now the mutation rate of the pathogen, we can infer calendar time from the evolution time, which is often depicted on the x-axis of a phylogenetic tree (Figure 1.7). This type of analysis is called phylodynamics. Also the place dimension can be combined with phylogeny, and has its own research field called phylogeography. This can be used to study how the evolution of pathogens relates to certain areas in the world. A good example is how the different seasonal influenza viruses have spread over the world [37] or how Zika virus has spread in the Americas [38]. Finally, phylogeny can be combined with the person dimension, for example, to study the genetic clustering of risk groups, such as has been done for hepatitis B by Hahné et al [39].

Increasing resolution and EDA in infectious disease epidemiology

As we have seen in the sections above, within each data dimension data resolution has increased over the years. However, there is not much attention paid to the fact that data is available at multiple resolution levels. There is scarcely any research on at what resolution level to present an epidemic curve as to best depicting the underlying data pattern. How to choose the bin width on the x-axis? And, is a histogram the best way of presenting this data? Similarly for geographical data. Surveillance data are often depicted on an incidence map, aggregating data at a regional level. But how to choose this region? Is there another way that is more objective and does more justice to the data and its pattern itself? Likewise for pathogen data types. When this data is available on a sequence level, it is almost automatically analyzed at the highest resolution scale. The risk here is to get lost in details whilst overlooking the bigger picture. We might miss trends in genotype occurrence, when focusing on sequences? Should we do both? Also, when combining data dimensions, how do we handle multiple resolution levels? At what level of resolution do we link data types? With this last question we touch on another issue. Not only has resolution increased within a data types, but also across data types. We do not only have data available on time and place, but also at a very detailed level at the pathogen dimension. With the increasing availability of multiple data dimensions, we can now combine them and study patterns in the combined data. But how to do this? We would have to compare distances in days (time), with kilometres (place) and mutations (pathogen). And if we succeed in this, would all data types then contribute to statistical evidence with equal weight? And if not, how do we weigh them? Wouldn't there be another solution?

In this thesis, we seek to answer these type of questions using a data science approach.

Data science

Data science involves principles, processes, and techniques for understanding phenomena via the extraction of knowledge from data [1]. Generally, a data science process involves the following steps. Raw data is collected for from the 'real world' and is then processed so that it is available in a database, file or other accessible format. Before anything then can be done with the data, they have to be cleaned. Cleaning steps typically consist of handling missing values, dealing with outliers and removing duplicates.

So far, the steps are very similar to the start of the data analysis in a traditional research process. However, the next steps are defined by the nature of the data science problem. The first step in solving the data science problem would be to perform exploratory data analysis to get a grip on the structure and patterns of the data. Besides traditional techniques for EDA as discussed in earlier paragraphs, a few special techniques are often applied in the data science process. First, as data science often deals with large highly dimensional datasets, dimension reduction techniques, such as principal component analysis (PCA) or multidimensional scaling (MDS) are often used to reduce the dataset to a more manageable form. Another frequently used EDA technique in data science is cluster analysis, such as K-means, to find patterns in the (reduced) datasets. An advantage of such clustering approaches is that they are independent on the unit of analysis.

After applying EDA it can be decided that more data needs to be collected, that more cleaning steps are required, that further analysis is required by performing machine learning, or that EDA has provided enough insight into the problem and that the results can be communicated through visualizations, reports or other data products. In Figure 1.8 shows an example of a data science pipeline as proposed by O'Neil and Schutt [2].

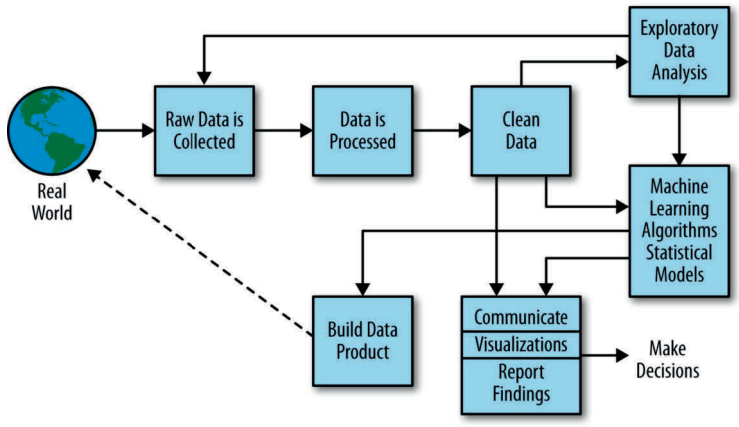

Fig 1.8. The data science process (adapted from O'Neil and Schutt, 2013[2]).

As described, the final stage of the data science process is concerned with how to communicate your findings to the intended audience. In the infectious disease domain, the audience will most likely be either the general public or public health professionals. Both type of audiences are most likely not familiar with analysis methods, so it will not do to just present the outcomes of the methods described in the previous paragraph. We have to package the results, and present them in a meaningful and understandable way. One such means is through visualization of the results. Data visualization is important in several earlier stages of the data science process: as part of exploratory data analysis and as part of evaluation of model performance. We now add another stage to this list: as part of communicating results. There is however a fundamental difference in the goal of visualization between the earlier stages, which is exploratory from nature, and in this stage, which is explanatory from nature. When the goal is to explain, there are several best practices to make effective visualizations. This is a popular field and many books have been written on this theme [40-46]. It will be beyond the scope of this thesis to discuss it further here.

What can we learn from a data science approach?

Several central data science concepts might be of use to solve questions pertaining to the increasing data resolution for infectious disease epidemiology. First, we can observe that many central techniques to data science are based on scale independent methods, for example PCA and K-means clustering. By using such scale independent methods, multiple data dimensions can be combined, and the intrinsic data structure is revealed. Such an approach might as well work for combining our data dimensions, and even for finding our optimal data resolution. There is however one drawback of such methods, and that is that the interpretation of the results is often hard to grasp.

That brings us to a second concept central to data science which is the use of visualization techniques. The use of data visualization is already important in EDA, but inevitable when it comes to visualizing results of a data science process. The use of visualization techniques in combination with scale independent methods might help us further on our quest.

One final concept central to data science is evaluation. At each step, outcomes and performances need to be evaluated in order to determine the following step in the process. Evaluation therefore makes sure that the data science process is iterative and non-linear by design. We have seen that depicting infectious disease data dimensions at one resolution scale might not suffice. Using a more flexible and non-linear approach might help us revealing the intrinsic patterns.

Aim and objectives of this thesis

In this thesis, we develop methods for EDA in infectious disease surveillance. Modern data development has posed interesting questions concerning patterns and structure of data, which so far have not gained the necessary attention. We develop methods to address issues with increasing resolution within and across data dimensions in infectious disease epidemiology. We use central concepts to the data science domain, such as scale independent methods, data visualization and process thinking, to bring EDA in infectious disease epidemiology to the next level.

The objective of Chapter 2 of this thesis is to use multiple data sources to unravel the mystery of rising hepatitis B incidence in certain parts of the Netherlands. This chapter is very much about collecting evidence to generate hypotheses, a core element of exploratory data analysis. In addition, we get ourselves acquainted with the all four data dimensions important in infectious disease epidemiology.

Chapter 3 and 4 of this thesis focus on a single data dimension, the place and time dimensions respectively, and address the increasing resolution within a data dimension. In Chapter 3 we address the issue of objectively presenting infectious disease incidence data on a map. Earlier in this introductory chapter, we already saw that quite some arbitrary choices regarding scale, classification and colour use heavily influence the appearance of your map. In Chapter 3, the objective is to develop a method for more objective and scale independent display of infectious disease data on a map, applied to incidence data of Q fever in the Netherlands and pertussis in Germany. Similarly, in Chapter 4, the objective is to develop a scale independent method for displaying infectious data in the time dimension. We show that by applying this method on a very large database of infectious disease notifications through ordering and clustering diseases, patterns in infectious disease dynamics are revealed and might lead to new hypotheses.

In Chapter 5, we use historical information on incubation periods to classify using a Bayesian model whether a person exposed to a case is still at risk of developing symptoms. This is important in the context of contact tracing of serious infectious diseases. We developed a data visualization based on this risk classification to aid public health professionals in decision making regarding follow-up of potential cases. Data dimensions addressed in this chapter are time and person.

In Chapter 6, we propose visualization tools to assess model performance of an earlier developed clustering algorithm using time, place and pathogen data dimensions. The model is an example of how to deal with increasing resolution across data dimensions. However, as it concerns an unsupervised clustering algorithm, no real truth exists to which

we can relate model performance. We developed an interactive tool in which clustering parameters can be varied in order to help public health professionals with the interpretation and to pick the optimal model with most plausible infectious disease clusters.

In the final chapter of this thesis (Chapter 7, general discussion), we reflect on whether a data science approach was useful in developing EDA methods for modern infectious disease surveillance.

References

1. Provost, F. and T. Fawcett, Data Science for Business. 2013: O'Reilly.
2. O'Neil, C. and R. Schutt, Doing Data Science. 2013: O'Reilly.
3. World Health Organisation. Public Health Surveillance. [cited 2019 26 April]; Available from: https://www.who.int/immunization/monitoring_surveillance/burden/vpd/en/.
4. Bijkerk, P., et al., State of Infectious Diseases in the Netherlands, 2015, Centre for Infectious Disease control RIVM, Editor. 2016: Bilthoven.
5. Tukey, J.W., Exploratory Data Analysis. 1977: Addison-Wesley Publishing Company.
6. Nightingale, F., Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army ... by Florence Nightingale. 1858: Harrison and Sons.
7. Snow, J., On the Mode of Communication of Cholera. 1855: John Churchill.
8. Wilkinson, L., et al., The Grammar of Graphics. 2005: Springer New York.
9. Wickham, H., ggplot2: Elegant Graphics for Data Analysis. 2016: Springer International Publishing.
10. Jebb, A.T., S. Parrigon, and S.E. Woo, Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 2017. 27(2): p. 265-276.
11. Behrens, J.T., Principles and procedures of exploratory data analysis. Psychological Methods, 1997. 2(2): p. 131-160.
12. Ware, C., Information Visualization: Perception for Design. 2013: Elsevier Science.
13. Good, I.J., The Philosophy of Exploratory Data Analysis. 1983. 50(2): p. 283-295.
14. Giesecke, J., Modern Infectious Disease Epidemiology, Third Edition. 2016: Taylor & Francis.
15. Nishiura, H., et al., The ideal reporting interval for an epidemic to objectively interpret the epidemiological time course. Journal of the Royal Society, Interface, 2010. 7(43): p. 297-307.
16. Cori, A., et al., A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. American Journal of Epidemiology, 2013. 178(9): p. 1505-1512.
17. Farrington, C.P., et al., A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. 1996. 159(3): p. 547-563.
18. Stroup, D.F., et al., Detection of aberrations in the occurrence of notifiable diseases surveillance data. Stat Med, 1989. 8(3): p. 323-9; discussion 331-2.
19. Hutwagner, L., et al., Comparing aberration detection methods with simulated data. Emerg Infect Dis, 2005. 11(2): p. 314-6.
20. Noufaily, A., et al., Comparison of statistical algorithms for daily syndromic surveillance aberration detection. Bioinformatics, 2019. 35(17): p. 3110-3118.
21. Noufaily, A., et al., An improved algorithm for outbreak detection in multiple surveillance systems. Stat Med, 2013. 32(7): p. 1206-22.
22. Bertin, J., Semiology of Graphics: Diagrams, Networks, Maps. 2011: ESRI Press.
23. Smith, C.M., et al., Spatial methods for infectious disease outbreak investigations: systematic literature review. Euro Surveill, 2015. 20(39).
24. Ribecca;, S. The Data Visualisation Catalogue. 2019 [cited 2019 19 November]; Available from: https://datavizcatalogue.com/index.html.
25. Pfeiffer, D.U., et al., Spatial Analysis in Epidemiology. 2008: OUP Oxford.
26. Hackert, V.H., et al., Q fever: single-point source outbreak with high attack rates and massive numbers of undetected infections across an entire region. Clin Infect Dis, 2012. 55(12): p. 1591-9.
27. Cuzick, J. and R. Edwards, Spatial Clustering for Inhomogeneous Populations. Journal of the Royal Statistical Society. Series B (Methodological), 1990. 52(1): p. 73-104.
28. Kulldorff, M., A spatial scan statistic. Communications in Statistics - Theory and Methods, 1997. 26(6): p. 1481-1496.
29. Kulldorff, M., et al., A space-time permutation scan statistic for disease outbreak detection. PLoS Med, 2005. 2(3): p. e59.
30. Kulldorff, M. and U. Hjalmars, The Knox Method and Other Tests for Space-Time Interaction. Biometrics, 1999. 55(2): p. 544-552.
31. JACQUEZ, G.M., A k NEAREST NEIGHBOUR TEST FOR SPACE–TIME INTERACTION. 1996. 15(18): p. 1935-1949.
32. Weinberger, D.M., R. Malley, and M. Lipsitch, Serotype replacement in disease after pneumococcal vaccination. Lancet, 2011. 378(9807): p. 1962-73.
33. Krone, M., et al., Increase of invasive meningococcal serogroup W disease in Europe, 2013 to 2017. Euro

surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 2019. 24(14): p. 1800245.
34. Bianchi, S., et al., Genetic characterisation of Measles virus variants identified during a large epidemic in Milan, Italy, March-December 2017. Epidemiol Infect, 2019. 147: p. e80.
35. Lemey, P., M. Salemi, and A.M. Vandamme, The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2009: Cambridge University Press.
36. Knol, M.J., et al., Temporal associations between national outbreaks of meningococcal serogroup W and C disease in the Netherlands and England: an observational cohort study. The Lancet Public Health, 2017. 2(10): p. e473-e482.
37. Bedford, T., et al., Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature, 2015. 523(7559): p. 217-20.
38. Metsky, H.C., et al., Zika virus evolution and spread in the Americas. Nature, 2017. 546(7658): p. 411-415.
39. Hahne, S., et al., Selective hepatitis B virus vaccination has reduced hepatitis B virus transmission in the Netherlands. PLoS One, 2013. 8(7): p. e67866.
40. Few, S., Show Me the Numbers: Designing Tables and Graphs to Enlighten. 2004: Analytics Press.
41. Few, S., Now You See it: Simple Visualization Techniques for Quantitative Analysis. 2009: Analytics Press.
42. Few, S., Information Dashboard Design. 2006: Oreilly & Associates Incorporated.
43. Cairo, A., The Functional Art: An introduction to information graphics and visualization. 2012: Pearson Education.
44. Cairo, A., The Truthful Art: Data, Charts, and Maps for Communication. 2016: Pearson Education.
45. Tufte, E.R., The Visual Display of Quantitative Information. 1983: Graphics Press.
46. Tufte, E.R., Envisioning information. 1991: Graphics Press.