

Universiteit
Leiden
The Netherlands

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

Soetens, L.C.

Citation

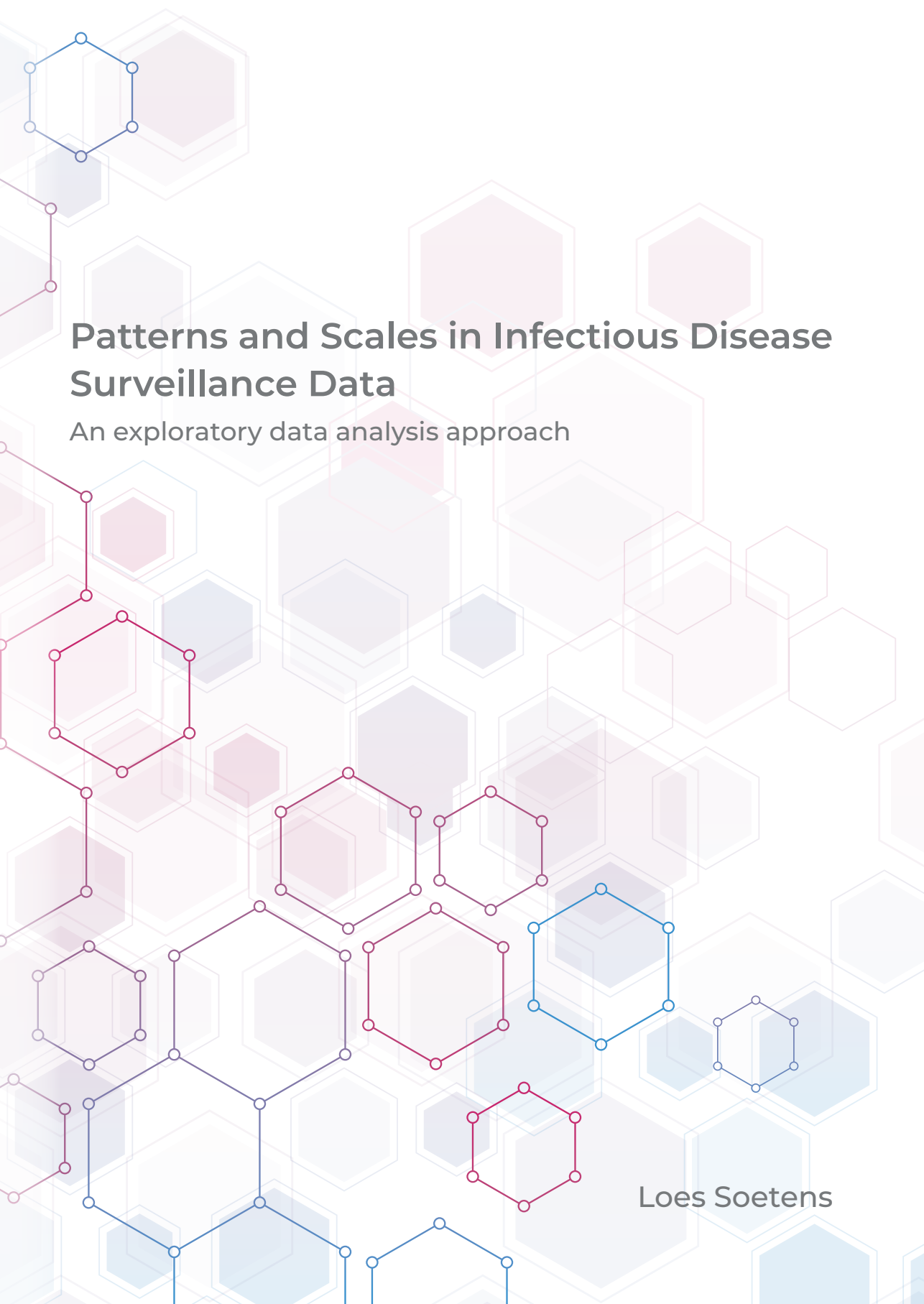
Soetens, L. C. (2021, December 15). *Patterns and scales in infectious disease surveillance data: an exploratory data analysis approach*. Retrieved from https://hdl.handle.net/1887/3247049

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3247049

Note: To cite this publication please use the final published version (if applicable).

Patterns and Scales in Infectious Disease Surveillance Data

An exploratory data analysis approach



Loes Soetens

PATTERNS AND SCALES IN INFECTIOUS DISEASE SURVEILLANCE DATA

An exploratory data analysis approach

Loes Soetens

Cover design: Jennifer Middleton
Cover illustratie: [pro100vector]/123RF
Layout: Ilse Modder | www.ilsemodder.nl
Printed: Gildeprint, Enschede | www.gildeprint.nl
ISBN: 978-94-6419-344-2



The publication of this thesis was financially supported by the National Institute for Public Health and the Environment (RIVM) and Leiden University.

Copyright © 2021 by L.C. Soetens.
All rights reserved. Any unauthorised reprint or use of this material is prohibited. No part of this thesis may be reproduced, stored or transmitted in any form or by any means, without written permission of the copyright owner.

PATTERNS AND SCALES IN INFECTIOUS DISEASE SURVEILLANCE DATA

An exploratory data analysis approach

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van de rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 15 december 2021
klokke 11.15 uur
door
Lucia Catharina Soetens
geboren te ‘s-Hertogenbosch
in 1985

Promotor
Prof. dr. J. Wallinga

Copromotor
Dr. S.J.M. Hahné (Rijksinstituut voor Volksgezondheid en Milieu)

Promotiecommissie
Prof. dr. S. le Cessie
Prof. dr. H. Putter
Prof. dr. M.A.B. van der Sande (Universiteit van Utrecht)
Prof. dr. C.J.P.A. Hoebe (Universiteit van Maastricht)

Table of contents

CHAPTER 1	General introduction	9
CHAPTER 2	Ongoing transmission of hepatitis B virus in rural parts of the Netherlands, 2009-2013	33
CHAPTER 3	Dot map cartograms for detection of infectious disease outbreaks	51
CHAPTER 4	Multi-scale analysis of infectious disease surveillance data	67
CHAPTER 5	Real-time estimation of epidemiologic parameters from contact tracing data	91
CHAPTER 6	Visual tools to assess the plausibility of algorithm-identified infectious disease clusters	121
CHAPTER 7	General discussion	147
APPENDIX	Summary	162
	Nederlandse samenvatting	166
	Dankwoord	168
	Curriculum vitae	170
	List of publications	171



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.

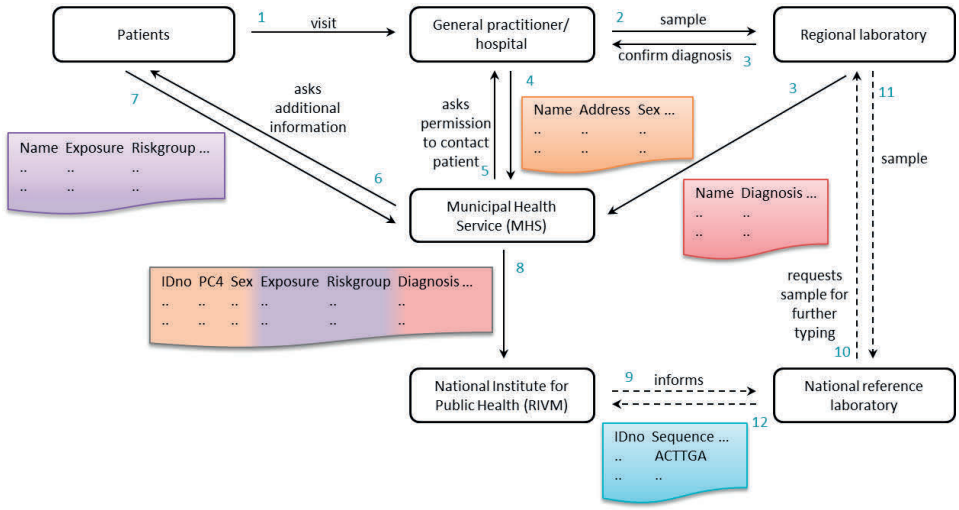


Figure 1.1. Information flow of notification process of notifiable infectious diseases in the Netherlands

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.

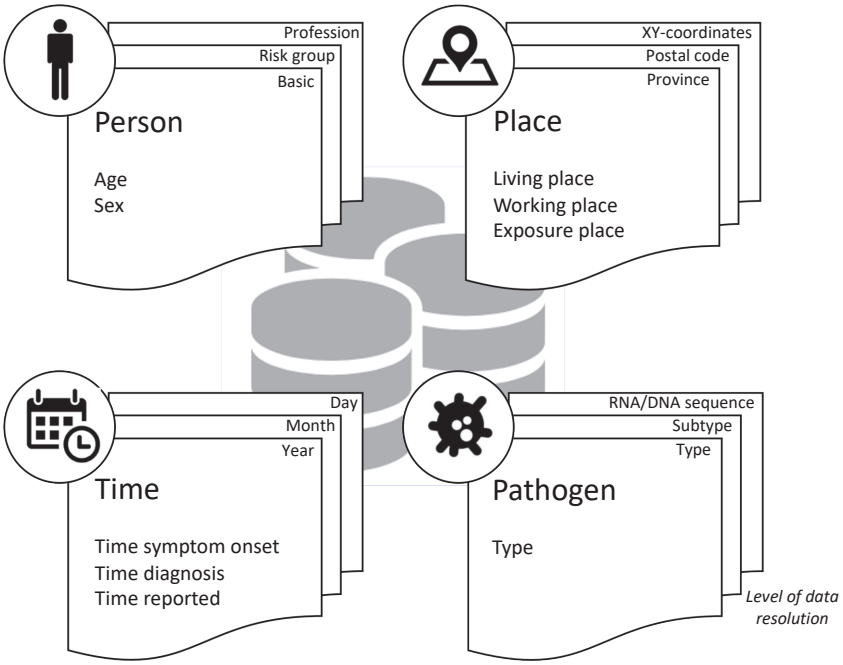


Fig. 1.2. Data types in infectious disease surveillance

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.

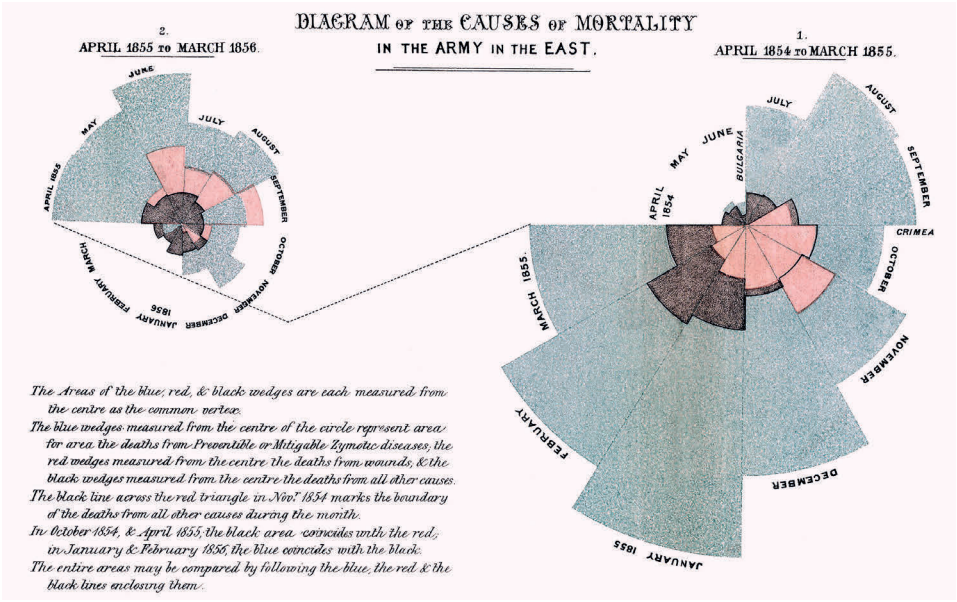


Figure 1.3. Polar diagram of Florence Nightingale (adapted from [6])

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.

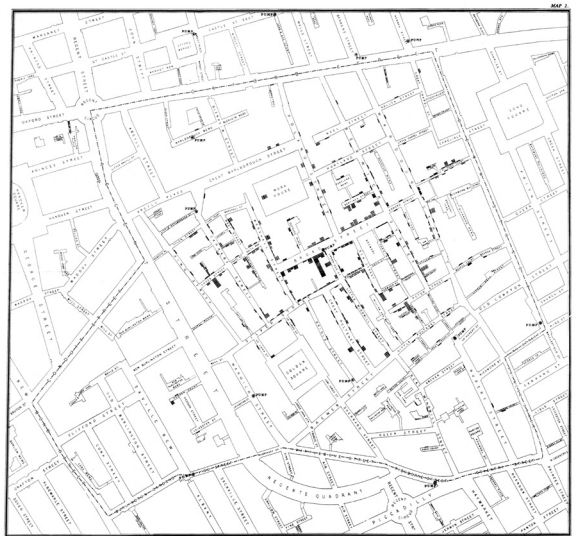


Figure 1.4. John Snow's map of London with cholera cases during the epidemic in 1854 (adapted from [7])

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.

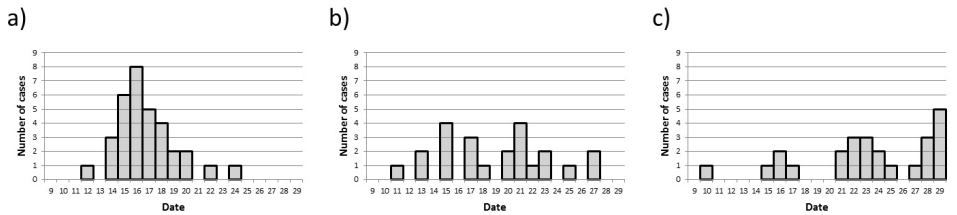


Figure 1.5. Typical epidemic curves of a point-source outbreak (a), continuous-source outbreak (b) and a person-to-person outbreak (c). (Adapted from Giesecke, 2016 [14])

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].

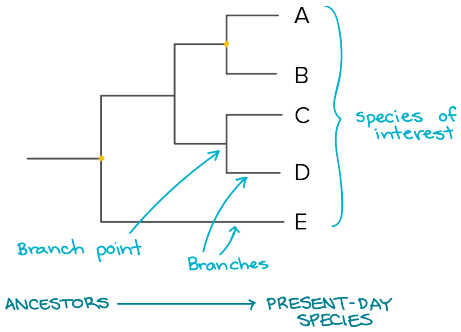


Figure 1.7. Example of phylogenetic tree

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].

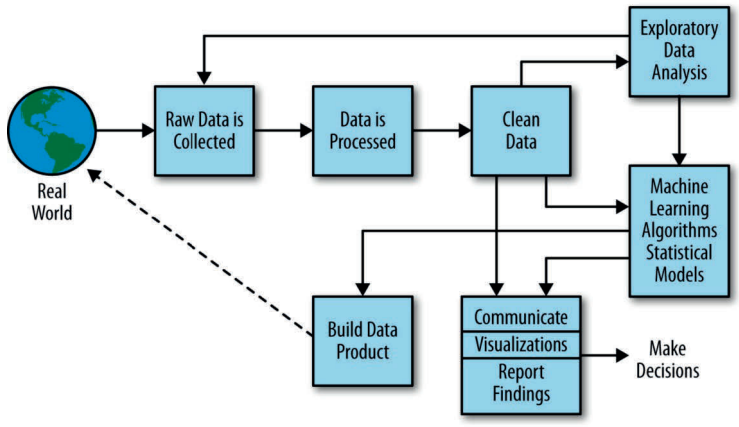


Fig 1.8. The data science process (adapted from O'Neil and Schutt, 2013[2]).

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.



2

Ongoing transmission of hepatitis B virus in rural parts of the Netherlands, 2009-2013

The contents of this chapter have been published in PLoS One:
Ongoing transmission of hepatitis B virus in rural parts of the Netherlands, 2009-2013
Loes Soetens, Birgit van Benthem, Anouk Urbanus, Jeroen Cremer, Kimberly Benschop, Ariene Rietveld, Erik I van Dijk, Susan JM Hahné

PLoS One. February 23 2015;10(2)

Abstract

Reported acute hepatitis B incidence in the Netherlands reached its nadir in 2013. However, regional signals about increased number of hepatitis B cases raised the question how hepatitis B incidence was distributed over the country. In this study, regional differences in hepatitis B epidemiology were investigated using epidemiological and molecular data.

Acute hepatitis B virus (HBV) infections, reported between 2009-2013, were included. If serum was available, a fragment of S and C gene of the HBV was amplified and sequenced. Regional differences in incidence were studied by geographical mapping of cases and cluster analysis. Regional differences in transmission were studied by constructing regional maximum parsimony trees based on the C gene to assess genetic clustering of cases.

Between 2009 and 2013, 881 cases were notified, of which respectively 431 and 400 cases had serum available for S and C gene sequencing. Geographical mapping of notified cases revealed that incidences in rural border areas of the Netherlands were highest. Cluster analysis identified two significant clusters ($p< 0.000$) in the South-western and North-eastern regions. Genetic cluster analysis showed that rural border areas had relatively large clusters of cases with indistinguishable sequences, while other regions showed more single introductions.

This study showed that regional differences in HBV epidemiology were present in the Netherlands. Rural border regions showed higher incidences and more ongoing transmission, mainly among MSM, than the more urban inland areas. Therefore, preventive measures should be enhanced in these regions.

Introduction

Hepatitis B virus (HBV) infection can cause a broad spectrum of disease outcomes from asymptomatic infection to severe complications. People with chronic HBV infection are at increased risk of developing hepatic decompensation, cirrhosis and hepatocellular carcinoma. Since the introduction of hepatitis B vaccination in the 1980s, the incidence of hepatitis B infection dropped tremendously worldwide [1,2]. However, in specific high-risk groups, the HBV prevalence and incidence can still be substantial [2].

In the Netherlands, reported acute hepatitis B incidence decreased since 2004 and reached its nadir in 2013 with 0.8 per 100,000 population [3]. This incidence was comparable with the overall acute hepatitis B incidence reported in the European Union/European Economic Area and with the incidence reported in the neighbouring country Germany in 2011 [4]. The most common route of transmission to acquire hepatitis B virus infection in the Netherlands is through sexual contact, mainly by men who have sex with men (MSM)[5], whereas in Europe in general the most common route of transmission is through heterosexual sexual contact [4].

In 2002, the Netherlands adopted a selective vaccination programme targeted at behavioural high-risk groups. The programme currently offers free vaccination to MSM and commercial sex workers (CSW). Vaccination takes place at Public Health Service (PHS) centres, sexually transmitted infections (STI) clinics and outreach locations. This selective vaccination programme has been proven effective, especially in reducing hepatitis B infections among MSM [5-7]. Since 2011, hepatitis B vaccination is part of the national immunisation programme, which means that all infants are offered vaccination against hepatitis B.

Although the national incidence of reported acute HBV infection in the Netherlands is decreasing, in the last years some regional PHS centres observed significant increases in the incidence, suggesting hepatitis B incidence was not distributed evenly throughout the country. This could have implications for the targeting of preventive and control measures. We therefore investigated regional differences in hepatitis B epidemiology by studying notifications, molecular data of their hepatitis B viruses and data of the vaccination programmes for behavioural high risk groups, in order to inform hepatitis B prevention.

Methods

Data collection

In the Netherlands, physicians and laboratories are obliged to notify acute hepatitis B infections. The case definition for reporting is a positive laboratory result for immunoglobulin M antibody to the hepatitis B core antigen (IgM anti-HBc) and/or hepatitis B surface antigen (HBsAg; the latter only after exclusion of hepatitis A and C viruses) in a person with an acute onset of symptoms compatible with acute hepatitis and with jaundice and/or increased serum aminotransferase. To study regional differences in the incidence of acute hepatitis B we used data from the National Register for Notifiable Diseases ('Osiris'). We used all cases diagnosed and registered between 2009 and 2013. The (molecular) epidemiology of cases from 2004-2010, was shown reported previously [5]. Of these cases, epidemiological information was collected, among which age, sex, postal code, date of diagnosis and (probable) mode of transmission. Data were extracted from the web-based notification register in March 2014. All information was available on an individual level.

To study regional differences in transmission, we used molecular typing data collected by the National Institute for Public Health and the Environment (RIVM). Since 2004, medical microbiology laboratories are asked to submit serum samples of patients diagnosed with acute hepatitis B to the laboratory of the RIVM. We used all available sequences of acute hepatitis B infections reported in 2009-2013 to conduct a genetic cluster analysis as described below. All sequences used in our analysis are deposited in GenBank (accession numbers: C-gene, KP243201 - KP243600; S-gene, KP243601 - KP244031).

To examine regional differences in preventive activities regarding hepatitis B infection, we included data of the national vaccination programme for behavioural high-risk groups in this study. The number of vaccinations and percentage of vaccinations performed by outreach activities, recorded between 2009-2013, were used. We also calculated the rate of first vaccinations per 100,000 men or women (depending on the risk group) for each region over 5 years. Data on population size by region from Statistic Netherlands [8] were used to calculate this rate. Data of the vaccination programme were available on regional level. Statistical differences in categorical variables across regions, spatio-temporal clusters or transmission routes were calculated using the Chi-squared test, and statistical differences in medians across regions or transmission routes were calculated using the Kruskal-Wallis test. All statistical analyses were performed using the SAS 9.3 statistical program.

Ethical approval was sought but deemed not necessary by the ethics committee of the Amsterdam Medical Centre. The reason for this was that residual sera collected for HBV diagnostic purposes were in this project used anonymously for HBV testing only. Informed consent by participants was therefore not part of the study procedures.

Geographic cluster analysis

To study regional differences in epidemiology of acute hepatitis B, we stratified statistical analyses by region (Figure 2.1). Furthermore, all cases were mapped using the geographic information system ArcGIS software, version 10.1. The postal code was used to identify the geographic location of cases, but data were aggregated and analysed on municipality level for privacy reasons. Incidences per municipality were calculated per 100,000 population. To identify geographic clusters of cases in the period 2009-2013, cluster analysis on municipality level was performed using the SaTScan software version 9.0 [9]. We used a discrete Poisson model, which assumes that the number of cases in each location is Poisson-distributed, and analysed the data with the space-time scan statistic to identify clusters in space and time. The space-time scan statistic uses cylindrical search windows in a space-time cube in order to seek significant spatio-temporal clusters. We set SaTScan to scan only for areas with high rates, that is for clusters.



Figure 2.1. Regions for sexual health care in the Netherlands. Purple areas indicate the more rural regions, green areas indicate the more urban regions.

Genotyping and phylogenetic analysis

The RIVM amplified and sequenced fragments of the S- and C-gene (656-nucleotide (nt) and 655-nt respectively) as described in previous studies [5,10,11]. The S-gene was used for typing [12], while the C-gene was used to study transmission chains; an earlier study [11] showed that the S-gene, a more conserved region, is more useful for distinguishing genotypes than

the C-gene. The latter, however, is preferable for identifying recent transmission chains, due to the relatively high mutation rate in this region. C-gene sequences were aligned using the software Bionumerics version 7.1 and were used to construct phylogenetic trees with the maximum parsimony method, to assess genetic clustering of cases. A cluster was defined as three or more cases with indistinguishable sequences. Patients with indistinguishable sequences were displayed in the same node in the maximum parsimony tree (the size of the node is relative to the number of indistinguishable sequences). The trees were categorized according to different geographic regions shown in Figure 2.1 and probable mode of transmission.

Results

Study population

Between 2009 and 2013, 881 cases were notified with acute hepatitis B in the Netherlands. The basic characteristics of the study population by region are shown in Table 2.1. The five-year incidence varied between 0.73 and 1.57 infections per 100,000 population across the regions, and was in all regions higher in men than in women. Across the regions, the percentage of men varied between 65.5% and 83.7%, the percentage of patients born in the Netherlands ranged from 65.1% to 92.3% and the median age varied between 32 years and 47 years. Considering the most probable mode of transmission, the country can be divided in regions where most patients got infected through heterosexual contact (Noord-Holland, Utrecht and Zuid-Holland Noord; in general the more urban areas) and regions where most patients got infected through male homosexual contact or were men with an unknown mode of transmission (Brabant/Zeeland, Zuid-Holland Zuid, Limburg, Noord and Oost; mainly the more rural areas of the Netherlands).

Spatio-temporal clusters

In Figure 2.2, the geographical distribution of the HBV incidence over time is shown. The highest incidences in the separate years and overall were seen in the north-eastern and southern regions of the country, while most large cities are in the central and western regions of the country. The space-time scan statistic identified two significant clusters. The first cluster was present between February 2009 and December 2010, and was located in the south-western region of the country (n= 96, p= 0.001). The second cluster was present between November 2010 and April 2013, and was located in the north-eastern region of the country (n= 39, p= 0.000). When comparing the patients inside and outside the spatio-temporal clusters (Table 2.2), it can be observed that the patients in the south-western spatio-temporal cluster did not differ from the patients outside the clusters with respect to sex, median age, ethnicity and transmission route (p> 0.05). However, the patients within the north-eastern spatio-temporal cluster were generally more often male (p= 0.021) and were more likely to be MSM or men with an unknown mode of transmission (p= 0.003) than patients outside the spatio-temporal clusters.

Table 2.1. Basic characteristics of acute hepatitis B patients by region of reporting, notified between 2009-2013 in the Netherlands

		Brabant/ Zeeland n=173 N(%)	Limburg n=52 N(%)	Noord n=135 N(%)	Noord-Holland n=145 N(%)	Oost n=129 N(%)	Utrecht n=45 N(%)	Zuid-Holland Noord n=63 N(%)	Zuid-Holland Zuid n=139 N(%)	P-value	Total n=881 N(%)
Incidence per 100,000*	Total	1.22	0.93	1.57	0.94	0.82	0.73	1.23	1.11	0.000	1.06
	Men	1.99	1.55	2.65	1.25	1.32	1.13	1.93	1.74	0.000	1.66
	Women	0.45	0.32	0.51	0.64	0.33	0.35	0.54	0.50	0.128	0.47
Sex	Men	141 (81.5)	43 (82.7)	113 (83.7)	95 (65.5)	103 (79.8)	34 (75.6)	49 (77.8)	107 (77.0)	0.012	685 (77.8)
	Women	32 (18.5)	9 (17.3)	22 (16.3)	50 (34.5)	26 (20.2)	11 (24.4)	14 (22.2)	32 (23.0)		196 (22.3)
Age	Median	42.5	45.0	45.0	38.0	47.0	32.0	38.0	45.0	<0.0001	43.0
	(IQR)	(33.0 - 51.0)	(34.5 - 57.5)	(30.0 - 54.0)	(28.0 - 47.0)	(31.0 - 60.0)	(25.0 - 45.0)	(27.0 - 52.0)	(34.0 - 57.0)		(30.0 - 53.5)
Ethnicity	Dutch	139 (80.4)	47 (90.4)	113 (83.7)	103 (71.0)	119 (92.3)	35 (77.8)	41 (65.1)	109 (78.4)	<0.0001	706 (80.1)
	Non-Dutch	34 (19.6)	5 (9.6)	22 (16.3)	42 (29.0)	10 (7.7)	10 (22.2)	22 (34.9)	30 (21.6)		175 (19.9)
Transmission route	Heterosexual men	32 (18.5)	8 (15.4)	19 (14.1)	35 (24.1)	22 (17.1)	13 (28.9)	19 (30.2)	33 (23.7)	0.004	181 (20.5)
	MSM	49 (28.3)	18 (34.6)	42 (31.1)	35 (24.1)	32 (24.8)	9 (20.0)	12 (19.1)	42 (30.2)		239 (27.1)
	Women - sex. contact	24 (13.9)	7 (13.5)	16 (11.9)	38 (26.2)	19 (14.7)	9 (20.0)	10 (15.9)	23 (16.6)		146 (16.6)
	Other	16 (9.3)	4 (7.7)	5 (3.7)	7 (4.8)	13 (10.1)	3 (6.7)	9 (14.3)	14 (10.1)		71 (8.1)
	Unknown-men	48 (27.8)	14 (26.9)	47 (34.8)	23 (15.9)	39 (30.2)	10 (22.2)	9 (14.3)	22 (15.8)		212 (24.1)
	Unknown-women	4 (2.3)	1 (1.9)	6 (4.4)	7 (4.8)	4 (3.1)	1 (2.2)	4 (6.4)	5 (3.6)		32 (3.6)
Number sequenced	S-gene	79 (45.7)	41 (78.8)	91 (67.4)	53 (36.6)	74 (57.4)	21 (46.7)	25 (39.7)	47 (33.8)	<0.0001	431 (48.9)
Genotype based on S-gene	A	62 (78.5)	29 (70.7)	80 (87.9)	27 (50.9)	56 (75.7)	11 (52.4)	12 (48.0)	31 (66.0)	0.000	308 (71.5)
	B	2 (2.5)	0 (0.0)	1 (1.1)	3 (5.7)	2 (2.7)	1 (4.8)	0 (0.0)	2 (4.3)		11 (2.6)
	C	4 (5.1)	3 (7.3)	3 (3.3)	2 (3.8)	4 (5.4)	3 (14.3)	3 (12.0)	5 (10.6)		27 (6.3)
	D	5 (6.3)	4 (9.8)	4 (4.4)	18 (34.0)	8 (10.8)	6 (28.6)	7 (28.0)	6 (12.8)		58 (13.5)
	E	5 (6.3)	2 (4.9)	2 (2.2)	2 (3.8)	4 (5.4)	0 (0.0)	2 (8.0)	2 (4.3)		19 (4.4)
	F	1 (1.3)	3 (7.3)	1 (1.1)	1 (1.9)	0 (0.0)	0 (0.0)	1 (4.0)	1 (2.1)		8 (1.9)

**The incidence was calculated as the average over the yearly incidences in 2009-2013*

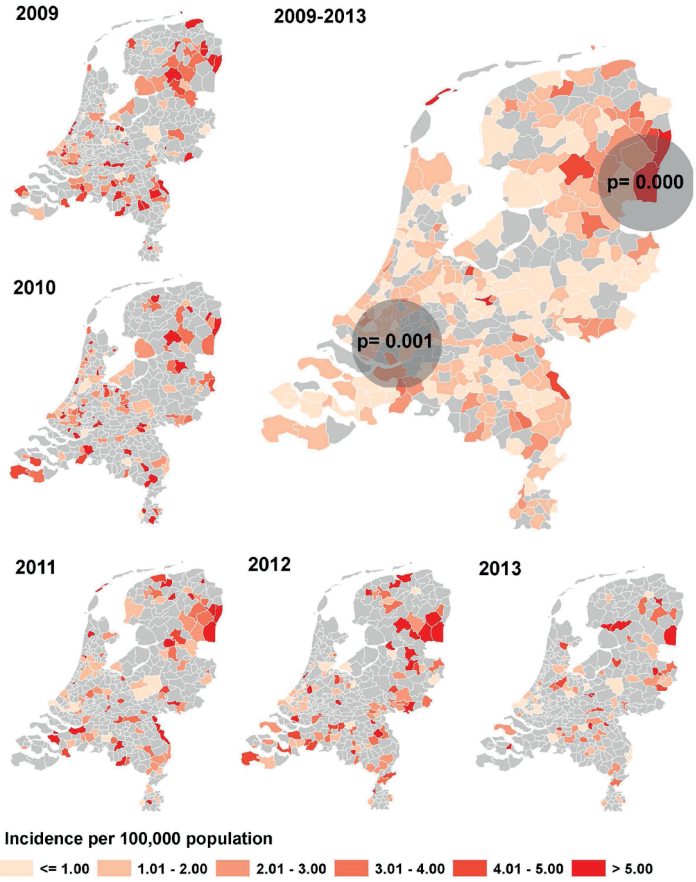


Figure 2.2. Incidence of acute HBV in the Netherlands, by year and overall incidence 2009-2013. The dark grey circles indicate regions with significant geographic clusters.

Table 2.2. Characteristics of patients within the spatio-temporal clusters compared to patients outside the spatio-temporal clusters

		Not in spatio-temporal cluster (ref) n= 746 n (%)	Southwestern spatio-temporal cluster n= 96 n (%)	p-value*	Northeastern spatio-temporal cluster n= 39 n (%)	p-value*
Sex	Men	570 (76.4)	79 (82.3)	0.1967	36 (92.3)	0.021
	Women	176 (23.6)	17 (17.7)		3 (7.7)	
Age	Median (IQR)	42 (30 - 53)	43 (29.5 - 53.5)	0.9371	46 (42 - 54)	0.0571
Ethnicity	Dutch	594 (79.6)	75 (78.1)	0.4033	37 (94.9)	0.0533
	Non-Dutch	126 (16.9)	15 (15.6)		1 (2.6)	
	Unknown	26 (0.1)	6 (6.3)		1 (2.6)	
Transmission route	Heterosexual men	155 (20.8)	24 (25.0)	0.0957	2 (5.1)	0.0025
	MSM	198 (26.5)	28 (29.2)		13 (33.3)	
	Women - sex. contact	127 (17.0)	16 (16.7)		3 (7.7)	
	Other	57 (7.6)	12 (12.5)		2 (5.1)	
	Unknown - men	177 (23.7)	16 (16.7)		19 (48.7)	
	Unknown - women	32 (4.3)	0 (0.0)		0 (0.0)	
Number sequenced	C-gene	343 (46.0)	28 (29.2)	0.0018	27 (69.2)	0.0046
Median cluster size (IQR)	C-gene	7 (1-54)	3.5 (1-51)	0.9012	51 (51-70)	<0.0001
Genetic cluster ≥ 3	C-gene	193 (56.3)	16 (57.1)	0.9285	23 (85.2)	0.0033

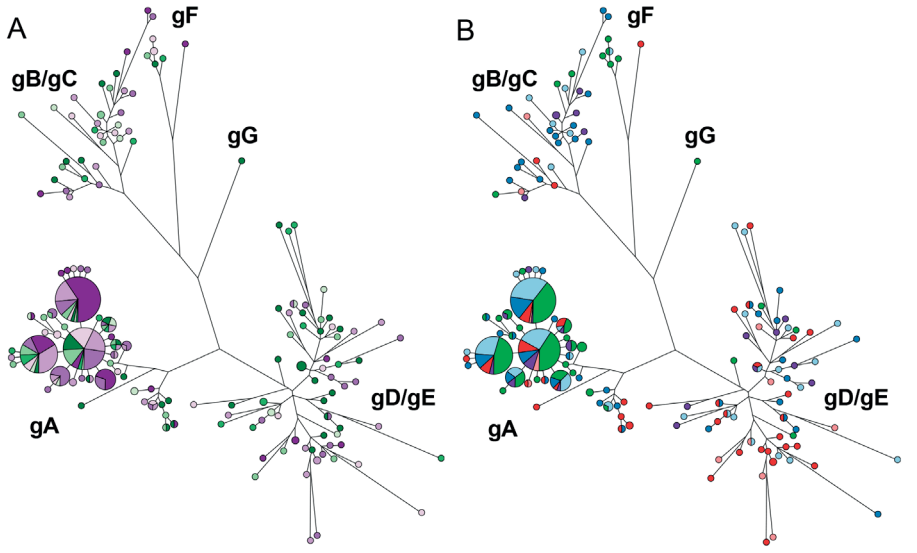
*Compared to patients not in a spatio-temporal cluster.

Genetic clusters

Of the 881 notified cases, 492 samples (56%) arrived at the RIVM for molecular typing. Reasons why samples did not arrive at the RIVM were refused cooperation by the patient (n= 101), insufficient or no material available in the local laboratory (n= 102), refused cooperation of the local laboratory (n= 50) and unknown (n= 136). In the end, 431 sequences (87.6%) of the S-gene and 400 sequences (81.3%) of the C-gene were successfully typed. The percentage samples sent in for sequencing differed by region (ranging from 33.8% in Zuid-Holland Zuid to 78.8% in Limburg, p< 0.0001, Table 2.1) and between the reported mode of transmission (ranging from 40.6% women with an unknown mode of transmission to 58.2% in MSM, p= 0.015). However, the transmission mode distribution of the received samples did not differ by region (p= 0.903), suggesting that the difference in received samples between the regions was not due to selection in a certain risk group, but was mainly due to external factors, such as mentioned above. Of all successfully typed S-gene sequences (n= 431), 71.5% was genotype A, 2.6% was genotype B, 6.3% was genotype C, 13.5% was genotype D, 4.4% was genotype E and 1.9% was genotype F (Table 2.1). Across all regions, genotype A was the most frequently reported genotype, followed by genotype D. The maximum parsimony trees based on C-gene of all HBV cases in the Netherlands, reported between 2009 and 2013, are shown in Figure 2.3. In Figure 2.3A, the cases are coloured by region of reporting and cases with indistinguishable sequences are clustered together in nodes. The percentage of patients present in a cluster of 3 or more individuals varied greatly between the regions, ranging from 31.3% to 78.2%

(p< 0.0001). Similarly, the median cluster size varied as well between the regions, ranging from 1 to 54 (p< 0.0001).

In Figure 2.3B the same maximum parsimony tree is shown as in Figure 2.3A, but now the tree is coloured by most probable mode of transmission. In this figure, it can be seen that cases in a cluster were more likely to be MSM and men with an unknown mode of transmission (higher percentage of samples present in a cluster of 3 or more samples and higher median cluster sizes).



Region	N	% samples in cluster ≥ 3	Median cluster size (n)
Noord	87	78.2	54
Oost	70	64.3	14
Brabant/Zeeland	72	70.8	51
Limburg	38	47.4	2
Utrecht	18	38.9	2
Zuid-Holland Zuid	40	47.5	2
Zuid-Holland Noord	25	36.0	1
Noord-Holland	48	31.3	1

Transmission route	N	% samples in cluster ≥ 3	Median cluster size (n)
MSM	128	76.6	51
Heterosexual men	73	43.8	2
Men - unknown	97	68.0	51
Women – sex. contact	57	29.8	2
Women - unknown	13	30.8	1
Other	30	50.0	2.5

Figure 2.3. Maximum parsimony trees of HBV cases (2009-2013) in the Netherlands based on C-gene sequences (N=400), by region (Figure 2.3A) and most probable mode of transmission (Figure 2.3B). (2.3A: purple= rural areas, green= urban areas; Genotypes: gA= genotype A, gB= genotype B, gC= genotype C, gD= genotype D, gE= genotype E, gF= genotype F, gG= enotype G; patients with indistinguishable sequences are displayed together in one node and the size of the node is relative to the number of indistinguishable sequences)

Genetic clustering within the spatio-temporal clusters

To assess if patients within a spatio-temporal cluster also showed higher genetic clustering rates, we compared the percentage of patients in a genetic cluster of 3 or more samples and the median cluster size of patients within a spatio-temporal cluster to patients outside a spatio-temporal cluster (Table 2.2). Compared to the patients outside a spatio-temporal cluster, patients within the south-western spatio-temporal cluster had significantly less samples available for sequencing ($p= 0.002$), but the percentage of patients in a cluster or the median cluster size did not significantly differ from those outside a spatio-temporal cluster. Patients within the north-eastern spatio-temporal cluster had significantly more samples available for sequencing compared to the patients outside the spatio-temporal cluster ($p= 0.005$). In addition, the percentage of patients in a cluster of three or more was significantly higher ($p= 0.003$) and the median cluster size was larger ($p< 0.0001$) within this spatio-temporal cluster than outside the spatio-temporal cluster.

Preventive activities

The number and location of first vaccinations (2009-2013) of the selective HBV vaccination programme for behavioural risk-groups are displayed in Table 2.3. Of all first vaccinations given within the HBV vaccination programme, most vaccinations were received by MSM, compared to CSW, varying from 54.3 to 80.7% of the total number of vaccinations across the regions. We calculated the number of first vaccinations received by MSM per 100,000 men for each region, which varied greatly across the regions: from 112 to 425 vaccinations per 100,000 men over 5 years. It is worth noticing that the region with the lowest vaccination rate among MSM reported the highest incidence among men (Table 2.1). We calculated the same rate for the first vaccinations received by CSW, which also varied across the regions but less than was seen among MSM: 45 to 147 vaccinations per 100,000 women. Most first vaccinations of MSM were given at the public health services or STI clinics (ranging from 67.3 to 89.0% across the regions). More variation between the regions was seen in the location of first vaccination for CSW; in this group the percentage vaccinations given at outreach locations varied between 30.0 and 77.0% across the regions.

Table 2.3. Number and location of first vaccinations of the selective HBV vaccination programme for behavioural risk-groups by region, The Netherlands, 2009-2013.

	Brabant/ Zeeland n=3,065 N(%)	Limburg n=1,899 N(%)	Noord n=1,756 N(%)	Noord-Holland n=8,453 N(%)	Oost n=3,899 N(%)	Utrecht n=1,201 N(%)	Zuid-Holland Noord n=2,194 N(%)	Zuid-Holland Zuid n=3,030 N(%)
No of 1st vaccinations								
MSM	2,426 (79.2)	1,222 (64.3)	954 (54.3)	6,449 (76.3)	3,033 (77.8)	831 (69.2)	1,431 (65.2)	2,446 (80.7)
CSW	639 (20.8)	677 (35.7)	802 (45.7)	2,004 (23.7)	866 (22.2)	370 (30.8)	763 (34.8)	584 (19.3)
No of 1st MSM vaccinations per 100,000 men*	171.6	219.6	111.8	424.6	194.9	138.3	281.7	198.6
No of 1st CSW vaccinations per 100,000 women*	45.0	119.6	93.0	128.4	54.7	58.9	147.0	46.1
Location of 1st vaccination MSM								
PHS/STI	1,886 (77.7)	1,088 (89.0)	729 (76.3)	4,935 (76.5)	2,508 (82.7)	708 (85.2)	963 (67.3)	1,772 (72.4)
Outreach	513 (21.1)	133 (10.9)	225 (23.6)	1,109 (17.2)	364 (12.0)	109 (13.1)	468 (32.7)	485 (19.8)
Other	27 (1.1)	1 (0.1)	1 (0.1)	405 (6.3)	160 (5.3)	14 (1.7)	0 (0.0)	189 (7.7)
Location of 1st vaccination CSW								
PHS/STI	379 (59.3)	466 (68.8)	289 (36.0)	272 (13.6)	360 (41.6)	82 (22.2)	217 (28.4)	286 (49.0)
Outreach	210 (32.9)	203 (30.0)	492 (61.3)	1,281 (63.9)	486 (56.1)	285 (77.0)	546 (71.6)	285 (48.8)
Other	50 (7.8)	8 (1.2)	21 (2.6)	451 (22.5)	20 (2.3)	3 (0.8)	0 (0.0)	13 (2.2)

MSM= men who have sex with men; CSW= commercial sex workers; PHS= public health centre; STI= sexually transmitted infections clinic
*Number of men or women is based on the average population 2009-2013 in that specific region

Discussion

We used epidemiological data from notified acute hepatitis B cases between 2009 and 2013 together with molecular data and data on preventive activities, to study the regional differences in incidence, transmission and preventive activities in the Netherlands. We found that in the rural border areas the incidence of acute HBV infection was higher, most patients got infected through male homosexual contact or were men with an unknown transmission mode, and that patients were genetically more clustered, suggesting more ongoing transmission in these areas. In the more inland and more densely populated areas of the Netherlands, heterosexual transmission was the most common transmission mode and HBV sequences of patients were genetically less clustered, suggesting more single introductions in these areas. In addition, two statistically significant spatio-temporal clusters were identified (one in the southwest and one in the northeast of the Netherlands), but only the patients in the north-eastern spatio-temporal cluster statistically differed from the other patients in the country with regard to age, sex, ethnicity and genetic clustering. Finally, considering preventive activities, we found that the lowest rate of number of vaccinations among MSM per 100,000 men was seen in the northern region, which also showed the highest incidence among men.

The higher incidences in the rural border areas, specifically in the northeast, suggest that transmission of acute hepatitis B is not yet controlled in these areas. We found that it is mainly a problem among MSM: more ongoing transmission and relatively low vaccination rates. There might be several reasons for the higher incidences and lower vaccination rates among MSM in these areas. First, it could be possible that it is harder for the public health services to reach MSM in these areas for vaccination, since it could be the case that in these areas more men did not disclose their homosexuality. We observed that in the more rural areas of the Netherlands the percentage of men with an 'unknown' transmission route is relatively higher compared to the urban areas. This could be an indication of lower disclosure rates in these areas. If men do not openly come out for their homosexual contacts, they might also not identify themselves as being gay. This could lead to a low perceived risk of acquiring acute hepatitis B infection, which was shown to be the most important reason for opting out of the vaccination programme in the Netherlands [13,14]. Another explanation for the higher incidences among MSM in the rural areas is that MSM in the rural areas might have different patterns of risk behaviour than MSM in urban areas [15]. Finally, since it concerns rural areas, the distance to reach the public health services to get a vaccination can be larger than in the urban areas. This can therefore be another barrier for MSM to get vaccinated, and therefore explain the higher incidences in these areas.

The major strength of this study is the comprehensive approach of addressing this research question, by making use of both epidemiological and molecular sequence data. These

different types of data are usually analysed separately, but we attempted to link and combine these data types and databases as much as possible, to get a complete overview of the hepatitis B situation in the Netherlands of the last five years.

There are several limitations of our study. First, although the notification of acute hepatitis B is obligatory in the Netherlands, it is likely that not all cases are included in this study since asymptomatic infections are not captured in the surveillance system. In addition, there is likely to be some underreporting, albeit less now that both laboratories and clinicians have to report cases. However, it can be expected that this missing of cases is randomly distributed throughout the country, which therefore will not influence our results. The second issue is that only around sixty per cent of all notified cases had a sample available for sequencing. We found that the percentage of samples available for sequencing was not distributed evenly across the regions, which could have been problematic if certain regions specifically selected certain cases to send in for sequencing. However, we could not detect this selection bias, and we can therefore assume that the sequenced samples reflect the patient populations in the regions. Finally, we calculated the rate of first vaccinations per 100,000 men or women (depending on the risk group) by region. This is a very crude rate since it assumes that the percentage of MSM or sex workers is distributed evenly throughout the regions, which is probably not true. Since it is not known how many MSM or CSW live in each region, our calculation provides currently the best possible estimate. In addition, MSM not always get their vaccination in the region they live, which could also influence this estimate.

Since 2011, hepatitis B vaccination in the Netherlands is part of the national vaccination programme for infants. It is expected that this will reduce the incidence of hepatitis B over what has been achieved by selective vaccination [6,16]. However, since most hepatitis B is transmitted sexually in the Netherlands, it will take several decades for this to be observed [6]. Therefore, it is of continued importance to put effort in further reducing transmission by improving or better targeting preventive activities. The results of this study made clear that the attention with regard to preventive activities should focus more on the MSM and not-yet-identified MSM in the rural areas. We recommend to further examine the reasons for the high incidences among MSM in these areas. Furthermore, it can be worthy to strengthen the outreach vaccination programme to overcome the distance barrier for vaccination. Finally, since the higher incidence regions are mostly located alongside the border with Germany, it might be useful to cooperate with our neighbouring country to further reduce the incidence in these areas.

To conclude, this study showed that regional differences in HBV epidemiology were present in the Netherlands. Rural border regions showed higher incidences and more ongoing transmission, mainly among MSM, than the more populated inland areas. Therefore, further preventive measures should focus more on these regions.

References

1. Zanetti AR, Van Damme P, Shouval D (2008) The global impact of vaccination against hepatitis B: a historical overview. Vaccine 26: 6266-6273.
2. Liaw YF, Chu CM (2009) Hepatitis B virus infection. Lancet 373: 582-592.
3. van Aar F, Koedijk, FDH, van den Broek, IVF, Op de Coul, ELM, Soetens, LC, Woestenberg, PJ, et al. (2014) Sexually transmitted infections, including HIV, in the Netherlands in 2013. Bilthoven: National Institute for Public Health and the Environment (RIVM).
4. ECDC (2013) Annual Epidemiological Report 2013. Reporting on 2011 surveillance data and 2012 epidemic intelligence data. Stockholm: European Centre for Disease Prevention and Control.
5. Hahne S, van Houdt R, Koedijk F, van Ballegooijen M, Cremer J, et al. (2013) Selective hepatitis B virus vaccination has reduced hepatitis B virus transmission in the Netherlands. PLoS One 8: e67866.
6. Xiridou M, van Houdt R, Hahne S, Coutinho R, van Steenbergen J, et al. (2013) Hepatitis B vaccination of men who have sex with men in the Netherlands: should we vaccinate more men, younger men or high-risk men? Sex Transm Infect 89: 666-671.
7. van Rijckevorsel G, Whelan J, Kretzschmar M, Siedenburg E, Sonder G, et al. (2013) Targeted vaccination programme successful in reducing acute hepatitis B in men having sex with men in Amsterdam, the Netherlands. J Hepatol 59: 1177-1183.
8. Statistics Netherlands StatLine. Available: http://statline.cbs.nl/statweb/?LA=en. Accessed 15 May 2014.
9. Kulldorff M. SaTScan, Software for the Spatial and Space-time Scan Statistics. Available: http://www.satscan.org/. Accessed 18 April 2014.
10. van Houdt R, Bruisten SM, Koedijk FD, Dukers NH, Op de Coul EL, et al. (2007) Molecular epidemiology of acute hepatitis B in the Netherlands in 2004: nationwide survey. J Med Virol 79: 895-901.
11. Boot HJ, Cremer J, Koedijk FD, van Ballegooijen WM, Op de Coul EL (2008) Improved tracing of hepatitis B virus transmission chains by phylogenetic analysis based on C region sequences. J Med Virol 80: 233-241.
12. Bartholomeusz A, Schaefer S (2004) Hepatitis B virus genotypes: comparison of genotyping methods. Rev Med Virol 14: 3-16.
13. Baars JE, Boon BJ, Garretsen HF, van de Mheen D (2011) The reach of a hepatitis B vaccination programme among men who have sex with men. Eur J Public Health 21: 333-337.
14. de Wit JB, Vet R, Schutten M, van Steenbergen J (2005) Social-cognitive determinants of vaccination behavior against hepatitis B: an assessment among men who have sex with men. Prev Med 40: 795-802.
15. Kakietek J, Sullivan PS, Heffelfinger JD (2011) You've got male: internet use, rural residence, and risky sex in men who have sex with men recruited in 12 U.S. cities. AIDS Educ Prev 23: 118-127.
16. Kretzschmar M, Mangen MJ, van de Laar M, de Wit A (2009) Model based analysis of hepatitis B vaccination strategies in the Netherlands. Vaccine 27: 1254-1260.



3

Dot map cartograms for detection of infectious disease outbreaks

The contents of this chapter have been published in Eurosurveillance:

Dot map cartograms for detection of infectious disease outbreaks: an application to Q fever, the Netherlands and pertussis, Germany

Loes Soetens, Susan Hahné, Jacco Wallinga

Eurosurveillance, June 29 2017, 22(26)

Abstract

Geographical mapping of infectious diseases is an important tool for detecting and characterising outbreaks. Two common mapping methods, dot maps and incidence maps, have important shortcomings. The former does not represent population density and can compromise case privacy, and the latter relies on pre-defined administrative boundaries. We propose a method that overcomes these limitations: dot map cartograms. These create a point pattern of cases while reshaping spatial units, such that spatial area becomes proportional to population size. We compared these dot map cartograms with standard dot maps and incidence maps on four criteria, using two example datasets.

Dot map cartograms were able to illustrate both incidence and absolute numbers of cases (criterion 1): they revealed potential source locations (Q fever, the Netherlands) and clusters with high incidence (pertussis, Germany). Unlike incidence maps, they were insensitive to choices regarding spatial scale (criterion 2). Dot map cartograms ensured the privacy of cases (criterion 3) by spatial distortion; however, this occurred at the expense of recognition of locations (criterion 4).

We demonstrate that dot map cartograms are a valuable method for detection and visualisation of infectious disease outbreaks, which facilitates informed and appropriate actions by public health professionals, to investigate and control outbreaks.

Introduction

Here we propose a method of mapping infectious disease data, the dot map cartogram, which displays the geographical locations of reported cases from routine surveillance or outbreak investigations, such that public health experts can visualise both absolute numbers and spatial trends in incidence of infection. The method is developed to address two major causes of misinterpretation in commonly used maps: masking patterns of disease by not taking into account the population density distribution, and masking patterns by categorisation of information (across space and incidence of disease). With the dot map cartogram, we address the problem, raised by a recent systematic review [1], that spatial methods are underutilised and used in only ca 0.4% of all published outbreak investigations.

The two most frequently used map types for spatial outbreak data are the dot map and choropleth incidence map [1]. In a dot map, every case is represented by a point on the map, showing absolute numbers of cases. Illustrating absolute numbers of cases reveals the size of the outbreak and identifies areas without cases. Dot maps do not account for the underlying geographical distribution of the population. As populations are usually heterogeneously distributed, important variations in incidence of infection can be masked. Another drawback of dot maps is that they may reveal too much information about the location of specific cases, by which the privacy of a case might be violated.

In a choropleth map, a quantitative attribute is displayed per spatial unit. For example, ordinal classes of incidence may be displayed by municipality, with the areas shaded according to their incidence value and a range of shading classes [2]. Choropleth maps are mostly used to display surveillance data. They are less suitable to display outbreak data when the interest is in exact locations of cases. Although this type of map gives a quick overview of where there are more cases than expected based on the background population, it has a major limitation: The appearance of this map is heavily influenced by arbitrary choices with regard to the classification system for the ordinal classes and the spatial unit used (the latter is also referred to as the Modifiable Areal Unit Problem, MAUP [3]). Because the data are categorised both across the quantitative attribute and across the geographical units, this can lead to a loss of information by masking important internal patterns and variations. Therefore, this type of map can be misleading. This is especially problematic when mapping rare events, which is often the case with infectious diseases.

Here we combine the advantages of both maps into one map type which we call dot map cartogram. This map type uses a diffusion-based algorithm [4], which creates contiguous or value-by-area cartograms. In these value-by-area cartograms, regions are enlarged or reduced relative to the number of individuals they contain. We apply this principle to traditional dot maps: we deform the map contours based on population size and

simultaneously deform the dot pattern. In this way, a dense point pattern in a big city will become more dispersed, whereas a dense point pattern in a rural area will become even more dense. A similar technique called density-equalising map projection (DEMP) has been pioneered before to describe the spatial distribution of cryptosporidiosis among AIDS patients in San Francisco [5]. This earlier study used a different density-equalising algorithm, and application was limited to local outbreak data. This method has the advantages that the density of the dots reflects the incidence, the number of dots represent the absolute number of cases, and arbitrary choices for scale of spatial unit and classification system are avoided.

In this study, we assessed whether the advantages of the dot map cartogram outweigh the potential disadvantages. One potential drawback is that the dot map cartogram may reveal too much information with regard to the privacy of the cases. Another potential drawback is that the spatial distortion makes it hard to recognise or localise the map topology. We created dot map cartograms using real life data and compared them to the traditional choropleth and dot maps regarding four criteria: (i) ability to show both absolute numbers and incidence of the disease, (ii) sensitivity to choices regarding spatial scale and classification system, (iii) ability to ensure the privacy of individual cases, and (iv) ability to recognise locations. The comparison was applied to the mapping of a point-source outbreak and to the mapping of the occurrence of a human-to-human transmissible disease.

Methods

Data

As an example of outbreak mapping to locate a source, we used data from the Q fever outbreak in the Netherlands in 2009. Q fever case reports in the Netherlands communicable disease notification system (Osiris) include the 4-digit postal code of residence. Cases with date of onset of illness in 2009 (n = 1,740) were extracted for this analysis. Population data and shapefiles (a data format which stores geometrical locations and metadata such as population size per geometrical location) on different spatial levels (4-digit postal code and municipality level) were accessed via Statistics Netherlands [5].

As an example of a human-to-human transmissible disease, we used pertussis notifications in Germany in 2015. Data on pertussis cases was made available by the Robert Koch Institute through the SurvStat@RKI 2.0 web portal [6], in which the district of residence is registered for every case. We extracted data on laboratory-confirmed pertussis cases with a date of diagnosis in 2015 (n = 9,017). The most recent population data and shapefiles for district and federal state level boundaries (2013) were accessed through the open data portal of the Federal Agency for Cartography and Geodesy Germany [7]. We assumed that the population size per district or federal state in 2015 was sufficiently similar to that in 2013.

Geographical representation

Dot map

In a dot map, every case is represented by a dot on their location of residence. As this would reveal the exact locations of the cases and harm their privacy, we chose to use a proxy location by assigning a random point location in the spatial unit of residence (4-digit postal code for the Netherlands and district for Germany) to every case. The 4-digit postal code areas in the Netherlands have a mean population of 4,119 persons (mean surface: 8.6 km^2), and the districts in Germany have a mean population of 200,914 persons (mean surface: 888.7 km^2). For the map of the Netherlands we have used the national RD ('rijksdriehoeksstelsel')-based projection and for Germany we used the Universal Transverse Mercator (UTM) 32 projection. Both are conformal map projections in which local angles are preserved and shapes are represented accurately and without distortion for small areas.

Choropleth incidence map

In a choropleth incidence map, disease incidence is displayed per spatial unit, using ordinal classes. We chose two spatial unit levels for each country: one at the same level as the patient data (spatial unit of residence as described above) and the other one level higher (municipality in the Netherlands or federal state in Germany). In addition, we categorised incidence into five ordinal classes, using two separate classification systems: the Jenks' natural breaks algorithm [8] which seeks to reduce the variance within classes and maximise

the variance between classes, and the quantile method in which equal numbers of spatial units are placed into each class [8]. The colour schemes for the incidence classes were based on previously established map colour palettes [9].

Dot map cartogram

Basic principle

We created cartograms by reshaping the spatial units such that their area was proportional to their population by applying the Gastner-Newman diffusion-based algorithm [4] without changing the underlying map topology [10]. The principle is that once the areas have been scaled to be proportional to their population, then population density is by definition the same for each area of visually the same size on the cartogram. To create dot map cartograms, the point patterns of cases (as described in the dot map section) were simultaneously transformed in the reshape process of the spatial units so that the number of cases per unit area reflected the incidence. In addition, to provide points of reference to interpret the dot map cartograms, capital municipality, provincial or state boundaries were simultaneously transformed, and an inset was added with the undistorted map. We used ScapeToad 1.2 software [11] to reshape the spatial units and the point patterns. From this programme, the transformed layers were exported as shapefiles. The exported files were imported into R statistical software to create the maps. The main R package used to create the maps was *ggplot2*. The R code for the construction of the dot map cartogram is provided elsewhere (https://github.com/lsoetens/DotMapCartogram).

Cartogram quality

As we used spatial units with a population size at a certain spatial level to deform the original map, the MAUP problem was inherited. However, the consequences of this problem can be reduced by relying on an objective measure to assess which spatial level we could best use for our transformations. We assessed the cartogram quality by the objective measures, i.e. the average and maximum normalised cartographic error [10]. The cartographic error is the most commonly used measure for distortion in the value-by-area realisation. We assumed an input map M that is partitioned into n regions with polygonal boundaries. For each region v, $a(v)$ denotes the area of v in M and the weight $w(v)$ is the desired area for the region based on the background population. The diffusion-based algorithm constructed the cartogram M^0, that is a deformation of M, in which $o(v)$ is the observed area of region v. The average normalised cartographic error e was calculated as:

$$e = \frac{1}{n}\sum_{v \in V} \frac{|o(v) - w(v)|}{\max\{o(v), w(v)\}}$$ with a range of [0,1]

and the maximum normalised cartographic error x was calculated as:

$$x = \max\left(\frac{|o(v) - w(v)|}{\max\{o(v), w(v)\}}\right)$$ with a range of [0,1]

No guidelines exist as to what constitutes acceptable cartogram quality; to be consistent with the values measured in a previous publication [10], we aimed for $e < 0.1$ and $x < 0.4$ in our cartograms. The performance of the algorithm, and subsequently the cartogram quality, depends on the distribution of the population density in the input file and the grid size on which the computation of the population density is based. Large discrepancies in the population density between regions impair the algorithm performance, while a finer grid will produce a higher quality cartogram, but may introduce distortion. We started with an input map at the same spatial unit of residence as the case data (i.e. 4-digit postal code for the Netherlands and district level for Germany) and assessed the average and maximum error of the cartograms. For the Netherlands, cartogram error was not satisfactory and we changed to the municipality level and accepted a maximum normalised cartogram error of $x > 0.4$; this was due to the very small population size of the municipalities on the West Frisian Barrier Islands. These areas should be smaller than presented in the created maps, but would no longer be visible. We used a 512 × 512 grid, which gave sufficient results; a higher-resolution grid of 1,024 × 1,024 is possible, but requires substantially greater computation times.

User requirements

To create the dot map cartograms, no specific geographical information system skills are needed. The only requirement is some basic knowledge of the R statistical software to be able to run the provided code (https://github.com/lsoetens/DotMapCartogram). Required software includes R studio and ScapeToad [11]. Both programmes are open source programmes, which can be downloaded for free. A detailed manual for creating the dot map cartograms is provided along with the code (https://github.com/lsoetens/DotMapCartogram).

Results

Geographical representation of point-source outbreaks: Q fever in the Netherlands

The maps of Q fever cases are compared in Figure 3.1. The dot map cartogram (Figure 3.1A) is based on the population size per municipality in the Netherlands in 2009 ($e = 0.08$, $sd = 0.09$; $x = 0.78$). It is compared with the dot map (Figure 3.1B) and the choropleth incidence maps (Figures 3.1C-E).

Ability to show both absolute numbers and incidence

The dot map cartogram is able to show both absolute numbers and incidence. Incidence is readily inferred from the choropleth incidence maps, but the exact size of the outbreak in absolute numbers is not apparent. The dot map showed a clear clustering pattern with several hotspots, but whether this is related to the population density (incidence) cannot be inferred from the dot map.

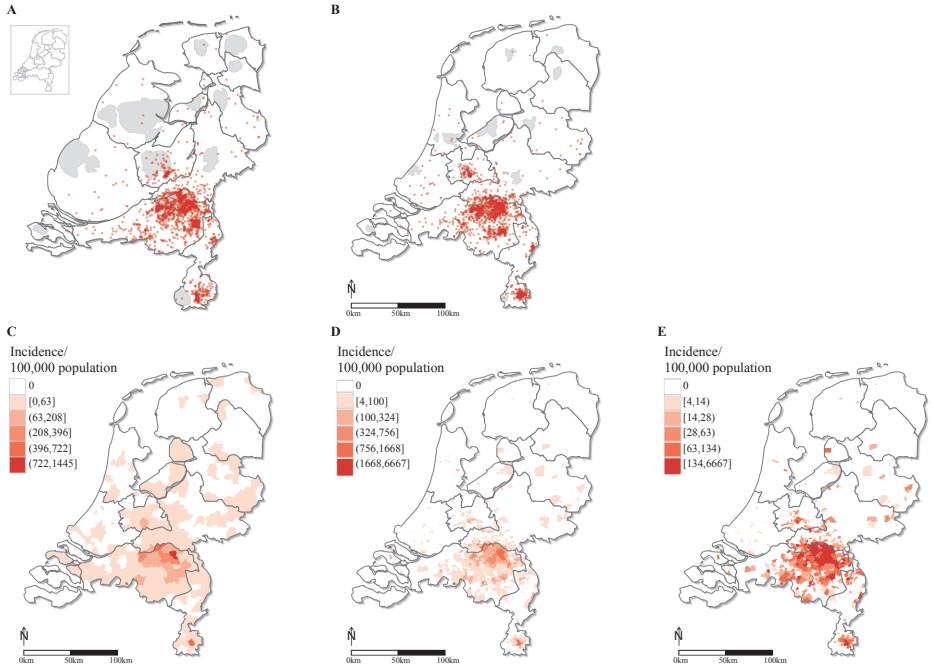


Figure 3.1. Spatial mapping of the Q fever outbreak in the Netherlands, 2009. A. Dot map cartogram. One red dot represents one case, shaded grey areas indicate the national and provincial capital municipalities. The geographical administrative units displayed as underlying layer are distorted as a result of building the cartograms. B. Dot map. One red dot represents one case, shaded grey areas indicate the national and provincial capital municipalities. C. Choropleth incidence map by municipality and Jenks' natural breaks classification system. D. Choropleth incidence map by 4-digit postal code and Jenks' natural breaks classification system. E. Choropleth incidence map by 4-digit postal code and the quantile classification system.

Sensitivity to choices regarding spatial scale and classification system

The dot map cartogram and dot map do not suffer from arbitrary classification issues. The choropleth incidence maps are sensitive to this; different spatial scale and classification systems result in highly variable maps. When only one of those maps is used, the geography of the outbreak could be misinterpreted. One would conclude that the outbreak is more widespread (Figure 3.1C) or more intense (Figure 3.1E) than is illustrated in Figure 3.1D.

Ability to ensure the privacy of cases

The dot map compromises case privacy by revealing exact locations of cases, while the choropleth incidence map protects privacy by aggregating case information. The dot map cartogram meets this criterion partially by deforming the underlying region.

Ability to recognise locations

The dot map cartogram is harder to read, and specific locations are difficult to recognise. This can be improved by providing locations of major towns or district boundaries. The choropleth incidence map and dot map outperform the dot map cartogram on this criterion.

Geographical representation of a human-to-human transmissible disease: pertussis in Germany

The maps for pertussis cases are compared in Figure 3.2. Figure 3.2A depicts the dot map cartogram based on the population size by district in Germany in 2013 ($e = 0.07$, sd = 0.06; $x = 0.29$); this map is contrasted with the corresponding dot map (Figure 3.2B) and choropleth incidence maps (Figures 3.2C-E).

Ability to show both absolute numbers and incidence

A much larger number of cases are illustrated in this example than for the Q fever data. This is immediately apparent from the dot map and the dot map cartogram, but less so from the choropleth incidence map. The clustering patterns on the dot map in several large cities and for example in the Ruhr district (in the west of Germany) disappear into an almost random pattern in the dot map cartogram, indicating that the clustering patterns on the dot map can be explained by the underlying population density in those areas. In contrast, in the south-east of Germany near Munich, a clustering pattern is present in both the dot map and the dot map cartogram, indicating that this cluster is not attributable to population density but reflects a higher incidence of pertussis notifications. Therefore, in this situation, both incidence and absolute number of cases are important in interpreting the map: they reveal the size of the problem and provide clustering of cases not attributable to the population density.

Sensitivity to choices regarding spatial scale and classification system

The choropleth incidence maps showed that the choice of classification system or spatial level results in highly variable maps. In this comparison, Figure 3.2D, based on the Jenks' natural

breaks classification system and the district level, is comparable to the dot map cartogram, showing the same areas with high incidence; however, there is no way to determine a priori which classification system and spatial level would result in the 'right' map.

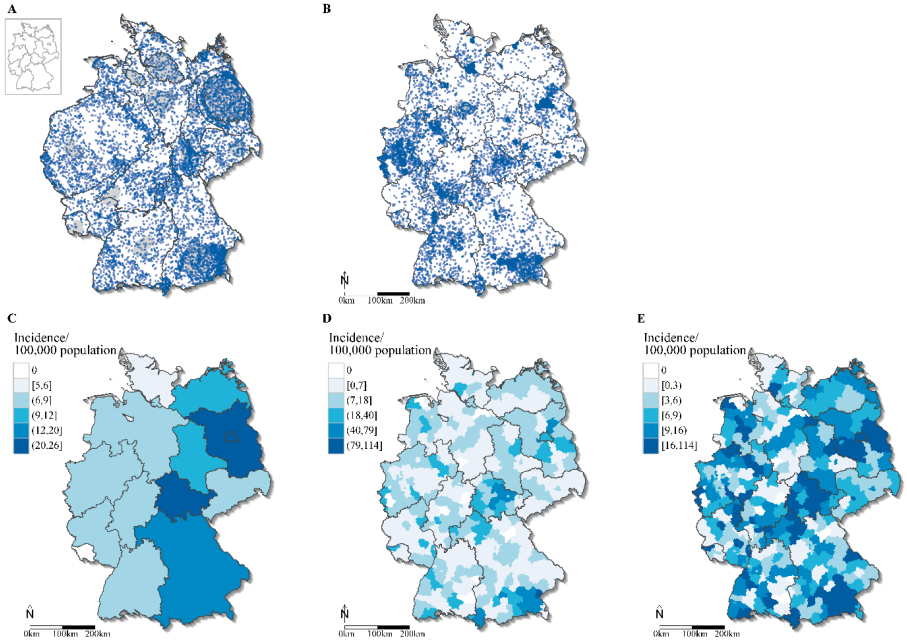


Figure 3.2. Spatial mapping of endemic pertussis in Germany, 2015. A. Dot map cartogram. One blue dot represents one case, shaded grey areas indicate the national and federal capital municipalities. The geographical administrative units displayed as underlying layer are distorted as a result of building the cartograms. B. Dot map. One blue dot represents one case, shaded grey areas indicate the national and federal capital municipalities.C. Choropleth incidence map by federal state and Jenks' natural breaks classification system. D. Choropleth incidence map by district and Jenks' natural breaks classification system. E. Choropleth incidence map by district and the quantile classification system.

Ability to secure the privacy of cases

All maps meet this criterion. In this dataset, the dot map does not compromise privacy because of the frequency of occurrence of the disease. As before, the dot map cartogram protects case privacy due to the deformation of the underlying regions, and the choropleth incidence map because information is aggregated.

Ability to recognise locations

The choropleth incidence map and dot map outperform the dot map cartogram on this criterion. Adding reference points in the dot map cartogram, such as the federal state and federal state municipality boundaries, can help in recognising locations.

Discussion

We have proposed the dot map cartogram for displaying spatial infectious disease data and illustrating both incidence and absolute numbers of cases. A similar technique has been suggested before [12], but to our knowledge it was never used to study clustering of point patterns at a national level in infectious disease epidemiology. We compared the dot map cartogram to the frequently used choropleth incidence map and the dot map [1]. The main advantages of the dot map cartogram over the other two is that it is able to simultaneously reveal epidemic patterns adjusted for the population distribution, and to unmask patterns that are hidden by aggregation and categorisation of information. The visual distortion of the dot map cartogram is a barrier to pinpointing the location of a case: this is a benefit in the field of infectious diseases because case privacy should be ensured in presenting surveillance data. In addition, dot map cartograms illustrate areas without cases, which are harder to discern by the use of choropleth incidence maps.

We did not address map types other than dot maps and incidence maps for this comparison. Smooth incidence maps are an alternative to the choropleth incidence maps, in which the incidence is smoothed across the spatial units. This technique was applied to the data from a previous study for the Q fever outbreak in the Netherlands in 2009 [13]. In illustrating hotspots, the dot map cartogram is comparable to the smooth incidence map. The advantage of the smooth incidence map is that it permits identification of the exact location of hot spots, since the map is not deformed. However, smooth incidence maps do not reveal areas without cases and it is hard to discern the absolute number of cases and the true scope and size of the outbreak. Many more mapping techniques exist, such as map types based on other interpolation techniques (such as inverse distance weighting, kriging and trend surface fitting [2]), incidence based value-by-area cartograms [14,15], and many others. Assessment of all available methods was not within the scope of this paper.

We used the Gastner and Newman diffusion-based algorithm [4] to construct the cartogram. Alternative algorithms are available to construct contiguous cartograms [16-26]. We chose the Gastner and Newman algorithm because it performs well compared to the alternative algorithms in terms of quantitative measures such as the cartographic error, shape error and topology error [10], and qualitative measures such as subjective preferences and task performance [27]. Recent studies have proposed quantitative [10] and qualitative [27] measures for cartogram generation techniques. The development of standardised performance measures will allow objective ranking and selecting of cartogram algorithms.
With the demonstration of dot map cartograms, we provide public health professionals with an alternative spatial method for outbreak analysis. Firstly, we expect that dot map cartograms minimise misinterpretation of the data. Secondly, as demonstrated in this study, dot map cartograms protect the privacy of cases. Thirdly, dot map cartograms do not

require expensive specialised GIS software, which facilitates use in settings with limited resources. The only requirements to construct dot map cartograms are free software, case reports with location data and access to population data as shapefiles. As there is a general trend for governments to make administrative population data publicly available, the latter is available for most countries. Finally, our method can easily be extended with information on covariates relevant for mapping: a relevant example in infectious disease surveillance or research is to colour the case dots for a specific attribute, such as age group, sex or vaccination status. Introducing other layers of detail or attributes further broadens the utility of this spatial method for use in infectious disease research and surveillance. However, as the technique is graphical rather than geographical, it does not replace a geographical information system explaining the impact of various geographical factors on the spread of a certain infectious disease.

References

1. Smith CM, LeComber SC, Fry H, Bull M, Leach S, Hayward AC. Spatial methods for infectious disease outbreak investigations: systematic literature review. Euro Surveill. 2015;20(39):30026.
2. Heywood I, Cornelius S, Carver S. An introduction to geographical information systems. New Jersey: Prentice-Hall; 2011. ISBN: 9780273722595.
3. Openshaw S. Ecological fallacies and the analysis of areal census data. Environ Plann A. 1984;16(1):17-31. https://doi.org/10.1068/a160017 PMID: 12265900
4. Gastner MT, Newman ME. From The Cover: Diffusion-based method for producing density-equalizing maps. Proc Natl Acad Sci USA. 2004;101(20):7499-504. https://doi.org/10.1073/pnas.0400280101 PMID: 15136719
5. Wijk- en buurtkaart 2008. [District and neighborhood maps 2008]. Den Haag: Statistics Netherlands. [Accessed: 20 Jan 2016]. Dutch. Available from: http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2009/2010-wijk-en-buurtkaart-2008.htm
6. SurvStat@RKI 2.0. Berlin: Robert Koch Institute. [Accessed: 3 Feb 2016]. Available from: https://survstat.rki.de/
7. Open Data - Freie Daten und Dienste des BKG. [Open data – free data and services of the BKG]. Leipzig: Federal Agency for Cartography and Geodesy (BKG). [Accessed: 15 Nov 2015]. German. Available from: http://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div?gdz_spr=deu&gdz_akt_zeile=5&gdz_anz_zeile=0&gdz_user_id=0
8. Brewer CA, Pickle L. Evaluation of methods for classifying epidemiological data on choropleth maps in series. Ann Assoc Am Geogr. 2002;92(4):662-81. https://doi.org/10.1111/1467-8306.00310
9. Harrower M, Brewer CA. ColorBrewer.org: An online tool for selecting colour schemes for maps. Cartogr J. 2003;40(1):27-37. https://doi.org/10.1179/000870403235002042
10. Alam MJ, Kobourov SG, Veeramoni S. Quantitative Measures for Cartogram Generation Techniques. Comput Graph Forum. 2015;34(3):351-60. https://doi.org/10.1111/cgf.12647
11. Andrieu D, Kaiser C, Ourednik A. ScapeToad: not just one metric. [Accessed: 20 Sep 2015]. Lausanne: Choros Laboratory, Available from: http://scapetoad.choros.ch/
12. Khalakdina A, Selvin S, Merrill DW, Erdmann CA, Colford JM Jr. Analysis of the spatial distribution of cryptosporidiosis in AIDS patients in San Francisco using density equalizing map projections (DEMP). Int J Hyg Environ Health. 2003;206(6):553-61. https://doi.org/10.1078/1438-4639-00245 PMID: 14626902
13. van der Hoek W, van de Kassteele J, Bom B, de Bruin A, Dijkstra F, Schimmer B, et al. Smooth incidence maps give valuable insight into Q fever outbreaks in The Netherlands. Geospat Health. 2012;7(1):127-34. https://doi.org/10.4081/gh.2012.111 PMID: 23242690
14. Dodd PJ, Sismanidis C, Seddon JA. Global burden of drug-resistant tuberculosis in children: a mathematical modelling study. Lancet Infect Dis. 2016;16(10):1193-201. https://doi.org/10.1016/S1473-3099(16)30132-3 PMID: 27342768
15. Dorling D, Barford A, Newman M. WORLDMAPPER: the world as you've never seen it before. IEEE Trans Vis Comput Graph. 2006;12(5):757-64. https://doi.org/10.1109/TVCG.2006.202 PMID: 17080797
16. Tobler WR. Democratic representation and apportionment: a continous transformation useful for districting. Ann N Y Acad Sci. 1973;219:215-20. https://doi.org/10.1111/j.1749-6632.1973.tb41401.x PMID: 4518429
17. Dougenik JA, Chrisman NR, Niemeyer DR. An algorithm to construct continuous area cartograms. Prof Geogr. 1985;37(1):75-81. https://doi.org/10.1111/j.0033-0124.1985.00075.x
18. Gusein-Zade SM, Tikunov VS. A new technique for constructing continuous cartograms. Cartogr Geogr Inf Syst. 1993;20(3):167-73. https://doi.org/10.1559/152304093782637424
19. Keim DA, North SC, Panse C. CartoDraw: a fast algorithm for generating contiguous cartograms. IEEE Trans Vis Comput Graph. 2004;10(1):95-110. https://doi.org/10.1109/TVCG.2004.1260761 PMID: 15382701
20. Keim DA, Panse C, North SC. Medial-axis-based cartograms. IEEE Comput Graph Appl. 2005;25(3):60-8. https://doi.org/10.1109/MCG.2005.64 PMID: 15943089
21. Kämper JH, Kobourov SG, Nöllenburg M. Circular-arc cartograms. 2013 IEEE Pacific Visualization Symposium (PacificVis). Sydney; 2013. http://dx.doi.org/10.1109/PacificVis.2013.6596121
22. Dorling D. Area cartograms: their use and creation. In: The map reader: theories of mapping practice and cartographic representation. Dodge M, Kitchin R, Perkins C, eds. Chichester: John Wiley & Sons; 2011. p. 252-60. http://dx.doi.org/10.1002/9780470979587.ch33
23. Edelsbrunner H, Waupotitsch R. A combinatorial approach to cartograms. J. Comput. Geom.1997;7(5-6):343-60.
24. House DH, Kocmoud CJ. Continuous cartogram construction. Proceedings of the IEEE Visualization. Research

Triangle Park; 18-23 Oct 1998. p. 197-204. https://dx.doi.org/10.1109/VISUAL.1998.745303
25. Schulman J, Selvin S, Merrill DW. Density equalized map projections: a method for analysing clustering around a fixed point. Stat Med. 1988;7(4):491-505. https://doi.org/10.1002/sim.4780070406 PMID: 3368676
26. Selvin S, Merrill D, Schulman J, Sacks S, Bedell L, Wong L. Transformations of maps to investigate clusters of disease. Soc Sci Med. 1988;26(2):215-21.
27. Nusrat S, Alam MJ, Kobourov SG. Evaluating cartogram effectiveness. IEEE Trans Vis Comput Graph. 2016. https://dx.doi.org/10.1109/TVCG.2016.2642109



4

Multi-scale analysis of infectious disease surveillance data

The contents of this chapter have been submitted for publication:
Multi-scale analysis of infectious disease surveillance data
Loes Soetens, Susan Hahné, Jacco Wallinga

Abstract

Surveillance systems collect information on many infections over long periods. To allow interpretation, these time series are summarized. Here we aim to identify the most informative summary statistics that capture dynamic patterns across scales in a multivariate time series of daily notifications for 45 infectious diseases from 2003 to 2018 in the Netherlands.

To put different flavors of summary statistics on the same footing, we calculate them as instances of the generalized Rényi entropy H_α, with H_0 corresponding to presence/absence, H_1 to epidemic intensity (Shannon entropy), and H_2 to the correlation sum. We then rank the order in which these statistics capture the dynamic variation across scales using Principal Component Analysis.

The Shannon entropy (H_1) is the most informative measure; the difference between presence/absence and correlation sum (H_0-H_2) is an informative orthogonal measure. We find a group of infections characterized by outbreaks, a group of infections continuously reported, and a group of infections intermittently reported. A few infections displayed a changing dynamic pattern: diphtheria and leptospirosis moved towards outbreaks, mumps and hantavirus moved away from outbreaks, and West Nile virus infection and yellow fever moved towards continuous reporting.

Our method for identifying changes in dynamic patterns of infectious diseases discerns disease groups sharing similar dynamic properties. By sorting diseases on recent change in dynamic properties, an informative overview is given of ongoing changes that might require enhanced disease control.

Introduction

Due to longstanding surveillance, databases containing case-based information on infectious disease notifications have expanded over the years. Epidemiologists use this information for outbreak detection, monitoring endemic disease trends, and the evaluation of intervention programs [1]. To achieve these aims, changes in dynamic patterns need to be identified; a change could be related to, for example, a starting outbreak, an increasing trend, or the introduction of an intervention program. To identify these changes, the number of disease notifications over time needs to be studied. This is often done separately for every disease, however, by doing so important relations between diseases dynamics might be missed. To allow for comparison of disease dynamics, they can be classified according to their dynamic properties. Dynamic properties can include the presence/absence of disease, and variation and autocorrelation in disease occurrence. When grouped, changes in these dynamic patterns might be more easily identified, enhancing infectious disease surveillance. The aim of this study is to provide a method for the identification of changes in dynamic patterns in multivariate infectious disease time series. To this end, we study the dynamic patterns of infectious diseases by using the complete database of the notifiable disease surveillance in the Netherlands, comprising 49 infectious diseases reported daily over sixteen years.

We use exploratory data visualisation to get an overview of large amounts of data [2]. So what measure can be used to visualize disease dynamics? The number of notifications can be visualized by plotting time series of incidence [3, 4]. However, the representation of such a time series is heavily dependent on the choice of the time aggregation scale. Other measures of disease dynamics discussed in literature concern measures to quantify the presence or absence of disease in a population [5-11]. These are often used to measure persistence of a pathogen in the population by counting either the number of or the total duration of extinctions in the population. The dynamics of measles have frequently served as a case study to investigate these patterns [6-9]. Another measure of disease dynamics is Shannon entropy [12]. Applied to infectious disease time series it can serve various purposes: as an early warning measure [13], to quantify epidemic intensity [14] or to forecast infectious disease incidence [15]. A final group of measures for capturing disease dynamics are measures of (auto)correlation, such as obtained with wavelet or spectral analysis. The power of these methods lies in their ability to describe how patterns change across aggregation scales [16]. This category has probably been the most frequently applied in the infectious disease domain [17-20].

All these methods often only cover one aspect of disease dynamics. To capture multiple aspects of disease dynamics, visualizing disease dynamics requires an easy to understand measure that is robust to aggregation scale and that can be applied to a large number of

infectious diseases. Here, we propose Rényi entropy as this measure [21]. Rényi entropy is a family of entropy measures that can be used to quantify diversity of phenomena, a so-called diversity index [22]. It is a generalisation of the earlier mentioned Shannon entropy. We chose this entropy family for various reasons. It is very straightforward to calculate; it only uses relative frequencies of notifications occurring at a specific timescale. We can therefore easily estimate the entropies along a continuum of aggregation levels. This has two advantages; we can examine how patterns change across these levels and separate members of the family all quantify another important aspect of time dynamics that may be related to measures mentioned in the previous paragraph, such as the presence/absence of notifications, uncertainty around the mean (Shannon entropy), and (auto)correlation. By calculating this single index, we can therefore capture multiple dimensions of disease dynamics.

We apply our approach to a large database with surveillance data of notifiable diseases in the Netherlands from 2003 to 2018, comprising 49 infectious diseases, ranging from frequently reported diseases, such as pertussis, to rare diseases, such as botulism. We visualize time series of all 49 diseases included in the database, and group and sort these time series according to their similarity. We study changes in the classification of diseases over time.

Methods

Data

We use all infectious disease notifications in the Netherlands between 2003-2018. The date of onset is used in our analysis; when this is missing we used date of laboratory confirmation; and if that was missing we used the date of diagnosis. We stratify the hepatitis B and C notifications into chronic and acute infections, and meningococcal infections into serogroups B, C, W and Y. If the hepatitis infection type (i.e., acute vs. chronic) or meningococcal serogroup is unknown or classified as 'other', the relevant cases are excluded. Cases with chronic hepatitis C infection are excluded as this infection was no longer notifiable from 2003. Cases with influenza A(H1N1)2009 infection are excluded as this infection was only notifiable for a short period of time (2009-2010). Dengue or chikungunya virus infections are excluded as these are only notifiable in the Dutch Caribbean areas. Only case-based notifications are included because a date of notification is needed for every record; this means exclusion of notifications of food-borne and MRSA clusters. Not all diseases were notifiable from the beginning of the study period (in 2003). For mumps, Hantavirus infection, Haemophilus influenzae type b (Hib) disease, pneumococcal disease, listeriosis and tuberculosis, cases are included from 2009 onwards; for group A streptococcal (GAS) disease cases are included from 2011 onwards; and for tularemia and zika virus infection, cases are included from 2017 onwards.

From time series to discrete frequency distributions

We consider time series (X) of notification counts. These time series can be aggregated at a level i (such as days, weeks, months or years). We can then calculate the relative frequencies p_i of notifications in the ith time window, transforming our time series in discrete frequency distributions. For example, if the time series is aggregated by week and in week 4 there are 10 notifications and the total number of notifications is 100, the relative frequency in week 4 is p_4= 10/100= 0.1. We transform the time series to discrete frequency distributions for every disease, at various aggregation levels (as discussed below). The Rényi entropy family can then be described by the following formula [21]:

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^\alpha\right)$$

where n represents the total number of time windows in the time series. The order α is indicating the family member and can vary from zero to infinity.

H(X) based on each level of α can be interpreted in its own right as a measure corresponding to a real phenomenon. In this study, we use particular instances or family members of the Rényi entropy:

- α = 0 is quantifying the presence/absence of notifications. If all time windows

contain notifications this is equal to the maximum entropy (H_{max}), which is defined as the logarithm of the time series length.

- $\alpha \to 1$, better known as Shannon entropy, is quantifying the fluctuations in notifications. When there are no fluctuations in the time series, i.e. a similar amount of notifications in every time window, this measure equals 1. When there is only one time window with notifications, this measure equals zero.
- $\alpha = 2$ captures the probability that two notifications, taken at random from the time series, come from the same time window of length τ. Specifically, H_2 is the integral over the (auto)correlation function for the time series, taken from 0 to τ. As such, it quantifies the autocorrelation in the occurrence of notifications. Again, for time series with a constant number of notifications throughout the series this measure will equal the maximum entropy.
- $\alpha \to \infty$, the entropy then equals the maximum p_i value in the time series, the proportional frequency of the most frequent time window. This is also called the minimum entropy.

The values of the Rényi entropy of order α, H_α, are ordered as $H_{max} \geq H_0 \geq H_1 \geq H_2 \geq H_{min}$ [21].

Analysis

Overview of the time series

To visualize the time series of diseases, we plot the aggregated number of notifications as a heat map, where rows represent diseases, and columns represent time windows. Diseases are ordered alphabetically. A color scheme indicates the number of notifications per time window. We map this scheme on a log scale as there is large variation in the incidence. Time windows without cases (log(0)) are left blank. We construct a heat map per week and per year, to show differences in patterns when aggregation scales vary. Second, we calculate the Rényi entropies of order 0, 1 and 2 and the maximum and minimum entropy based on the time series of every disease for various aggregation levels over the complete inclusion range. We aggregate the time series by the following time window sizes: day, week, month, 3 months, 6 months, 9 months, year, 1.5 years, 2 years, 3 years and 5 years. We plot the results in a multi panel plot, in which the values of the entropies are depicted against the natural logarithm of the aggregation level (in days) for every infectious disease.

Grouping of infectious diseases

To identify groups of infectious diseases with similar dynamic patterns, we perform principal component analysis on the Rényi entropies of order 0, 1 and 2, on the complete time series of all diseases at various aggregation levels. As principal component analysis is applied to only three entropy measures, we are able to interpret these components by examining the contributions of the individual entropies to the principal components (PC) across the various aggregation levels. This is unusual in principal component analysis, which often involves

large numbers of variables, making the interpretation of the components impossible. The change of the contributions across aggregation levels will then give us an assessment of the robustness of the entropy values across aggregation levels. The results of the principal component analysis are plotted in a scatterplot per aggregation level to assess which diseases are closest to each other and therefore share similar time dynamics.

Change in dynamic patterns over time

To examine the change in dynamic patterns for the various diseases, entropies of order 0, 1 and 2 are calculated per disease over three-year moving windows, shifting by year. The aggregation level is based on the previous analyses. We use three-year periods to be able to take into account multiyear outbreaks and seasonal effects. The results are plotted in a graph with the interpretation of the PC in terms of entropies on the x- and y-axis and a path through time per disease, showing the change in dynamic patterns over time. Diseases with considerable dynamic change in their patterns will therefore show much variation and hence a long path, while for relatively stable diseases the path will be much shorter. Finally, we compare the dynamic change in the last two time periods (t_1= 2015-2017 versus t_2= 2016-2018) between diseases, to assess most recent dynamics. We define dynamic change as the Euclidian distance in entropies between (x_{t1}, y_{t1}) and (x_{t2}, y_{t2}) (the length of the path between these time periods). We sort diseases in the heatmap of disease incidence according to dynamic change in four different panels: one for disease incidence, one showing the decreasing dynamic change, and two with the x- and y- components (the interpretation of the PC in terms of entropies) to help with the interpretation of the dynamic change. In Appendix B, we repeat this analysis for two different time periods.

Results

In the period 2003 – 2018, 210,715 infectious disease cases were notified in the Netherlands. 196,779 cases were included, distributed over 45 infectious diseases (Table 4.1). Pertussis cases represent the vast majority (56.6%) of included cases.

In Figure 4.1a the time series of weekly notifications per infectious disease are shown on a log10 color scale. We observe several diseases with large outbreaks in the past 15 years, such as measles, rubella and Q fever; diseases with a stable number of cases, such as pertussis and chronic hepatitis B; diseases with a more irregular pattern, such as hepatitis A; diseases with a strong seasonal component, such as legionellosis; and very rare diseases, such as trichinosis, rabies and botulism. In Figure 4.1b the same time series are aggregated per year. Here, seasonal patterns are lost.

For every infectious disease we have calculated Rényi entropies of order 0, 1 and 2 and the maximum and minimum entropy based on time series at various aggregation levels (Figure 4.2). At a daily aggregation level, a wide gap exists between the curves of H_{max} and H_0 for diseases such as brucellosis (BRUC), rubella (RUBE) and meningococcal C, W and Y disease (MCOCC, MCOCW, MCOCY). This means that these diseases have many days without reported cases. At a 5-year aggregation level, we can only observe a small gap between these curves for very rare diseases, such as rabies (RABI) and anthrax (ANTH). We can also observe diseases, such as pertussis (PERT), tuberculosis (TBC), chronic hepatitis B (HEPBCH) and shigella (SHIG), for which all orders of the Rényi entropies are close to the H_{max} across all aggregation levels. These are diseases with a continuous reporting of cases (H_0), with little fluctuations around the mean number of reported cases (H_1) and low autocorrelation in the number of reported cases (H_2). Finally, we have a group of diseases, such as measles (MEAS), Q fever (QFEV), rubella (RUBE) and meningococcal W disease (MCOCW), which show a relatively large gap between the curves of order 1 and 2 and the curve of order 0, especially at monthly and yearly aggregation levels. These are diseases for which if cases are reported, large fluctuations around the mean number of reported cases (H_1) and a large variation in the number of reported cases throughout the time series are shown. We defined this as a typical pattern for outbreak diseases.

Table 4.1. Overview of notifiable disease cases 2003 – 2018 included in this study

Disease	Disease short	Start year	Number of notifications
Anthrax	ANTH	2003	2
Botulism	BOTU	2003	21
Brucellosis	BRUC	2003	81
Cholera	CHOL	2003	39
Creutzfeldt-Jakob disease	CJD	2003	345
Variant Creutzfeldt-Jakob disease	CJDV	2003	3
Diphtheria	DIPH	2003	16
Hantavirus infection	HANTA	2009	239
Hepatitis A	HEPA	2003	3,717
Hepatitis B acute	HEPBAC	2003	3,489
Hepatitis B chronic	HEPBCHR	2003	21,679
Hepatitis C acute	HEPCAC	2003	846
Invasive group A streptococcal disease	IGAS	2011	1,610
Invasive Haemophilus influenzae type b infection	IHIB	2009	253
Human infection with zoonotic influenza virus	INFL	2009	1
Invasive pneumococcal disease (in children 5 years or younger)	IPNEU	2009	456
Legionellosis	LEGI	2003	6,875
Leptospirosis	LEPT	2003	843
Listeriosis	LIST	2009	1,152
Malaria	MALA	2003	4,360
Meningococcal disease, serogroup B	MCOCB	2003	2,233
Meningococcal disease, serogroup C	MCOCC	2003	158
Meningococcal disease, serogroup W	MCOCW	2003	276
Meningococcal disease, serogroup Y	MCOCY	2003	168
Measles	MEAS	2003	3,246
Mumps	MUMPS	2009	2,181
Paratyphoid A fever	PARA	2003	263
Paratyphoid B fever	PARB	2003	351
Paratyphoid C fever	PARC	2003	24
Pertussis	PERT	2003	111,360
Psittacosis	PSIT	2003	1,018
Q fever	QFEV	2003	4,704
Rabies	RABI	2003	4
Rubella	RUBE	2003	595
Shigella	SHIG	2003	7,795
STEC/enterohemorragic E.coli infection	STEC	2003	6,605
Tuberculosis	TBC	2009	9,298
Tetanus	TETA	2009	15
Trichinosis	TRIC	2003	3
Tularemia	TUL	2017	5
Typhoid fever	TYPH	2003	427
Viral hemorrhagic fever	VHF	2003	2
West Nile virus infection	WNV	2009	5
Yellow fever	YFV	2003	3
Zika virus infection	ZIKA	2017	13
Total			**196,779**

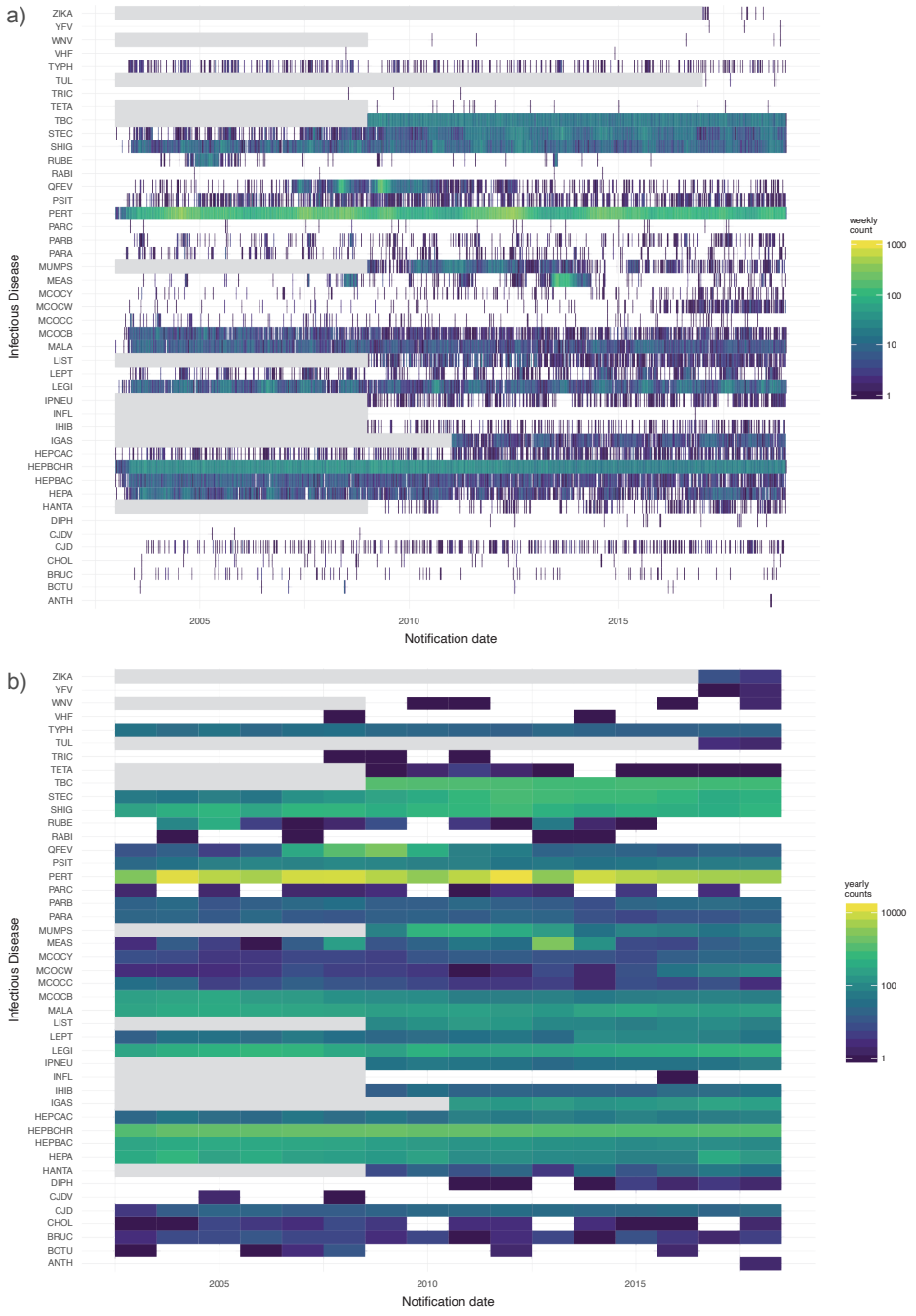


Figure 4.1. A first visualization of the notification data for 45 infectious diseases in the Netherlands, 2003 – 2018, aggregated by week (a) and by year (b). Grey areas indicate periods in which diseases were not yet notifiable.

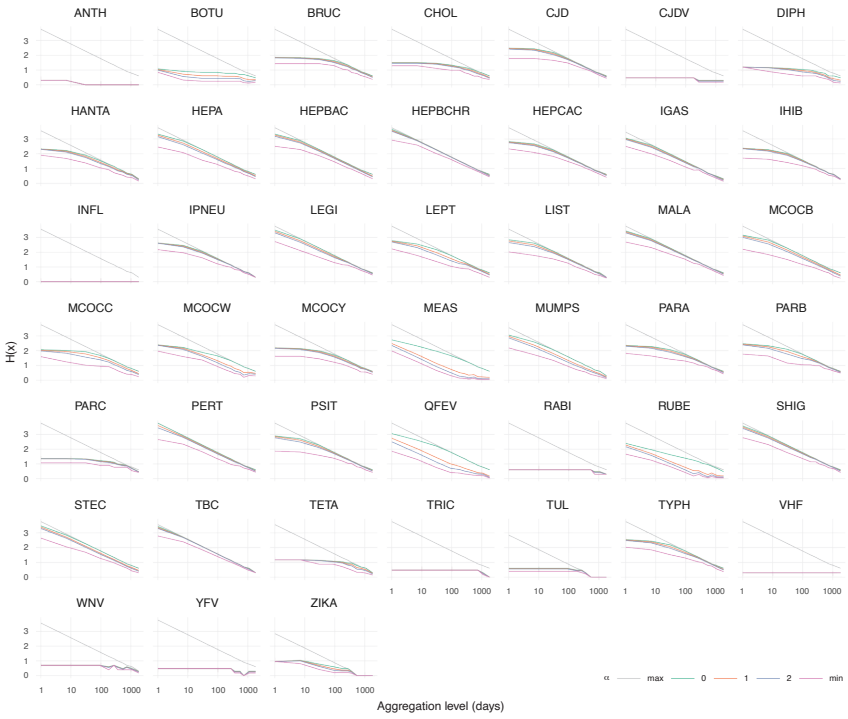


Figure 4.2. Rényi entropy for different orders of α against the aggregation level in days of the time series of notifications of 45 infectious diseases in the Netherlands, 2003-2018.

Principal Component Analysis reveals that dynamic patterns of infectious diseases can be characterized by just two variables, of which the most important variable is the intensity of fluctuations (H_1), and the lesser important one the sensitivity of this intensity to a change in order α (H_0-H_2) (Appendix A). Together, these two variables explain almost 100% of the variance in H_0, H_1 and H_2 of the diseases. Three groups of diseases are identified based on these properties: those characterized by outbreaks (high H_0-H_2, low H_1), those which are continuously reported (low H_0-H_2, high H_1), and those which are intermittently reported (low H_0-H_2, low H_1).

To examine the change in dynamics over time, entropies are calculated for different periods (Figure 4.3). Of interest are diseases that were formerly rare but over time are more constantly reported (shifting from bottom left to bottom right corner in Figure 4.3). Examples of such diseases are West Nile virus infection (WNV) or yellow fever (YFV). Also, diseases moving in the other direction on the same axis can be of interest: from being continuously reported to more fluctuations in their time series. This always goes together with movement along the other axis and indicates a group of diseases that show more fluctuations, less autocorrelation and few intervals with no reported cases in their time

series over the years. These patterns might possibly indicate (starting) outbreaks (diseases that shift from the bottom upwards in Figure 4.3). Examples of such diseases are hepatitis A (HEPA), leptospirosis (LEPT), and diphtheria (DIPH). A final group of diseases are those with a large range in Figure 3. These diseases show large fluctuations over time, and are known for large past outbreaks. Examples are rubella (RUBE), measles (MEAS) and Q fever (QFEV). Diseases with constant dynamics over time include acute and chronic hepatitis B infection (HEPBAC and HEPBCHR), tuberculosis (TBC), psittacosis (PSIT) and meningococcal B infection (MCOCB).

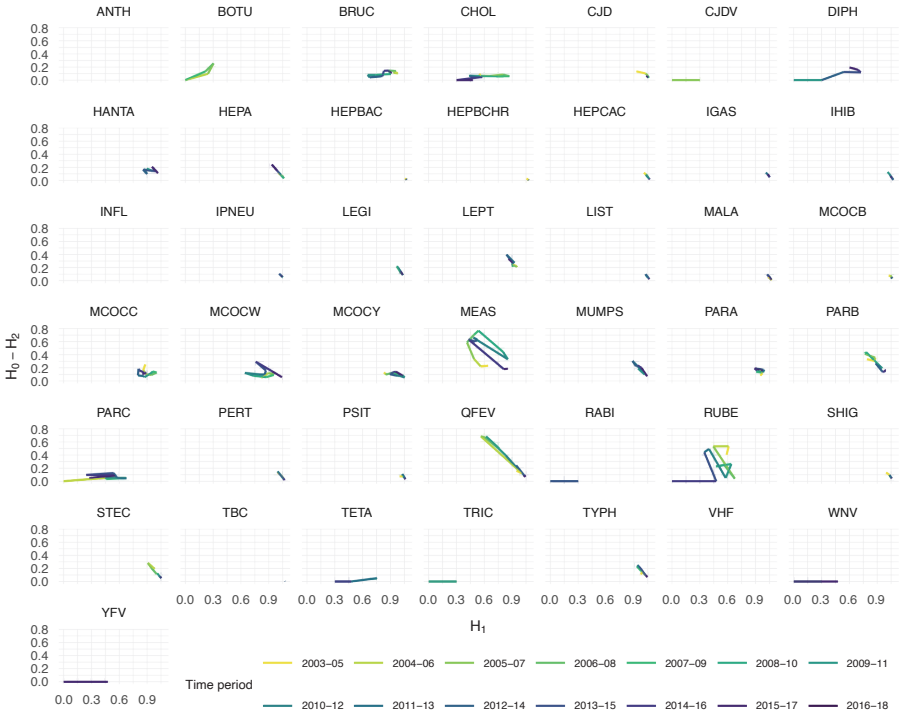


Figure 4.3. Change in entropy dynamics over time per infectious disease. A rolling window of three years, shifting per year, was applied on quarterly aggregated time series.

To get an overview of recent changes in dynamic patterns, we zoom in on the last two time periods presented in Figure 4.3 (purple). We can compare the change in dynamics between 2015-2017 and 2016-2018 by computing dynamic distance (the length of the path in Figure 4.3) between these periods and sort the heat map according to this distance (Figure 4.4). The largest change in dynamics between these periods was for West Nile virus infection (WNV) and yellow fever virus infection (YFV). This is due to an increase in H_1 entropy (see right panels of Figure 4.4). Previously, we observed that diseases with a decreasing H_1 and increasing H_0-H_2 difference possibly indicate starting outbreaks. We observe this pattern also in diphtheria (DIPH), Creutzfeldt Jakob disease (CJD) and leptospirosis

(LEPT). We observe an opposite pattern, a decrease in outbreak potential (decrease in H_0-H_2 distance, increase in H_1 distance), for mumps, meningococcal W disease (MCOCW), hantavirus infection (HANTA), hepatitis A (HEPA) and meningococcal Y disease (MCOCY) (amongst others). For most diseases, little change in dynamics has taken place between the periods 2015-2017 and 2016-2018. This is completely different when we compare two other periods, 2010-2012 and 2011-2013 (Figure B1, Appendix B). Here we can see large changes in dynamics for measles (MEAS) and rubella (RUBE), due to an increase in outbreak potential (H_0-H_2 distance) and a decrease in H_1 entropy, indicating more fluctuating time series.



Figure 4.4. Change in disease dynamics according to entropy measures between period t_1 (2015-2017) and t_2 (2016-2018) across notifiable infectious diseases in the Netherlands. The quarterly aggregated time series of these periods shown as a heat map (first panel) are sorted by decreasing dynamic distance between these periods (second panel).The third and fourth panel show the distance in H_0-H_2 and distance in H_1 between t_1 and t_2 for all diseases. Light grey areas indicate diseases for which no data was reported as the disease was not yet notifiable (first panel) or that the entropies could not be estimated due to no reported cases in one or both periods (second, third and fourth panel).

Discussion

We captured the intrinsic dynamic properties of 45 notifiable infectious diseases between 2003 and 2018 in the Netherlands, using a measure based on the relative incidence of infectious diseases. We have used Rényi entropy, a measure for which the outcome is consistent across timescales, to classify and order diseases by their dynamic properties.

Principal Component Analysis identified the most important quantity describing variation in dynamic patterns of diseases: H_1 (Shannon entropy), which gives a measure of intensity of fluctuations in the time series, and H_0-H_2, describing the difference in the entropy values, indicating the sensitivity of the intensity in fluctuations to a change in order.

Our aim was to provide a method for the identification of changes in dynamic patterns in multivariate infectious disease time series. Recent changes requesting further analysis by infectious disease epidemiologists include the more continuously reporting of West Nile virus and the more fluctuating reporting rate of diphtheria notifications in the Netherlands in 2018. In 2018, the incidence of West Nile virus infection was relatively high in South European countries [23-25], causing more imported cases in the Netherlands. Diphtheria was acquired abroad. Traveling or immigration from high incidence countries, such as Venezuela [26] (largest neighbouring country of the Kingdom of the Netherlands) or Indonesia [27] (former Dutch colony), might explain the more fluctuating pattern for this disease.

In this study, we have shown that Rényi entropy can be calculated at multiple aggregation scales and that the ranking of diseases on entropy measures is quite similar across different aggregation levels. The analysis does therefore not require a choice on the timescale of analysis, which is a major advantage in the case that no aggregation level is a priori of interest. Another advantage of this measure is that it is very easy to calculate, giving a low computational burden even for large datasets. As it relies on relative frequencies, diseases with low numbers can also be analysed. A change in the dynamics in the notifications of diphtheria in 2018 could therefore be observed, an issue that would otherwise probably have gone unnoticed.

So how can the visualisation of Rényi entropy be used in daily practice? We would advise to use this alongside regular surveillance activities, to check on a regular basis whether any important changes have occurred in disease dynamics. Hidden intrinsic patterns, that currently go unnoticed in regular surveillance activities (for example with rare diseases) can then be revealed and further investigated. Entropy might also be useful as a measure for early warning of outbreaks as suggested by Brett et al [13]. Our research has additionally demonstrated that change in patterns can be detected, but whether this can be done in a timely manner, as is requested for early warning, needs further research.

This method can be applied to any infectious disease surveillance data, as many developed countries have similar surveillance systems yielding similar data [3, 19, 28, 29], it would also be interesting to compare grouping of diseases according to their dynamic properties across countries. If similar patterns exist, new hypotheses on disease dynamics might be generated and typical patters for diseases might be revealed, as in Graham et al [30]. Finally, the analysis can also take into account age, sex, specific risk groups, or geography, as was done by Dalziel et al [14], to reveal and compare dynamic patterns between these subgroups.

To conclude, we have introduced a multi-scale measure for representing the dynamic properties of infectious diseases. By applying this measure on notifications of 45 infectious diseases in the Netherlands from 2003 -2018, we were able to discern disease groups sharing similar dynamic properties. In addition, by sorting diseases by the amount of change in dynamic properties, an informative overview is given of ongoing changes that might require the attention of epidemiologists and allow for the generation of hypotheses on disease dynamics.

References

1. Giesecke J. Modern Infectious Disease Epidemiology, Third Edition: Taylor & Francis; 2016.
2. Tukey JW. Exploratory Data Analysis: Addison-Wesley Publishing Company; 1977.
3. van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. The New England journal of medicine. 2013;369(22):2152-8.
4. Gibney KB, Cheng AC, Hall R, Leder K. An overview of the epidemiology of notifiable infectious diseases in Australia, 1991-2011. Epidemiology and infection. 2016;144(15):3263-77.
5. Bjornstad ON, Grenfell BT. Noisy clockwork: time series analysis of population fluctuations in animals. Science (New York, NY). 2001;293(5530):638-43.
6. Keeling MJ, Grenfell BT. Understanding the persistence of measles: reconciling theory, simulation and observation. Proceedings Biological sciences. 2002;269(1489):335-43.
7. Bartlett MS. Measles Periodicity and Community Size. Journal of the Royal Statistical Society Series A (General). 1957;120(1):48-70.
8. Grenfell BT, Bjørnstad ON, Finkenstädt BF. DYNAMICS OF MEASLES EPIDEMICS: SCALING NOISE, DETERMINISM, AND PREDICTABILITY WITH THE TSIR MODEL. 2002;72(2):185-202.
9. Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model. Ecological Monographs. 2002;72(2):169-84.
10. Nieddu GT, Billings L, Kaufman JH, Forgoston E, Bianco S. Extinction pathways and outbreak vulnerability in a stochastic Ebola model. Journal of the Royal Society, Interface. 2017;14(127).
11. Vredenburg VT, Knapp RA, Tunstall TS, Briggs CJ. Dynamics of an emerging disease drive large-scale amphibian population extinctions. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(21):9689-94.
12. Shannon CE. A Mathematical Theory of Communication. 1948;27(3):379-423.
13. Brett TS, Drake JM, Rohani P. Anticipating the emergence of infectious diseases. Journal of the Royal Society, Interface. 2017;14(132).
14. Dalziel BD, Kissler S, Gog JR, Viboud C, Bjornstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science (New York, NY). 2018;362(6410):75-9.
15. Scarpino SV, Petri G. On the predictability of infectious disease outbreaks. Nature communications. 2019;10(1):898.
16. Levin SA. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. 1992;73(6):1943-67.
17. Holdsworth AM, Kevlahan NK, Earn DJ. Multifractal signatures of infectious diseases. Journal of the Royal Society, Interface. 2012;9(74):2167-80.
18. van Wijhe M, Tulen AD, Korthals Altes H, McDonald SA, de Melker HE, Postma MJ, et al. Quantifying the impact of mass vaccination programmes on notified cases in the Netherlands. Epidemiology and infection. 2018;146(6):716-22.
19. Grenfell BT, Bjørnstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature. 2001;414(6865):716-23.
20. Soetens LC, Boshuizen HC, Korthals Altes H. Contribution of seasonality in transmission of Mycobacterium tuberculosis to seasonality in tuberculosis disease: a simulation study. American journal of epidemiology. 2013;178(8):1281-8.
21. Renyi A, editor On Measures of Entropy and Information. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics; 1961 1961; Berkeley, Calif.: University of California Press.
22. Tóthmérész B. Comparison of different methods for diversity ordering. 1995;6(2):283-90.
23. Burki T. Increase of West Nile virus cases in Europe for 2018. The Lancet. 2018;392(10152):1000.
24. Holt E. West Nile virus spreads in Europe. The Lancet Infectious Diseases. 2018;18(11):1184.
25. Haussig JM, Young JJ, Gossner CM, Mezei E, Bella A, Sirbu A, et al. Early start of the West Nile fever transmission season 2018 in Europe. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2018;23(32).
26. Paniz-Mondolfi AE, Tami A, Grillet ME, Márquez M, Hernández-Villena J, Escalona-Rodríguez MA, et al. Resurgence of Vaccine-Preventable Diseases in Venezuela as a Regional Public Health Threat in the Americas. Emerging infectious diseases. 2019;25(4):625-32.

27. Tosepu R, Gunawan J, Effendy DS, Ahmad LOAI, Farzan A. The outbreak of diphtheria in Indonesia. Pan Afr Med J. 2018;31:249-.
28. Enki DG, Garthwaite PH, Farrington CP, Noufaily A, Andrews NJ, Charlett A. Comparison of Statistical Algorithms for the Detection of Infectious Disease Outbreaks in Large Multiple Surveillance Systems. PloS one. 2016;11(8):e0160759.
29. Enki DG, Noufaily A, Garthwaite PH, Andrews NJ, Charlett A, Lane C, et al. Automated biosurveillance data from England and Wales, 1991-2011. Emerging infectious diseases. 2013;19(1):35-42.
30. Graham M, Winter AK, Ferrari M, Grenfell B, Moss WJ, Azman AS, et al. Measles and the canonical path to elimination. 2019;364(6440):584-7.

Appendix A

PRINCIPAL COMPONENT ANALYSIS RESULTS ACROSS AGGREGATION LEVELS

To identify groups of infectious diseases with similar dynamic patterns, we perform principal component analysis on the Rényi entropies of order 0, 1 and 2, based on the complete time series of all diseases at various aggregation levels. As principal component analysis is applied to only three entropy measures, we are able to interpret these components by examining the contributions of the individual entropies to the principal components (PC) across the various aggregation levels. This is unusual in principal component analysis, which often involves large numbers of variables, making the interpretation of the components impossible. The change of the contributions across aggregation levels will then give us an assessment of the robustness of the entropy values across aggregation levels. The results of the principal component analysis are plotted in a scatterplot per aggregation level to assess which diseases are closest to each other and therefore share similar time dynamics. It can be seen in Table A1 and Figure A1 that the contribution of individual entropies to the three principal components is very consistent across time scales: H_0, H_1, and H_2 contribute equally to principal component 1 (PC1), H_0 for 60% and H_2 for 40% to principal component 2 (PC2) and H_1 for 60% and H_2 for 40% to principal component 3 (PC3). Together PC1 and PC2 explain almost 100% of the variance in H_0, H_1, and H_2 of the diseases. Based on these contributions and since we only have three quantities in the analysis, we can interpret the principal components: PC1 can be interpreted as the mean of H_0, H_1, and H_2. As $H_0 \geq H_1 \geq H_2$, the mean of these entropies will be almost similar to H_1, the Shannon entropy. This justifies the use of Shannon entropy to characterize dynamic patterns by e.g. Dalziel et al [14]. PC2 can be interpreted as the difference between H_0 and H_2, describing the variance in entropies. Substituting principal components with the entropy terms yields similar figures (Figure A1 and A2). This indicates that the variation between diseases can be explained by (a combination of) the individual entropies and that this is robust across different aggregation levels of the time series.

In Figure A2 the results of the principal component analysis are shown for different aggregation levels. On the x-axis H_1 is depicted, the approximation of principal component 1, and on the y-axis, the difference between H_0 and H_2 is shown, the approximation of principal component 2. A considerable proportion of infectious diseases cluster together in the bottom right corner, indicating diseases with a high mean of the entropies (H_1) and low H_0-H_2 . These are continuously reported diseases with little fluctuations in their time series. Examples of such diseases are tuberculosis (TBC) and chronic hepatitis B (HEPBCHR). If we move from the bottom right corner to the bottom left corner, we move from diseases with little fluctuations in their time series, but which are less continuously reported, to the most extreme situation with one reported case ever of human infection with zoonotic influenza virus (INFL). If we move from the bottom right corner upwards, we come across

diseases with larger fluctuations in their time series, such as mumps, rubella, measles and Q fever, indicating outbreaks. The ratios between diseases across aggregation levels do not differ too much: with increasing aggregation level, the variation in H_1 between diseases decreases and the variation between diseases in H_0-H_2 peaks at the 3-month aggregation level after which it decreases again (Figure A1). It can also be observed that with increasing aggregation level, diseases are arranged on parallel diagonals. These diagonals represent periodicity. When considering the five year aggregation level, ZIKA and INFL are on the diagonal of one period of reported data, and for example, MCOCW and HEPBAC on the diagonal of four (all) periods of reported data.

Table A1. Relative contributions of H_0, H_1 and H_2 to the principal components at various aggregation levels.

		Contribution (%)		
Aggregation level	**Alpha**	**PC1**	**PC2**	**PC3**
day	0	35.42	53.61	10.97
	1	33.31	1.05	65.63
	2	31.27	45.34	23.39
week	0	35.75	58.23	6.02
	1	33.18	3.85	62.98
	2	31.07	37.93	31.00
month	0	35.41	60.33	4.27
	1	33.33	5.49	61.18
	2	31.26	34.19	34.55
3 months	0	33.88	62.54	3.58
	1	33.73	5.83	60.44
	2	32.39	31.62	35.98
6 months	0	31.97	64.82	3.21
	1	34.31	5.72	59.98
	2	33.72	29.47	36.81
9 months	0	31.67	65.60	2.73
	1	34.44	6.28	59.27
	2	33.89	28.12	38.00
year	0	30.61	66.83	2.56
	1	34.74	6.17	59.09
	2	34.65	27.00	38.35
18 months	0	31.50	65.78	2.72
	1	34.58	6.28	59.14
	2	33.92	27.94	38.14
2 years	0	31.37	65.81	2.82
	1	34.60	6.09	59.32
	2	34.03	28.10	37.86
3 years	0	31.72	65.93	2.36
	1	34.64	6.95	58.40
	2	33.64	27.12	39.24
5 years	0	38.40	59.03	2.57
	1	32.57	9.06	58.37
	2	29.02	31.91	39.07

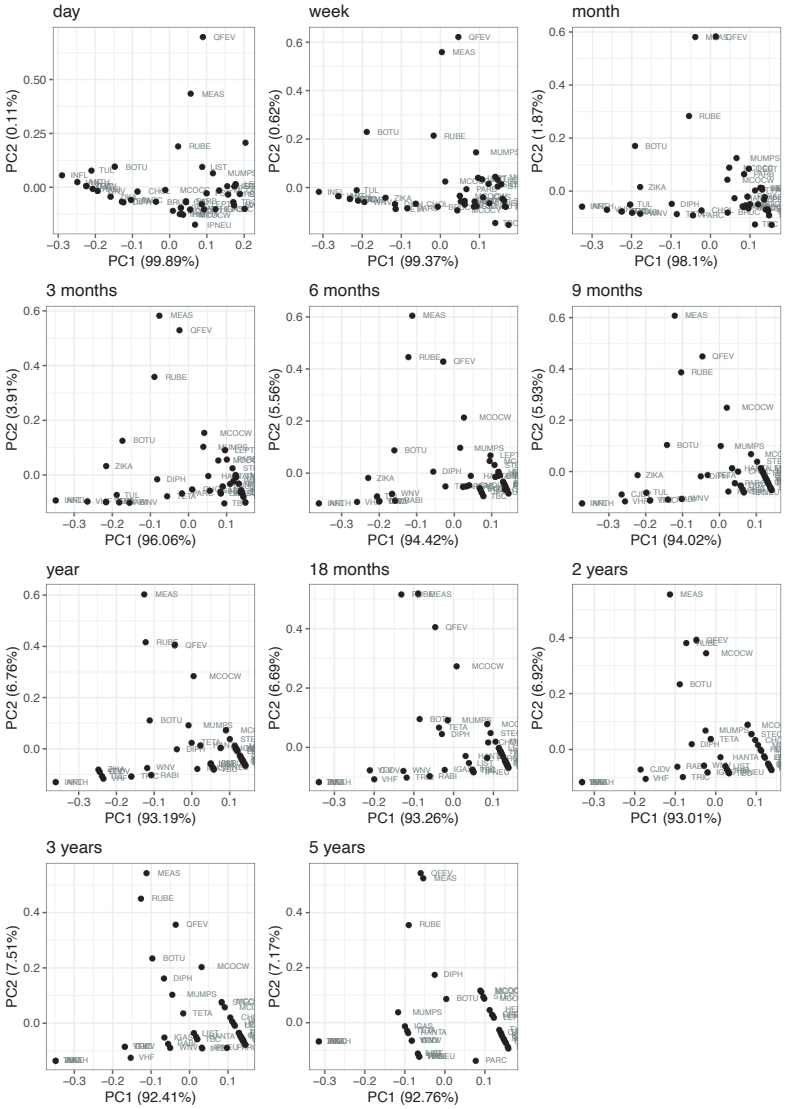


Figure A1. Results of the principal component analysis on H_0, H_1 and H_2 at various aggregation levels.

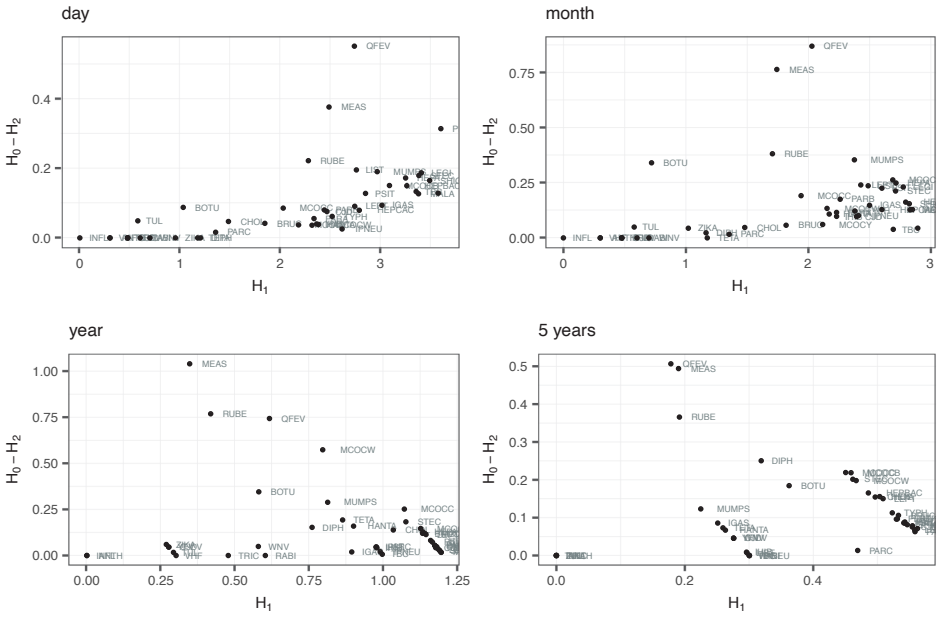


Figure A2. Results of the principal component analysis of H_0, H_1 and H_2 across different aggregation levels and for all diseases.

Appendix B

CHANGE IN DISEASE DYNAMICS ACCORDING ENTROPY MEASURES BETWEEN PERIOD T_1 (2010-2012) AND T_2 (2011-2013)

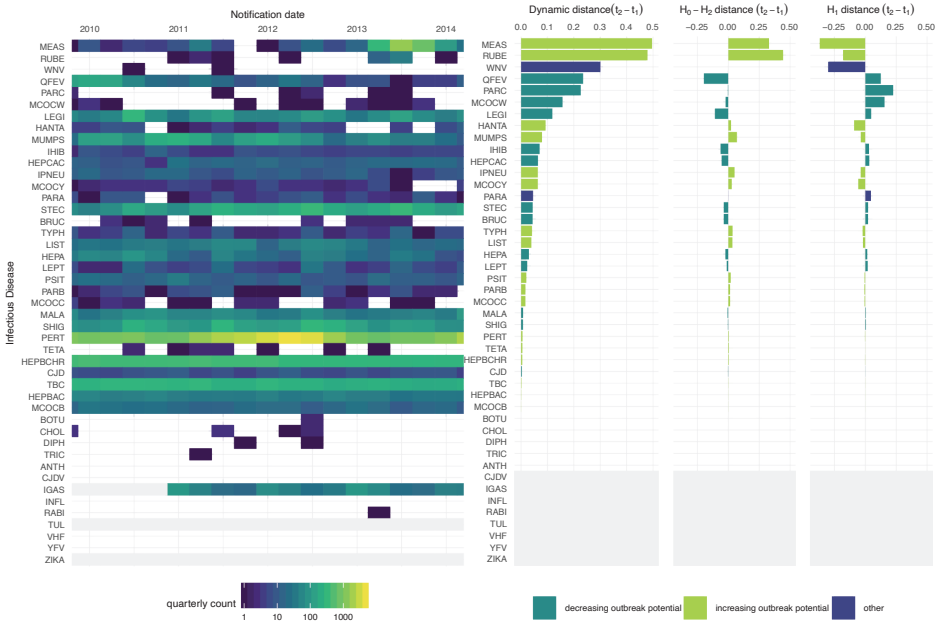


Figure B1. Change in disease dynamics according entropy measures between period t_1 (2010-2012) and t_2 (2011-2013) for all notifiable infectious diseases in the Netherlands. The quarterly aggregated time series of these periods shown as a heat map (first panel) are sorted by decreasing dynamic distance between these periods (second panel).The third and fourth panel show the distance in H_0-H_2 and distance in H_1 between t_1 and t_2 for all diseases. Light grey areas indicate diseases for which no data was reported as the disease was not yet notifiable (first panel) or that the entropies could not be estimated due to no reported cases in one or both periods (second, third and fourth panel).



5

Real-time estimation of epidemiologic parameters from contact tracing data

The contents of this chapter have been published in *Epidemiology*:
Real-time estimation of epidemiologic parameters from contact tracing data during an emerging infectious disease outbreak
Loes Soetens, Don Klinkenberg, Corien Swaan, Susan Hahné, Jacco Wallinga

Epidemiology, March 2018, 29(2):230-236

Abstract

Contact tracing can provide accurate information on relevant parameters of an ongoing emerging infectious disease outbreak. This is crucial to investigators seeking to control such an outbreak. However, crude contact tracing data are difficult to interpret and methods for analyzing these data are scarce. We present a method to estimate and visualize key outbreak parameters from contact tracing information in real time by taking into account data censoring.

Exposure type-specific attack rates and the reproduction number *R(t)* are estimated from contact tracing data by using maximum likelihood estimation to account for censored data. The attack rates reflect, in the context of contact tracing, the specificity of the contact definition; a higher value indicates relatively efficient contact tracing. The evolution of *R(t)* over time provides information regarding the effectiveness of interventions. To allow a real-time overview of the outbreak, the attack rates and the evolution of *R(t)* over time are visualized together with the case–contact network and epicurve. We applied the method to a well-documented smallpox outbreak in the Netherlands, to demonstrate the added value. Our method facilitates the analysis of contact tracing information by quickly turning it into accessible information, helping outbreak investigators to make real-time decisions to more effectively and efficiently control infectious disease outbreaks.

Introduction

In an ongoing emerging infectious disease outbreak, contact tracing can provide accurate real-time information on relevant parameters, which is needed by public health professionals to control an outbreak [1-4]. Especially in an emerging infectious disease outbreak, active daily (or even hourly) monitoring of contacts provides a wealth of information. First, data on how many contacted individuals (hereafter: contacts) remain under surveillance can indicate what control efforts are possibly required in the short term. Second, data on exposure type-specific attack rates (ARs) can help in targeting control efforts to contacts at high risk. We defined the exposure type-specific AR as the proportion of infected individuals with a certain exposure type among the total number of traced individuals with that exposure type. Third, estimation of the effective reproduction number *R(t)*, defined as the expected number of secondary cases per primary case infected at time *t* [5], allows evaluation of infection transmissibility and the effect of current control measures.

However, real-time use of contact tracing data by public health professionals is limited, despite publication of several methods for inferring key parameters from outbreak data [6-13]. One problem is that contact tracing data contains a lot of information on confirmed cases and their contacts through various exposure routes at different time-points or intervals. Using all that information requires an overview. In addition, many methods focus on just a single parameter [7-10], often the effective reproduction number. To obtain all key outbreak parameters, a public health professional has to search through a wide range of methods, which is quite cumbersome and time-consuming. A third issue is that real-time analysis should account for the fact that the data are typically censored: at the time of analysis, the outcome (whether a contact is infected or not) is unknown for many exposed persons [9]. Censoring has a consequence for accurate estimation of outbreak parameters, such as ARs, which may be underestimated if traced contacts later appear to be cases. As far as we know, only few studies take such censoring into account [9, 10]. A final issue is that published methods require specific software that is either hard to obtain or requires advanced programming skills [8, 12, 13].

To improve public health decision-making during an emerging infectious disease outbreak, we developed a tool that uses contact tracing information to estimate outbreak parameters in real time while correcting for right censoring of the data (i.e. taking into account that contacts in follow-up may become cases in the near future in the real-time estimation of outbreak parameters). Our tool provides a comprehensive visualization of the outbreak, including the course of the outbreak and contacts under surveillance. The tool is easy to use by public health professionals as it is available in a user-friendly web interface. The tool is specifically aimed at analyzing contact tracing data of emerging infectious disease outbreaks, as contact tracing is most relevant during such outbreaks where there is the

greatest need for such knowledge about the ongoing outbreak. Our tool is aimed at serious infectious disease outbreaks, for which active daily contact tracing is in place, whereby all cases are identified. In this paper, we demonstrate this tool, using contact tracing data from a well-documented smallpox outbreak in the Netherlands in 1951.

Materials and methods

Data

Data for the tool should be entered in a spreadsheet, with a row for each case or contact and the following information in the columns:

- unique identifier (ID)
- gender
- infection status: is it a case or a contact?
- ID of source
- exposure type
- exposure date (first day)
- exposure duration
- date of disease onset

In this paper, we use the word “contact” to mean an individual who has come into contact with a case and has not yet been shown to be a case himself and “exposure” to mean the event of contact between the two individuals. Additional information that should be obtained is the incubation period distribution (mean and standard deviation) and the generation interval distribution (mean and standard deviation). The generation interval refers to the time between symptom onset of successive cases in a chain of transmission. The data saved as a csv-file can be uploaded into a user-friendly web interface for easy use of the tool (https://github.com/lsoetens/Contactviz) or loaded via the R code provided at the same place, using the R statistical software.

For illustration of the tool, we used data from a smallpox outbreak in Tilburg, The Netherlands, in 1951 [14, 15]. In this outbreak, 52 cases were identified, with a date of symptom onset ranging between 23 February and 25 May 1951. The outbreak was detected on 24 April and by 29 April control measures, such as additional mass vaccination and isolation of cases and contacts, had been implemented. No individual details on vaccination status were available, but the vaccination coverage in Tilburg in 1951 was approximately 88% [14, 15]. The cases occurring before April 29 were traced back and their contacts identified. These efforts led to a very complete documentation of this outbreak. For almost all cases (n=50), it is known by whom or how they were most likely infected. Contacts of cases were also registered and monitored. The data are given in eTable 5.1; http://links.lww.com/EDE/B290. The exposure type was categorized as several types: family/household, friend/neighbor, work-related, and medical/religious setting. For some family/household cases or contacts, the date and duration of exposure were not known exactly. We therefore assumed that the duration was 21 days forward from the symptom-onset date of the related case. If the related case was put in isolation, we assumed contact exposure from the symptom-onset date of the related case to the isolation date. To demonstrate our tool in this paper, we analyzed the data as

they would have been observed on 30 May 1951. For the remainder of the paper we will call this date “the current date”. Ethical approval was deemed not necessary, as the data was already publicly available and does not allow identification of individual persons.

Monitoring period

An important decision in outbreak control concerns which contacts need follow-up and which can be set aside. For each contact *i* we define the monitoring period as the interval from the 5-percentile time-point to the 95-percentile time-point of the effective incubation period. The effective incubation period $f_{eff,i}(t)$ is defined as the total period from a first exposure to symptom onset. For any infected individual, the effective incubation period is the sum of the time from first exposure to effective transmission, and from effective transmission to symptom onset. The distribution of the effective incubation period is the convolution of the uniform distribution from first exposure to last exposure with the lognormal distribution that describes the incubation period distribution. In our method, repeated exposures are merged to one exposure interval, which is assumed to be uniformly distributed. For smallpox, the lognormal distribution that best fits the observed incubation periods has a median of 13.0 days and a dispersion of 1.13 days [16]. We denote the corresponding cumulative probability distribution function as $F_{eff,i}(t)$ and the corresponding survival distribution as $S_{eff,i}(t)= 1- F_{eff,i}(t)$. In the visual output, the monitoring period is displayed as a horizontal black line, and it is shown only for those contacts that are still monitored, i.e. with $S_{eff,i}$(current date) > 0.05.

Probability of infection for a contact

Each contact of exposure type *x* has a probability π_x of being infected and a probability of $(1- \pi_x)$ of not being infected. If not infected, the contact does not develop any symptoms. If infected, the contact will develop symptoms, and the time from first exposure to symptom onset will follow the effective incubation period. This defines a mixture distribution of a contact *i* of type *x* for its time to infection: $f_{x,i}(t) = \pi_x f_{eff,i}(t)$. We denote the corresponding cumulative probability distribution function as $F_{x,i}(t) = \pi_x F_{eff,i}(t)$ and the corresponding survival distribution as $S_{x,i}(t) = 1 - \pi_x F_{eff,i}(t)$. For each contact *i* who has a link of exposure type *x* to its linked case *j* and has not developed any symptoms up to time *t* since exposure, the probability that it is infected is calculated by Bayes’ rule as

$$\Pr(\text{infected}|\text{no symptoms at } t) = \frac{\Pr(\text{no symptoms at } t|\text{infected})\Pr(\text{infected})}{\Pr(\text{no symptoms at } t)} \quad (1)$$

which is, based on the definitions above, given by

$$p_{i,j,x} = \frac{\hat{\pi}_x S_{eff,i}(t)}{S_{x,i}(t)} \quad (2)$$

We used maximum likelihood estimation to estimate π_x. More details can be found in Appendix A.

Attack rate by exposure type

The exposure type-specific AR, π_x, is defined as the probability of a contact being infected, among contacts of exposure type *x*. In general, the exposure type-specific ARs will tell public health professionals the type of exposure through which contacts are at the highest risk of developing symptoms. In the context of contact tracing, the AR measures the true positive rate, or precision, of the contact definition. A higher value indicates a more efficient contact tracing. The AR is separately estimated for every exposure type *x* having a total number of cases and contacts greater than 5.

Reproduction number

By evaluating the course of the effective reproduction number *R(t)* during the outbreak, public health professionals can assess changes in the effective transmissibility of the disease [7, 11]. There are various ways of estimating *R(t)* over time [7, 11, 17], but most assume that the links between the cases, who infected whom, are not known. The *R(t)* then needs to be inferred indirectly, whereas when using contact tracing data, the *R(t)* can be estimated directly. For each case *j* we calculate the expected number of secondary cases. The secondary cases consist of confirmed secondary cases, contacts that are or might be infected but have not yet shown symptoms, and orphans. Orphans are infected individuals for whom the infector's identity is unknown, though by assumption the infector is one of the known cases. To calculate the first term on the righthand side, we count the number of confirmed secondary cases of case *j* ($D_j(t)$). Confirmed secondary cases are defined as contacts that were followed up as a result of exposure to a case, and eventually developed symptoms within their monitoring period. To calculate the second term, we sum over the contacts of case *j* and weigh each contact by his/her probability of being infected by case *j* ($p_{i,j,x}(t)$). To calculate the third term, we sum over the number of orphans and weigh each by his/her probability of being infected by case *j* [17].

$$R_j(t) = D_j + \Sigma_i \Sigma_x p_{i,j,x}(t) + \Sigma_k \frac{w(t_k - t_j)}{\Sigma_l w(t_k - t_l)} \qquad (3)$$

Here, the index *i* refers to all contacts, the index *k* to all 'orphan' cases, and the index *l* to all contacts that could have infected the *k*th orphan case. The function *w* gives the generation interval distribution, where the generation interval is defined as the duration between symptom onset of an infected individual and symptom onset of his/her infector. For smallpox, we take the generation interval to follow a gamma distribution with a mean and variance of 17.7 days [18].

For any given period of time, we can calculate the average value of the reproduction number for cases who developed symptoms within that period. Of interest here are the period from start of the outbreak to the time the first case was detected (including all primary cases with a date of onset of disease before 24 April 1951 and all secondary cases with an exposure

date before 24 April 1951) and the period from implementation of control measures to the current date (including all cases with an exposure date between 29 April 1951 and 30 May 1951). To visualize the time course of the average reproduction number, we calculate for each time t a running average over a time window $(t - \tau, t)$ of length τ. We choose τ equal to the generation interval for the specific infection. For smallpox it is 17.7 days [18]. We estimate 95% nonparametric bootstrap percentile confidence intervals [11] around the reproduction numbers, which reflect the uncertainty in whether contacts in follow-up will become cases and the uncertainty regarding possible links to orphan cases. The intervals are estimated by repeating the following steps a thousand times: first, draw from a binomial distribution with probability $p_{i,j,x}$ for every contact in follow up whether it is going to be a case or not and for every possible link to an orphan case if this a true link or not; second, sum the total number of secondary cases per case (this includes confirmed secondary cases, contacts in follow up and orphan cases); and third, compute the mean of the total number of secondary cases per case. By repeating this a thousand times, a thousand means are estimated, from which the 2.5% and 97.5% percentiles define the 95% confidence interval around the *R(t)* in a certain time interval.

Algorithm for layout of the time-oriented transmission tree

To provide an overview of the links between cases and contacts and the evolution of the outbreak over time, we adapted the Reingold-Tilford tree-drawing algorithm [19] to obtain a time-oriented transmission tree (see further details in Appendix B). In this node-link tree, cases and contacts are represented as nodes, and the exposures between cases and contacts are represented as links. The nodes of the cases are placed according to the first day of symptom onset, and the nodes of the contacts are placed according to the lower level of the monitoring period.

Model assumptions

Underreporting

The model assumes that all cases and contacts have been observed. However, this is most often not the case. This underreporting might therefore introduce bias to our model. As intensive daily contact tracing is in place during an emerging infectious disease outbreak, we assume that the number of missed secondary cases is minimal and subsequently that this underreporting has a limited effect on the estimation of the effective reproduction number. Underreporting does not play a role in the estimation of the attack rates, as we have defined the attack rate as the proportion of infected individuals among the total number of traced individuals, instead of susceptibles.

Reporting delay

The model does not take into account reporting delay, which may introduce bias in our estimates of the effective reproduction number and the attack rates. As intensive contact

tracing is in place, we assumed that reporting delays are minimal as all cases and contacts are under strict monitoring.

Introductions versus orphans

The model assumes there was only one introduction (index case) in the community, and that cases not reported as a contact by anyone else (i.e. orphans), have their infector among the observed cases. We considered it more likely that a case with an unknown source originates from another observed case than that a new introduction of the emerging infectious disease occurred in the community.

Simulation study

To allow assessing the accuracy of the estimated attack rates and effective reproduction number, and to assess the performance of our tool for a larger outbreak, we performed a simulation study. We used the outbreaker package in R to simulate an outbreak three times as large as the smallpox outbreak in Tilburg, with similar parameter settings. We compared the estimates of our model that use information on cases in follow-up to correct for right censoring with reference estimates obtained by including all present and future information on the infections and their contacts (no right censoring) and with naïve estimates obtained by ignoring any current information on the contacts (no correction for right censoring). A detailed description of the simulation study can be found in Appendix C.

Visualization

To provide an overview of the outbreak, we display the contact tracing data in a multi-panel plot (Figure 5.1) using the above-mentioned parameters and algorithms. It includes a large plot with three horizontal panels plus a fourth panel at lower right.

In the large plot, the bottom panel contains an epidemic curve of the outbreak so far. The epidemic curve is displayed as a histogram of the number of cases per three days from the date of symptom onset of the first case to the current date. We take as a rule of thumb: the width of the bars in the epidemic curve is usually one quarter of the length of the incubation period.

The middle panel, which is vertically aligned with the epidemic curve, contains a line graph of the course of $R(t)$ with its 95% CI during the outbreak. In addition, for every case the observed number of secondary cases $D_j(t)$ is displayed by date of symptom onset of case j as a scatterplot, in which the size of the dots is proportional to the frequency of the number of secondary cases.

The top and main panel, vertically aligned with those below, contains the time-oriented transmission tree. We have used the algorithm described above to place the nodes

and links. The nodes representing cases in the tree are displayed as squares (men) or as diamonds (women) at the date of symptom onset. The nodes representing contacts in follow-up are displayed as dark red bars reflecting their monitoring period. In this way, it can easily be seen which contacts are still at risk of developing symptoms, i.e., where the period overlaps the current date. Nodes are labeled with the case or contact ID number. The links between nodes are depicted as colored lines reflecting the various exposure types (family/household, friend/neighbor, medical/religious, work); the line types differ for links between cases (solid) and case–contact links (dashed).

A solid vertical bar intersects these three panels to indicate the current date. The time-period after the current date is colored grey to indicate that the outcomes are still unknown for that period. Finally, the plot is annotated with dashed vertical lines that indicate important dates, such as the day when the first case was discovered and the day when control measures were installed. With the addition of such annotations, it can easily be seen if the situation has changed after a certain time-point.

The fourth panel, lower right, contains a bar chart representing the exposure type-specific ARs and their 95% CIs. The bar colors match the line colors in the network plot. Exposure types are displayed only when their total number of cases and contacts is greater than 5. Finally, above the fourth panel is information containing some summary statistics of the outbreak so far (number of cases/contacts, number of contacts in follow-up or discarded, and maximum number of secondary cases per case) as well as important dates and explanation of the above-mentioned features.

The R 3.1.0 statistical software is used to construct this multi-panel plot; main packages used are *igraph*, *ggplot* and *shiny*.

Results

An overview of the outbreak situation at 30 May 1951 in a multi-panel plot is given in Figure 5.1. The runtime of the analysis of this outbreak at this date was 10.92 seconds. An overview of the outbreak at an earlier time, the 29th of April 1951, can be found in Appendix D.

Between 23 February and 30 May 1951, 52 smallpox cases including 28 men and 24 women were identified, with the highest number of cases reported at the beginning of May, as seen in the epidemic curve (bottom left panel, Figure 5.1). For three cases (ID 1, 34, 50), no source could be identified. By May 30, known as the current date, 22 contacts were in follow-up (i.e., within their surveillance period) and 174 contacts were discarded, as their surveillance was over. The latter are not shown to prevent crowding of the plot. If no new cases and contacts present, follow-up of all contacts was to end by mid-June 1951.

The time-oriented transmission tree (top left panel, Figure 5.1) shows that most cases were infected through household/family contact (31/52) (red lines), followed by friend/neighbor contact (9/52) (blue lines). The exposure type-specific ARs are depicted in the bar chart in the lower right panel of Figure 5.1. The AR was highest among those infected by work-related contact (AR= 42%; 95% CI: 9 - 81%) and by medical contact (AR= 31%; 95% CI: 5 – 71%).

The evolution of *R(t)* during the outbreak (left middle panel, Figure 5.1) shows several peaks in March and April, after which the *R(t)* declines to zero at the beginning of May. At the end of May, a slight increase in the *R(t)* is observed, indicating probable secondary cases among the contacts still in follow-up. The maximum number of secondary cases per case was 13 by case 35, who could be identified as the main spreader in this outbreak thus far. The transmission tree and the course of *R(t)* over time indicate that before the first case was discovered and additional control measures were implemented, the *R(t)* was decreasing. The *R(t)* before the first case was detected on 24 April was 2.42 (95% CI: 2.4, 2.53), and the *R(t)* after additional control measures were installed was 0.83 (95% CI: 0.43, 1.29). The *R(t)* before outbreak detection is slightly lower than the reproduction numbers reported in previous studies [20-22], however, when taking into account that the population in this outbreak was only partially susceptible due to childhood vaccination, the findings are quite comparable.

The simulation study in Appendix C shows that a larger outbreak can be studied in a similar way as the outbreak used in this study. The proposed estimators for the attack rate and effective reproduction number that include information on contact tracing to account for right censoring of the data, come closer to the actual values in comparison with another (naïve) estimator that does not correct for right censoring. The runtime of the analysis of the simulated outbreak was 27 seconds on a standard laptop computer (Intel Core i5 processor).

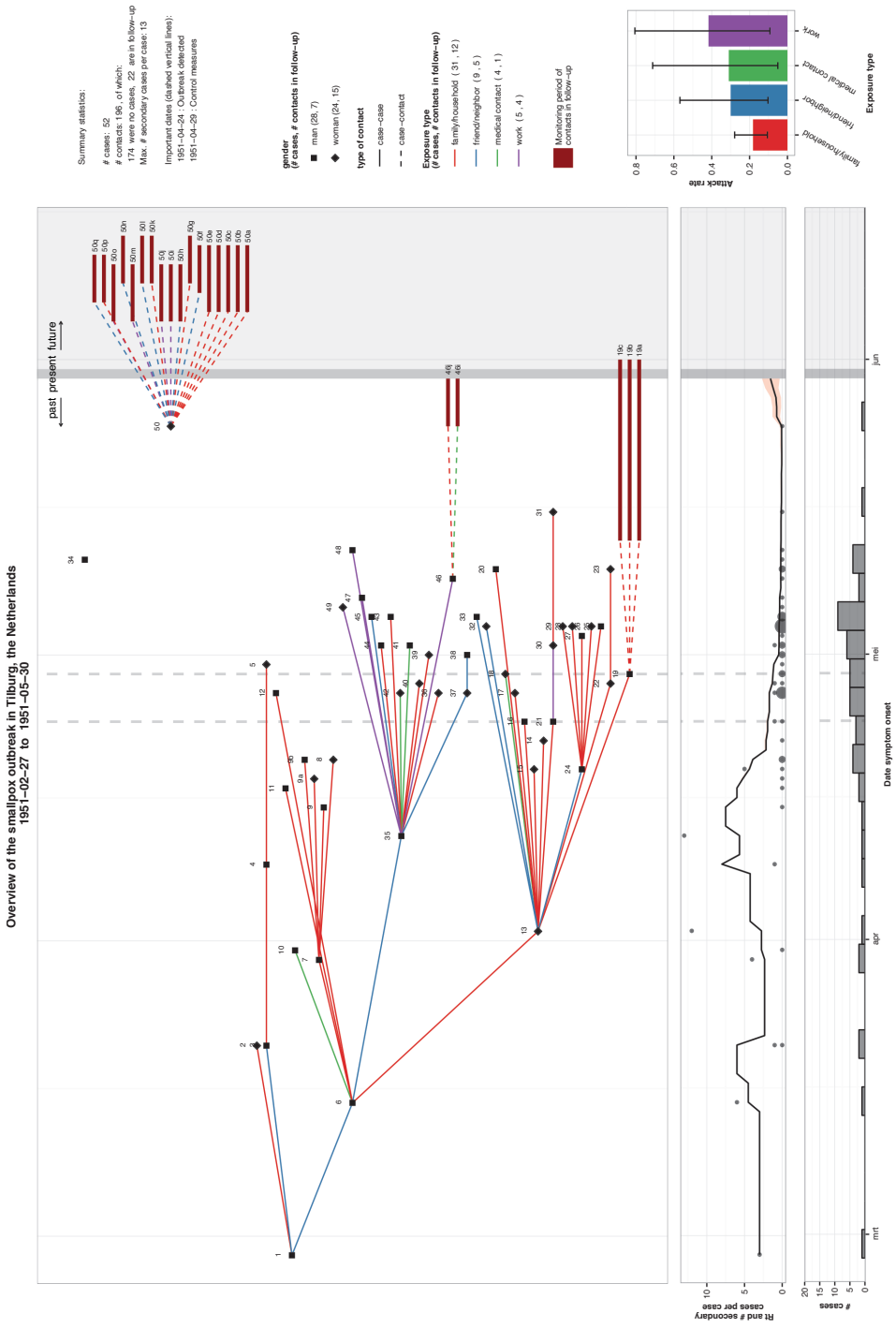

Overview of the smallpox outbreak in Tilburg, the Netherlands
1951–02–27 to 1951–05–30
past present future
Summary statistics:
cases: 52
contacts: 196, of which:
174 were no cases, 22 are in follow–up
Max. # secondary cases per case: 13
Important dates (dashed vertical lines):
1951–04–24 : Outbreak detected
1951–04–29 : Control measures
gender
(# cases, # contacts in follow–up)
man (28, 7)
woman (24, 15)
type of contact
case–case
case–contact
Exposure type
(# cases, # contacts in follow–up)
family/household (31 , 12)
friend/neighbor (9 , 5)
medical contact (4 , 1)
work (5 , 4)
Monitoring period of contacts in follow–up
Attack rate
Exposure type
family/household
friend/neighbor
medical contact
work
Rt and # secondary cases per case
cases
Date symptom onset
mrt
apr
mei
jun

Figure 5.1. Screenshot of the contact tracing visualization tool with an overview of contact tracing information during an outbreak of smallpox in Tilburg, the Netherlands, using information as of May 30, 1951. Top panel: Node-link diagram of the (possible) transmission events between cases and contacts over time. The cases are shown as symbols (square=men, diamond=women, circle= unknown) by date of onset of symptoms. The contacts with Seff,i(current date) > 0.05 are displayed as a horizontal dark red bar indicating their monitoring period. The links between cases (case-case links, solid) and between cases and contacts in follow-up (case–contact links, dashed) are depicted as lines and colored by exposure type. Middle panel: Course of R(t) during the outbreak and observed number of secondary cases per primary case. The course of the average case reproduction number, R(t) (with 95% nonparametric bootstrap percentile confidence intervals), was estimated using a left-sided moving average with a time-window of 18 days (the generation interval). The observed numbers of secondary cases per primary case are displayed as black dots, with the size of the dots being proportional to the frequency of the number of secondary cases. Bottom panel: The epidemic curve of the outbreak per three days by date of onset of symptoms. Bar chart: The exposure type-specific attack rates with their confidence intervals. Only exposure types with >5 cases/contacts are displayed.

Discussion

We have proposed a tool and methods to accurately infer outbreak parameters in real time from contact tracing data in order to enhance decision-making and, subsequently, the efficacy and efficiency of outbreak control. The tool provides an indication of the work-load of outbreak investigators (the number of contacts in follow-up) and gives an impression of the effective transmissibility of the disease during the outbreak (*R(t)*). Visualization of this information shows clearly which contacts are still at risk of developing symptoms and which are no longer at risk and can therefore be discarded, reducing the work-load. The estimation and visualization of exposure type-specific ARs helps public health practitioners to identify the exposure types through which contacts are at the highest risk of developing symptoms. Contact definitions can then be made more specific during the outbreak. In our example of smallpox, we surprisingly found that work-related and medical contacts yielded the highest AR, although family/household contact is usually more intense. Finally, the information available through contact tracing allows assessment of interventions. In the smallpox outbreak, it was observed that before the first case was discovered and the implementation of additional control measures, the *R(t)* was decreasing. This decrease might be attributed to the intervention, as cases identified just before its installation (up to 18 days, the generation interval) were less likely to cause secondary cases. The course of *R(t)* is likely to be consistent with effectiveness of the control measure.

Our approach has three main strengths. First and foremost is its ability to analyze the data in real time. Incubation period distributions are used to estimate when a contact is most likely to develop symptoms, providing an up-to-date overview of the contacts in follow-up and the related work-load for outbreak investigators. We added to the real-timeliness of the tool by taking the censored data into account. This involved estimating the probability of infection for a contact, using maximum likelihood estimation for those contacts still in follow-up to estimate the AR and *R(t)* during the course of the outbreak. The AR is often obtained by using aggregated numbers for total number of cases and contacts at a single time-point. Contacts in follow-up are usually neglected and left out of the analysis. However, the contacts in follow-up can still become cases, and the ARs might therefore be underestimated. Also, *R(t)* is often underestimated, as potential secondary cases are not taken into account. By accounting for the censoring of contacts still in follow-up, we could make our AR and *R(t)* estimations more accurate (as we have shown in our simulation study in Appendix C).

The second strength of our approach is that it provides a comprehensive overview of the outbreak at a glance. This graph is easily brought to table when outbreak management decisions are discussed, making it easier to use by outbreak investigators compared to other more software-oriented tools [8, 12, 13]. The tool is provided with a user-friendly interface, making the use of the tool relatively straightforward. Value has been shown for all

elements of the visual, such as the epidemic curve, the contact network, and the statistics; but having them readily available in one visual output with the latest results is novel. Our combination of elements offers a great advantage over previous methods focused on only one element of the outbreak analysis [7-9].

The third strength is that the output is presented visually. We specifically aimed for this, as the human mind can acquire more information through vision than through all of the other senses combined [23]. Besides fast and efficient processing, visualization has several other advantages [23], allowing the viewer to comprehend large amounts of data and perceive patterns. Problems with the data can become more easily apparent. In our smallpox example, it is immediately clear that the generation time between case 6 and 12 is improbably long and most likely due to a registration error or missed case in between. Finally, visualization facilitates hypothesis generation. For these reasons, we support the use of visuals instead of tables and text to communicate contact tracing information for decision- making. In addition, the visual output bridges the gap between the more software oriented-tools [8, 12, 13] and public health professionals whose programming skills are not always advanced.

Several issues regarding use of our tool in daily practice should be kept in mind. First, the visual is only as good as the input information. For this study, we used historical data of a well-documented outbreak. We are aware that during an outbreak, registration of all cases and contacts in a tidy way might not be the first priority. However, we expect that our approach to visualize the valuable information for decision-making provides an incentive for outbreak investigators to organize the data in the required format, as it will be time-saving in the end. On the other hand, since this method is so dependent on data quality, it can be used to trace errors of data entry as well. Second, the visual layout can be adapted to deal with very large outbreaks with numerous contacts, such as the pertussis outbreak in the Netherlands in 2012 [24], infectious diseases with a potentially long incubation time, such as tuberculosis [25], or other than human-to-human transmissible diseases. For very large outbreaks, the tool could be adapted by splitting the outbreak into smaller clusters. For other than human-to-human transmissible diseases, like environmental point-source outbreaks, the tool could be adapted by estimating the attack rates and transmission tree by environmental source. The estimates of the effective reproduction number are not relevant in such a situation.

There are various ways in which the tool can be improved upon in future studies. We chose to use the incubation period distribution and generation time distribution from the literature. One could also try to estimate these parameters from the data itself, with as a major advantage a higher accuracy of the estimates for that specific outbreak. However, the estimates would be very unstable at the beginning of the outbreak, as only limited

data is available. It could be a possible direction for future development of the tool to combine prior information from literature with a growing amount of data from an ongoing outbreak. Second, outbreaks are typically described by the time, pathogen type, place, and person characteristics. This tool was mainly focused on the dimensions of time and person. Future work on how to add the dimensions of pathogen type and place to this layout would be useful. A final minor point for improvement is the tree plotting algorithm, for which we adapted a proven algorithm to optimize the use of available space (Appendix B), but which in some cases results in crossing lines.

To conclude, we have proposed a tool to visualize contact tracing information to increase timely access to information on the characteristics of an emerging outbreak. The tool obtains and visualizes accurate outbreak parameters by taking into account the censoring of the data. The resulting information can help public health professionals to get a real-time overview of the outbreak and subsequently make better decisions in order to control the outbreak. Further research should focus on improving the network layout algorithm and enhancing the visual with data on pathogen type and place that are obtained during the outbreak.

References

1. Eames KT, Keeling MJ. Contact tracing and disease control. Proc Biol Sci. 2003;270(1533):2565-71.
2. Klinkenberg D, Fraser C, Heesterbeek H. The effectiveness of contact tracing in emerging epidemics. PLoS One. 2006;1:e12.
3. Kiss IZ, Green DM, Kao RR. Disease contact tracing in random and clustered networks. Proc Biol Sci. 2005;272(1570):1407-14.
4. Bonacic Marinovic A, Swaan C, van Steenbergen J, Kretzschmar M. Quantifying reporting timeliness to improve outbreak control. Emerg Infect Dis. 2015;21(2):209-16.
5. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLoS One. 2007;2(8):e758.
6. Jewell CP, Roberts GO. Enhancing Bayesian risk prediction for epidemics using contact tracing. Biostatistics (Oxford, England). 2012;13(4):567-79.
7. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505-12.
8. Obadia T, Haneef R, Boelle PY. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. BMC Med Inform Decis Mak. 2012;12:147.
9. Ghani AC, Donnelly CA, Cox DR, Griffin JT, Fraser C, Lam TH, et al. Methods for estimating the case fatality ratio for a novel, emerging infectious disease. Am J Epidemiol. 2005;162(5):479-86.
10. Cauchemez S, Boelle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis. 2006;12(1):110-3.
11. Hens N, Calatayud L, Kurkela S, Tamme T, Wallinga J. Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. Am J Epidemiol. 2012;176(3):196-203.
12. Jombart T, Aanensen DM, Baguelin M, Birrell P, Cauchemez S, Camacho A, et al. OutbreakTools: a new platform for disease outbreak analysis using the R software. Epidemics. 2014;7:28-34.
13. Höhle M. An R package for the monitoring of infectious diseases. Computational Statistics. 2007;22(4):571-82.
14. Sas GJ. Klinische en epidemiologische waarnemingen tijdens de pokkenepidemie te Tilburg in 1951: Universiteit Leiden; 1954.
15. Ministerie van sociale zaken en volksgezondheid. Rapport omtrent de pokkenepidemie te Tilburg in 1951. Staatsdrukkerij- en uitgeverijbedrijf, 's-Gravenhage: 1953 33565.
16. Sartwell PE. The incubation period and the dynamics of infectious disease. Am J Epidemiol. 1966;83(2):204-6.
17. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004;160(6):509-16.
18. Vink MA, Bootsma MC, Wallinga J. Serial intervals of respiratory infectious diseases: a systematic review and analysis. Am J Epidemiol. 2014;180(9):865-75.
19. Reingold EM, Tilford JS. TIDIER DRAWINGS OF TREES. IEEE Transactions on Software Engineering. 1981;SE-7(2):223-8.
20. Gani R, Leach S. Transmission potential of smallpox in contemporary populations. Nature. 2001;414(6865):748-51.
21. Eichner M, Dietz K. Transmission potential of smallpox: estimates based on detailed data from an outbreak. American journal of epidemiology. 2003;158(2):110-7.
22. Riley S, Ferguson NM. Smallpox transmission and control: spatial dynamics in Great Britain. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(33):12637-42.
23. Ware C. Information Visualization: Perception for Design: Elsevier Science Morgan Kaufmann; 2004.
24. van der Maas NA, Mooi FR, de Greeff SC, Berbers GA, Spaendonck MA, de Melker HE. Pertussis in the Netherlands, is the current vaccination strategy sufficient to reduce disease burden in young infants? Vaccine. 2013;31(41):4541-7.
25. Borgdorff MW, Sebek M, Geskus RB, Kremer K, Kalisvaart N, van Soolingen D. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. Int J Epidemiol. 2011;40(4):964-70.

Appendix A

MAXIMUM LIKELIHOOD ESTIMATION FOR PROBABILITY ON INFECTION

Notation and data

By a contact we mean an individual that has been identified in contact tracing, irrespective of the fact whether this individual is infected or not. We use the subscript $i = 1, 2, \ldots, n$ to label the contacts. Contacts are categorized by their type (household, work etc). We use the subscript x to denote type.

By an orphan we mean an infected individual for which the infector hasn't been traced.
For each contact i we have observation on time of start of exposure, $t_{start,i}$, and the end of exposure $t_{end,i}$. We have a censoring indicator d_i which is 0 if the contact has not shown any symptoms at the current time t, and which is 1 if the contact has shown symptoms before the current time t. For a contact i who has shown symptoms before the current time t ($d_i = 1$), we denote the time of symptom onset as $t_{sym,i}$. For a contact i who has not shown symptoms the current time, we denote the current time as censoring time for $t_{sym,i}$. For each contact i we calculate the time from start of exposure to the time of symptom onset / censoring time:

$$t_i = t_{sym,i} - t_{start,i} \tag{1}$$

Exposure and incubation period

The incubation period is defined as the time from exposure to symptom onset. For many infectious diseases this distribution is well described by a lognormal distribution. For smallpox the best fitting lognormal distribution to observed incubation periods has a median of 13 days and a dispersion of 1.13 (Sartwell, 1966). We denote this distribution as f_{inc}:

$$f_{inc}(t) \sim lognormal(ln(median),\ ln(dispersion)^2) \tag{2}$$

We denote the cumulative distribution as F_{inc} and the corresponding survival distribution as $S_{inc} = 1 - F_{inc}$. For most cases we do not have a single moment of exposure, but a period from first exposure to last exposure. We assume that effective transmission from the infector to the infected can occur with equal probability during the period from first to last exposure. We denote this exposure distribution as f_{exp}.

$$f_{exp,i}(t) \sim uniform(t_{start,i},\ t_{end,i}) \tag{3}$$

We define an effective incubation period as the total period from a first exposure to symptom onset. For any infected individual, the effective incubation period is the sum of the time from first exposure to effective transmission, and from effective transmission to symptom onset. The distribution of the effective incubation period is the convolution of the uniform

distribution from first exposure to last exposure with the lognormal distribution that describes the incubation period distribution. We denote the effective incubation period distribution as

$$f_{eff,i}(t) = f_{inc}(t) * f_{exp,i}(t) \quad (4)$$

We denote the corresponding cumulative probability distribution function by $F_{eff,i}(t)$, and the corresponding survival distribution as $S_{eff,i}(t) = 1 - F_{eff,i}(t)$.

Probability of infection for a contact

Each contact of type x has a probability π_x of being infected and a probability $(1- \pi_x)$ of not being infected. If not infected, the contact does not develop any symptoms. If infected, the contact will develop symptoms and the time from first exposure to symptom onset will follow the effective incubation period. This defines a mixture distribution of a contact of type x for its time to infection:

$$f_{x,i}(t) = \pi_x f_{eff,i}(t) \quad (5)$$

$$F_{x,i}(t) = \pi_x F_{eff,i}(t) \quad (6)$$

with a corresponding survival distribution

$$S_{x,i}(t) = 1 - \pi_x F_{eff,i}(t) \quad (7)$$

We can calculate the type-specific hazard of symptom onset as

$$h_{x,i}(t) = \frac{\pi_x f_{eff,i}(t)}{1-\pi_x F_{eff,i}(t)} \quad (8)$$

Likelihood equation for probability of infection

The likelihood contribution of the ith contact with symptom onset/censoring time t_i and indicator d_i is

$$L_i(\pi_x|t_i, d_i) = [h_{x,i}(t_i)]^{d_i} S_{x,i}(t_i) \quad (9)$$

The corresponding contribution to the log likelihood is

$$\ell_i(\pi_x|t_i, d_i) = d_i \ln h_{x,i}(t_i) + \ln S_{x,i}(t_i) \quad (10)$$

The total log likelihood is

$$\ell(\pi_x|t,d) = \Sigma_i \ell_i(\pi_{x,i}|t_i, d_i) = \Sigma_i d_i \ln h_{x,i}(t_i) + \Sigma_i \ln S_{x,i}(t_i) \quad (11)$$

substituting the definitions for $h_{x,i}$ and $S_{x,i}$ and simplifying gives

$$\ell(\pi_x|t,d) = c + \Sigma_{i:d_i=1} \ln \pi_x + \Sigma_{i:d_i=0} \ln (1 - \pi_x F_{eff,i}(t_i)) \quad (12)$$

where c is a constant. For practical purposes we might simplify this equation further. We have n contacts, of which a number of n_s contacts showed symptoms, a number of n_f contacts are in follow up as they might be incubating infection, and a number of n_a contacts that showed no symptoms during their monitoring period.

$$\ell(\pi_x|t,d) = c + \mathrm{n_s} \ln (\pi_x) + n_a \ln(1 - \pi_x) + \Sigma_{i=1}^{i=n_f} \ln (1 - \pi_x F_{eff,i}(t_i)) \quad (13)$$

We can check that when the epidemic is over, and there are no contacts incubating infection, this reduces to the binomial likelihood for proportion symptomatic.

Estimator for probability of infection

The maximum likelihood estimate for the probability of contact of type x, π_x, to be infected can be found by maximizing the log-likelihood.

$$\hat{\pi}_x = argmax_{\pi x} \ell(p_{x,i}|t,d)) \quad (14)$$

The 95% confidence interval for probability of contact of type x to be infected, π_x, is obtained as a profile likelihood. We include values of π_x where the deviance $2(\ell(\hat{\pi}_x) - \ell(\pi_x)$ is smaller than 3.841, the critical value for Pearson's chi-square test at a level of significance of 0.05 on one degree of freedom.

Probability of infection for a contact of type *x* that has not developed any symptoms up to time *t*

For each contact i which has a link of type x to its infector case j, the probability that it is infected at time t is given by

$$p_{i,j,x} = \frac{\hat{\pi}_x S_{eff,i}(t)}{S_{x,i}(t)} \quad (15)$$

Numerical considerations

We approximate the cumulative effective incubation period distribution $F_{eff,i}(t)$ by drawing a sample from the uniform distribution, drawing a sample from the lognormal distribution, and adding these to obtain a sample from the effective incubation period distribution. We repeat this 10,000 times, and rank the samples. We obtain a close approximation to the probability

density of the effective incubation period distribution, $F_{eff,I}(t)$. We evaluate the log-likelihood for values of π_x ranging from 0 to 1. For these values we identify first the maximum likelihood value, and then the values that lie within the 95% confidence interval around the maximum likelihood value.

Appendix B

THE LAYOUT-ALGORITHM OF THE TRANSMISSION TREE

We adapt the Reingold-Tilford tree drawing algorithm to get a time-oriented transmission tree. In this appendix, we explain the adaptations step by step, and we illustrate the adaptations with a hypothetical tree structure.

We start with placing the cases and contacts in a standard node-link diagram (Fig. B1A). We apply the Reingold-Tilford tree drawing algorithm on this diagram (Fig. B1B). This algorithm is based on a few principles: 1) nodes at the same generation of the tree should lie along a straight line (blue dotted lines), and the straight lines defining the generations should be parallel, 2) a left son should be positioned to the left of its father and a right son to the right, and a father should be centered over its sons. This algorithm results in a generation-based transmission tree. The next step is rotating the positions of the nodes such that cases in the same generation are aligned at the same horizontal position (Fig. B1C). We could shift the horizontal position of a case towards the left or the right such that this horizontal position reflects time of symptom onset and accordingly we could shift the horizontal position of a contact such that it becomes the time window for the 90%-interval of sympom onset; unfortunately, this would result in several crossing links (Fig. B1D). We introduce a trick to reduce the number of crossing links: before we reposition the nodes according to calendar time, we add ‘dummy offspring’ in such a way that all branches of the transmission tree have at least one node left in the last generation of the transmission tree (Fig. B1E). In the example, we created three generations of ‘dummy offspring’ for nodes E, F and H, which are in the same and last generation as nodes I, J, K, L and M of the transmission tree. We now reposition the nodes according to calendar time (Fig. B1F), and remove the ‘dummy offspring’ (Fig. B1G).

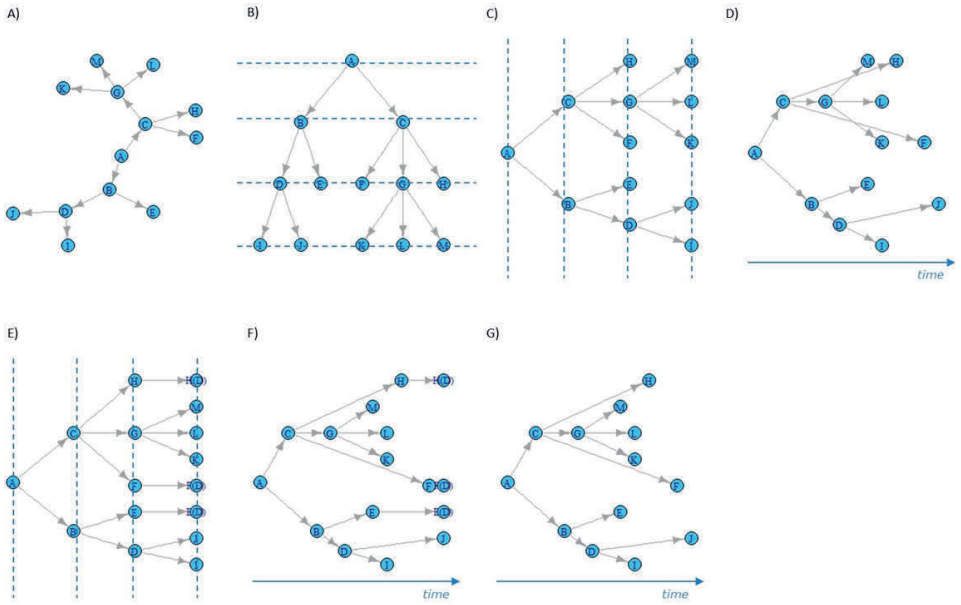


Figure B1: Steps in creating the layout of the transmission tree. A) Place the cases and contacts in a standard node-link diagram. B) Apply the Reingold-Tilford algorithm on the data. C) Align the generations vertically. D) Reposition the horizontal position of the nodes according to calendar time. E) To reduce the number of crossing links, add 'dummy offspring' to the vertically aligned transmission tree. F) Reposition the nodes and 'dummy offspring' according to time. G) Remove the'dummy offspring'.

Appendix C
SIMULATION STUDY

In order to assess the accuracy of the estimator of the attack rates (AR) and the effective reproduction numbers we performed a simulation study. We used the R package 'Outbreaker' and the simOutbreak function to generate an outbreak of about three times the size of the outbreak used in the main paper. We generated an outbreak in a population of 150 individuals for an infection with a basic reproduction number of 3, starting with one infectious case on March 1, 2015. We simulated contacts by linking each simulated case to a random number of contacts, this number was drawn from a discrete uniform distribution *Unif*{3, 10}. We simulated the exposure dates, durations of exposure and exposure routes for each contact. The code for generating the data is given at the end of this appendix.

An overview of the simulated outbreak for May 30, 2015 is shown in Figure C1. The epidemic curve, the time course of the reproduction number and the histogram with attack rates present a summary of the simulated data that are available at that point in time. The network gives a clear picture of how the infection was spreading at the start of the outbreak, although the more than 300 contacts in follow-up appear crowded.

The proposed estimator of the overall attack rate uses information on cases in follow-up to correct for right censoring. We compare the performance of this proposed estimator to two alternative estimators. One alternative estimator, the reference estimator, is obtained by including all present and future information on the infections and their contacts (no right censoring). Another estimator, the naive estimator, is obtained by ignoring any current information on the contacts (no correction for right censoring). Figure C2 shows the comparison between the proposed estimator of the overall attack rate and the two alternatives. One can see that the estimate of the overall attack rate is very uncertain when only a few cases have been observed. As more cases are observed, the confidence interval shrinks and closes in on the actual value without right censoring. This actual value is covered by the confidence interval throughout the epidemic. The proposed estimator is consistently better than the naive estimator that ignores information on contacts in follow-up and that does not correct for right censoring of the data.

The proposed estimator of the effective reproduction number also uses information on cases in follow- up to correct for right censoring. We again compare the performance of this proposed estimator to two alternative estimators. Again a reference estimator, that is obtained by including all present and future information on the infections and their contacts (no right censoring), and a naive estimator, that is obtained by ignoring any current information on the contacts (no correction for right censoring). Figure C3 shows the comparison between the proposed estimator of the effective reproduction number and the

two alternatives at four points in time during the outbreak. One can see that the estimate of the effective reproduction number is very uncertain when only a few cases have been observed at the beginning of the outbreak. As more cases are observed, the confidence interval shrinks. The proposed estimator is performing considerably better than the naive estimator that consistently underestimates the effective reproduction number in the most recent time intervals.

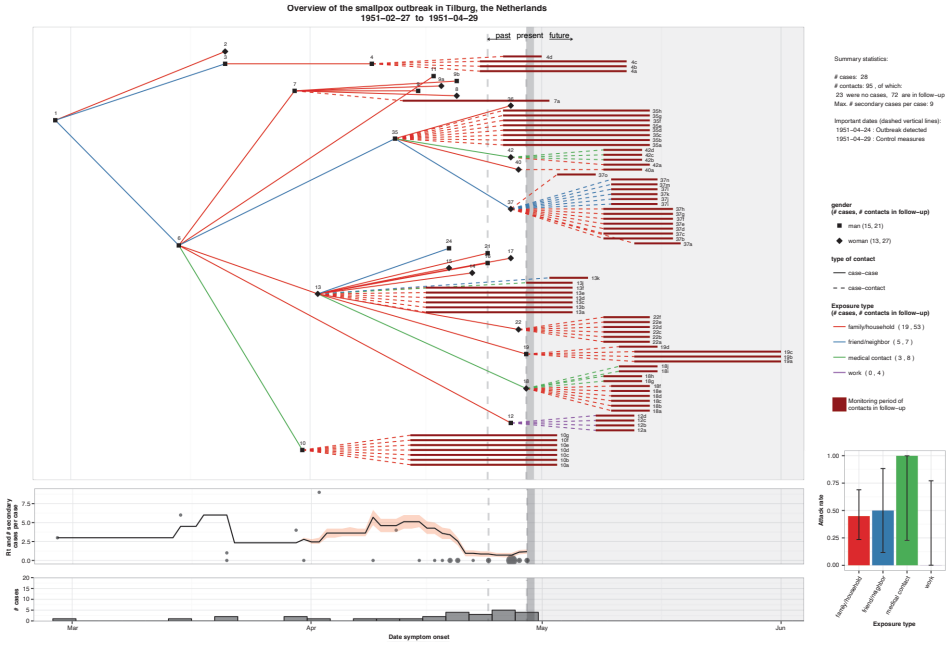


Figure C1. Overview of the simulated outbreak at May 30, 2015

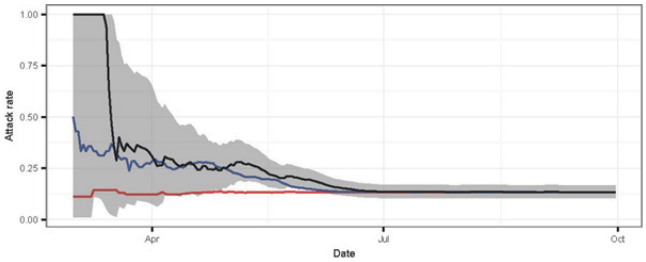


Figure C2. Estimators of the overall attack rates over time. Black: estimator of the overall attack rate over time that uses information on cases in follow-up to correct for right censoring (with 95% CI in grey); red: naive estimator of the overall attack rate over time that is obtained by ignoring any current information on the contacts (no correction for right censoring); blue: reference estimator of the overall attack rate over time that is obtained by including all present and future information on the infections and their contacts (no right censoring).

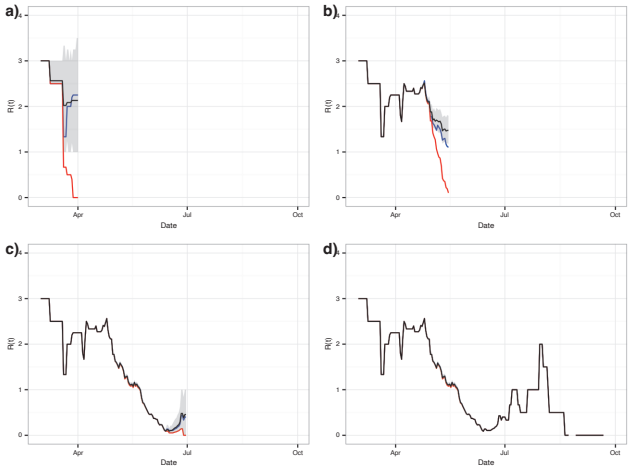


Figure C3: Estimators of effective reproduction numbers over time. a) Situation at 1 April 2015, b) Situation at 15 May 2015, c) Situation at 30 June 2015, d) Situation at 30 September 2015. Black: estimator of the effective reproduction number that uses information on cases in follow-up to correct for right censoring (with 95% CI in grey); red: naive estimator of the effective reproduction number that is obtained by ignoring any current information on the contacts (no correction for right censoring); blue: reference estimator of the effective reproduction number that is obtained by including all present and future information on the infections and their contacts (no right censoring).

```
R code to generate simulated dataset :
# Simulation outbreak
library(outbreaker)
# Simulate data using the outbreaker package
set.seed(10)
dat <- list(n=0)
## simulate data with at least 20 cases
while(dat$n < 20 ) {
dat <- simOutbreak(R0 = 3,  infec.curve = c(dgamma(seq(from =1 ,  to= 40 ,  by=1) ,
17.7)) , n.hosts=  150, duration = 250)}
simdat<- data.frame(ID= dat$id, IDSource=  dat$ances ,  date= dat$onset)
# add additional data for cases
simdat$Type<- c ("Case")
route <- c (rep ("Family/Household" , 31 ), rep("Friend/neighbor", 9),
rep("Medical", 4), rep("Work", 5)) simdat$Exposuretype <- sample(route,
length(simdat$Type), replace= T)
expdurfam <- rgamma (1000, 4)
simdat$Exposureduration <- c ( )
for(i in 1: length(simdat$Exposuretype)){
if(simdat$Exposuretype[i]== "Family/Household") {
simdat$Exposureduration[i]<- round(sample(expdurfam, 1), 0) } else {
simdat$Exposureduration[i]<- 1}}
incperdist<- rlnorm(1000, log(13), log(1.13))
simdat$Expdate<- c( )
for (i in 1: length(simdat$date)){
simdat$Expdate[i]<- simdat$date[i] - round(sample(incperdist, 1), 0)-
simdat$Exposureduration[i]}
#add random number (between 3 and 10) of contacts to every case
for (i in 1 : length(simdat$ID)){
ncontacts <- sample(c(3:10), 1)
id <- c(paste(c(rep(i, ncontacts)), letters[1: ncontacts], sep=""))
source <- c(rep(i, ncontacts))
datesample <- sample(c(-3:5), ncontacts, replace=T)
expdate <- simdat$date[i] + datesample
route <- c(rep("Family/Household", 12), rep("Friend/neighbor", 5), rep("Medical",
1), rep("Work", 4))
exposuretype <- sample(route, ncontacts, replace=T)
exposureduration  <- c()
for (j in 1: length(exposuretype)){
if (exposuretype[j] == "Family/Household"){
exposureduration[j] <- round(sample(expdurfam, 1), 0) } else {
exposureduration[j]<-1}}
contact.df <- data.frame(ID= id, IDSource= source, date= rep(NA, ncontacts),
Type= c(rep("Contact", ncontacts)), Exposuretype = exposuretype, Exposureduration=
exposureduration, Expdate= expdate)
simdat <- rbind(simdat, contact.df)
}
simdat$Gender <- sample(c("man", "woman"), length(simdat$Type), replace=T)
simdat$DOD <- as.Date("2015/03/01", origin= "1970-01-01") + simdat$date
simdat$Exposuredate <- as.Date("2015/03/01", origin= "1970-01-01") + simdat$Expdate
```

Appendix D

OVERVIEW OUTBREAK AT DATE OF IMPLEMENTATION OF CONTROL MEASURES

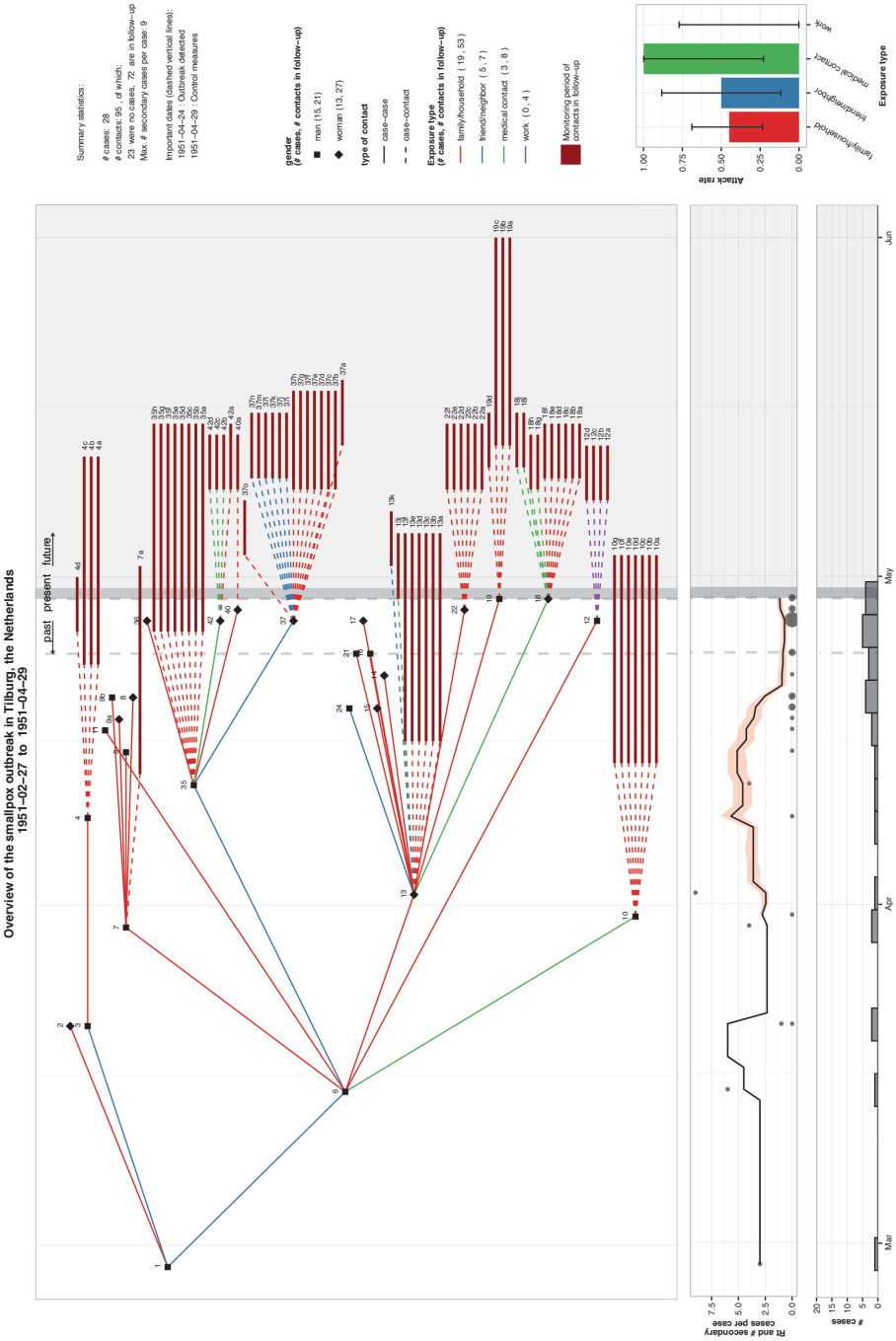


Figure D1: Overview of the smallpox outbreak in Tilburg, the Netherlands at April 29, 1951



6

Visual tools to assess the plausibility of algorithm-identified infectious disease clusters

The contents of this chapter have been published in *Eurosurveillance*:

Visual tools to assess the plausibility of algorithm-identified infectious disease clusters: an application to mumps data from the Netherlands dating from January 2009 to June 2016

Loes Soetens, Jantien A. Backer, Susan Hahné, Rob van Binnendijk, Sigrid Gouma, Jacco Wallinga

Eurosurveillance, March 2019, 24(12)

Abstract

With growing amounts of data available, identification of clusters of persons linked to each other by transmission of an infectious disease increasingly relies on automated algorithms. We propose cluster finding to be a two-step process: first, possible transmission clusters are identified using a cluster algorithm, second, the plausibility that the identified clusters represent genuine transmission clusters is evaluated.

We developed tools to visualise: (i) clusters found in dimensions of time, geographical location and genetic data; (ii) nested sub-clusters within identified clusters; (iii) intra-cluster pairwise dissimilarities per dimension; (iv) intra-cluster correlation between dimensions. We applied our tools to notified mumps cases in the Netherlands with available disease onset date (January 2009 – June 2016), geographical information (location of residence), and pathogen sequence data (n = 112). We compared identified clusters to clusters reported by the Netherlands Early Warning Committee (NEWC).

We identified five mumps clusters. Three clusters were considered plausible. One was questionable because, in phylogenetic analysis, genetic sequences related to it segregated in two groups. One was implausible with no smaller nested clusters, high intra-cluster dissimilarities on all dimensions, and low intra-cluster correlation between dimensions. The NEWC reports concurred with our findings: the plausible/questionable clusters corresponded to reported outbreaks; the implausible cluster did not.

Our tools for assessing automatically identified clusters allow outbreak investigators to rapidly spot plausible transmission clusters for mumps and other human-to-human transmissible diseases. This fast information processing potentially reduces workload.

Introduction

Individual case data originating from routine infectious disease surveillance more and more also include genetic sequence information. With increasing availability of different types of data (e.g. geographical data, time, genetic sequence), each adding their own dimension, and quantities of data rising, transmission-cluster identification of infectious diseases progressively relies on automated algorithms. A transmission cluster can be defined as several cases of an infectious disease which are connected by transmission of this disease from one person to another. A transmission chain is then defined as a series of cases connected by transmission events. Much work has been done on developing algorithms to identify transmission clusters of cases using large datasets [1]. Existing algorithms focus on cluster identification in time [2-9], in space or space-time [10-12], in genetics [13-15], or by combining all three data dimensions [16-18].

A major challenge with clustering algorithms is to balance specificity and sensitivity. If an algorithm lacks specificity, it finds clusters of cases even though there are no transmission events that link them. If it lacks sensitivity, the algorithm does not find genuine transmission chains. To be on the safe side, most algorithms have a high sensitivity at the expense of specificity and as a result also identify clusters of cases that are not genuine transmission clusters. We therefore propose cluster detection using algorithms as a two-step process: (i) detecting possible clusters of infectious diseases with an algorithm and (ii) assessing the plausibility that an identified cluster represents a transmission cluster.

While there has been much work on the first step, little research attention has been paid to methods for improving the plausibility assessment. Currently, identified clusters are usually assessed by epidemiologists who assess information and verify it through communicating with the municipal health services (MHS). This can be quite labour intensive, especially if there are many identified clusters stretching across multiple regions. Only recently, a study has been published that introduced a framework for computing epidemiological concordance of microbial subtyping data of *Campylobacter jejuni* [19]. Epidemiological cluster cohesion is based on time, geographical location, and environmental source distances with adjustable weights. This method requires the computation of a disease specific source distance matrix, making it difficult to apply generically. To our knowledge no further tools are available for careful plausibility assessment of automatically detected clusters.

In order to develop such tools, general characteristics for discriminating transmission clusters from non-transmission clusters have to be identified. We propose to assess the variation of clusters in their time, geographical location and genetic profile. The variation on these dimensions can be visualised by projecting cases on an epidemic curve, map,

and phylogenetic tree, respectively, as well as by estimating the relative distance between clustered cases on these respective dimensions and comparing the distance to the inter-case distances from non-clustered cases. It is assumed that clustered cases will have smaller inter-case distances on these respective dimensions than non-clustered cases. However, there are exceptions: an outbreak may show large variation in time between the occurrence of cases (single persistent source, e.g. typhoid [20]), large variation in geographical distances between cases (initial cases travel large distances, e.g. severe acute respiratory syndrome (SARS) [21]), or include large genetic sequence variation in the pathogen causing the outbreak (fast mutating strains, e.g. Ebola [22, 23]). In order to settle several of these exceptions, intra-cluster correlation between the data dimensions time, geographical location and genetics can be used as another discriminatory characteristic. In genuine transmission clusters, variation on one dimension tends to be correlated with the other dimension. For example, with tuberculosis, cases within a genuine transmission cluster with a larger genetic distance, also tend to have a larger distance in time [24].

The largest hurdle to effectively use algorithms in outbreak investigations is the interpretation of their output, rather than the application of the algorithms themselves. As visualisation techniques support fast processing of large amounts of information, developing tools for visually assessing the plausibility of transmission clusters identified through statistical algorithms may help outbreak investigators [25]. Moreover if data are available in a timely fashion, this may allow pointing outbreak investigators to the most plausible signals first, which, when time is scarce, may facilitate task prioritisation. Finally, outbreak information, such as what is obtained with various available tools (e.g. typical cluster size, typical inter-case distance and correlate estimates between dimensions for a specific disease), might contribute to our current understanding of transmission model parameters [26, 27].

To apply and assess the tools that we develop, we use mumps notification and sequence data reported between 2009 and mid-2016 in the Netherlands. We specifically chose mumps in the Netherlands as it has been intensively studied over the past few years, with comprehensive documentation available [28-33]. In the Netherlands, mumps is a notifiable disease and symptomatic cases are reported to the MHS by physicians and/or laboratories. Cases are either notified when there is a laboratory confirmation or when there is an established epidemiological link with a confirmed case. In case of laboratory confirmation, the national reference laboratory aims to obtain material from regional laboratories for further sequencing. Sequencing provides information on the circulating genotypes and helps to assess whether there is endemic circulation or new introductions of mumps viruses in the country.

Currently, epidemiologists mainly rely on epidemic curves (time data dimension) to detect mumps outbreaks. We set out to assess whether combining geographical location,

time and genetic information can contribute to mumps cluster identification. We use an existing clustering algorithm which can take into account these three data dimensions [16, 17] (hereafter: the time-place-type clustering algorithm), and which is already in use to identify outbreaks of various diseases in the Netherlands, such as meningococcal W disease, meticillin-resistant *Staphylococcus aureus* (MRSA) [17] and echovirus type 6 [34]. We develop and validate visual tools to determine whether identified clusters with this algorithm represent transmission clusters.

Methods

Data

In this study, we include all notified mumps cases in the Netherlands who were diagnosed between 1 January 2009 and 31 May 2016. Notification criteria for mumps include more than one related symptom (i.e. acute onset of painful swelling of the parotid or other salivary glands, orchitis, or meningitis) and laboratory confirmation of infection or an epidemiologic link to a laboratory-confirmed case. The notification criteria did not change during the study period.

For our analysis, for each case we require data on three factors. The first is the disease onset date, which is collected during routine surveillance. The second is the geographical location. The geographical location can be any location that is most relevant for the transmission pattern of the disease under study. For pragmatic reasons, this is usually the location of residence of the case, but this might also be a working address or other place visited. In this study, we used more specifically the latitude and longitude of location of residence of the cases. In the Netherlands, cases' four digit postal code of residence is collected during routine surveillance. We take the centroids of the four digit postal codes and use its latitude and longitude as the input variables for the algorithm. The third required factor consists of the sequences of the small hydrophobic (SH) gene (316 bp), the haemagglutinin/neuraminidase (HN) gene (1,749 bp), and the fusion (F) gene (1,617 bp). Sequences of these three genes are used in combination to distinguish between different mumps genotypes [35], and clusters within genotype G [33]. Since the algorithm that we use is only able to handle cases with data on all three dimensions, cases with missing data on one of the three required factors are excluded from our analysis.

Cluster algorithm

In order to find infectious disease transmission clusters using time, geographical location, and genetic information, Ypma et al. [16] have developed an algorithm to combine pairwise distances between cases on all three data dimensions into one metric. The algorithm sorts cases by relatedness on all three dimensions and subsequently defines a relative distance for all possible pairs of cases reflecting the number of cases found in between the two cases. The relative distances (dissimilarities) for each dimension are calculated, and the combined dissimilarity (d_{combi}) between every pair of cases is then defined as the product of the separate dimension dissimilarities. Next, the cases are joined to form a hierarchical tree of related cases, based on d_{combi}, using single-linkage clustering. For every cluster in the tree, statistical significance of each cluster given its height and cluster size is calculated using permutation. More details on the algorithm are presented in Appendix A.

To demonstrate the tools in this paper, we choose a p value < 0.001 as cut-off level for significance of clusters and consider only the clusters that are not nested within other

identified clusters (hereafter 'highest unnested clusters') at that cut-off level. Since the p value cut-off level is an arbitrary choice, we add flexibility to the tool, by allowing setting cut-offs for p value, maximum tree-height and maximum cluster size.

Assessing the plausibility of clusters representing transmission events

We have developed four tools to assess the plausibility that clusters identified by the time-place-type clustering algorithm represent transmission events. Below we describe each of these four tools in detail.

Overview visualisation of the clusters in the time, geographical location, and genetic dimensions

To assess the variation in time, geographical location and genetic profile of the identified clusters, we visualise the distribution over time by projecting clusters on an epidemic curve (a histogram showing the distribution of cases over time). The distribution across geographical location is visualised by projecting cases coloured according to their cluster membership on a proportional symbol map, in which the point size is proportional to the number of cases at that location. In the interactive version of the tool, this map is replaced by an interactive dot map, which allows for zooming. The distribution across the genetic dimension is visualised by projecting clusters on an arbitrarily rooted maximum likelihood phylogenetic tree. Only the significant highest unnested clusters are visualised, using different colours for every cluster.

Hierarchical clustering tree to visualise the nesting of clusters

Identified clusters can be nested within larger clusters. The structure of the nesting provides valuable information on the strength of the clusters, for example, a cluster that contains several significant clusters at a lower nesting level is stronger than a cluster with no significant clusters at a lower nesting level. Therefore, the structure of the nesting is visualised by providing a hierarchical clustering tree of related cases, a dendrogram, based on d_{combi}. The significant highest unnested clusters are visualised by colouring the end-nodes, and all significant clusters are visualised using black dots at the significant internal nodes.

Intra-cluster pairwise dissimilarity per dimension

To determine the impact of each dimension on d_{combi} (time, geographical location or genetics), we calculate for every significant highest unnested cluster the pairwise dissimilarities per dimension (time, geographical location, genetics, and combined). The pairwise dissimilarities are a measure for intra-cluster variance for the different dimensions. The median dissimilarity is defined as the median of the pairwise dissimilarities (d_{time}, d_{geo}, d_{gen} and d_{combi}) per cluster. We visualise these pairwise dissimilarities using notched boxplots [36]. In a notched box plot, the notches extend 1.58 x interquartile range (IQR) / sqrt(n),

which roughly corresponds to a 95% tolerance interval (assuming a normal distribution). The notches are then used to compare medians, i.e. non-overlapping notches for two different dimensions suggest that the medians are significantly different.

Intra-cluster correlations between the different dimensions

We visualise the intra-cluster correlation of the pairwise dissimilarities between the different dimensions. The intra-cluster correlation provides information on the internal cohesion of a significant cluster, for example, if cases within a cluster are close in time (small d_{time}), are also close in geographical space (small d_{geo}). In addition, the intra-cluster correlation coefficient between the separate dimensions and the combined dimension informs us on the contribution of each dimension to the combined dimension. For every significant highest unnested cluster, we compute the Spearman rank correlation coefficient *(r*) between the pairwise dissimilarities of all dimensions and its p-value [37]. We visualise the strength and direction of the correlation coefficients per cluster using a matrix layout. In the interactive version of the tool, one can hover over the matrices to allow for the correlation coefficients and p-values to pop up.

Epidemiological validation

We use epidemiological information to check the validity of the identified significant highest unnested clusters. The gold standard for confirming transmission links is the presence of an epidemiological link between cases. However, this information is only available for a very small subset of mumps cases and is only described in free text fields, which is difficult to analyse. We therefore use mumps outbreaks described in the reports of the Netherlands Early Warning Committee (NEWC) as gold-standard-identified clusters and assess whether these outbreaks correspond to clusters identified with the algorithm [16]. The clusters that do not correspond with the reported outbreaks are considered false positives. In addition, we check whether the identified clusters are described in the literature.

Analysis with only two dimensions

Since the algorithm cannot handle missing data and since genetic data are often delayed or missing, in Appendix B we show results when using time and geographical location data only.

Availability

All analyses are performed in R 3.4.3 [38]. We have developed an R package *ClusterViz* containing an R shiny app to allow users to interactively set parameter values such as cut-offs for p values, tree heights, and cluster sizes. The R package can be downloaded from a github page (https://github.com/lsoetens/ClusterViz). A demo file for testing the tool is also available with this package (as described below). Considering the genetic data used in this study, all F gene, SH gene and HN gene sequences are submitted to the GenBank database

and are available with the accession numbers KJ125045–51, KJ125053–9, KJ125061–7, and KU756625-930.

Ethical statement

Due to privacy concerns, data on date of diagnosis and geographical location are not published in any public database. Thus we have slightly obfuscated the time and geographical location data and have added the data file as a demo file to the R package. In accordance with Dutch law, no informed consent was required for this study using anonymised routine surveillance data.

Results

Between 1 January 2009 and 30 June 2016, 2,039 cases of mumps were reported in the Netherlands. A sequenced sample of the SH, HN, and F gene was available for 118 (5.8%) of the cases. Of the 118 cases with sequenced data, six had missing geographical data. Therefore, 112 (5.5%) cases were included in the analyses. These cases were mainly male (n = 65; 58.0%) and had a median age of 24 years (IQR: 20–27 years). In this study period, 14 mumps related signals were reported by the NEWC (Table 6.1).

Figures 6.1 to 6.4 represent output from our tool. The algorithm identifies 10 clusters with $p < 0.001$ of which five are nested (Figure 6.1). After collapsing the nested clusters into their parent clusters, five significant highest unnested clusters remain. Of those five highest unnested clusters, cluster 2 (blue, n = 9), 3 (green, n = 12) and 4 (pink, n = 13) contain smaller clusters which are also significant, whereas cluster 1 (red, n = 3) and 5 (orange, n = 28) are not supported by other significant clusters at a lower nesting level.

Table 6.1. Summary of all mumps outbreak reports by the Netherlands Early Warning Committee between January 2009–May 2016 (n=14)

No	Date reported	Reported by	Covering time period	N cases in report	Age range (years)	Remark/ source	Cluster number according to current study
1	09 Apr	RIVM	Aug 2007–Apr 2009	171	NR	Start of nationwide mumps epidemic	4
2	12 Feb	RIVM	Dec 2009–Feb 2012	1,264	NR	Overview of nationwide mumps epidemic	NL
3	12 Apr	GGD Gelderland - Midden	Mar 2012	22	15–26	Party	5
4	12 Jul	GGD Hollands - Noorden	Jul 2012	3	6–8	School	NL
5	12 Aug	GGD Zaanstreek - Waterland	Jul 2012–Aug 2012	21	16–48	Unknown	5
6	13 Feb	Utrecht	Feb 2013	8	NR	Unknown	5
7	13 Jun	GGD Hollands Noorden	Jun 2013	11	23–29	Unknown	NL
8	13 Nov	GGD Zaanstreek - Waterland	Sep 2013–Nov 2013	16	4–47	All living in Volendam	3
9	13 Nov	GGD Groningen	Sep 2013–Nov 2013	13	17–36	Students	NL
10	14 Feb	GGD Zaanstreek - Waterland, GGD Haaglanden	Feb 2014	3	25–30	Work in healthcare setting	NL
11	15 Apr	GGD Haaglanden	Mar 2015–Apr 2015	5	NR	Sports club	2
12	15 Jun	GGD Haaglanden	Apr 2015–Jun 2015	NR	NR	Students linked to school and earlier cluster at sports club	2
13	16 Mar	GGD Brabant Zuidoost	Feb 2016–Mar 2016	6	18–40	Carnival	1
14	16 Apr	GGD Hart voor Brabant	Mar 2016–Apr 2016	6	17–23	Friends/party	1

GGD: Gemeentelijke GezondheidsDienst (Municipal Health Service); NL: no link to a cluster; NR: not reported; RIVM: Rijksinstituut voor Volksgezondheid en Milieu (National institute for public health and the environment).

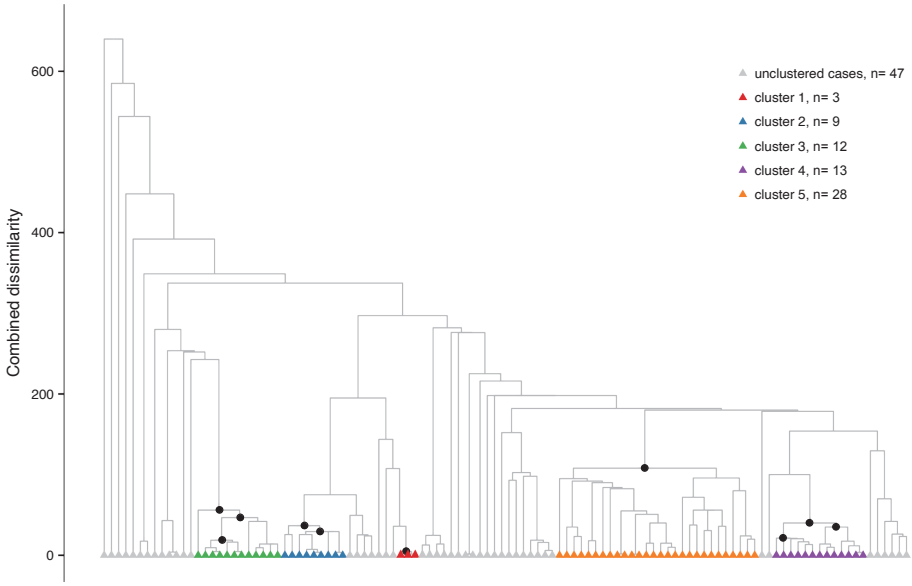


Figure 6.1. Hierarchical clustering tree of the combined dissimilarities of all dimensions[a] for cases of mumps in the Netherlands, January 2009–May 2016 (n=112 cases)
[a] All dimensions include time, geographical location or genetic dimensions.
The colours represent the cases belonging to significant highest unnested clusters. The black dots respresent significant clusters (p<0.001) at all nesting levels.

To assess the plausibility of the clusters for a specific disease, we focus on the variation within clusters across the time, geographical location or genetic dimension. The clusters show differences in how the cases and samples are distributed over time (Figure 6.2a), geographical location (Figure 6.2b), and sequence space (Figure 6.2c). Compared with clusters 4 and 5, cluster 1, 2 and 3 are very compact on all three dimensions (time, geographical location, and genetics). While cluster 4 is relatively concentrated in time and geographical location, it is distributed across two branches of the phylogenetic tree. For mumps this makes it less plausible that all cases belong to the same transmission chain, as the mumps virus is characterised by a very low mutation rate [39]. For each of the two clusters nested within cluster 4 in the hierarchical tree, cases are located on two branches of the phylogenetic tree, suggesting that also the nested clusters contain substantial genetic disparity. Cluster 5 is quite dispersed on all three dimensions (time, geographical location, and genetics), making this cluster very implausible.

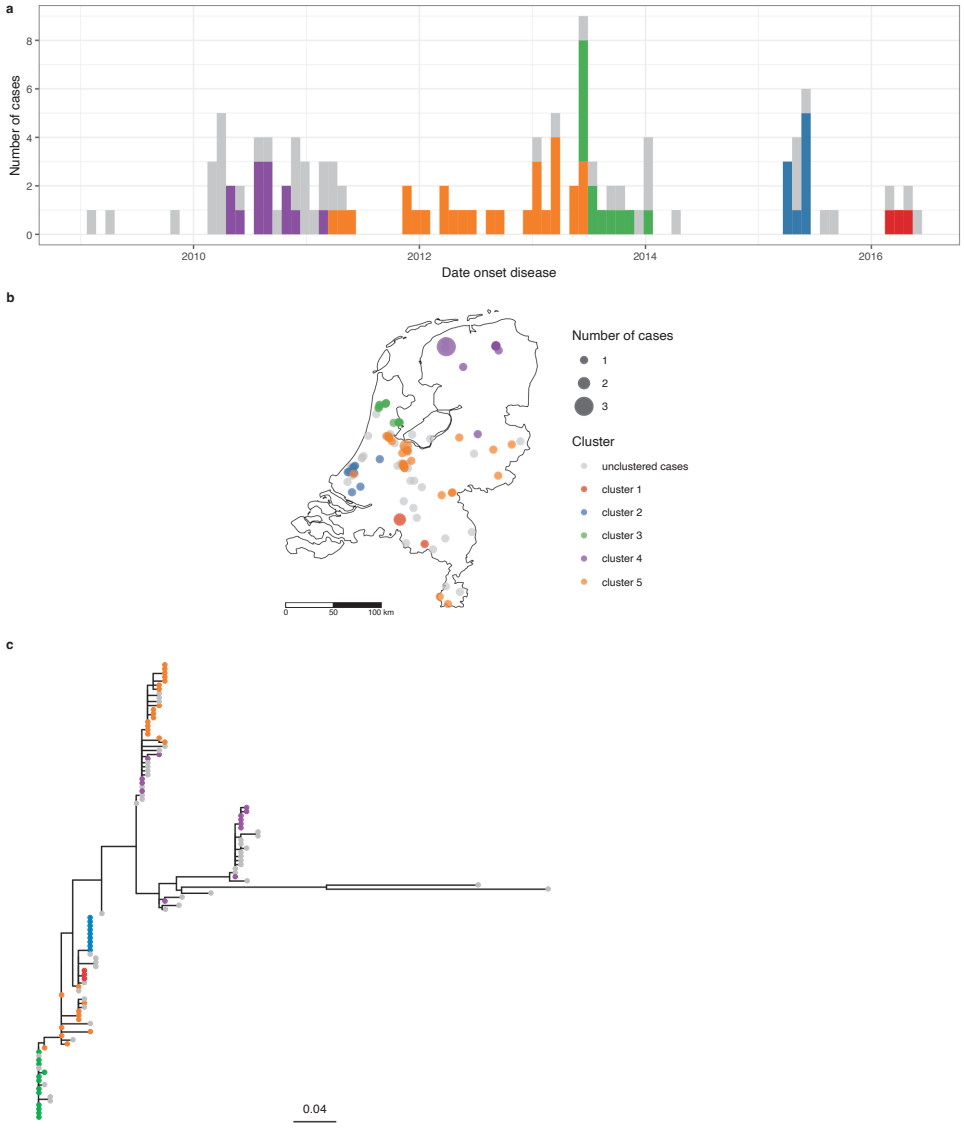


Figure 6.2. Identified clusters of mumps with cases projected on (a) an epicurve (time), (b) maps of the Netherlands (geographical location), and (c) an arbitrarily rooted maximum likelihood phylogenetic tree of the pathogen sequences (genetics), Netherlands, January 2009–May 2016 (n=112 cases). The colours indicate the significant highest unnested clusters identified with the time-place-type algorithm (cluster 1 is red, 2 blue, 3 green, 4 pink, 5 orange); the unclustered cases are depicted in grey. The scale represents the number of nucleotide substitutions per site.

We estimate and visualise for every significant highest unnested cluster the pairwise dissimilarities per data dimension (Figure 6.3). We find that the median pairwise dissimilarity is significantly lower on the combined dimension in all clusters when compared with the

combined pairwise dissimilarity in the unclustered cases. Of the five clusters, cluster 1 has the lowest median pairwise dissimilarities on the three individual dimensions and their combination and cluster 5 has the highest intra-cluster variance on the three individual dimensions and their combination.

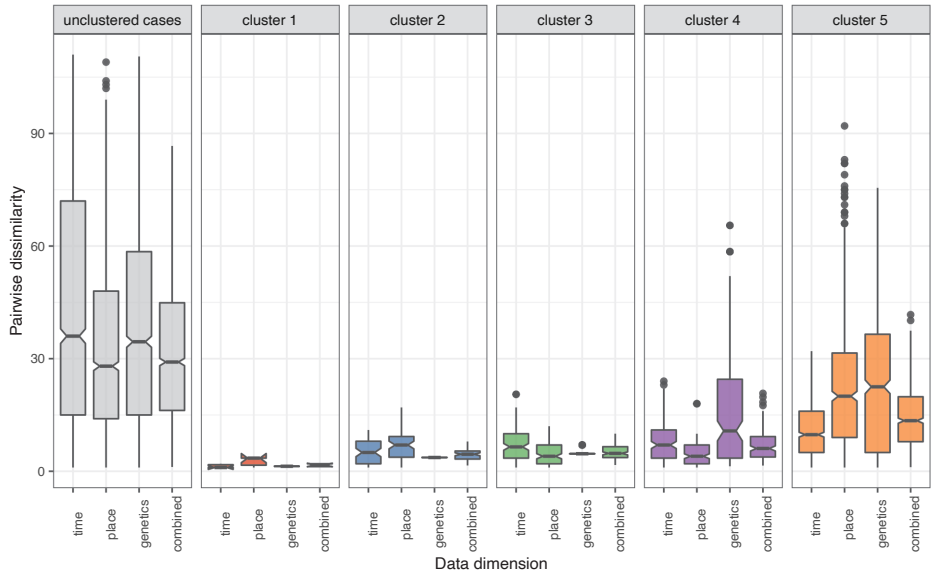


Figure 6.3. Notched boxplots of the pairwise dissimilarities per dimension and cluster of mumps cases, Netherlands, January 2009–May 2016 (n=112 cases)
IQR: interquartile range.
*For the unclustered cases and for each cluster (1–5), the first column in each figure shows the pairwise dissimilarities in time, the second pairwise dissimilarities in geographical location, the third pairwise dissimilarities on genetics, and the fourth shows the combined pairwise dissimilarity. The boxes' height represents the IQR. The horizontal lines in the notches represent the median of the pairwise dissimilarities and the lower and upper bounds of the notches can be interpreted as the 95% tolerance interval. The upper whisker extends from the box to the largest value no further than 1.5 * IQR from the box. The lower whisker extends from the box to the smallest value at most 1.5 * IQR of the box. Black dots indicate outliers.*

We visualise the intra-cluster Spearman rank correlation coefficient (r) of the pairwise dissimilarities between the different dimensions (Figure 6.4). When looking at the correlation coefficients between the data dimensions time, geographical location and genetics, we can see that many correlation coefficients either cannot be estimated due to zero variance (identical sequences) on the genetics dimension (cluster 1 and 2) or are not statistically significant ($p > 0.05$). Only in cluster 3 the time dimension is significantly correlated with the geographical location ($r = 0.4$) and genetics ($r = 0.5$) data dimension, and in cluster 4 and 5 the time dimension is correlated with the genetics dimension only ($r = 0.3$ and $r = 0.2$ respectively). When then looking at the contribution of the individual data dimensions to the combined dimension, we can see that in cluster 1, 2 and 3, the dimension of time and

geographical location contribute equally and strongly to the combined dimension (r={0.9, 0.6, 0.9}), and in cluster 4 and 5 the dimension of genetics contributes the most information to the combined dimension (r={0.8, 0.7}).

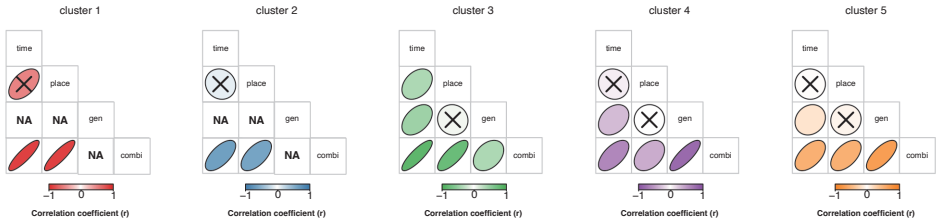


Figure 6.4. Matrix plots of the Spearman intra-cluster correlation between pairwise dissimilarities of all four dimensions for cluster 1 – 5
Combi: combined; gen: genetics; place: geographical location; NA: correlation cannot be estimated. The width of the ellipse indicates the strength of the Spearman rank correlation coefficient r, while the slope of the ellipse indicates the direction of r (positive slope, positive r; negative slope, negative r). The crossed-out ellipses indicate insignificant correlations (p>0.05), and an NA indicates a correlation could not be estimated due to zero variation on one or both dimensions.

As a measure of validity, we have assessed whether mumps outbreaks described in the reports of the NEWC correspond to clusters identified with the time-place-type algorithm. Clusters 1–4 are easily linkable to reported mumps outbreaks of the NEWC (Table 6.1). Given its time, period, and the spatial distribution, cluster 1 corresponds to outbreaks 13 and 14, cluster 2 corresponds to outbreaks 11 and 12, cluster 3 corresponds to outbreak 8, and cluster 4 corresponds to outbreak 1. Cluster 5 is the only identified cluster to which no clear reported outbreaks can be linked. Outbreaks 3, 5, and 6 might together possibly compose cluster 5. In addition to the NEWC reports, clusters 2 and 3 are described in references [33] and [32], respectively.

Finally, analysis using time and geographical location data only (Appendix B) shows that our visual plausibility tools can also be used when data are only available for two dimensions. The cases included in the main analysis are representative for the total notified mumps cases from 2013 onwards, as the shape of the epidemic curves is comparable. However, before 2013 the shapes of the epicurves differ: in the main analysis the large peaks in 2010, 2011 and 2012 cannot be observed. In 2013–16, we identify six clusters using only two dimensions that are similar to those identified using three dimensions, we miss only three minor clusters. In the period before 2013, nine additional clusters are identified in the time-place analysis, of which three are very large (n > 40). The lesser plausible pink (cluster 4) and orange (cluster 5) clusters from the main analysis fall in the less representative period before 2013, so it might be due to unrepresentative sequencing in this period that transmission cluster detection with this algorithm is more difficult.

Discussion

In this paper, we have introduced tools in order to assess the plausibility of transmission clusters. In the mumps case study, five significant clusters are identified, several of which also contain nested clusters. In assessing the plausibility of these significant clusters, the tools that we have developed all point in the same direction: clusters 1 (red), 2 (blue), and 3 (green) can be considered highly plausible; cluster 4 (pink) has moderate plausibility as the sequences related to it span across two branches of the phylogenetic tree; and cluster 5 (orange) has low plausibility. Compared with the other clusters, cluster 5 shows a relatively dispersed pattern across time, geographical location, and genetics; contains no nested clusters; shows relatively high intra-cluster dissimilarity on all dimensions; and shows the lowest intra-cluster correlation between all four dimensions. In our epidemiological validation, no clear reported outbreak can be linked to cluster 5. In contrast, the other four identified clusters are easily linkable to a reported outbreak.

The major advantage of our tools is that we use visualisation techniques to improve assessment of plausibility. Human vision supports fast processing of information [25], allowing for quick decision-making and this can therefore facilitate work for outbreak investigators. Besides fast processing, visualisation also allows for disease-specific characteristics in the assessment of the plausibility. While the first step in cluster detection, which identifies possible transmission clusters, can be done by algorithms, as it is a very generic process, the second step needs disease-specific considerations which cannot easily be incorporated in an algorithm. For example, in the case study our tools show that cluster 4 (pink) and cluster 5 (orange) span across multiple branches of the phylogenetic tree. A mumps expert knows that the mumps virus mutation rate is very low, which decreases the plausibility that these clusters represent unique transmission clusters.

An important aspect of our study is that only 5.5% of notified mumps cases had sufficient genetic information to be included. There are several reasons for this. First, mumps notification does not require laboratory confirmation in the Netherlands, but can also be based on the presence of an epidemiological link to a confirmed case. For these cases no material is available for testing and sequencing. Second, the obtained material is not always suitable for typing; viral loads can be low, which often result in failed sequencing. Third, we specifically chose to include only cases with an available sequenced sample of the SH, HN and F gene. Instead we could also have included cases with a sequenced sample of the SH gene only, as this would have resulted in a higher number of included cases. Nevertheless an earlier study [33] showed that the SH gene alone did not provide sufficient resolution for finding transmission clusters, whereas the combination of the three genes did. Since we aim to find transmission clusters here, including only sequenced samples of the SH gene was not an option. Because of these reasons, it is highly likely that the identified clusters in

this study are actually larger or that some clusters are completely missed by the algorithm, as only one or two cases of a cluster might have a laboratory confirmation. This might explain the clusters reported by the NEWC (report 4, 7, 10 and 11), which were not identified by our tool (Table 6.1).

Our approach can handle incomplete data, such as cases with missing sequences, by performing a partial data analysis as in Appendix B. Complete data analysis is shown in the main text. Further work can focus on extending the algorithm to allow for missing data on one or more dimensions. Especially considering that genetics information will often be missing if sequences are not available, one could replace sequence information by a categorical variable with pathogen subtype or other lower resolution indication of the pathogen type. Cases with a similar pathogen (sub)type would then have a distance of zero versus a distance of >0 to cases with another (sub)type. The (sub)type information, however, should still have sufficient resolution to be able to contribute to transmission cluster detection. Similarly, if geographical location information is not available on the latitude/longitude level, one can think of lower resolution solutions. In this study, we use the centroids of the four digit postal codes as geographical location information. The information should have sufficient resolution to be informative. The tool is not limited by the number or type of dimensions. The addition or reduction of dimensions only requires small adaptions in our code and it is therefore straightforward to use our tools in combination with other algorithms, for example, space-time algorithms [10]. By increasing or decreasing the number of dimensions in the algorithm, the relative weight of the included dimensions decreases or increases as well, respectively. Depending on the quality of the data from the additional or removed sources, this may not be desirable. We have specifically chosen not to put weights on the separate dimensions, as determining the size of the weights is a very arbitrary decision. Instead, in the current study, we would rather interpret a cluster, which was primarily identified on the geographical location dimension, as less plausible, as the geographical location data are considered quite unreliable for mumps in the Netherlands. Indeed, information on place of residence (geographical location) is of questionable accuracy as mumps mainly occur among students who often have more than one living address (near the university and their parents' address). It is then often not clear if the students actually live on the reported address at the time of the outbreak. On the other hand, if we would have had reliable geographical location information, other or more clusters might have been identified that now go undetected. Similarly, for mumps it is very unlikely that a transmission cluster is spread across multiple branches in the phylogenetic tree, so for mumps the genetic dimension might have more relevance than e.g. the geographical one. Instead of putting a weight on this dimension however, we considered cluster 4 and 5 as less plausible. Further work could investigate ways of determining the size of the weights, based either on the quality of the data (as discussed here) or on the type and transmission routes of the disease under investigation. A final issue regarding the internal correlation plots is that they are

more useful when cluster sizes are larger. If the algorithm detects very small clusters ($n < 4$), we suggest to rely on the other tools to determine whether an identified cluster is plausible. To conclude, our proposed tools for assessing plausibility of automatically identified clusters in time, geographical location and genetic dimensions can help outbreak investigators to focus on the most plausible clusters first. Timely availability of data are a prerequisite for this. In addition, using visual tools allow for fast and efficient information processing, which facilitates work. Mumps serves as an example in this study, but the algorithm can be transferred to other human-to-human transmissible diseases.

References

1. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2012;175(1):49-82.
2. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. Journal of the Royal Statistical Society Series A (Statistics in Society). 1996;159(3):547-63.
3. Hutwagner LC, Maloney EK, Bean NH, Slutsker L, Martin SM. Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks. Emerging infectious diseases. 1997 Jul-Sep;3(3):395-400.
4. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in medicine. 1999 Dec 30;18(24):3463-78.
5. Stroup DF, Williamson GD, Herndon JL, Karon JM. Detection of aberrations in the occurrence of notifiable diseases surveillance data. Statistics in medicine. 1989;8(3):323-9.
6. Nobre FF, Stroup DF. A Monitoring System to Detect Changes in Public Health Surveillance Data. International Journal of Epidemiology. 1994;23(2):408-18.
7. Stern L, Lightfoot D. Automated outbreak detection: a quantitative retrospective analysis. Epidemiology and infection. 1999 Feb;122(1):103-10.
8. Bédubourg G, Le Strat Y. Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study. PloS one. 2017;12(7):e0181227.
9. Salmon M, Schumacher D, Höhle M. Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance. 2016. [R; surveillance; outbreak detection; statistical process control]. 2016 2016-05-18;70(10):35 %J Journal of Statistical Software.
10. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS medicine. 2005 Mar;2(3):e59.
11. Watkins RE, Eagleson S, Veenendaal B, Wright G, Plant AJ. Disease surveillance using a hidden Markov model. BMC medical informatics and decision making. 2009 Aug 10;9:39.
12. Hossain MM, Lawson AB. Space-time Bayesian small area disease risk models: development and evaluation with a focus on cluster detection. Environmental and ecological statistics. 2010 Mar 1;17(1):73-95.
13. Ragonnet-Cronin M, Hodcroft E, Hue S, Fearnhill E, Delpech V, Brown AJ, et al. Automated analysis of phylogenetic clusters. BMC bioinformatics. 2013 Nov 6;14:317.
14. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology. 2014 Apr;10(4):e1003537.
15. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart TJBB. outbreaker2: a modular platform for outbreak reconstruction. [journal article]. 2018 October 22;19(11):363.
16. Ypma RJ, Donker T, van Ballegooijen WM, Wallinga J. Finding evidence for local transmission of contagious disease in molecular epidemiological datasets. PloS one. 2013;8(7):e69875.
17. Donker T, Bosch T, Ypma RJ, Haenen AP, van Ballegooijen WM, Heck ME, et al. Monitoring the spread of meticillin-resistant Staphylococcus aureus in The Netherlands from a reference laboratory perspective. The Journal of hospital infection. 2016 Aug;93(4):366-74.
18. Cori A, Nouvellet P, Garske T, Bourhy H, Nakoune E, Jombart T. A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. PLoS computational biology. 2018 Dec 17;14(12):e1006554.
19. Hetman BM, Mutschall SK, Thomas JE, Gannon VPJ, Clark CG, Pollari F, et al. The EpiQuant Framework for Computing Epidemiological Concordance of Microbial Subtyping Data. Journal of clinical microbiology. 2017 May;55(5):1334-49.
20. Keddy KH, Sooka A, Ismail H, Smith AM, Weber I, Letsoalo ME, et al. Molecular epidemiological investigation of a typhoid fever outbreak in South Africa, 2005: the relationship to a previous epidemic in 1993. Epidemiology and infection. 2011 Aug;139(8):1239-45.
21. Yu IT, Li Y, Wong TW, Tam W, Chan AT, Lee JH, et al. Evidence of airborne transmission of the severe acute respiratory syndrome virus. The New England journal of medicine. 2004 Apr 22;350(17):1731-9.
22. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nature genetics. 2013 Jul;45(7):784-90.

23. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (New York, NY). 2014 Sep 12;345(6202):1369-72.
24. Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar IJBM. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. [journal article]. 2016 March 23;14(1):21.
25. Ware C. Information Visualization: Perception for Design: Elsevier Science; 2004.
26. Jansen VA, Stollenwerk N, Jensen HJ, Ramsay ME, Edmunds WJ, Rhodes CJ. Measles outbreaks in a population with declining vaccine uptake. Science (New York, NY). 2003 Aug 8;301(5634):804.
27. De Serres G, Gay NJ, Farrington CP. Epidemiology of transmissible diseases after elimination. American journal of epidemiology. 2000 Jun 1;151(11):1039-48; discussion 49-52.
28. Sane J, Gouma S, Koopmans M, de Melker H, Swaan C, van Binnendijk R, et al. Epidemic of mumps among vaccinated persons, The Netherlands, 2009-2012. Emerging infectious diseases. 2014 Apr;20(4):643-8.
29. Gouma S, Sane J, Gijselaar D, Cremer J, Hahne S, Koopmans M, et al. Two major mumps genotype G variants dominated recent mumps outbreaks in the Netherlands (2009-2012). The Journal of general virology. 2014 May;95(Pt 5):1074-82.
30. Ladbury G, Ostendorf S, Waegemaekers T, van Binnendijk R, Boot H, Hahne S. Smoking and older age associated with mumps in an outbreak in a group of highly-vaccinated individuals attending a youth club party, the Netherlands, 2012. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2014 Apr 24;19(16):20776.
31. Greenland K, Whelan J, Fanoy E, Borgert M, Hulshof K, Yap KB, et al. Mumps outbreak among vaccinated university students associated with a large party, the Netherlands, 2010. Vaccine. 2012 Jun 29;30(31):4676-80.
32. Whelan J, van Binnendijk R, Greenland K, Fanoy E, Khargi M, Yap K, et al. Ongoing mumps outbreak in a student population with high vaccination coverage, Netherlands, 2010. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2010 Apr 29;15(17).
33. Gouma S, Cremer J, Parkkali S, Veldhuijzen I, van Binnendijk RS, Koopmans MPG. Mumps virus F gene and HN gene sequencing as a molecular tool to study mumps virus transmission. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2016 Nov;45:145-50.
34. Monge S, Benschop K, Soetens L, Pijnacker R, Hahne S, Wallinga J, et al. Echovirus type 6 transmission clusters and the role of environmental surveillance in early warning, the Netherlands, 2007 to 2016. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2018 Nov;23(45).
35. Jin L, Orvell C, Myers R, Rota PA, Nakayama T, Forcic D, et al. Genomic diversity of mumps virus and global distribution of the 12 genotypes. Reviews in medical virology. 2015 Mar;25(2):85-101.
36. McGill R, Tukey JW, Larsen WA. Variations of Box Plots. The American Statistician. 1978;32(1):12-6.
37. Best DJ, Roberts DE. Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. Journal of the Royal Statistical Society Series C (Applied Statistics). 1975;24(3):377-9.
38. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, 2017.
39. Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. Journal of Molecular Evolution. [journal article]. 2002 February 01;54(2):156-65.

Appendix A

ALGORITHM TO FIND INFECTIOUS DISEASE CLUSTERS USING TIME, PLACE AND GENETIC INFORMATION

In order to find infectious disease clusters using time, place, and genetic information, Ypma et al [1] have developed an algorithm to combine into one metric pairwise distances between cases on all three data dimensions. The algorithm sorts cases by relatedness on all three dimensions and subsequently defines a relative distance for all possible pairs of cases reflecting the number of cases found in between the two cases. For the time dimension (d_{time}), this is defined as the number of cases with a disease onset date between the disease onset dates of the two cases. For the place (d_{geo}) and genetic (d_{gen}) dimensions, cases are sorted by Euclidean distance and number of point mutations respectively to define pairwise relative distances. The relative distances (dissimilarities) for each dimension are calculated, the combined (time, place, genetic) dissimilarity (d_{combi}) between every pair of cases is then defined as the product of the separate dimension dissimilarities: $d_{combi} = d_{time} \times d_{geo} \times d_{gen}$. Next, the cases are joined to form a hierarchical tree of related cases, based on d_{combi}, using single-linkage clustering, i.e. a bottom-up approach that sequentially connects cases with the smallest d_{combi}. We use single-linkage clustering as it very well resembles the chain- like structure of transmission clusters [1]. The result of this step is a tree in which the height represents d_{combi}.

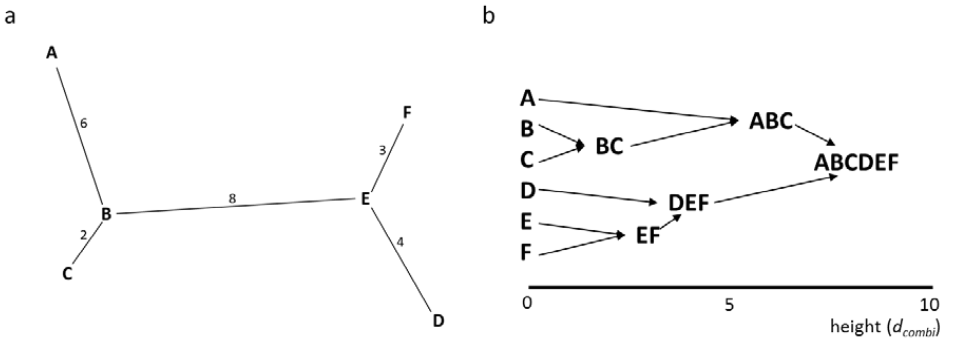


Figure A1. Graphical representation of hierarchical clustering (adapted from Ypma et al [1]). Cases (A-E) are sequentially connected by smallest distance d_{combi} (a) to form a hierarchical clustering tree in which height represents d_{combi} (b)

Next, for every cluster (i.e., dissection) in the tree, statistical significance of each cluster given its height and cluster size is calculated using permutation as follows: (i) the relative dissimilarities between pairs of cases in each dimension are permuted (random sampling without replacement), (ii) d_{combi} for every pair of combinations is computed, (iii) hierarchical clustering is repeated, (iv) for every cluster in the original tree, it is recorded whether a cluster of same height and size exists in the permuted tree, and (v) steps i to iv were

repeated 10,000 times.

Finally, the p-value for each cluster is calculated as the number of times a cluster of the same height and at least the same size is found in the permuted trees divided by 10,000.

To demonstrate the tools in this paper, we choose a p-value < 0.001 as cut off level for significance of clusters and consider only the clusters that are not nested within other identified clusters (hereafter “highest unnested clusters”) at that cut off level. However, the p-value cut off level is an arbitrary choice, and we can imagine that the users of the tool would like to experiment with other choices. We therefore add flexibility to the tool, by allowing setting cut-offs for p-value, maximum tree-height and maximum cluster size.

References

1. Ypma RJ, Donker T, van Ballegooijen WM, Wallinga J. Finding evidence for local transmission of contagious disease in molecular epidemiological datasets. PloS one. 2013;8(7):e69875

Appendix B

ANALYSIS WITH TIME AND GEOGRAPHICAL LOCATION DIMENSIONS ONLY

In the main paper we describe the analysis and results when including three data dimensions: time, place and genetic data. As genetic (sequence) information is often lagging behind or missing and to demonstrate the use of the algorithm with only two dimensions, we perform a similar analysis in this appendix with time and place data only.

We have made some minor adaptations to the algorithm: instead of multiplying with three dimensions, we multiply with only two dimensions.

Between January 1st 2009 and June 30th 2016, 2,039 cases of mumps were notified in the Netherlands. Of those, 103 cases have missing information on location of residence and therefore 1,936 cases (94.9%) are included in this analysis (compared to 5.5% in the main article).

When repeating the analysis with similar settings as in the main article (p-value cut-off <0.001, clusters at highest nesting level), the results show only one large cluster encompassing almost all cases (n= 1,748, 90.3%). Plausibility of this cluster is very low given most indicators: it shows a very dispersed pattern across time and place, shows a high and almost indistinguishable intra-cluster variance on all dimensions compared to the unclustered cases, and shows no intra-cluster correlation between the time and geographic dimensions. Since this cluster has many underlying significant clusters at a lower nesting level, we decide to use different cut-off levels for the maximum tree height of the “highest unnested clusters”. We gradually lower this maximum height from 100% downwards, until we obtain clusters that are plausible considering our indicators. This is at a maximum tree height of 16% (height = 132) of the original tree height (height = 825), and these results are shown in Figure B1.

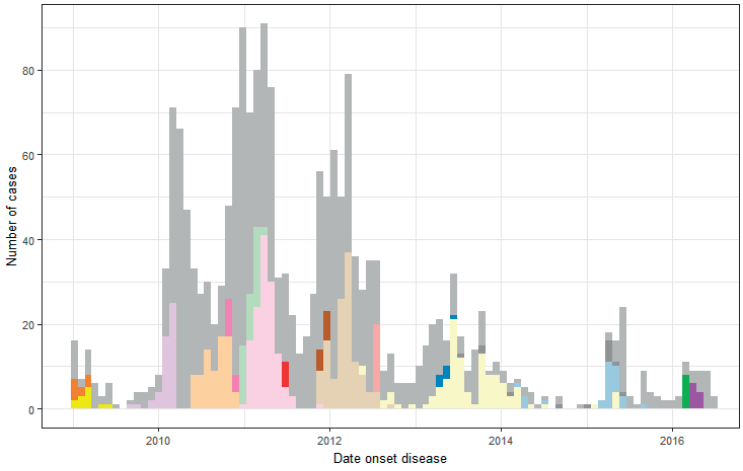
With these settings 17 possible transmission clusters are detected of various sizes (clustersize range: 6-134). These 17 clusters are all quite plausible given most indicators, although the intra-cluster correlation coefficient between time and place dimensions is relative low or nonexistent for all clusters.

The identified clusters in the main article can also be observed in this analysis. The red cluster from the main article corresponds with cluster 3, blue with cluster 11, green with cluster 15, pink more or less with cluster 14 and orange more or less with cluster 16. The number of cases attributed to the clusters identified in the main article is larger. We notice that the lesser plausible clusters in the main article (pink and orange) do not agree completely to cluster 14 and 16 in this analysis. We can also notice that, when we compare

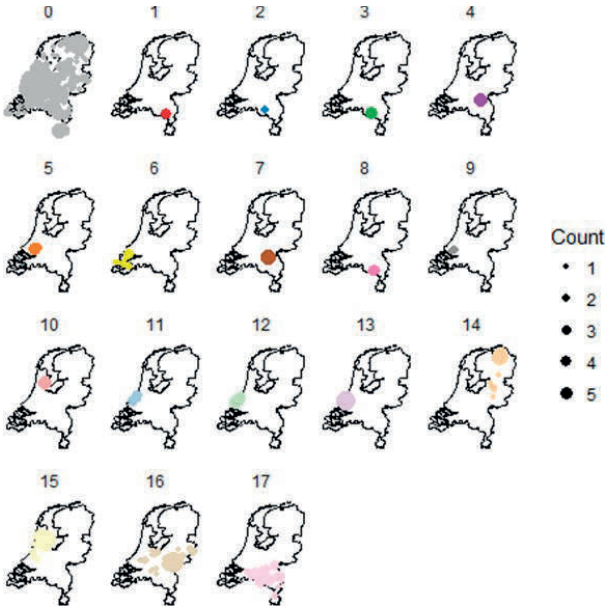
the epidemic curve with the epidemic curve in the main article, from 2013 onwards the pattern is roughly similar, which might indicate fairly representative sequencing in that time period, but is quite different in the period before 2013. We can see some large peaks in 2010, 2011 and 2012, which are not depicted in the epidemic curve of the main article. The lesser plausible pink and orange cluster fall in this period before 2013, so it might be due to unrepresentative sequencing in this period that transmission cluster detection is difficult. This might also be the reason that the quite large (n> 40) clusters 12, 13 and 17 are not picked up in the main article.

With this additional study we show that it is quite well possible to use these plausibility tools with two dimensions only. We have also discovered that the cases included in the main analysis are quite representative for the total notified mumps cases from 2013 onwards, and that in this period similar clusters are detected if only two dimensions would have been used. We conclude that biased sequencing can influence cluster detection with this algorithm, so its therefore always important to aim for a representative sample of the total population for sequencing. In addition, this analysis also shows that even with a small representative sequenced sample of the total population, a great amount of cluster information is captured.

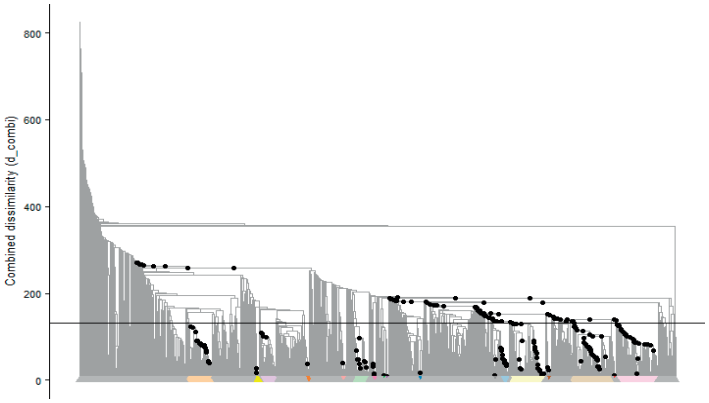
a)



b)

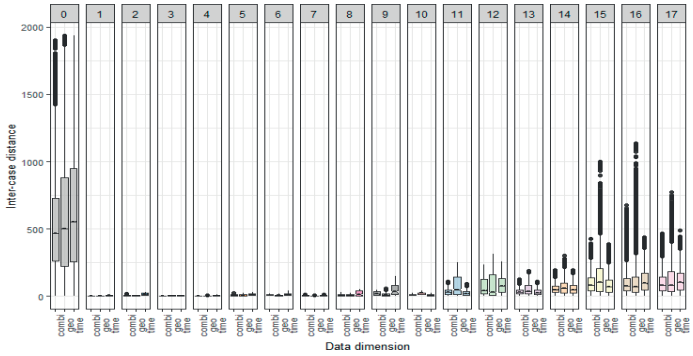


c)



Tree height cut-off at 16%

d)



e)

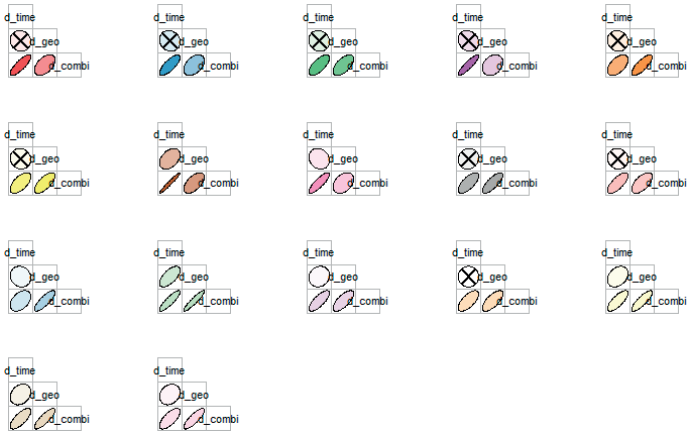


Figure B1. Results of the cluster analysis using a p-value cut-off of 0.001 and only considering the clusters at the highest nesting level with a maximum tree height of 16% of the total tree height. a) epidemic curve, b) map of unclustered and clustered cases, c) hierarchical clustering tree, d) notched boxplots of the inter-case distance per dimension and cluster, e) matrix plot of the Spearman intra-cluster correlation coefficients.



7

General discussion

The objective of this thesis was to develop and improve new methods for exploratory data analysis (EDA) in infectious disease surveillance. The increasing availability of infectious disease data poses interesting questions concerning patterns and structure in data, which so far have received little attention. In this thesis, we propose methods to analyze data available at multiple resolution levels and to combine different data types in infectious disease surveillance.

Regarding data availability at multiple resolution levels, we proposed a new method for a more objective and more scale-independent display of disease incidence data on a map (Chapter 3). We used dot map cartograms to simultaneously display absolute as well as relative numbers of disease incidence. We showed its added value by applying the method to incidence data of Q fever in the Netherlands and pertussis in Germany. Also for the time data type we proposed a method for data analysis at multiple resolution levels (Chapter 4). We used entropy measures at various aggregation scales to identify intrinsic data patterns in a large database of all infectious disease notifications in the Netherlands since 2003. We showed that these intrinsic data patterns could be used to characterize and sort infectious diseases by their dynamic properties, which enhances surveillance outputs.

The second theme addressed in this thesis focuses on combining different data types in infectious disease epidemiology. We used the time, place, person and pathogen data types to form hypotheses for the increasing incidence of hepatitis B in certain parts of the Netherlands (Chapter 2). We studied the time and person data types in contact tracing data and developed an interactive tool that shows whether a person exposed to a case is still at risk for developing symptoms (Chapter 5). And finally, we improved a previously developed method for using time, place and pathogen data types to identify infectious disease clusters (Chapter 6). We applied this method to identify clusters of mumps cases in the Netherlands. Here, we developed an additional interactive visualization tool to assess model performance of this algorithm, in which clustering parameters can be varied in order to help public health professionals with the interpretation and selection of the model with most plausible clusters according to expert knowledge.

In developing these methods, we use central data science techniques, such as considering the data problem within a process pipeline, starting with data collection and ending with a data product, and techniques for combining different data types.

The above-mentioned main themes will be further discussed in this chapter. The first theme is the use of multiple resolution levels within one data type, and the second theme concerns the combination of multiple data types. This is followed by a paragraph on ethical considerations, and a paragraph with recommendations. This chapter is finalized with a conclusion regarding EDA use in infectious disease surveillance.

Using multiple resolution levels withing one data type

One of the issues regarding the increasing availability of data in the infectious disease domain is that data, especially for the pathogen data type, can now be obtained at an increasingly detailed level of resolution, which can then be aggregated to a wide range of resolutions. So far, little attention has been paid to the question how to make a choice between options for these multiple resolution levels within a data type.

In Chapters 3 and 4, we addressed the issue of how to deal with the increasing availability of infectious disease data at multiple resolution levels for the place and time data types, and we sought the answer in developing methods and applying measures that combine or study effects across resolution scales. Intrinsic data patterns unrelated to aggregation levels are then revealed. In Chapter 3, we used visualization techniques to develop dot map cartograms, and in Chapter 4 we proposed Rényi entropy as a robust measure to summarize time data of infectious diseases. Visualization techniques and entropy measures are often both used in the data science domain. There are other ways to study data availability at multiple resolution levels. One approach is to simply apply traditional mapping methods or time visualizations at multiple resolution scales. However, when the outcome depends on the resolution scale and this scale is chosen arbitrarily, then the outcome is likely to be an artefact. When intrinsic data patterns unrelated to aggregation levels can be studied, it is then more likely that any identified pattern indicates interesting dynamic changes instead of aggregation artefacts.

In this thesis, we have focused on how data science techniques can offer a solution for resolution and scaling issues in infectious disease epidemiology. These scaling issues are however not limited to the infectious disease domain. In systems biology and ecology, it is widely recognized that the response of an organism or species can only be observed at a specific scale. It is therefore very important to choose the appropriate scale for addressing a certain hypothesis. This was emphasized in one of the most influential papers to this field ‘On the problem of pattern and scale in ecology’ by Simon Levin in 1992 [1]. Methods to address issues of scale are therefore numerous in this field; ranging from identifying power law relations and fractal dimensions to wavelet transform analysis. The power of these methods lies in their ability to describe how patterns change across scales [1]. In Chapter 4 of this thesis, we refer to these methods handling scale issues, and we use entropy to study how patterns change across scales. Although ecology and infectious disease epidemiology cover similar data types (time and place), in contrast to ecology, hardly any attention has been given to scale decisions for analysis in the infectious disease domain. Wavelet analysis and related methods have been used to study patterns in infectious disease data [2-5]; however, this work has so far not resulted in widescale awareness regarding scale issues in this field. There is much to be learned from fields such as ecology regarding this topic,

which can offer a good starting point for further research in this area. Also, further research into barriers to adoption of these methods in infectious disease epidemiology would be useful. Besides ecology, similar methods are also used in other fields involving hierarchical systems in time and place, such as in geography and demography. For example, entropy and power laws can be used to understand the structure of cities [6] and population growth in urban areas [7].

Combining multiple data types

The second main theme addressed in this thesis is how to deal with increasing data availability across data types. In this thesis we discern four data types for infectious disease epidemiology: time, place, person and pathogen. To obtain, combine and analyze these data types in order to discover patterns and generate hypotheses from the information is not straightforward, and, as highlighted earlier, lacks the necessary attention in infectious disease epidemiology. We have seen that with joint analysis of multiple data types in exploratory infectious disease epidemiology, other questions arise at various stages in the data science pipeline.

First, there is a challenge in combining data types in the data collection phase. Data of different types are often not present in the same database. In Chapter 2 and in Chapter 6 epidemiological data and pathogen data were obtained from different databases. Case records have to be matched one by one. For this task, it would have been very helpful if the same unique patient identifier was available across systems. Currently this is not possible in the Netherlands due to privacy regulations. Accurate linking of records is important, as the quality of the input data determines the quality of the data product at the end of the process. In addition, in the same chapters, we observed that data of all data types are not available for all patients. In particular, pathogen type data are often lacking or delayed. If all data types are jointly analyzed, this can result in substantial data loss (sometimes even 95%, see Chapter 6). Solutions for missing data on one type might be sought in the analysis phase, as shown in this thesis, but more effort should preferably be put into measuring these missing data. In infectious disease epidemiology, this would mostly mean increasing the availability of the pathogen data type.

Second, there is the challenge of actually combining data types for analysis. We have seen in Chapter 2 that with two-way combinations of data types we can still rely on traditional epidemiological methods, such as space-time clustering methods [8, 9] for time-place combinations and phylogenetic analysis [10] for pathogen-time, pathogen-place and pathogen-person combinations. However, with three-way combinations, we had to use a scale-independent method to combine distances in time, geographic location and genetics. As for missing data on one or more dimensions, flexible adaption of the techniques to include two instead of three data types, as shown in Chapter 6, is one solution.

The final step in the data science pipeline is the development of data products. A challenge here is how to present results of combined data type analysis in such a way that it is most useful for the recipient. We have learned (and shown in Chapters 5 and 6) that data visualization techniques take an important position here, and we have used these techniques to develop interactive data products. By using these products in daily practice, we can acquire more information on the user-friendliness of the tools, and can subsequently continually improve upon these products.

Ethical considerations

With the implementation of the European privacy regulations [11], the General Data Protection Regulation (GDPR), in May 2018, a lot of public attention in the Netherlands has been devoted to how data are stored, collected and disseminated. This regulation illustrates the importance of careful treatment of (personal) data, including infectious disease data. Given the increasing resolution and availability of infectious disease data, we have to be aware of not infringing on the privacy of the patient. Data collection for surveillance purposes is arranged by a special law in the Netherlands, the public health act [12]. This means that informed consent of the patient is not needed, and that the data collection is not overseen by an ethical board. Therefore, surveillance data have to be analyzed anonymously, so that patients cannot be identified and privacy is protected. However, this is a rather grey area, as the question "can the patient be identified?" depends on the type and amount of available information. For example, if we know that a patient lives in a certain four-digit postal code area, we would say that the patient is not identifiable, as the area is often too large. However, he or she might be identifiable if the area is very sparsely populated, a very rare disease is involved, or if other patient characteristics are known as well. Therefore, with increasing resolution of infectious disease data by combining more and more data types, we increase the risk of identifying the patient and therefore infringing on his or her privacy. This is a risk we should avoid when collecting information. Also, not all data types that we have discussed pose an equal risk of disclosing identifiable information. Disease onset date is generally not regarded as identifiable information, whereas a patient's home address is. Similarly, a pathogen's genetic code cannot be used to identify a patient, whereas a combination of personal characteristics, such as age, sex, and risk group can be identifiable. The place data type definitely poses the greatest risk for infringing on the privacy of the patient. Therefore, it is a great advantage that in our method for displaying infectious disease data on a map in Chapter 3, the visual distortion of the dot map cartogram makes it hard to exactly pinpoint the location of a case, which consequently protects the patient's privacy. By proposing a method that also protects privacy, we align with the public debate concerning data and privacy issues.

Recommendations

Based on the research presented in this thesis, we can make several recommendations and propose directions for further research. We will structure these along the steps in the data science process.

Data collection

Good quality data is the cornerstone of all further steps in the surveillance process. When multiple data systems are combined to enhance surveillance goals, a few challenges lie ahead.

First, we have seen in our studies in Chapters 2 and 6 that records across systems had to be matched one by one to the most likely record, as a central unique identifier was not in place. Such an identifier would reduce linking errors and therefore enhance data quality. In the Netherlands, this problem is especially present when combining epidemiological data with laboratory data, as these originate from different systems.

We have shown in this thesis that it is important to combine these data sources to find patterns in transmission and to form hypotheses regarding unexplained phenomena. We therefore recommend to find a solution for more valid linking of records between these systems. Matching all data based on the personal identification number (burgerservicenummer, BSN) would be the ultimate solution. This is done in other European countries, such as Denmark and Sweden; however, this is not possible in the Netherlands due to how the privacy regulations are implemented in Dutch law. Data merging can be achieved in the Netherlands by making use of a trusted third party, such as Statistics Netherlands; however, this is rather cumbersome for surveillance purposes which involves ongoing data collection, and thus would require ongoing data merging. Estonia goes one step further and has a complete e-government system, including health data, in which personal data can easily be transferred from one party to another using block chain technology. Further research could be focused on how to be able to combine all these data at an individual level, while still respecting a patient's privacy and following Dutch and European law. New technologies might be the answer for this problem.

Second, pathogen data from regional laboratories are most often obtained and stored for clinical diagnosis rather than for surveillance or research purposes. This means that data are normally not standardized between laboratories, and major effort is required to combine data from multiple laboratories and/or epidemiological data in a structured manner. For example, each laboratory might use a different written form of the pathogen name. Luckily, this is rapidly improving in recent years thanks to large data management projects for data standardization between laboratories using international standards, such

as the Dutch 'Eenheid van Taal' project [13]. The next challenge in this direction is how to systematically collect, exchange and store pathogen data at a high resolution level (such as whole genome sequences) for surveillance purposes, as these data are stored in quite large files, which will rapidly build up as more data are collected. In Chapter 6 of this thesis, we showed that only a small proportion of all epidemiologically confirmed mumps cases had a sequenced sample available. This was largely due to the fact that positive samples from the regional laboratory are not systematically sent to the national reference laboratory for sequencing. How to improve data collection and management for this type of data at a national and international level, especially for surveillance purposes, should be a direction for future research and policy making. This would require a closer look as to how best to incorporate regional and national laboratories in the surveillance process.

Another challenge is to combine surveillance data from multiple countries to examine patterns across countries. This is especially important for a small country such as the Netherlands, as infectious diseases do not remain within international boundaries. Trends or outbreaks occurring in Belgium, Germany and the UK are relevant to assess the infectious disease risks in the Netherlands. In other words, it is important to not only increase resolution scale of the time and place data types, but also to extend this resolution scale across a larger area (Europe, the world). Currently, data sharing is often achieved through personal contact with a researcher or public health officer at another national institute. Ideally, these data would be provided in an open source platform for use by other countries. Naturally, these data should then be aggregated, to protect personal information. A good example of such a platform is the Survstat application developed by the Robert Koch Institute in Germany [14]. We used data from this platform in Chapter 3, to illustrate the application of dot map cartograms in other countries. A recommendation for the European Centre for Infectious Disease Control (ECDC) and for the national institutes for public health (including the Netherlands) would be to stimulate development of such platforms for timely data sharing at a high resolution level within European boundaries, and to ultimately combine these data to allow for pattern detection across borders.

Finally, in this thesis we have focused on four infectious disease data types, obtained during standard surveillance processes. However, many other data sources are available that could possibly enhance surveillance. For example, antenna or gps data collected via mobile phones of cases would greatly improve the resolution of spatial location data. If we could use this source to track a patient's past whereabouts , it might be possible to use these data for source finding in outbreak investigations [15-18]. Currently, we are running a pilot project to examine the usefulness of this type of data in relation to legionella source finding. Other sources one could think of include social media data, internet search terms data (Google Trends or Wikipedia), medicine registries, general practitioner registries, and weather data. Of course, future research can be directed at examining the added value

of these data sources for surveillance, but more importantly research should be directed at how to incorporate these sources – if valuable – in the surveillance process flow. This is quite a challenge as most of these data sources comprise unstructured data (data that cannot be captured in a tabular format), which does not easily lend itself to be processed in a structured manner, as is required for surveillance purposes. It would require advanced data science skills to obtain and include such data sources.

Data analysis

In this thesis, we have developed methods for exploratory data analysis in infectious disease epidemiology. We have focused on methods that are mostly scale-independent for the combination of multiple data types and for the assessment of patterns across resolution levels. In this, we heavily relied on exploratory data visualization techniques. So how can we translate our findings into recommendations for daily practice of infectious disease surveillance?

In this thesis we have explicitly addressed the fact that infectious diseases dynamics relevant for disease control occur at a certain resolution level, which might not necessarily be the resolution level of the available data. As stated before, this is an issue that ecologists are well aware of, but one that gets little attention in epidemiology. However, it is important to take resolution level into account when searching for patterns in the data; patterns might easily be missed when searching at the wrong scale. So why is this awareness lacking in infectious disease epidemiology? The key might be education. Most infectious disease epidemiologists have had a general training in epidemiology or medicine. In many epidemiology domains, for example chronic disease epidemiology, it is assumed that observations are independent of each other. In infectious disease epidemiology we are sure that observations are dependent. Infections are clustered in time, place and at the genetic sequence level. Hence, scale issues are relevant.

Infectious disease epidemiology is only a small field and exploratory infectious disease epidemiology is even smaller. This fact is reflected in the training programs, where the most attention is devoted to confirmatory data analysis techniques. It would be a great start if in special training programmes for infectious disease epidemiologists, such as the field epidemiology training program of the ECDC or in-house training programs at public health departments, exploratory data analysis and visualisation techniques for place and pathogen type data, as well as time data, are incorporated. Issues of scale could be addressed in these courses, hopefully increasing awareness. Knowledge regarding scale issues would lead to a better understanding of infectious disease data and additionally infectious disease dynamics. Moreover, such awareness should decrease the use of arbitrary scale choices in representing these data, leading to better research and surveillance. It would be a bridge too far to massively train infectious disease epidemiologists in adopting scale-independent

methods, because interpretation of these methods is not that straightforward. Future research could be directed at how to improve interpretation of these methods; visualization techniques might play a central role.

Further research on this topic can contribute to increasing awareness of scalability issues. Here, we have focused on scale-independent methods, as this approach is often used in data science. However, that is not to say that this is the only way to address scalability. As discussed above, there is much to be learned from other fields, such as ecology and demography, or even from forensic police, in handling and analyzing time, place and pathogen type data. Finally, next to scalability there are other issues in exploratory data analysis for infectious disease epidemiology that warrant further attention from the research community. For example, how can non-structured data sources (other than the data traditionally collected through surveillance) be analyzed? How can these data sources be related to existing surveillance sources? And, how can missing data on one or more data types be dealt with in the analysis?

Data products

The final step in the data science process is the creation of data products. Data products can be seen as a product for communication of the findings to the intended audience. Examples of such products are visualizations or interactive tools. We have developed interactive tools for contact tracing and cluster validation in Chapters 5 and 6 of this thesis. Data products are extremely useful for bridging gaps between disciplines and for translating research findings. In Chapter 5, we bridged a gap between epidemiologists and policy makers with an interactive tool for decision making in a contact tracing context, and in Chapter 6, between statisticians and epidemiologists with an interactive tool to assess plausibility of automatically identified transmission clusters. Luckily, this view is becoming increasingly adopted within the research community. First, the software for developing such tools is becoming more accessible; for instance such software is often made available in easy-to-use open source packages, such as R shiny. For infectious disease epidemiology, developed tools can be shared on a central platform: the R epidemics consortium (RECON) [19]. Second, more and more scientific journals require sharing of data and code in accessible formats. This is one of the strengths of the data science approach, which needs to be further adopted in the infectious disease domain.

Data process

Finally, and most importantly, data science is characterized by process thinking: it starts with obtaining and merging data and ends with a data product for communicating results. Through regarding data questions as a process, the main advantage is that hurdles between different steps in the process are easily identified and can subsequently be overcome. Especially in Chapters 5 and 6, we have shown that such an inclusive process approach

delivers very comprehensive solutions. Currently in epidemiology, methodological development is very focused on one step in the data process: the analysis phase. However, what we need is a broader focus with the inclusion of methods development for data collection and merging, as well as methods development for creating data products. This can be achieved by reorganizing epidemiology training programmes, by including courses on data management and the creation of data products. Besides being beneficial for infectious disease epidemiology, this would also be applicable to other domains such as genetic epidemiology, where large amounts of data have to be handled as well. This will deliver more broadly schooled epidemiologists, in that they will become experts in data management, data analysis and data visualization. In addition, close collaboration in teams covering these expert fields would enhance this process thinking. Currently, mainly experts in epidemiology or in data management are represented in infectious disease departments. Broadening these departments with data visualization experts and data scientists would be essential to face future data developments. Organizing infectious disease epidemiology departments at regional and national levels accordingly would be the necessary step to maximise the benefit of exploratory infectious disease epidemiology for infectious disease control.

Conclusion

The objective of this thesis was to develop and improve upon new methods for exploratory data analysis in infectious disease epidemiology, as an answer to upcoming questions concerning patterns and structure in data. We have used a data science approach to develop scale-independent methods for combining data of different data types, and for methods for revealing intrinsic data patterns within a data type. Also, the use of visualization techniques for data exploration and developing data products has greatly contributed to the generation of hypotheses and for bridging gaps between various research professions. Comprehensive exploratory data analysis has enormous potential to provide answers to current and future data challenges concerning patterns and structure in infectious disease data. Research into novel methods for exploratory data analysis should be pursued, as the data challenges are only becoming greater. However, exploratory data analysis should be considered within the larger surveillance process, as its success heavily depends on the quality of the input data and on how its findings are communicated and used. By considering the larger surveillance process, gaps between different phases in the process can be identified and provide keys for improving surveillance as well as suggestions for further research to take exploratory infectious disease epidemiology to the next level. Within this thesis, this leads to the following recommendations: 1) to increase data collection and integrate data management of pathogen type data within the surveillance process to increase availability of this data type and to handle the resulting future data challenges, and 2) to expand epidemiology training with data management and data visualization/data product creation courses, and to organise public health departments in closely collaborating teams covering the expert fields of data management, data science, epidemiology and data visualization, to best contribute to the provision of evidence for infectious disease control.

References

1. Levin SA. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. 1992;73(6):1943-67.
2. van Wijhe M, Tulen AD, Korthals Altes H, McDonald SA, de Melker HE, Postma MJ, et al. Quantifying the impact of mass vaccination programmes on notified cases in the Netherlands. Epidemiology and infection. 2018;146(6):716-22.
3. Grenfell BT, Bjørnstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature. 2001;414(6865):716-23.
4. Holdsworth AM, Kevlahan NK, Earn DJ. Multifractal signatures of infectious diseases. Journal of the Royal Society, Interface. 2012;9(74):2167-80.
5. Dalziel BD, Kissler S, Gog JR, Viboud C, Bjornstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science (New York, NY). 2018;362(6410):75-9.
6. Chen Y, Jiang B. Hierarchical Scaling in Systems of Natural Cities. 2018;20(6):432.
7. Rozenfeld HD, Rybski D, Andrade JS, Jr., Batty M, Stanley HE, Makse HA. Laws of population growth. Proc Natl Acad Sci U S A. 2008;105(48):18702-7.
8. Kulldorff M. A spatial scan statistic. Communications in Statistics - Theory and Methods. 1997;26(6):1481-96.
9. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS medicine. 2005;2(3):e59.
10. Lemey P, Salemi M, Vandamme AM. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing: Cambridge University Press; 2009.
11. European Commission. 2018 reform of EU data protection rules [Available from: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.
12. Wet Publieke Gezondheid: Overheid.nl; [Available from: https://wetten.overheid.nl/BWBR0024705/2019-07-01.
13. CH vG, PA V, M S, den Blv, J M, den Blv, et al. Eenheid van Taal in de Nederlandse zorg : Van eenduidige informatie-uitwisseling tot hulpmiddel voor betere zorg: Rijksinstituut voor Volksgezondheid en Milieu RIVM; 2018.
14. Faensen D, Claus H, Benzler J, Ammon A, Pfoch T, Breuer T, et al. SurvNet@RKI--a multistate electronic reporting system for communicable diseases. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2006;11(4):100-3.
15. Panigutti C, Tizzoni M, Bajardi P, Smoreda Z, Colizza V. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. Royal Society open science. 2017;4(5):160950.
16. Finger F, Genolet T, Mari L, de Magny GC, Manga NM, Rinaldo A, et al. Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proc Natl Acad Sci U S A. 2016;113(23):6421-6.
17. Sacks JA, Zehe E, Redick C, Bah A, Cowger K, Camara M, et al. Introduction of Mobile Health Tools to Support Ebola Surveillance and Contact Tracing in Guinea. Global health, science and practice. 2015;3(4):646-59.
18. Eisenkraft A, Afriat A, Hubary Y, Lev R, Shaul H, Balicer RD. Using Cell Phone Technology to Investigate a Deliberate Bacillus anthracis Release Scenario. Health security. 2018;16(1):22-9.
19. RECON: R Epidemics Consortium; [Available from: https://www.repidemicsconsortium.org/.



A

Summary
Nederlandse samenvatting
Dankwoord
Curriculum vitae
List of publications

Summary

Over the past decades the amount of digital data in the world is increasing exponentially. The increasing amount and complexity of data has implications for 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 infected persons. Advances in sequencing technology have resulted in an enormous increase in resolution of data on genetic information of infectious pathogens. Advances in data management and data storage capacity allow for connecting all these data at an individual level. Taken together, these developments have led to an increase in both complexity and resolution of infectious disease data. This gives rise to new questions: 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? And at what scales should we look for patterns in surveillance data? These questions concern the structure of data and patterns in data. These issues are typically addressed in exploratory data analysis (EDA). EDA has gained little attention in traditional infectious disease epidemiology research, but has a central place in data science. Data science aims to maximise the amount of information derived from data by studying it intensively, until its contents, structure, patterns and pitfalls are fully understood. The objective of this thesis is to develop and improve new methods for exploratory data analysis in infectious disease surveillance. In this thesis, we propose methods to analyze data available at multiple resolution levels and to combine different data types in infectious disease surveillance. In doing so, we use techniques central to data science.

In Chapter 2 we introduce and use four data types in infectious disease epidemiology: time; place; person; and pathogen. We generate hypotheses for the increasing incidence of hepatitis B in certain parts of the Netherlands. In Chapter 3, we focus on the place data type and we propose a new method (dot map cartograms) for a more objective and more scale independent display of disease incidence data on a map. We use dot map cartograms to simultaneously display absolute as well as relative numbers of disease incidence. We show its added value by applying the method to incidence data of Q fever in the Netherlands and pertussis in Germany. In Chapter 4, we focus on the time data type and used scale independent entropy measures to identify intrinsic data patterns in a large database of all infectious disease notifications in the Netherlands since 2003. We show that these intrinsic data patterns can be used to characterize and order infectious diseases by their dynamic properties, which enhances surveillance goals. So far, all these developed methods deal with disease incidence data. In Chapter 5 we use another data source in infectious disease epidemiology: data collected during contact tracing in serious infectious diseases. We use a Bayesian model to calculate the probability whether a person exposed to a case is still at risk of developing symptoms. By visualizing the risk classifications, we develop an informative

tool to aid public health professionals in decision-making. In Chapter 6, we improve an existing method that uses time, place and pathogen data types for identifying infectious disease clusters. We develop an interactive visualization tool to assess the appropriateness of the outcome and to asses model performance of this algorithm. Clustering parameters can be varied in order to help public health professionals with the interpretation and picking the model with most plausible clusters according to expert knowledge.

Our research focuses around two main themes: the use of multiple resolution levels within one data type (Chapter 3 and 4) and the combination of multiple data types (Chapter 2, 5 and 6).

Regarding the use of multiple resolution levels within one data type, we have sought the answer in developing methods (dot map cartograms) or applying measures (Rényi entropy) that combine or study effects across resolution scales. A major advantage of these methods is that intrinsic data patterns are revealed regardless of aggregation level. This allows identification of intrinsic data patterns relevant to epidemiological phenomena, rather than of data patterns dependent upon choices in aggregation.

In the combination of data types, we came across several challenges. First, regarding the data collection phase, data on the four data types is often not present in the same database, and unique identifiers may be lacking. It requires advanced data management skills to perform proper matching of these data types on an individual level. Also, not all data on all data types might be available. Pathogen type data in particular are often lacking or delayed. If all data types are jointly analyzed, this will result in substantial data loss. Solutions for missing data on one type might be sought in the analysis phase, as shown in this thesis, but preferably, more efforts should be put in completion of data. Second, there is the challenge of actually combining data types for analysis. We find that with two-way combinations of data types we can still rely on traditional epidemiological methods, however, with 3-way combinations, we have to use a scale independent method for combining distances in time, geographic location and genetics. As for missing data on one or more dimensions, flexible adaption of the techniques to include two instead of three data types is one solution. A final challenge is presentation of combined data types. We show that data visualization techniques are important, and we use these techniques to develop interactive data products. For use in daily practice, we can acquire more information on user-friendliness of the tools and subsequently continuously improve upon these products. Comprehensive exploratory data analysis has enormous potential to provide answers to current and future data challenges concerning patterns and structure in infectious disease data. Research into novel methods for exploratory data analysis should be pursued, as the data challenges are only becoming greater. However, exploratory data analysis should be considered within the larger surveillance process, as its success heavily depends

on the quality of the input data and on how its findings are communicated and used. By considering the larger surveillance process, gaps between different phases in the process can be identified and provide keys for improving surveillance as well as suggestions for further research to take exploratory infectious disease epidemiology to the next level. Within this thesis, this leads to the following recommendations: 1) to increase data collection and integrate data management of pathogen type data within the surveillance process to increase availability of this data type and to handle the resulting future data challenges, and 2) to expand epidemiology training with data management and data visualization/data product creation courses, and to organise public health departments in closely collaborating teams covering the expert fields of data management, data science, epidemiology and data visualization, to best contribute to the provision of evidence for infectious disease control.

Samenvatting

De laatste tientallen jaren is de hoeveelheid digitale data wereldwijd exponentieel toegenomen. Deze toegenomen hoeveelheid van en complexiteit in de data heeft implicaties voor alle wetenschapsgebieden, inclusief het domein van de infectieziekten epidemiologie. Dankzij automatische en digitale surveillance programma's, is het gemakkelijker geworden om grote hoeveelheden aan data over geïnfecteerde personen te verzamelen. Ontwikkelingen in de sequentie technologie hebben geresulteerd in een enorme toename in resolutie van data over genetische informatie van infectieuze pathogenen. Ontwikkelingen in data management en data opslag capaciteit maken het verbinden van al deze data op individueel patiënt niveau mogelijk. Samen hebben deze ontwikkelingen geleid tot een toename in complexiteit en resolutie van infectieziekten data. Dit leidt tot een nieuw soort vragen: Hoe kunnen we uitbraken detecteren en onderzoeken wanneer we niet alleen informatie hebben over de tijd en plaats van patiënten, maar ook informatie over de genetisch structuur van de pathogenen? Hoe moeten we al deze data typen combineren? En op wat voor schaal moeten we kijken naar patronen in surveillance data? Deze vragen gaan over de structuur van en patronen in de data. Dit zijn typisch onderwerpen die worden aangesproken in exploratieve data analyse (EDA). EDA heeft weinig aandacht gekregen in onderzoek van het traditionele infectieziekten epidemiologie domein, maar heeft een centrale plaats in data science. Data science heeft als doel om de hoeveelheid aan informatie die afgeleid wordt uit data te maximaliseren door deze intensief te bestuderen, totdat de inhoud, structuur, patronen en valkuilen volledig begrepen zijn. Het doel van deze thesis is om nieuwe methoden te ontwikkelen en verbeteren voor exploratieve data analyse in infectieziekten epidemiologie. In deze thesis introduceren we methoden om data te analyseren op meerdere resolutie niveaus en methoden voor het combineren van verschillende data typen in infectieziekten epidemiologie. Hierbij gebruiken we technieken die centraal staan in de data science.

In hoofdstuk 2 introduceren we vier data typen in de infectieziekten epidemiologie: tijd, plaats, persoon en pathogeen. We genereren hypothesen voor de toegenomen hepatitis B incidentie in een aantal regio's van Nederland. In hoofdstuk 3 richten we ons op het plaats data type en introduceren we een nieuwe methode (een stippen-kaart cartogram) voor een meer objectieve en schaal onafhankelijke weergave van ziekte incidentie op een kaart. We gebruiken stippen-kaart cartogrammen om tegelijkertijd absolute en relatieve aantallen van ziekte incidentie weer te geven. We laten de toegevoegde waarde van deze methode zien door deze toe te passen op de incidentie data van Q koorts in Nederland en kinkhoest in Duitsland. In hoofdstuk 4 richten we ons op het tijd data type en gebruiken we schaal onafhankelijke entropie maten om intrinsieke data patronen te identificeren in een grote database met alle infectieziekten meldingen in Nederland sinds 2003. We laten zien dat deze intrinsieke data patronen gebruikt kunnen worden om infectieziekten te karakteriseren

en ordenen aan de hand van hun dynamische eigenschappen, wat uiteindelijk bijdraagt aan de doelen van infectieziekten surveillance. Tot nu toe gaan al deze methoden over het omgaan met ziekte incidentie data. In hoofdstuk 5 gebruiken we een andere data bron in de infectieziekten epidemiologie: data zoals verzameld in contactopsporing bij ernstige infectieziekten uitbraken. We gebruiken een Bayesiaans model om de kans uit te rekenen dat een blootgesteld persoon nog symptomen gaat ontwikkelen. Door het visualiseren van risico classificaties hebben we een informatieve tool ontwikkeld om publieke gezondheidszorgprofessionals te helpen bij het beslissen over uitbraakbeperkende maatregelen. In hoofdstuk 6 verbeteren we een al bestaande methode die gebruik maakt van tijd, plaats en pathogeen data typen voor het identificeren van infectieziekten clusters. We ontwikkelen een interactieve visualisatie tool voor het beoordelen van de correctheid van de geïdentificeerde clusters en de model performance van het algoritme. Cluster parameters kunnen gevarieerd worden om publieke gezondheidszorg professionals te helpen met de interpretatie en het kiezen van het model met de meest plausibele clusters gegeven kennis van experts.

Ons onderzoek richt zich op twee hoofdthema's: het gebruik van meerdere resolutie levels binnen een data type (hoofdstuk 3 en 4) en het combineren van meerdere data types (hoofdstuk 2, 5 en 6).

Wat betreft het gebruik van meerdere resolutie levels binnen een data type, hebben we het antwoord gezocht in het ontwikkelen van methoden (stippen-kaart cartogrammen) of het toepassen van maten (Rényi entropie) die effecten over resolutieschalen heen combineren of bestuderen. Een enorm voordeel van deze methoden is dat intrinsieke data patronen worden onthuld ongeacht het aggregatie niveau. Dit maakt het mogelijk om intrinsieke patronen te identificeren die kenmerkend zijn voor epidemiologische fenomenen, in plaats van patronen die afhankelijk zijn van keuzes in aggregatie niveaus.

Bij het combineren van data typen kwamen we verschillende uitdagingen tegen. Allereerst met betrekking tot de dataverzamelingsfase; data van de vier verschillende data typen is vaak niet in eenzelfde database aanwezig en unieke identificatienummers ontbreken meer dan eens. Het vereist daarom vergaande data management vaardigheden om deze data typen op een individueel niveau goed te matchen. Ook komt het voor dat niet alle data typen beschikbaar zijn. Pathogeen type data in het bijzonder ontbreekt vaak of is vertraagd. Wanneer alle data typen gezamenlijk worden geanalyseerd, leidt het beperkt beschikbaar zijn van een data type tot substantieel data verlies. Oplossingen voor missende data van één type kunnen gezocht worden in de analyse fase, zoals getoond in dit proefschrift, maar bij voorkeur wordt er meer inspanning gestopt in het completeren van de data. Ten tweede, de uitdaging van het daadwerkelijk combineren van de verschillende data typen voor analyse. We hebben ontdekt dat met tweeweg combinaties van data typen

we nog steeds kunnen terugvallen op traditionele epidemiologische methoden, maar dat met drie- of vierweg combinaties we schaal onafhankelijke methoden moeten gebruiken om afstanden in tijd, plaats en genetica te combineren. Wat betreft ontbrekende data op een of meer dimensies, is flexibele aanpassing van de technieken om twee in plaats van drie data typen op te nemen een oplossing. Een laatste uitdaging is de presentatie van gecombineerde data typen. We laten zien dat datavisualisatie technieken belangrijk zijn en we gebruiken deze technieken om interactieve dataproducten te ontwikkelen. Voor gebruik van deze producten in de dagelijkse praktijk hebben we meer informatie nodig over de gebruiksvriendelijkheid van de tools, zodat we deze producten continu kunnen blijven verbeteren.

Uitgebreide exploratieve data-analyse heeft een enorm potentieel om antwoorden te bieden op huidige en toekomstige data uitdagingen met betrekking tot patronen en structuur in infectieziektegegevens. Er moet meer onderzoek gedaan worden naar nieuwe methoden voor exploratieve data-analyse, aangezien de uitdagingen op dit vlak alleen maar groter worden. Exploratieve data-analyse staat echter niet op zichzelf en dient beschouwd te worden als onderdeel in het grotere surveillanceproces, aangezien het succes van goede exploratieve data analyse ervan sterk afhangt van de kwaliteit van de data en van de manier waarop de bevindingen ervan worden gecommuniceerd en gebruikt. Door rekening te houden met het grotere surveillanceproces, kunnen hiaten tussen verschillende fasen in het proces worden geïdentificeerd en aanknopingspunten bieden voor het verbeteren van de surveillance, evenals suggesties voor verder onderzoek om de exploratieve infectieziekten epidemiologie naar een hoger niveau te tillen. Binnen dit proefschrift leidt dit tot de volgende aanbevelingen: 1) data verzameling van het pathogeen data type moet worden uitgebreid en geïntegreerd binnen het surveillanceproces om de beschikbaarheid van dit data type te vergroten en de daaruit voortvloeiende toekomstige data uitdagingen aan te pakken, en 2) epidemiologie training moet worden uitgebreid met cursussen voor data management en datavisualisatie/creëren van data producten, en public health afdelingen moeten worden georganiseerd in nauw samenwerkende expert teams op het gebied van data management, data science, epidemiologie en datavisualisatie, om zo goed mogelijk bij te dragen aan het leveren van informatie voor infectieziektebestrijding.

Dankwoord

Tot slot, het dankwoord, het meest gelezen deel van dit proefschrift.

Allereerst, mijn promotor, Jacco. Bedankt voor het enorme vertrouwen wat je altijd in me hebt gehad. Naast de inhoudelijke kennis die je altijd graag met me deelde, heb ik me ook altijd enorm gesteund gevoeld in de totstandkoming van dit proefschrift, met wat jij noemt een 'typisch Loes-gehalte'.

Ook Susan, mijn co-promotor, wil ik graag bedanken. Naast je inhoudelijke kennis en expertise, wil ik je met name bedanken voor de mental support die ik van je heb gekregen tijdens dit traject. Als ik er even doorheen zat, kon ik altijd bij je terecht, dat was voor mij enorm waardevol.

Van alle co-auteurs wil ik specifiek Don en Jantien bedanken. Naast de vele koffies en lunches die we samen genoten hebben, kon ik altijd bij jullie binnenlopen voor vragen over ingewikkelde modellen, statistische vraagstukken, visualisaties en andere relevante (en minder relevante) zaken. Ook wil ik hier graag mijn 'native-english-speaker'- (ex-)collega's, Rolina, Scott en Lisa, bedanken voor het doorlezen en verbeteren van mijn manuscripten.

Mijn paranimfen, Marianne en Annika. Marianne, Mar, we gaan al lang terug. Ik ben enorm dankbaar voor jouw vriendschap en vind het een enorme steun dat je tijdens mijn promotie, maar eigenlijk altijd, naast me staat. We delen al jaren lief en leed en hoop dat we dat nog jaren blijven doen. Annika, ook wij gaan inmiddels al een hele tijd terug. Bedankt voor alle koffie/bier/thee/wijn/lunch-momentjes van de afgelopen jaren en voor dat je nu naast me staat.

Wanneer je een tijdje over je proefschrift doet zijn er ook heel veel (ex-)EPI-collega's te bedanken. Met stip op nummer 1 voor koffie, afleiding en andere niet-werkgerelateerde zaken is daar Chantal. Onze koffiemomenten zijn wat teruggeschroefd nu je in Aberdeen zit, maar het digitale coronatijdperk biedt mogelijkheden. Chantal, enorm bedankt voor je mental support. Joram, wat een geluk dat ik met jou een kamer kon delen. Het was altijd gezellig en vind het ontzettend jammer dat je weg bent. Janneke, Renske, Jeroen, ook met jullie heb ik veel tijd doorgebracht in de pantry, U tweede verdieping. Het was fijn om met jullie af en toe een wandelingetje te maken, koffie of borreltje te drinken en/of ander vorm van afleiding te hebben. Dan is er nog een hele rits aan (ex-)collega's die in meer of mindere mate hebben bijgedragen aan proefschrift, al dan niet in de vorm van afleiding. Bedankt!

Naast collega's zijn er natuurlijk ook nog vrienden en familie die ik graag wil bedanken voor de ondersteuning en afleiding de afgelopen jaren. Jennifer, Jen, wat ben ik blij dat ik jou heb leren kennen tijdens mijn studietijd. De vele borrels en tripjes staan nog in mijn geheugen gegrift. Bedankt voor je creatieve blik op mijn proefschrift, het ontwerpen van de cover en de vele gezellige momenten de afgelopen jaren. Valérie, Val, je bent de enige van mijn vriendinnen die naast het dankwoord waarschijnlijk ook de rest van mijn proefschrift wil lezen. Het is altijd lachen met jou en ik hoop dat we dat dan ook nog lang kunnen doen. Graag wil ik ook mijn familie bedanken. Ieder afzonderlijk hebben jullie een belangrijke rol gespeeld, die ik zeer op prijs heb gesteld. Pap en mam, Anne en Giel, bedankt voor jullie steun. Aly en Nico, bedankt dat jullie het mogelijk maken dat wij ons kunnen blijven ontwikkelen door zoveel zorg voor onze boefjes op te vangen. Dat heeft zeker bijgedragen aan dit proefschrift! Dan is er nog één persoon, Bas, bedankt voor alles. Als allerlaatste, Liza en Robin. Jullie inhoudelijke bijdrage aan dit proefschrift laat nog even op zich wachten, maar wat betreft afleiding scoren jullie een 10!

Curriculum vitae

Loes Soetens was born on March 22nd in 1985 in 's-Hertogenbosch, the Netherlands. She studied Pedagogical Sciences at the Radboud University in Nijmegen, where she obtained her Bachelor's degree in 2007. In 2006, she changed track and started studying Biomedical Sciences at the Radboud University in Nijmegen. She obtained her Bachelor's degree in 2009 and her Master of Science degree (specialization Epidemiology) in 2011. She did her final master's internship and thesis at the National Institute for Public Health and the Environment (RIVM) on the seasonality of tuberculosis. She continued working at the RIVM after her graduation as an epidemiologist in various research groups (healthcare infections, sexually transmitted infections). After a short secondment to the World Health Organization in 2013, she started her PhD project on exploratory infectious disease epidemiology in 2014. In this project, she combined epidemiology with data visualization and data science, to study current data challenges in infectious disease epidemiology. This project was finished in January 2020. In 2018 she decided to broaden her knowledge on data science processes and started a Master on Business Process Management and IT (specialization Data Science Management) at the Open University. She obtained her Master of Science degree in October 2020. After finishing her PhD project she continued her work at the RIVM at the surveillance department. She is currently working on projects improving surveillance processes and coordinates the notifiable infectious disease surveillance.

List of publications

de Gier, B., et al. (2021). Vaccine effectiveness against SARS-CoV-2 transmission and infections among household and other close contacts of confirmed cases, the Netherlands, February to May 2021. Euro Surveill 26(31).

McDonald, S. A., et al. (2021). Risk factors associated with the incidence of self-reported COVID-19-like illness: data from a web-based syndromic surveillance system in the Netherlands. Epidemiol Infect 149: e129.

Overbosch, F., et al. (2020). Public health response to two imported, epidemiologically related cases of Lassa fever in the Netherlands (ex Sierra Leone), November 2019. Euro Surveill 25(15).

Soetens, L., et al. (2019). Visual tools to assess the plausibility of algorithm-identified infectious disease clusters: an application to mumps data from the Netherlands dating from January 2009 to June 2016. Euro Surveill 24(12).

Fitzner, J., et al. (2018). Revision of clinical case definitions: influenza-like illness and severe acute respiratory infection. Bull World Health Organ 96(2): 122-128.

Monge, S., et al. (2018). Echovirus type 6 transmission clusters and the role of environmental surveillance in early warning, the Netherlands, 2007 to 2016. Euro Surveill 23(45).

Soetens, L., et al. (2018). Real-time Estimation of Epidemiologic Parameters from Contact Tracing Data During an Emerging Infectious Disease Outbreak. Epidemiology 29(2): 230-236.

Koek, M. B. G., et al. (2017). Adhering to a national surgical care bundle reduces the risk of surgical site infections. PLoS One 12(9): e0184200.

Soetens, L., et al. (2017). Dot map cartograms for detection of infectious disease outbreaks: an application to Q fever, the Netherlands and pertussis, Germany. Euro Surveill 22(26).

van Aar, F., et al. (2017). Outbreaks of syphilis among men who have sex with men attending STI clinics between 2007 and 2015 in the Netherlands: a space-time clustering study. Sex Transm Infect 93(6): 390-395.

Soetens, L. C., et al. (2015). Chlamydia test results were associated with sexual risk behavior change among participants of the Chlamydia screening implementation in The Netherlands. Sex Transm Dis 42(3): 109-114.

Soetens, L. C., et al. (2015). Ongoing transmission of hepatitis B virus in rural parts of the Netherlands, 2009-2013. PLoS One 10(2): e0117703.

Xiridou, M., et al. (2015). Public health measures to control the spread of antimicrobial resistance in Neisseria gonorrhoeae in men who have sex with men. Epidemiol Infect 143(8): 1575-1584.

Soetens, L. C., et al. (2013). Contribution of seasonality in transmission of Mycobacterium tuberculosis to seasonality in tuberculosis disease: a simulation study. Am J Epidemiol 178(8): 1281-1288.

