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

Soetens, L.C.

A

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.