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

Soetens, L.C.

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/.