



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).

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.